

# Convergence and error control of consistent PINNs for elliptic PDEs

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

V1: 06/13/24; V2:10/4/24

## 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. 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 total number of point samples, under all Besov class model assumptions for the right hand side and boundary data.

We then turn to the study of numerical procedures and analyze whether a proposed numerical procedure (nearly) achieves the optimal recovery error. In particular, we analyze numerical methods which generate the numerical approximation to  $u$  by minimizing specified data driven loss functions over a set  $\Sigma$  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 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. We prove that minimization of the new loss  $\mathcal{L}^*$  over restricted neural network spaces  $\Sigma$  provides an optimal recovery of the solution  $u$ , provided that the optimization problem can be numerically executed and  $\Sigma$  has sufficient approximation capabilities. We also analyze variants of  $\mathcal{L}^*$  which are more practical for implementation. Finally, numerical examples illustrating the benefits of the proposed loss functions 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  $\Omega$  be a bounded domain in the Euclidean space  $\mathbb{R}^d$ ,  $d \geq 2$ , and  $\bar{\Omega}$  be its closure in  $\mathbb{R}^d$ . Consider the elliptic Dirichlet problem

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

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

**A1:**  $\Omega$  is a Lipschitz (graph) domain, i.e.,  $\Omega$  is an open connected set in  $\mathbb{R}^d$  with a Lipschitz boundary in the sense described in [67];

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

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

**A4:** the diffusion coefficient  $a$  satisfies the coercivity condition

$$0 < M_0 \leq a(\mathbf{x}) \leq M, \quad \mathbf{x} \in \Omega, \quad (1.2)$$

for some constants  $M_0, M > 0$ . 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. [32]) implies that (1.1) has a unique solution  $u \in H^1(\Omega)$  which satisfies

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

where the constants  $c, C$  depend on  $\Omega$ ,  $M_0$  and  $M$ <sup>1</sup>. Note that assumptions **A2-A3** could be replaced with other smoothness assumptions on  $f$  and  $g$ , respectively, which would result in a theory similar to the one described in this paper.

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. We will comment on the extension of our results to more general domains  $\Omega$  in §11.

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

$$\|u - v\|_{H^1(\Omega)} \asymp \|f + \Delta v\|_{H^{-1}(\Omega)} + \|g - v\|_{H^{1/2}(\partial\Omega)}, \quad (1.4)$$

with absolute constants in the equivalence. Hence, the problem of finding the minimum of the function  $\mathcal{L}_T$  (which we call theoretical loss function),

$$\mathcal{L}_T(v) := \|f + \Delta v\|_{H^{-1}(\Omega)} + \|g - v\|_{H^{1/2}(\partial\Omega)}, \quad (1.5)$$

over the whole of  $H^1(\Omega)$  will have  $u$  as its unique solution, that is,

$$u = \operatorname{argmin}_{v \in H^1(\Omega)} \mathcal{L}_T(v). \quad (1.6)$$

## 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  $\Omega$  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. We remark that our approach can in principle be applied to other error norms as well. In fact, the case of  $X = H^{1/2}(\Omega)$  has been considered in a similar way in [78].

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 (A)FEMs, this complexity can be

---

<sup>1</sup>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  $c_1, c_2 > 0$ .

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. [6]). 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 filling manifolds and thus avoid unstable numerical methods, see [10]. 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  $f, g, \Omega$  are known/given to us. One setting is to assume that we can query (ask questions about)  $f, g, \Omega$ , 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  $\bar{\Omega}$ . In order for these to make sense, we assume that  $f$  and  $g$  are continuous functions, i.e.,  $f \in C(\bar{\Omega})$ ,  $g \in C(\partial\Omega)$ . 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  $\Sigma_n$  is the underlying space used to create  $\hat{u}$  then the accuracy of the numerical recovery will depend on  $n$  and the choice of  $\Sigma_n$ . 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  $n \rightarrow \infty$ . This issue is discussed in §9. We remark that when neural network based methods such as PINNs are used to perform this numerical approximation, a critical component of the numerical algorithm becomes the optimization method used to train the neural network. This issue, which is specific to neural networks, is addressed in detail in §10.

In going further, we consider problem (1.1) under the following assumptions:

**B1:**  $\Omega = (0, 1)^d$ ,  $d \geq 2$ ;

**B2:**  $f \in C(\bar{\Omega}) \subset H^{-1}(\Omega)$ ;

**B3:**  $g \in H^{1/2}(\partial\Omega) \cap C(\partial\Omega)$ ;

**B4:**  $a \equiv 1$ .

Concerning the sense in which we know  $\Omega$ , 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  $\Omega$  is to first find a polyhedral domain  $\hat{\Omega}$  that approximates  $\Omega$ , solve the corresponding PDE on  $\hat{\Omega}$ , and then analyze the effect of the approximation of  $\Omega$  by  $\hat{\Omega}$ .

## 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  $\Sigma_n$  denote the set of outputs of a neural

network with  $n$  parameters and fixed architecture. Once  $\Sigma_n$  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  $\bar{\Omega}$ , thereby providing the values

$$\mathbf{f} = (f_1, \dots, f_{\tilde{m}}), \quad f_i := f(\mathbf{x}_i), \quad i = 1, \dots, \tilde{m}; \quad \mathbf{g} = (g_1, \dots, g_{\bar{m}}), \quad g_i := g(\mathbf{z}_i), \quad i = 1, \dots, \bar{m}, \quad (1.7)$$

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

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

The performance of such a numerical method will depend on the numbers  $\tilde{m}, \bar{m}$ , 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  $\Sigma_n$  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

$$\hat{u} \in \operatorname{argmin}_{S \in \Sigma_n} \mathcal{L}(S), \quad (1.8)$$

where

$$\mathcal{L}(S) := \left[ \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} [\Delta S(\mathbf{x}_i) + f(\mathbf{x}_i)]^2 \right]^{1/2} + \left[ \frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} [S(\mathbf{z}_i) - g(\mathbf{z}_i)]^2 \right]^{1/2}. \quad (1.9)$$

Note that this is equivalent to the square root of the original PINNs loss [59]. Later in the paper we will consider weighted versions of  $\mathcal{L}$ . The appropriateness of the loss  $\mathcal{L}$  and how  $\tilde{m}, \bar{m}$ , and the data sites  $(\mathcal{X}, \mathcal{Z})$  should be chosen are important issues and are one of the focal points of this paper. We remark that an analysis of other methodologies for solving (1.1) using neural networks, such as the deep Ritz method [73] and finite neuron method [74], can be found in [51, 49, 66, 56, 28, 26].

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  $(\mathcal{X}, \mathcal{Z})$ . Such methods were studied in the last century (see e.g. [2, 41, 38]) 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  $\bar{\Omega}$  and  $\partial\Omega$ , respectively. Secondly, to have any hope of proving a priori guarantees on convergence or rates of convergence as  $\tilde{m}, \bar{m}, n \rightarrow \infty$ , we need at a minimum that  $f$  comes from a compact subset  $\mathcal{F}$  of  $C(\bar{\Omega})$  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 [25, 47], and has recently become extremely popular with the introduction of PINNs [59]. Despite the increasing usage of PINNs for numerically solving a wide range of PDEs (see for example, [43, 11, 7, 52] and the references therein), a satisfactory analysis of the convergence and performance of these methods has not yet been given. We refer to [19] and the references therein for an overview of the existing approaches. Some partial progress towards this goal has recently been made, see for instance [31, 54, 63, 64, 78].

For example, in [63], convergence is proved in the  $C(\bar{\Omega})$ -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 [47, 48, 4], this approach is unable to rigorously handle arbitrary boundary conditions on general domains.

Further, in [64] and [78], 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 [3, 68, 27] for instance).

An investigation of the PINNs method for solving a variety of equations, obtaining upper bounds on the rate of convergence with respect to the number of collocation points, has been obtained in [54]. In [50], a mini-max analysis for various neural network PDE formulations (with built in zero boundary conditions) is given for the Poisson equation *with zero boundary values* under the assumption that noisy values of the right hand side of the PDE are measured at random points.

Our work differs from the previous approaches in the following important ways:

- we consider general Sobolev and Besov smoothness assumptions on the RHS and boundary data (instead of just the Hilbert space setting as is typical in prior work) and consider error in the energy norm  $H^1(\Omega)$  (in contrast to the weaker norms often considered in prior work); Besov and Sobolev regularity is the standard staple of formulating regularity theorems for PDEs;
- we obtain matching upper and lower bounds on the theoretically optimal convergence rate in terms of the number of collocations points;
- we show that a properly designed loss function can be used to obtain an a posteriori error estimate on the solution error for the elliptic PDE;
- based upon our analysis, we propose new loss functions, including a loss function which correctly balances the boundary and domain error in the original PINNs formulation, and numerically demonstrate their improved performance on problems with (relatively) low regularity.

In particular, our analysis concentrates on 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  $\tilde{m}$  and  $\bar{m}$  to be able to reach a target accuracy  $\varepsilon$  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  $\Sigma_n$  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 guarantee small values of the solution approximation error, up to the optimal recovery rate ?

We 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 consider the error in the  $H^1(\Omega)$ -norm, and treat the situation where the boundary values are enforced through the loss function. We assume 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  $\bar{\Omega}$  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. We also discuss there 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  $B_q^s(L_p(\Omega))$  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 content 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 [13, 14, 45, 58]. The paper [69], which studies optimal sampling with respect to the negative Sobolev norm  $H^{-1}(\Omega)$ , 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 [69]. 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  $g$ , see the discussion at the beginning of §4. Our analysis there 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 then further 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  $\Sigma_n$  will approximate well  $u$ , provided  $\Sigma_n$  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 (1.7)). An important consequence of the analysis in §5 is that the widely used loss function  $\mathcal{L}$  may not be an appropriate discretization of the theoretical loss  $\mathcal{L}_T$ . We introduce the loss function  $\mathcal{L}^* : H^1(\Omega) \rightarrow \mathbb{R}$  given by

$$\mathcal{L}^*(v) := \left[ \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} |f(\mathbf{x}_i) + \Delta v(\mathbf{x}_i)|^\gamma \right]^{1/\gamma} + \left[ \frac{1}{\tilde{m}^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{\tilde{m}} \frac{|[g-v](\mathbf{z}_i) - [g-v](\mathbf{z}_j)|^2}{|\mathbf{z}_i - \mathbf{z}_j|^d} \right]^{1/2} + \left[ \frac{1}{\tilde{m}} \sum_{j=1}^{\tilde{m}} |g(\mathbf{z}_j) - v(\mathbf{z}_j)|^2 \right]^{1/2}, \quad (1.10)$$

and show that it correctly discretizes an upper bound on the theoretical loss  $\mathcal{L}_T$ . Here  $\gamma$  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 (1.4). 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  $\Sigma_n$  has suitable approximation properties. The a priori guarantee provides a near optimal recovery whenever  $\Sigma_n$  provides sufficiently good approximation of the elements in the solution model class  $\mathcal{U}$ .

In §8, we analyze a variety of other possible loss functions which provide the same a posteriori error guarantees as the loss  $\mathcal{L}^*$ . We call the use of any of these loss functions *consistent PINNs*. If the set  $\Sigma_n$  is chosen large enough, all these losses will provide a near best optimal recovery (in terms of number of data points) of the solution  $u$ . We prove that for any function  $v \in H^1(\Omega)$ , the losses suggested in §8 always provide an upper bound of the true error  $\|u - v\|_{H^1(\Omega)}$  up to the optimal recovery rate for the solution model class  $\mathcal{U}$ , provided that the approximant  $v$  can be controlled in the  $\|\cdot\|_{\mathcal{U}}$ -norm (see Theorem 7.2). This means that these losses 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  $\varepsilon$ .

However, proving that a particular numerical method of optimization (such as gradient descent) converges to a minimizer (or near minimizer) of the loss is a serious issue in optimization, which we do not address

here.

We next develop in §9 the properties needed of a set  $\Sigma_n$  so that minimizing the loss  $\mathcal{L}^*$  over  $\Sigma_n$  results in a near optimal recovery of  $u$ . It turns out that a certain restricted approximation property, guaranteed to hold when  $\Sigma_n$  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  $\Sigma_n$ , provided  $n$  is large enough.

Finally, in §10, we numerically test the loss functions described in §8 on a two-dimensional Poisson equation. These results show that using consistent PINNs provides a significant improvement in the solution error, especially for problems with low regularity.

## Acknowledgments

We are grateful to Peter Binev, Albert Cohen, Wolfgang Dahmen, and Jinchao Xu for many insightful conversations about the material in this paper. We also thank the referees for their careful reading of the manuscript and their insightful comments/suggestions which helped improve the 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. For the range  $s > 0$ , and  $0 < p, q \leq \infty$ , the Besov space  $B_q^s(L_p(\Omega))$ ,  $\Omega = (0, 1)^d$ ,  $d \geq 2$ , 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 [23, 21, 22] 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 [20].

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

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

where

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

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  $B_q^s(L_p(\Omega))$  is defined as the set of all functions in  $L_p(\Omega)$  for which

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

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  $B_q^s(L_p(\Omega))$  by adding  $\|f\|_{L_p(\Omega)}$  to it. An equivalent (quasi-)semi-norm is given by

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

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

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

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  $B_{q_2}^s(L_p(\Omega)) \subset B_{q_1}^s(L_p(\Omega))$ , i.e., these spaces get smaller as  $q$  gets smaller. Thus, all  $B_q^s(L_p(\Omega))$ ,  $q > 0$ , are contained in  $B_\infty^s(L_p(\Omega))$  once  $s$  and  $p$  are fixed. The effect of  $q$  in the definition of the Besov spaces is subtle. We remark that when  $p = q$  and  $s$  is not an integer, the space  $B_p^s(L_p(\Omega))$  is equivalent to the Sobolev space  $W^s(L_p(\Omega))$  (see [1]).

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

$$B_p^s := B_p^s(\Omega) := B_\infty^s(L_p(\Omega)) \quad (2.6)$$

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

$$\omega_r(f, t)_p \leq Mt^s, \quad t > 0, \quad (2.7)$$

and the smallest  $M$  for which (2.7) is valid is the (quasi-)semi-norm  $|f|_{B_p^s}$ . 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  $B_p^s$  is not a Lipschitz space when  $s$  is a positive integer. In view of (2.5), we have that a function  $f$  is in  $B_p^s$  if and only if

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

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

$$H^s(\Omega) = B_2^s(L_2(\Omega)) \subset B_\infty^s(L_2(\Omega)) =: B_2^s(\Omega).$$

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  $H_0^1(\Omega)$  is the collection of functions in  $H^1(\Omega)$  which vanish on the boundary of  $\Omega$ , 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  $H_0^1(\Omega)$  and is equipped with the norm

$$\|f\|_{H^{-1}(\Omega)} := \sup_{\|v\|_{H_0^1(\Omega)}=1} \langle f, v \rangle_{H^{-1} \times H_0^1}. \quad (2.9)$$

There are two equivalent ways to describe the space  $H^{1/2}(\partial\Omega)$ . The first is as the Besov space  $B_2^{1/2}(L_2(\partial\Omega))$ . Here, one needs the concept of Besov spaces on manifolds. In our case, the boundary of  $\Omega$  is quite simple since it is the union of the faces of  $\Omega$ . This definition gives the norm in  $H^{1/2}(\partial\Omega)$  through the modulus of smoothness. Using the averaged modulus of smoothness, see (12.3) in the Appendix, §12.1, one finds that the semi-norm of  $g \in H^{1/2}(\partial\Omega)$ , is

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

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  $\bar{\Omega}$  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

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

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 [34]). 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  $\bar{\Omega}$  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  $\bar{\Omega}$ . Consider the simplicial (Kuhn-Tucker) decomposition of  $\bar{\Omega}$  into simplices  $T^0$ ,

$$\bar{\Omega} = \bigcup \bar{T}^0. \quad (2.12)$$

Given any one of these simplices  $T^0$ , the number  $n_r$  of grid points from  $G_r$  in the simplex  $\bar{T}^0$  is the same as the dimension of  $\mathcal{P}_r := \mathcal{P}_r^d$ , where  $\mathcal{P}_r^d$  denotes the linear space of algebraic polynomials of order  $r$  (total degree  $r-1$ ), namely,

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

Let us consider the closed simplex

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

It is well known that polynomial interpolation using elements of  $\mathcal{P}_r$  at the grid points in  $G_r \cap \bar{T}_0$  is well posed (see e.g. Section 4 in [61] or Chapter 2 in [8]). 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  $G_r \cap \bar{T}_0$ . Then, the operator

$$L_{T_0}(f) := \sum_{\mathbf{x}_j \in G_r \cap \bar{T}_0} f(\mathbf{x}_j) \phi_j, \quad (2.13)$$

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

$$\|L_{T_0}(f)\|_{C(\bar{T}_0)} \leq \Lambda_r \|f\|_{C(\bar{T}_0)}, \quad \text{where } \Lambda_r := \left\| \sum_{j=1}^{n_r} |\phi_{j, T_0}(\mathbf{x})| \right\|_{C(\bar{T}_0)}, \quad (2.14)$$

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

For each  $k \geq 0$ , we define  $\mathcal{D}_k = \mathcal{D}_k(\Omega)$  to be the partition of  $\Omega$  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 the tensor product grid  $G_{k,r}$  of  $\bar{\Omega} = [0, 1]^d$ ,

$$G_{k,r} := \{\mathbf{x}_1, \dots, \mathbf{x}_{\tilde{m}}\} \subset \bar{\Omega}, \quad \tilde{m} = (r2^k)^d, \quad (2.15)$$

with spacing  $h = 2^{-k}(r-1)^{-1}$ . We fix  $k \geq 0$ ,  $I \in \mathcal{D}_k$ , and a 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  $F_T : \bar{T}_0 \rightarrow \bar{T}$  is an affine map from the reference simplex  $\bar{T}_0$  to  $\bar{T}$ . Notice that  $\phi_{j,T}$ ,  $j = 1, \dots, n_r$ , are the Lagrange polynomial basis for  $\mathcal{P}_r$ , corresponding to the points in  $G_{k,r} \cap \bar{T}$ , and thus we obtain the projector  $L_T$  onto  $\mathcal{P}_r$  which satisfies

$$\|L_T(f)\|_{C(\bar{T})} \leq \Lambda_r(T) \|f\|_{C(\bar{T})}, \quad \text{where} \quad \Lambda_r(T) := \left\| \sum_{j=1}^{n_r} |\phi_{j,T}(\mathbf{x})| \right\|_{C(\bar{T})}, \quad T \subset I \in \mathcal{D}_k, \quad k \geq 0. \quad (2.16)$$

Here  $\Lambda_r = \Lambda_r(T)$ , i.e.  $\Lambda_r(T)$  does not depend on  $T$  since for  $\mathbf{x} \in \bar{T}$  we have  $\phi_{j,T}(\mathbf{x}) = \phi_{j,T_0}(F_T^{-1}(\mathbf{x})) = \phi_{j,T_0}(\mathbf{y})$ , where  $\mathbf{y} = F_T^{-1}(\mathbf{x}) \in \bar{T}_0$ . 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

$$\|f - L_T(f)\|_{C(\bar{T})} \leq \|f - P\|_{C(\bar{T})} + \|L_T(f - P)\|_{C(\bar{T})} \leq (1 + \Lambda_r) \|f - P\|_{C(\bar{T})}. \quad (2.17)$$

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

$$S_k^*(f) := S_{k,r}^*(f) := \sum_{T \in \mathcal{T}_k} L_T(f) \chi_T, \quad (2.18)$$

where  $\mathcal{T}_k$  is the collection of all simplices arising from the (Kuhn-Tucker) decomposition of the dyadic cubes in  $\mathcal{D}_k(\Omega)$  and  $\chi_T$  is the characteristic function of  $\bar{T}$ . The function  $S_k^*(f)$  is well defined and continuous on  $\bar{\Omega}$  since the number of points from  $G_{k,r}$  that are on a shared face of any two simplices 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, §12.5.1) describes the approximation accuracy of  $S_k^*$ .

**Theorem 2.1.** *Let  $s > 0$ ,  $0 < p \leq \infty$ , 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 (2.18). Then, for any  $f \in B_p^s(\Omega) = B_\infty^s(L_p(\Omega))$  and any  $\tau \geq p$ , we have*

$$\|f - S_k^*(f)\|_{L_\tau(\Omega)} \leq C \|f\|_{B_p^s(\Omega)} 2^{-k(s-d/p+d/\tau)}, \quad (2.19)$$

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. The following theorem, see the Appendix, §12.5.2 for its proof, holds.

**Theorem 2.2.** *Let  $d \geq 2$ . If  $f \in B_q^s(L_p(\Omega))$  with  $s > d/p$  and  $0 < p \leq 2$ , and  $S_k^*(f)$  is as defined in (2.18), then*

$$\|f - S_k^*(f)\|_{H^1(\Omega)} \leq C \|f\|_{B_q^s(L_p(\Omega))} 2^{-k(s-1-d/p+d/2)}, \quad k \geq 0. \quad (2.20)$$

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  $C(\bar{\Omega})$ . For this, we can use the following theorem proved in the Appendix, §12.4.1.

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

$$\omega_r(\tilde{f}, t)_{C(\bar{\Omega})} \leq C \|f\|_{B_p^s(\Omega)} t^{s-d/p}, \quad t > 0. \quad (2.21)$$

### 3 Optimal recovery

In this section, we study the question of whether the information  $(\mathbf{f}, \mathbf{g})$  at the *fixed* data sites

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

is sufficient to determine  $u$  to a prescribed accuracy  $\varepsilon$  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  $\tilde{m}, \bar{m}$ , the positions of the points from  $\mathcal{X}$  and  $\mathcal{Z}$ , and the smoothness assumptions we make on  $f$  and  $g$ . Thus, we require that  $f, g$  are in certain Besov spaces that compactly embed into  $C(\bar{\Omega})$  and  $C(\partial\Omega)$ , respectively, and are such that  $u$  lies in a compact subset of  $H^1(\Omega)$ .

Before presenting the results on optimal recovery in 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 [13, 14, 69] 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 [69] and ours given below is in how the optimal recovery is obtained/proven. The optimal recovery method in [69] 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 (see for instance [30]). 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 exposition below useful and the OR community can just do a very light reading of this section. The statements in this section are all proved in detail in the appendix.

We shall assume that

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

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

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  $B_{\bar{q}}^{\bar{s}}(L_{\bar{p}}(\Omega))$  with  $\bar{s} > d/\bar{p}$  to ensure continuity. We also place the restrictions  $0 < \bar{p} \leq 2$  in order to simplify the presentation that follows. In other words, our model class assumptions on the boundary function  $g$  take the form

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

Then  $\mathcal{G}$  is a compact subset of both  $H^{1/2}(\partial\Omega)$  and  $C(\partial\Omega)$ . The latter follows because  $\bar{\mathcal{B}}$  embeds compactly into  $C(\bar{\Omega})$  since  $\bar{s} > d/\bar{p}$ . When taking the trace this compactness is preserved. The former embedding follows since the stated conditions also imply that  $\bar{s} > 1 + (d/\bar{p} - d/2)$  as  $d \geq 2$ , which implies that  $\bar{\mathcal{B}}$  embeds into  $H^1(\Omega)$ . Taking the trace will then be compact in  $H^{1/2}(\partial\Omega)$ . Note that  $\bar{s} > 1$ .

Note that, in general, the parameters  $s, p, q$  used in defining  $\mathcal{F}$  are different from  $\bar{s}, \bar{p}, \bar{q}$  used in defining  $\mathcal{G}$ . These assumptions imply that the function  $u$  we want to numerically approximate is an element of the set

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

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

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

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

$$\mathcal{G}_{\text{data}} := \mathcal{G}_{\text{data}(g)} := \{\tilde{g} \in \mathcal{G} : \tilde{g}_i = g_i, i = 1, \dots, \bar{m}\}. \quad (3.4)$$

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

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

or equivalently

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

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

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

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

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

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

is then the optimal recovery error. We are interested in bounds for  $R(\mathcal{U}_{\text{data}(u)})_X$ , 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

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

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  $m = |\mathcal{X}| + |\mathcal{Z}| = \tilde{m} + \bar{m}$  of data sites, we can ask for the optimal error we can achieve under such a budget restriction. Accordingly, we define

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

Similarly,

$$R^*(\mathcal{F}, \mathcal{X})_X := \sup_{f \in \mathcal{F}} R(\mathcal{F}_{\text{data}(f)})_X \quad \text{and} \quad R_{\tilde{m}}^*(\mathcal{F})_X := \inf_{\mathcal{X} \subset \bar{\Omega}, |\mathcal{X}| = \tilde{m}} R^*(\mathcal{F}, \mathcal{X})_X, \quad \tilde{m} \geq 1,$$

where the latter is the minimum error we can achieve over all the possible choices of data sites  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_{\tilde{m}}\}$ . We can similarly ask the question of how well can we recover  $g$  from the data  $\mathbf{g}$ . 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  $\mathcal{B} = B_q^s(L_p(\Omega))$  with  $0 \leq q, p \leq \infty$  and  $s > \frac{d}{p}$ . Recall, see (2.15), the tensor product grid  $G_{k,r}$  of  $\bar{\Omega} = [0, 1]^d$  for  $r \geq 2$ , with spacing  $h = 2^{-k}(r-1)^{-1}$  (the role of  $r$  will be made clear below). We will prove that, see Theorem 3.1,

$$R^*(\mathcal{F}, G_{k,r})_X \lesssim \tilde{m}^{-\alpha_X}, \quad \tilde{m} = (r2^k)^d.$$

Moreover, we show that for any choice of data sites  $\mathcal{X} \subset \bar{\Omega}$  with  $|\mathcal{X}| = \tilde{m}$ , we have

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

and therefore  $R_{\tilde{m}}^*(\mathcal{F})_X \asymp \tilde{m}^{-\alpha_X}$ . The proof of this theorem is presented in the Appendix, §12.6.1.

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

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

$$R_m^*(\mathcal{F})_X \asymp \tilde{m}^{-\alpha_C}, \quad \tilde{m} \geq 1, \quad \text{where} \quad \alpha_C := \frac{s}{d} - \frac{1}{p}, \quad (3.8)$$

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

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

$$R_m^*(\mathcal{F})_X \asymp \tilde{m}^{-\alpha_\tau}, \quad \tilde{m} \geq 1, \quad \text{where} \quad \alpha_\tau := \frac{s}{d} - \left[ \frac{1}{p} - \frac{1}{\tau} \right]_+, \quad (3.9)$$

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

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

$$R_m^*(\mathcal{F})_X \asymp \tilde{m}^{-\alpha_H}, \quad \tilde{m} \geq 1, \quad \text{where} \quad \alpha_H := \frac{s-1}{d} - \left[ \frac{1}{p} - \frac{1}{2} \right], \quad (3.10)$$

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

(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

$$R_m^*(\mathcal{F})_X \asymp \tilde{m}^{-\alpha_{-1}}, \quad \tilde{m} \geq 1 \quad \text{where} \quad \alpha_{-1} := \frac{s}{d} - \left[ \frac{1}{p} - \frac{1}{\delta} \right]_+ \quad \text{and} \quad \frac{1}{\delta} := \frac{1}{2} + \frac{1}{d}, \quad (3.11)$$

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

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

$$\tilde{m}^{-\alpha_{-1}} \lesssim R_m^*(\mathcal{F})_X \lesssim \log(\tilde{m}) \tilde{m}^{-\alpha_{-1}}, \quad \tilde{m} \geq 1, \quad (3.12)$$

with  $\alpha_{-1}$  and  $\delta$  as in (3.11).

Moreover, when  $\tilde{m} = (r2^k)^d$ , 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 extension of the piecewise polynomial

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

described in (2.18), where  $\mathcal{T}_k = \mathcal{T}_k(\Omega) = \bigcup T$  is the collection of all simplices arising from the (Kuhn-Tucker) decomposition of the dyadic cubes in  $\mathcal{D}_k(\Omega)$ , the  $\chi_T$ 's are the characteristic functions of  $T$ , and the  $L_T(f)$ 's are the polynomials of order  $r$  ( $L_T(f) \in \mathcal{P}_r^d$ ) gotten by interpolating  $f$  at the data points  $\bar{T} \cap G_{k,r}$ .

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  $f \in \mathcal{F} = U(B_q^s(L_p(\Omega)))$  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

$$\|f\|_{H^{-1}(\Omega)} \leq \begin{cases} C(d) \|f\|_{L_\delta(\Omega)}, & d \geq 3, \quad \text{where } \delta = \frac{2d}{d+2}, \\ \frac{C}{\tau-1} \|f\|_{L_\tau(\Omega)}, & d = 2, \quad \tau > 1, \end{cases} \quad (3.13)$$

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  $\bar{\mathcal{B}} := B_{\bar{q}}^{\bar{s}}(L_{\bar{p}}(\Omega))$ , where we fix  $\bar{s}, \bar{p}, \bar{q}$  and assume that  $\bar{s} > d/\bar{p}$  so that we are sure that the functions in  $\bar{\mathcal{B}}$  are continuous on  $\bar{\Omega}$ . 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  $\bar{\mathcal{B}}$  compactly embeds into  $H^1(\Omega)$ .

We define the model class

$$\mathcal{G} := \{g : g = Tr(v) : \|v\|_{\bar{\mathcal{B}}} \leq 1\}, \quad (3.14)$$

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  $\bar{\Omega} = [0, 1]^d$ , see (2.15). We let

$$\bar{G}_{k,r} := \{\mathbf{z}_i : i = 1, \dots, \bar{m}\} = G_{k,r} \cap \partial\Omega \quad (3.15)$$

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

$$\bar{m} = \#(\bar{G}_{k,r}) = 2d[r2^k]^{d-1} \asymp 2^{k(d-1)}, \quad (3.16)$$

with constants of equivalency depending only on  $r$  and  $d$ .

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

$$\bar{S}_k(g) := \sum_{i=1}^{\bar{m}} g(\mathbf{z}_i) \bar{\phi}_i, \quad (3.17)$$

where each  $\bar{\phi}_i$  is the trace of the Lagrange element  $\phi_i$  centered at  $\mathbf{z}_i \in \bar{G}_{k,r}$ . Note that if  $v \in \bar{\mathcal{B}}$  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 (2.18), then we have

$$Tr(S_k^*(v)) = \bar{S}_k(g). \quad (3.18)$$

This follows from the fact that each Lagrange interpolant  $\phi_i$  centered at  $\mathbf{x}_i \in G_{k,r} \setminus \bar{G}_{k,r}$  vanishes at the faces of the simplex from the corresponding Kuhn-Tucker decomposition that contains  $\mathbf{x}_i$ . Note that  $\bar{S}_k(g)$  does not depend on  $v$ . According to Theorem 2.2,  $S_k^*(v) \in H^1(\Omega)$  since  $v$  belongs to unit ball of  $\bar{\mathcal{B}}$ . Then, (3.18) gives that  $S_k(g) \in H^{1/2}(\partial\Omega)$ .

The proof of the following result is provided in the Appendix, §12.6.2.

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

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

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

To understand the exponent  $\beta$  in (3.19) better, let  $v$  be a function in  $U(B_{\bar{q}}^{\bar{s}}(L_{\bar{p}}(\Omega)))$  whose trace is  $g$ , and rewrite  $\beta$  as

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

Here 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

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

under a given total budget  $m = \tilde{m} + \bar{m}$  of data observations. Recall that  $\mathcal{F} := U(B_q^s(L_p(\Omega)))$ , where  $s > d/p$  and  $0 < q, p \leq \infty$ ,  $\mathcal{G} := Tr(U(B_{\bar{q}}^{\bar{s}}(L_{\bar{p}}(\Omega))))$ , where  $\bar{s} > \frac{d}{\bar{p}} \geq 1$  since  $d \geq 2$  and  $0 < \bar{p} \leq 2$ ,  $0 < \bar{q} \leq \infty$ , and

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

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  $d \geq 3$ , or  $d = 2$  and  $p > 1$ , we have

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

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

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

where

$$\alpha = \frac{s}{d} - \left[ \frac{1}{p} - \left( \frac{1}{2} + \frac{1}{d} \right) \right]_+, \quad \beta = \frac{\bar{s} - 1 - (\frac{d}{\bar{p}} - \frac{d}{2})}{d - 1}, \quad (3.23)$$

and all constants of equivalence are independent of  $m$ .

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

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

and therefore

$$R^*(\mathcal{U}, \mathcal{X}, \mathcal{Z})_{H^1(\Omega)} \asymp R^*(\mathcal{F}, \mathcal{X})_{H^{-1}(\Omega)} + R^*(\mathcal{G}, \mathcal{Z})_{H^{1/2}(\Omega)}. \quad (3.24)$$

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

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

On the other hand, using the same theorems and (3.24), we obtain

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

and the proof is completed.  $\square$

**Remark 3.4.** Consider  $\Omega = (0, 1)^d$  and  $\mathcal{F}, \mathcal{G}, \mathcal{U}$  as above. Let us assume that we have a fixed budget of  $\tilde{m}$  data sites  $\mathcal{X}$  from  $\overline{\Omega}$  and a fixed number  $\overline{m}$  of data sites  $\mathcal{Z}$  on the boundary  $\partial\Omega$  and denote by

$$R_{\tilde{m}, \overline{m}}^*(\mathcal{U})_{H^1(\Omega)} := \inf_{\mathcal{X} \subset \overline{\Omega}, \mathcal{Z} \subset \partial\Omega, |\mathcal{X}|=\tilde{m}, |\mathcal{Z}|=\overline{m}} R^*(\mathcal{U}, \mathcal{X}, \mathcal{Z})_{H^1(\Omega)}.$$

It follows from (3.24) that

$$R_{\tilde{m}}^*(\mathcal{F})_{H^{-1}(\Omega)} + R_{\overline{m}}^*(\mathcal{G})_{H^{1/2}(\partial\Omega)} \asymp R_{\tilde{m}, \overline{m}}^*(\mathcal{U})_{H^1(\Omega)}. \quad (3.25)$$

Theorem 3.1 and Theorem 3.2 then yield

$$\tilde{m}^{-\alpha} + \overline{m}^{-\beta} \lesssim R_{\tilde{m}, \overline{m}}^*(\mathcal{U})_{H^1(\Omega)} \lesssim \begin{cases} \log(\tilde{m})\tilde{m}^{-\alpha} + \overline{m}^{-\beta}, & d = 2, 0 < p \leq 1, \\ \tilde{m}^{-\alpha} + \overline{m}^{-\beta}, & d = 2, 1 < p \leq \infty \quad \text{or} \quad d \geq 3, \end{cases} \quad (3.26)$$

where  $\alpha, \beta$  are as in (3.23).

Now suppose instead that we are given a fixed total budget  $m$  and  $d \geq 2$ . Then we can obtain the optimal rate  $R_m^*(\mathcal{U})_{H^1(\Omega)}$  by choosing  $\tilde{m} \asymp \overline{m} \asymp m/2$ .

### 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  $\tilde{m}, \overline{m}, m$ , 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, 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 superfluous 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  $\mathcal{B} = B_q^s(L_p(\Omega))$  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  $U(B_\infty^\alpha(L_\delta(\Omega)))$  and, in addition,  $\mathcal{F}$  embeds into  $C(\overline{\Omega})$ .

#### Largest Model Classes for $f$ :

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

$$\mathcal{F} := U(\mathcal{B}), \quad \mathcal{B} = B_\infty^s(L_p(\Omega)), \quad \text{with } p \geq \delta \text{ and } s > \frac{d}{p} \quad (3.27)$$

has the optimal recovery rate

$$R_m^*(\mathcal{F})_{H^{-1}(\Omega)} \asymp m^{-s/d}, \quad m \geq 1, \quad (3.28)$$

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

$$\mathcal{F} := U(\mathcal{B}), \quad \mathcal{B} = B_\infty^s(L_1(\Omega)), \quad \text{with } s > 2, \quad (3.29)$$

has recovery rate

$$m^{-s/2} \lesssim R_m^*(\mathcal{F})_{H^{-1}(\Omega)} \lesssim (1 + \log(m))m^{-s/2}, \quad m \geq 1. \quad (3.30)$$

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

$$\mathcal{G} := Tr(U(\overline{\mathcal{B}})), \quad \overline{\mathcal{B}} = B_\infty^{\bar{s}}(L_2(\Omega)), \quad \text{with } \bar{s} > d/2. \quad (3.31)$$



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

$$R_m^*(\mathcal{G})_{H^{1/2}(\partial\Omega)} \asymp m^{-(\bar{s}-1)/(d-1)}, \quad m \geq 1. \quad (3.32)$$

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  $g = \text{Tr}(v)$  in  $H^{1/2}(\Omega)$ .

**Remark 3.5.** *Because of the above remarks, in going further, we always take  $\mathcal{F}$  as in (3.27) and  $\mathcal{G}$  as in (3.31).*

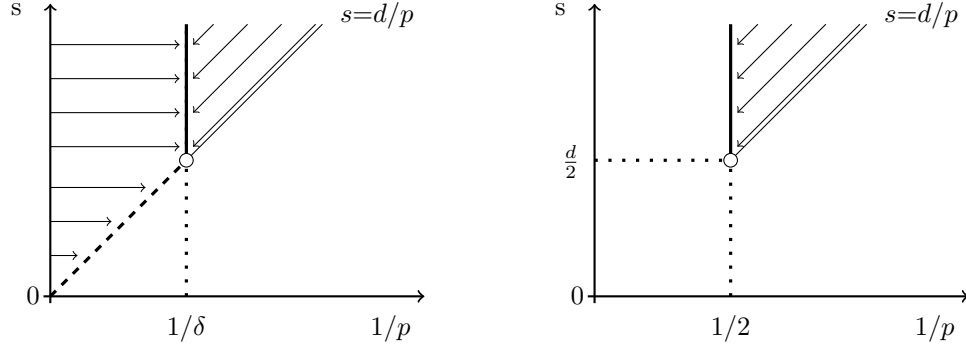


Figure 3.1: The spaces  $B_\infty^s(L_p(\Omega))$  are represented by the point  $(1/p, s)$  in the first quadrant. The arrows indicate embedding with a slope of  $d$  for the oblique lines. In particular, the spaces above the line  $s = d/p$  embed into  $C(\bar{\Omega})$ . (Left) The largest model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

**Model Classes for  $u$ :** If we use one of the model classes from (3.27) for  $\mathcal{F}$  and one of the model classes from (3.31) 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  $\tilde{m} \asymp m/2$  points in  $\bar{\Omega}$  and  $\bar{m} \asymp m/2$  points on the boundary  $\partial\Omega$ .

## 4 Numerical algorithms based on linear approximation

Recall that we consider the case when  $\Omega = (0, 1)^d$ ,  $d \geq 2$ ,  $\mathcal{F} := U(\mathcal{B})$ , where  $\mathcal{B} = B_q^s(L_p(\Omega))$  with  $0 \leq q, p \leq \infty$  and  $s > \frac{d}{p}$ , and  $\mathcal{G} = \{\text{Tr}(v) : \|v\|_{\bar{\mathcal{B}}} \leq 1\}$ , where  $\bar{\mathcal{B}} = B_{\bar{q}}^{\bar{s}}(L_{\bar{p}}(\Omega))$  with  $\bar{s} > d/\bar{p}$ ,  $0 < \bar{p} \leq 2$ ,  $0 < \bar{q} \leq \infty$ . For these right hand sides and boundary values, we consider the solution set

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

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  $u, u' \in \mathcal{U}$  and satisfy the same data, namely  $u, u' \in \mathcal{U}_{\text{data}}$ , then they are close in the  $H^1(\Omega)$  norm. More precisely, if  $u^*$  is the center of a Chebyshev ball  $B(\mathcal{U}_{\text{data}})$ , then

$$\|u - u'\|_{H^1(\Omega)} \leq \|u - u^*\|_{H^1(\Omega)} + \|u^* - u'\|_{H^1(\Omega)} \leq 2R_m^*(\mathcal{U})_{H^1(\Omega)}. \quad (4.1)$$

However, our analysis does not give a numerical algorithm that takes the data and creates an approximation  $\hat{u}$  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 §8.

We assume in this section that the data sites form tensor product grids, that is,  $\mathcal{X} = G_{k,r}$  and  $\mathcal{Z} = \bar{G}_{\bar{k},r} := G_{\bar{k},r} \cap \partial\Omega$ . We know such data sites provide a near optimal recovery. We have shown that if we use the data  $\mathbf{f}$  at  $G_{k,r}$ , then we can numerically construct a continuous piecewise polynomial interpolant  $\hat{f}$  to  $f$  at the data sites which is a near optimal (in terms of the budget  $\tilde{m} = (r2^k)^d$ ) approximation to  $f$  in the  $H^{-1}(\Omega)$  norm. Similarly, we can numerically construct a  $\hat{g}$  from the data  $\mathbf{g}$  which is a near optimal recovery of  $g$  in the  $H^{1/2}(\partial\Omega)$  norm in terms of the budget  $\bar{m} \asymp 2^{k(d-1)}$ . The following lemma holds.

**Lemma 4.1.** *Let  $\tilde{u}$  be the solution to (1.1) with right hand side  $\hat{f} := S_k^*(f)$  and boundary value  $\hat{g} := \bar{S}_k(g)$ . Then*

$$\|u - \tilde{u}\|_{H^1(\Omega)} \lesssim R_{\tilde{m}, \bar{m}}^*(\mathcal{U})_{H^1(\Omega)}, \quad \text{where } \tilde{m} \asymp 2^{kd}, \quad \bar{m} \asymp 2^{k(d-1)}. \quad (4.2)$$

*If instead, we are given a fixed total budget  $m$ , then*

$$\|u - \tilde{u}\|_{H^1(\Omega)} \lesssim R_m^*(\mathcal{U})_{H^1(\Omega)}, \quad \text{provided } \tilde{m} \asymp \bar{m} \asymp m/2. \quad (4.3)$$

*Proof.* We have that  $\hat{f} := S_k^*(f) \in C(\bar{\Omega}) \cap H^1(\Omega)$ , see Theorem 3.1, and  $\hat{g} := \bar{S}_k(g) \in H^{1/2}(\partial\Omega) \cap C(\partial\Omega)$ , see Theorem 3.2. Note that we cannot use (4.1) to bound  $\|u - \tilde{u}\|_{H^1(\Omega)}$  since  $\hat{f} \notin \mathcal{F}$  and  $\hat{g} \notin \mathcal{G}$  (they do not have the necessary smoothness) and therefore  $\tilde{u} \notin \mathcal{U}$ . In this case, in view of (1.4) with  $v = \tilde{u}$ , we have

$$\|u - \tilde{u}\|_{H^1(\Omega)} \lesssim \|f - \hat{f}\|_{H^{-1}(\Omega)} + \|g - \hat{g}\|_{H^{1/2}(\partial\Omega)} \lesssim R_m^*(\mathcal{F})_{H^{-1}(\Omega)} + R_m^*(\mathcal{G})_{H^{1/2}(\partial\Omega)} \lesssim R_{\tilde{m}, \bar{m}}^*(\mathcal{U})_{H^1(\Omega)},$$

where we have used (3.25). This proves (4.2). In view of Remark 3.4, given  $m$  data sites, the solution  $\tilde{u}$  that corresponds to the choice of  $\tilde{m} \asymp m/2$  points on  $\bar{\Omega}$  and  $\bar{m} \asymp m/2$  points on  $\partial\Omega$  satisfies (4.3) and thus is a near optimal recovery of  $u$  in the  $H^1(\Omega)$  norm.  $\square$

Note that the constructions of  $\hat{f}$  and  $\hat{g}$  are numerical and we can control their complexity. In order to do that for  $\tilde{u}$ , 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

In order to streamline the discussion using FEM, we make the simplifying but unnecessary assumption that the data tensor product grids  $G_{k,r}$  and  $\bar{G}_{\bar{k},r}$  are matching, i.e.  $k = \bar{k}$ . For  $n \geq k$ , we let  $\mathcal{T}_n$  be the Kuhn-Tucker subdivision of  $\Omega$  associated with the tensor grid  $G_{n,r}$  and  $V_n$  be the spaces of continuous piecewise polynomials of degree  $r$  subordinate to  $\mathcal{T}_n$ . We use Lemma 4.1 to derive the following statement about FEMs.

**Lemma 4.2.** *Let  $\tilde{u}$  be the solution to (1.1) with right hand side  $\hat{f} := S_k^*(f)$  and boundary value  $\hat{g} := \bar{S}_k(g)$ . For  $n \geq k$ , let  $\hat{u}_n$  be the Galerkin projection of  $\tilde{u}$  onto  $V_n$ . Then, we have*

$$\|u - \hat{u}_n\|_{H^1(\Omega)} \lesssim R_{\tilde{m}, \bar{m}}^*(\mathcal{U})_{H^1(\Omega)} + \varepsilon_n(\mathcal{U})_{H^1(\Omega)}, \quad \text{where } \varepsilon_n(\mathcal{U})_{H^1(\Omega)} := \text{dist}(\mathcal{U}, V_n)_{H^1(\Omega)}. \quad (4.4)$$

*Proof.* It follows from (4.2) that

$$\|u - \hat{u}_n\|_{H^1(\Omega)} \leq \|u - \tilde{u}\|_{H^1(\Omega)} + \|\tilde{u} - \hat{u}_n\|_{H^1(\Omega)} \lesssim R_{\tilde{m}, \bar{m}}^*(\mathcal{U})_{H^1(\Omega)} + \|\tilde{u} - \hat{u}_n\|_{H^1(\Omega)}. \quad (4.5)$$

Note that since  $n \geq k$ , we have that  $\hat{g} \in \text{Tr}(V_n)$ . In particular,  $\hat{u}_n$  and  $\tilde{u}$  have the same boundary data (and are defined with the same right hand side  $\hat{f}$ ). Consequently,  $\hat{u}_n$  is a quasi-best approximation to  $\tilde{u}$  in  $V_n$ ,

$$\|\tilde{u} - \hat{u}_n\|_{H^1(\Omega)} \lesssim \inf_{v \in V_n} \|\tilde{u} - v\|_{H^1(\Omega)}. \quad (4.6)$$

Invoking (4.2), we find that for every  $v \in V_n$ ,

$$\|\tilde{u} - v\|_{H^1(\Omega)} \leq \|\tilde{u} - u\|_{H^1(\Omega)} + \|u - v\|_{H^1(\Omega)} \lesssim R_{\tilde{m}, \bar{m}}^*(\mathcal{U})_{H^1(\Omega)} + \|u - v\|_{H^1(\Omega)},$$

and hence

$$\inf_{v \in V_n} \|\tilde{u} - v\|_{H^1(\Omega)} \lesssim R_{\tilde{m}, \bar{m}}^*(\mathcal{U})_{H^1(\Omega)} + \inf_{v \in V_n} \|u - v\|_{H^1(\Omega)} \leq R_{\tilde{m}, \bar{m}}^*(\mathcal{U})_{H^1(\Omega)} + \varepsilon_n(\mathcal{U})_{H^1(\Omega)}.$$

This, together with (4.6), gives

$$\|\tilde{u} - \hat{u}_n\|_{H^1(\Omega)} \lesssim R_{\tilde{m}, \bar{m}}^*(\mathcal{U})_{H^1(\Omega)} + \varepsilon_n(\mathcal{U})_{H^1(\Omega)}. \quad (4.7)$$

Combining (4.5) and (4.7) yields the desired estimate (4.4).  $\square$

Let us now discuss the statement of Lemma 4.2. We have determined the optimal recovery rates  $R_{\tilde{m}, \bar{m}}^*(\mathcal{U})_{H^1(\Omega)}$ , see (3.26) for the class  $\mathcal{U}$ . The term  $\varepsilon_n$  will tend to zero as  $n \rightarrow \infty$  and we would obtain the optimal recovery error on the right side of (4.4) if  $n$  is chosen suitably large. The actual decay rate of  $\varepsilon_n$  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. [12, 42, 55] and the references therein) and thereby obtain concrete bounds on the decay rate  $\varepsilon_n$ . 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 for solving (1.1) with right side  $\hat{f}$ , and boundary value  $\hat{g}$ , then the decay rate for  $\varepsilon_n(\mathcal{U})$  can be improved (see [6] 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. In the case when  $\mathcal{X} = G_{k,r}$ ,  $|G_{k,r}| = \tilde{m} = [r2^k]^d$ , see (2.15) and  $\mathcal{Z} = \bar{G}_{\bar{k},r}$ ,  $|\bar{G}_{\bar{k},r}| = \bar{m} = 2d[r2^{\bar{k}}]^{d-1}$ , see (3.15), we know that the simplicial interpolation operator  $S_k^*$  defined in (2.18) provides a near optimal recovery for  $\mathcal{F}$ . Namely,

$$\|f - S_k^*(f)\|_{H^{-1}(\Omega)} \lesssim R_{\tilde{m}}^*(\mathcal{F})_{H^{-1}(\Omega)}, \quad \tilde{m} = [r2^k]^d. \quad (4.8)$$

We write

$$S_k^*(f) = \sum_{i=1}^{\tilde{m}} f(\mathbf{x}_i) \phi_i, \quad (4.9)$$

where the  $\phi_i$  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  $\phi_i^*$  be the solution to (1.1) with right side  $\phi_i$  and zero boundary conditions.

Similarly, the interpolant  $\bar{S}_{\bar{k}}(g)$  to  $g \in \mathcal{G}$  used for optimal recovery, see (3.17), can be written as

$$\bar{S}_{\bar{k}}(g) = \sum_{i=1}^{\bar{m}} g(\mathbf{z}_i) \psi_i, \quad \bar{m} = 2d[r2^{\bar{k}}]^{d-1}. \quad (4.10)$$

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

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

$$\tilde{u} := \sum_{i=1}^{\tilde{m}} f(\mathbf{x}_i) \phi_i^* + \sum_{i=1}^{\bar{m}} g(\mathbf{z}_i) \psi_i^* \quad (4.11)$$

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

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

$$\hat{u} := \sum_{i=1}^{\tilde{m}} f(\mathbf{x}_i) \hat{\phi}_i + \sum_{i=1}^{\bar{m}} g(\mathbf{z}_i) \hat{\psi}_i, \quad \text{with} \quad \tilde{m} \asymp \bar{m} \asymp m/2, \quad (4.12)$$

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  $\bar{s}$  are all less than a fixed number  $s_0$ .

## 5 Numerical methods based on optimization

Another method of finding a near optimal approximation to  $u$  from the given data is through optimization, such as the one used in PINNs. The starting point of this approach is the fact that the unique solution  $u \in H^1(\Omega)$  to (1.1) is the unique minimizer of the theoretical loss  $\mathcal{L}_T(u)$ , see (1.6). However, the minimization in (1.6) is over all  $v \in H^1(\Omega)$ . To make this approach more amenable to numerical implementation, we take this minimization over a set  $\Sigma = \Sigma_n$ , where  $\Sigma$  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

$$S_\Sigma \in \{\operatorname{argmin}_{S \in \Sigma} \mathcal{L}_T(S)\}. \quad (5.1)$$

We want to understand how close such an  $S_\Sigma$  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  $S_\Sigma$  to be a good approximation to  $u$ , then we need  $\Sigma$  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  $S_\Sigma$  from (5.1) and any  $S \in \Sigma$ , we have from (1.4) that

$$\|u - S_\Sigma\|_{H^1(\Omega)} \lesssim \mathcal{L}_T(S_\Sigma) \lesssim \mathcal{L}_T(S) \lesssim \|u - S\|_{H^1(\Omega)}. \quad (5.2)$$

Since (5.2) holds for all  $S \in \Sigma$ , we have that

$$\|u - S_\Sigma\|_{H^1(\Omega)} \lesssim \inf_{S \in \Sigma} \|u - S\|_{H^1(\Omega)} \lesssim E(\mathcal{U}, \Sigma). \quad (5.3)$$

The discussion that we have just given shows that if we solve the continuous minimization problem (1.6) over  $\Sigma$  instead of all of  $H^1(\Omega)$ , then the solution will get as close to  $u$  as the efficiency of  $\Sigma$  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  $\Sigma$ , 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  $\Sigma$  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  $\mathcal{L}_T$  by these discrete norms leads to a loss  $\mathcal{L}^*$  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  $\bar{\mathcal{B}}$ .

As it will be seen in §7, the new loss  $\mathcal{L}^*$  is based on a discretization of the  $\|\cdot\|_{L_\gamma(\Omega)}$  (for a special value of  $\gamma$ ) and  $\|\cdot\|_{H^{1/2}(\partial\Omega)}$  norms. These discretizations have to satisfy the following two requirements:

- use the given information  $(\mathbf{f}, \mathbf{g})$  at the grids (data sites)  $G_{k,r} \subset \bar{\Omega}$  and  $\bar{G}_{\bar{k},r} = G_{\bar{k},r} \cap \partial\Omega$ , respectively.
- the errors incurred for functions in  $\mathcal{F}$  and  $\mathcal{G}$  be comparable with the optimal recovery rates for these classes.

We remark that the discretization of integral norms has been studied in many other papers. For example, in [66] the discretization of the variational form of the Poisson equation for functions in the Barron space (or variation space corresponding to shallow ReLU<sup>k</sup> networks) is considered. This analysis is done for both random data sites and a fixed grid. There are also papers that prove existence of data sites  $\mathcal{X}$  for which the continuous and discrete  $L_\tau$  norms are equivalent for elements of certain linear spaces, see [16, 15]. To the best of our knowledge, there are no results in the literature which treat the discretization of the  $H^{1/2}$ -norm that are comparable to what we do in §6.2.

### 6.1 A discrete $L_\tau$ norm

In this section, we study the discretization of the  $L_\tau$  norm for any  $1 \leq \tau \leq \infty$ . Let  $\mathcal{F} = U(B_q^s(L_p(\Omega)))$ ,  $s > d/p$ , be a model class assumption on  $f$ . We consider the uniform grid  $G_{k,r} \subset \bar{\Omega}$ , see (2.15), consisting of  $\tilde{m}$  points,  $\tilde{m} = |G_{k,r}| = \lceil r2^k \rceil^d$ ,  $k \geq 0$  and  $r > \max(s, 1)$ . For any continuous function  $f$  and any  $1 \leq \tau < \infty$ , we define

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

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

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

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

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

with  $\alpha_\tau := \frac{s}{d} - \left(\frac{1}{p} - \frac{1}{\tau}\right)_+$ . Recall that  $\tilde{m}^{-\alpha_\tau}$  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  $1 \leq \tau < \infty$ . A simple modification of the proof handles the case  $\tau = \infty$ . We use the simplicial interpolation operator  $S_k^*$  given in (2.18). Let  $\mathcal{T}_k := \mathcal{T}_k(\Omega) = \bigcup T$  be the set of simplices that make up the simplicial decomposition of the cubes in  $\mathcal{D}_k(\bar{\Omega})$ . Then, from (2.18) we have

$$\|S_k^*(f)\|_{L_\tau(\Omega)}^\tau = \sum_{T \in \mathcal{T}_k} \|L_T(f)\|_{L_\tau(T)}^\tau, \quad (6.2)$$

where  $L_T(f)$  is the polynomial in  $\mathcal{P}_r^d$  that interpolates the data  $\{f(\mathbf{x}_i) : \mathbf{x}_i \in \bar{T}\}$ . By equivalence of norms on  $\mathcal{P}_r^d$  (the linear space of algebraic polynomials of  $d$  variables and order  $r$  (total degree  $r - 1$ )), rescaling arguments, and the fact that we have  $d!$  simplices in the Kuhn-Tucker decomposition of a cube,

$$\|L_T(f)\|_{L_\tau(T)} \asymp \left[ \frac{1}{\tilde{m}} \sum_{\mathbf{x}_i \in \bar{T}} |f(\mathbf{x}_i)|^\tau \right]^{1/\tau}, \quad (6.3)$$

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

$$\|S_k^*(f)\|_{L_\tau(\Omega)} \asymp \left[ \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} |f(\mathbf{x}_i)|^\tau \right]^{1/\tau} = \|f\|_{L_\tau(\Omega)}^*. \quad (6.4)$$

For  $f \in \mathcal{B} = B_q^s(L_p(\Omega))$ , the piecewise polynomial interpolant  $S_k^*(f)$  provides a near optimal recovery of  $f$  from the data  $\mathbf{f}$ , that is, see Theorem 2.1,

$$\|f - S_k^*(f)\|_{L_\tau(\Omega)} \leq \begin{cases} C|f|_{B_p^s(\Omega)} 2^{-k(s-d/p+d/\tau)} = C\|f\|_{\mathcal{B}} \tilde{m}^{-\alpha_\tau}, & \tau \geq p, \\ \|f - S_k^*(f)\|_{L_p(\Omega)} \leq C|f|_{B_p^s(\Omega)} 2^{-ks} = C\|f\|_{\mathcal{B}} \tilde{m}^{-\alpha_\tau}, & \tau < p, \end{cases} \quad (6.5)$$

where we have used that  $\tilde{m} = [r2^k]^d$ . It follows then from (6.4) that

$$\|f\|_{L_\tau(\Omega)} - \|S_k^*(f)\|_{L_\tau(\Omega)} \leq C\|f\|_{\mathcal{B}} \tilde{m}^{-\alpha_\tau}, \quad (6.6)$$

and the proof is completed.  $\square$

## 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 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  $\mathcal{Z} := \bar{G}_{k,r} = \{\mathbf{z}_j\}_{j=1}^{\bar{m}} = G_{k,r} \cap \partial\Omega$ , see (3.15), be the grid points on the boundary  $\partial\Omega$ . The number of grid points in  $\mathcal{Z}$  is  $\bar{m} = 2d[r2^k]^{d-1}$ .

We fix a model class  $\mathcal{G} = Tr(U(\bar{\mathcal{B}}))$  of the form (3.31). The optimal recovery rate for this class is

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

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  $v \in \bar{B}$ , we use the trace norm

$$\|g\|_{Tr(\bar{B})} := \inf_{Tr(v)=g} \|v\|_{\bar{B}}. \quad (6.8)$$

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

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

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  $\Omega$  (see [34]). We shall use (6.9) through out this section.

Let  $\bar{S}_k$  be as defined in (3.17). We have proven in Theorem 3.2 that

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

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

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

We concentrate therefore on finding a discrete  $H^{1/2}(\partial\Omega)$  norm for the functions  $\bar{S}_k(g)$ ,  $g \in \mathcal{G}$ . Note that  $\bar{S}_k(g)$  is completely determined by the values of  $g$  at the data sites  $\bar{G}_{k,r}$ . For each  $k \geq 0$ , we define  $\mathcal{D}_k(\partial\Omega) := \partial\Omega \cap \mathcal{D}_k(\bar{\Omega})$  to be the partition of  $\partial\Omega$  into dyadic cubes  $I$  of side length  $2^{-k}$ . Given  $k \geq 0$ , let  $\mathcal{T}_k := \mathcal{T}_k(\partial\Omega)$  denote the collection of all  $d-1$  dimensional simplices 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  $\bar{S}_k$  is a projector onto  $V^r(\mathcal{T}_k)$ .

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

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

and prove that 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*

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

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  $\bar{T} \cup \bar{T}'$ . A function  $S \in X$  satisfies  $\mathcal{N}(S) = 0$  if and only if  $S$  is constant on  $\bar{T}$ ,  $\bar{T}'$  and, additionally,  $S$  is constant on  $\bar{T} \cup \bar{T}'$  when  $\bar{T}$  and  $\bar{T}'$  touch. A similar statement holds for the semi-norm  $\mathcal{N}'$ . It follows that for any pair  $T, T'$ , we have

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

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 simplices 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:**  $\bar{T} \cap \bar{T}' \neq \emptyset$ .

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:**  $\bar{T} \cap \bar{T}' = \emptyset$ . 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  $S \in V^r(\mathcal{T}_k(\partial\Omega))$

$$c_2 \left[ \int_{T \times T'} |S(\mathbf{z}) - S(\mathbf{z}')|^2 d\mathbf{z} d\mathbf{z}' \right]^{1/2} \leq \left[ \frac{1}{\bar{m}^2} \sum_{\substack{\mathbf{z} \neq \mathbf{z}' \in \mathcal{Z} \\ \mathbf{z} \in \bar{T}, \mathbf{z}' \in \bar{T}'}} |S(\mathbf{z}) - S(\mathbf{z}')|^2 \right]^{1/2} \leq C_2 \left[ \int_{T \times T'} |S(\mathbf{z}) - S(\mathbf{z}')|^2 d\mathbf{z} d\mathbf{z}' \right]^{1/2}. \quad (6.14)$$

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  $r, d$ .  $\square$

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

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

with

$$\beta = \frac{\bar{s} - 1}{d - 1}, \quad (6.16)$$

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

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

and

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

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

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

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

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  $\bar{m}$  data sites  $\mathcal{Z} := \bar{G}_{k,r}$ , we define

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



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

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

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

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

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

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

where  $\beta$  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  $\alpha > \beta$ .  $\square$

## 7 A discrete loss function with error control

In section §4, we have introduced several numerical methods for finding a near optimal recovery to the solution  $u \in \mathcal{U}_{\text{data}}$  from the given data. In this section, we analyse an alternative approach that is based on optimization (such as PINNs). We consider a discrete loss function  $\mathcal{L}^*$  which is a surrogate for the theoretical loss function  $\mathcal{L}_T$ , see (1.5). 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  $\mathcal{X} = G_{k,r} \subset \bar{\Omega}$  and  $\mathcal{Z} = \bar{G}_{\bar{k},r}$  be the data sites of  $\partial\Omega$ . Note that  $\bar{k}$  could be different from  $k$ . This fixes  $\tilde{m} = \#(G_{k,r})$  and  $\bar{m} := \#(\bar{G}_{\bar{k},r})$  for the remainder of this section.

**Remark 7.1.** *In this and the following two sections, the results and their proofs are simplest in the case  $d \geq 3$ . When  $d = 2$ , log factors appear. For this reason, we state the results in correct form (including the log factors) for all cases  $d \geq 2$ , but give the proofs only for  $d \geq 3$ . 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  $\gamma$ ,

$$\gamma = \begin{cases} \frac{2d}{d+2}, & d \geq 3, \\ 1 + [\log(\tilde{m})]^{-1}, & d = 2, \end{cases} \quad (7.1)$$

and let  $\|\cdot\|_{H^{1/2}}^*$  be the discrete  $H^{1/2}(\partial\Omega)$  introduced in §6.2, see (6.11), (6.18), and (6.20). Given the data vectors  $(\mathbf{f}, \mathbf{g})$ , we introduce the discrete loss function

$$\mathcal{L}^*(v) := \begin{cases} \|f + \Delta v\|_{L_\gamma(\Omega)}^* + \|g - v\|_{H^{1/2}(\partial\Omega)}^*, & d \geq 3, \\ [1 + \log(\tilde{m})] \|f + \Delta v\|_{L_\gamma(\Omega)}^* + \|g - v\|_{H^{1/2}(\partial\Omega)}^*, & d = 2, \end{cases} \quad (7.2)$$

which is defined whenever  $v$  and  $\Delta v$  are continuous on  $\bar{\Omega}$ .

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 [5], 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 (3.27) and  $\mathcal{G} = Tr(U(\bar{\mathcal{B}}))$  of (3.31).

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 (3.27), the uniform optimal recovery rate for  $\mathcal{F}$  in  $H^{-1}(\Omega)$  is  $\asymp \tilde{m}^{-s/d}$ ,  $\tilde{m} \geq 1$  where  $s$  is the excess regularity of  $\mathcal{F}$  in  $L_\delta(\Omega)$ . For a model class  $\mathcal{G}$  of the form (3.31), its uniform optimal recovery rate in  $H^{1/2}(\partial\Omega)$  is  $\asymp \bar{m}^{-(\bar{s}-1)/(d-1)}$ , 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  $\max\{\tilde{m}^{-\frac{s}{d}}, \bar{m}^{-(\bar{s}-1)/(d-1)}\}$ . We introduce the following notation for a function  $v$

$$\|v\|_{\mathcal{U}} := \max\{\|\Delta v\|_{\mathcal{B}}, \|Tr(v)\|_{Tr(\bar{\mathcal{B}})}\}. \quad (7.3)$$

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  $v \in \bar{\mathcal{B}}$ .

**Theorem 7.2.** *Let  $u$  be the solution to (1.1) with  $f \in \mathcal{F} = U(\mathcal{B})$  of (3.27) and  $g \in \mathcal{G} = Tr(U(\bar{\mathcal{B}}))$  of (3.31). Given the data  $(\mathbf{f}, \mathbf{g})$  of  $f$  and  $g$  at grid data sites  $(G_{k,r}, \bar{G}_{k,r})$  with  $|G_{k,r}| = \tilde{m}$  and  $|\bar{G}_{k,r}| = \bar{m}$ , consider the functional  $\mathcal{L}^*$ , defined in (7.2). If  $v$  is any continuous function in  $H^1(\Omega)$ , then*

$$\|u - v\|_{H^1(\Omega)} \lesssim \mathcal{L}^*(v) + [1 + \|v\|_{\mathcal{U}}] \mathcal{R}_{\mathcal{U}}(\tilde{m}, \bar{m}), \quad (7.4)$$

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

$$\mathcal{R}_{\mathcal{U}}(\tilde{m}, \bar{m}) := \begin{cases} \max\{\tilde{m}^{-\frac{s}{d}}, \bar{m}^{-\frac{\bar{s}-1}{d-1}}\}, & d \geq 3, \\ \max\{\log(\tilde{m})\tilde{m}^{-\frac{s}{d}}, \bar{m}^{-(\bar{s}-1)}\}, & d = 2. \end{cases} \quad (7.5)$$

*Proof.* We only prove this for  $d \geq 3$ . From (1.4), we have

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

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  $\|g\|_{Tr(\bar{\mathcal{B}})} \leq 1$ , we complete the proof of the theorem.  $\square$

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 loss  $\mathcal{L}^*$  so that it is independent of  $\tilde{m}$ . This choice would lead to the actual optimal recovery rates  $\max\{\tilde{m}^{-\frac{s}{2}}, \bar{m}^{-(\bar{s}-1)}\}$  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. This means that when we try to obtain a good approximation to  $u$  (in  $H^1(\Omega)$ ) through optimization, then either the norm  $\|\cdot\|_{\mathcal{U}}$  will have to participate in the loss function as a regularization term, see [5], or the approximant has to come from a set with good restricted approximation properties, see Theorem 9.1 and Remark 9.2. In both of these situations there does not appear to exist an optimization procedure with a priori guarantees when using deep neural networks, although using shallow networks a priori bounds can be achieved using a greedy algorithm [66]. In practice, however, once an approximant  $v$  is obtained (say, as a result of a minimization of a loss that does not involve  $\|v\|_{\mathcal{U}}$ ), we can in principle use (7.4) to compute an upper bound for the quality of approximation  $\|u - v\|_{H^1(\Omega)}$ , provided that we can compute an (a posteriori) estimate on  $\|v\|_{\mathcal{U}}$ . In fact, similar methods for certifying PINNs solutions have been studied before (see for instance [29] and the references therein).*

**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  $\Delta S$ . If at any stage of the optimization procedure we have an  $S \in \Sigma_n$ , then we can use (7.4) 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  $\tilde{m}$  and  $\bar{m}$ , the losses  $\mathcal{L}$  and  $\mathcal{L}^*$  are equivalent up to a constant depending upon the parameters  $\tilde{m}$  and  $\bar{m}$ . 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 independent on  $m$  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  $\Sigma_n$  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  $\Sigma_n$  in the  $H^1(\Omega)$  but rather by a form of restricted approximation which involves  $\|\cdot\|_{\mathcal{U}}$  [5].

## 8 Implications for Loss Function Design

In this section, we discuss the implications of Theorem 7.2 on the design of loss functions for the solution of elliptic PDEs using PINNs. Theorem 7.2 implies that the loss function  $\mathcal{L}^*$  given in (7.2) gives an a posteriori control on the error up to the optimal recovery rate. However, there are numerous factors that need to be taken into account when designing an appropriate loss function, including the ease of training and implementation. For this reason, directly optimizing the loss  $\mathcal{L}^*$  may not be the best choice in practice.

We observe that any loss function  $\mathcal{L}$  which upper bounds  $\mathcal{L}^*$  also gives a posteriori error control via Theorem 7.2. With this in mind, we consider the following variants of PINNs for elliptic PDEs. We consider the case  $d \geq 3$ . Similar discussion holds when  $d = 2$ , taking into account the form of  $\mathcal{L}^*$  in (7.2):

- The original PINNs formulation with an adjustable weight  $\lambda > 0$  on the boundary terms,

$$\mathcal{L}_{sq,\lambda}(v) := \left[ \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} [\Delta v(\mathbf{x}_i) + f(\mathbf{x}_i)]^2 \right] + \left[ \frac{\lambda}{\bar{m}} \sum_{i=1}^{\bar{m}} [v(\mathbf{z}_i) - g(\mathbf{z}_i)]^2 \right]. \quad (8.1)$$

Note that the case  $\lambda = 1$  corresponds to the original formulation of PINNs [59] and is equivalent to the (square of the) loss  $\mathcal{L}$ , see (1.9). Numerous heuristic strategies for choosing the weight  $\lambda$  have been explored in a variety of works, see [18, 71, 72].

- The loss function  $\mathcal{L}^*$  in (7.2) with exponent  $\tau \geq 1$

$$\begin{aligned} \mathcal{L}_{sq,\tau}^*(v) := & \left[ \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} |f(\mathbf{x}_i) + \Delta v(\mathbf{x}_i)|^\tau \right]^{2/\tau} + \left[ \frac{1}{\bar{m}^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{\bar{m}} \frac{|[g-v](\mathbf{z}_i) - [g-v](\mathbf{z}_j)|^2}{|\mathbf{z}_i - \mathbf{z}_j|^2} \right] \\ & + \left[ \frac{1}{\bar{m}} \sum_{j=1}^{\bar{m}} |g(\mathbf{z}_j) - v(\mathbf{z}_j)|^2 \right]. \end{aligned} \quad (8.2)$$

When  $\tau = \gamma$  given in (7.1), this is equivalent to the (square of the) loss  $\mathcal{L}^*$ .

Note that all of these loss functions are squared, which is typically done to allow for easier optimization (the notation  $sq$  here is a shorthand for squared).

We first observe that the  $L_\gamma$ -norm is bounded by the  $L_\tau$ -norm for any  $\tau \geq \gamma$  on a finite measure space, such as the bounded domain  $\Omega$ , or the normalized empirical measure at the data points  $\mathcal{X}$ . This means that

$$(\mathcal{L}^*)^2 \lesssim \mathcal{L}_{sq,\tau}^* \quad \text{for any } \tau \geq \gamma, \quad (8.3)$$

and thus Theorem 7.2 gives an a posteriori bound on the solution error in terms of  $\mathcal{L}_{sq,\tau}^*$ . This means that we can switch to a different norm  $L_\tau$  with  $\tau \geq \gamma$  on the domain term in the loss  $\mathcal{L}^*$  and retain an a posteriori error guarantee.

The original PINNs loss  $\mathcal{L}_{sq,\lambda}$  with fixed, independent on  $\bar{m}$ ,  $\lambda$  is not an upper bound (with constant independent of  $\bar{m}$  or  $\tilde{m}$ ) on the loss  $(\mathcal{L}^*)^2$ , and thus does not give an a posteriori error estimate via Theorem 7.2. This is due to the fact that the discrete  $L_2$ -norm is not an upper bound for the discrete  $H^{1/2}$ -norm on the boundary used in  $\mathcal{L}^*$ . However, by letting  $\lambda$  depend upon the number  $\bar{m}$  of the boundary collocation points, one can correct this. Indeed, it follows from (8.3) with  $\tau = 2 > \gamma = \frac{2d}{d+2}$  that

$$(\mathcal{L}^*)^2 \lesssim \mathcal{L}_{sq,2}^* \lesssim \mathcal{L}_{sq,\lambda(\bar{m})}, \quad \text{where } \lambda(\bar{m}) := \bar{m}^{1/(d-1)},$$

and the constants in  $\lesssim$  are independent of  $\bar{m}$ . The last inequality is easily proved by observing that

$$\frac{1}{\bar{m}^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{\bar{m}} \frac{|u(\mathbf{z}_i) - u(\mathbf{z}_j)|^2}{|\mathbf{z}_i - \mathbf{z}_j|^d} \leq \frac{2}{\bar{m}^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{\bar{m}} \frac{|u(\mathbf{z}_i)|^2 + |u(\mathbf{z}_j)|^2}{|\mathbf{z}_i - \mathbf{z}_j|^d} \lesssim \bar{m}^{-\frac{d-2}{d-1}} \sum_{i=1}^{\bar{m}} |u(\mathbf{z}_i)|^2. \quad (8.4)$$

**Remark 8.1.** The value  $(\mathcal{L}_{sq,\lambda(\bar{m})}(v))^{1/2}$  can be used in Theorem 7.2 instead of  $\mathcal{L}^*(v)$ . This indicates how to **correctly weight** the boundary and domain terms depending on the number of collocation points in the original PINNs formulation, i.e. to use the loss

$$\mathcal{L}_{sq,\lambda(\bar{m})}(v) = \left[ \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} [\Delta v(\mathbf{x}_i) + f(\mathbf{x}_i)]^2 \right] + \left[ \bar{m}^{(2-d)/(d-1)} \sum_{i=1}^{\bar{m}} [v(\mathbf{z}_i) - g(\mathbf{z}_i)]^2 \right]. \quad (8.5)$$

All of these alternative loss functions provide an upper bound for the loss  $\mathcal{L}_{sq,\gamma}^*$ . Consequently, if the (restricted) optimization problem were solved exactly, the loss  $\mathcal{L}_{sq,\gamma}^*$  would reach a lower value than these alternatives and would give the best a posteriori error bound up to OR rates. However, in practice, the optimization of neural networks introduces numerous complications, and it is not a priori clear which loss function is best. In §10, we present numerical experiments comparing the practical performance of these loss functions.

Finally, we remark that our analysis of loss functions is performed with respect to the  $H^1$ -error. This same analysis can be performed for different error norms as well (see for instance [78] for the  $H^{1/2}$ -norm), but we leave this for future work.

## 9 A numerical optimal recovery algorithm

Let  $\mathcal{U}$  be the model class (3.3), determined by  $\mathcal{F}$  and  $\mathcal{G}$ , where  $\mathcal{F}$  is a maximal model class (3.27), (3.29) and  $\mathcal{G}$  is a maximal model class (3.31). 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  $\Sigma_n$ . Let  $M \geq 1$  and let  $\tilde{u}$  be such that

$$\tilde{u} \in \{\operatorname{argmin}_{\|v\|_{\mathcal{U}} \leq M} \mathcal{L}^*(v)\}. \quad (9.1)$$

Clearly,  $u$  is a solution to this problem since  $\|u\|_{\mathcal{U}} \leq 1$  and  $\mathcal{L}^*(u) = 0$ . Thus,  $\mathcal{L}^*(\tilde{u}) = 0$  and it follows from (7.4) that

$$\|u - \tilde{u}\|_{H^1(\Omega)} \lesssim M \mathcal{R}_{\mathcal{U}}(\tilde{m}, \bar{m}), \quad m \geq 1, \quad (9.2)$$

with the implied constant depending only on  $r$  and  $d$ . In other words,  $\tilde{u}$  provides a near optimal recovery when  $d \geq 3$  and a near optimal recovery up to a  $\log m$  factor when  $d = 2$ . However, the minimization problem (9.1) 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  $\Sigma_n$  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

$$\Sigma_n(M) := \{S \in \Sigma_n : \|S\|_{\mathcal{U}} \leq M\} \subset \Sigma_n, \quad (9.3)$$

and consider the solution  $\widehat{S}$  to the minimization problem

$$\widehat{S} \in \{ \operatorname{argmin}_{S \in \Sigma_n(M)} \mathcal{L}^*(S) \}. \quad (9.4)$$

In order to present our results, we denote by

$$E(v, \Sigma_n(M)) := \begin{cases} \inf_{S \in \Sigma_n(M)} (\|\Delta v - \Delta S\|_{L_\gamma(\Omega)} + \|Tr(v - S)\|_{H^{1/2}(\partial\Omega)}), & d \geq 3, \\ \inf_{S \in \Sigma_n(M)} ((1 + \log(\widetilde{m}))\|\Delta v - \Delta S\|_{L_\gamma(\Omega)} + \|Tr(v - S)\|_{H^{1/2}(\partial\Omega)}), & d = 2, \end{cases} \quad (9.5)$$

the error of simultaneously approximating  $v$  and  $\Delta v$ . We remark that in the case  $d = 2$  and  $1 < p \leq \infty$ , 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 9.1.** *Let  $\Sigma_n$  be a set in  $H^1(\Omega)$  and let  $\Sigma_n(M)$  be defined by (9.3) for a fixed value  $M \geq 1$ . If  $\widehat{S}$  is any function from (9.4) then*

$$\|u - \widehat{S}\|_{H^1(\Omega)} \lesssim M\mathcal{R}_{\mathcal{U}}(\widetilde{m}, \overline{m}) + E(u, \Sigma_n(M)), \quad (9.6)$$

with the constant in  $\lesssim$  depending only on  $r, d$ . Moreover, if the approximation set  $\Sigma_n(M)$  is such that  $E(\mathcal{U}, \Sigma_n(M)) \leq C\mathcal{R}_{\mathcal{U}}(\widetilde{m}, \overline{m})$  for some  $C$  depending only on  $r, d$ , then  $\widehat{S}$  provides a uniform near optimal recovery (up to logarithmic factors 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 (7.4) for  $\widehat{S}$  to obtain

$$\|u - \widehat{S}\|_{H^1(\Omega)} \lesssim \mathcal{L}^*(\widehat{S}) + (1 + M)\mathcal{R}_{\mathcal{U}}(\widetilde{m}, \overline{m}). \quad (9.7)$$

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

$$\mathcal{L}^*(\widehat{S}) \lesssim E(u, \Sigma_n(M)) + M\mathcal{R}_{\mathcal{U}}(\widetilde{m}, \overline{m}), \quad (9.8)$$

which together with (9.7) proves the theorem.  $\square$

**Remark 9.2.** *Theorem 9.1 guarantees that  $\widehat{S}$  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. Note that the quantity  $E(\mathcal{U}, \Sigma_n(M))$  determines how well the elements of the set  $\Sigma_n(M)$  simultaneously approximate the Laplacians (in the  $L_\gamma(\Omega)$  norm) and the traces (in  $H^{1/2}(\partial\Omega)$  norm) of the elements in  $\mathcal{U}$ . Thus, the question is whether  $\Sigma_n$  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 [24] and the references therein). When error is measured in the  $L_p$ -norm, optimal rates have been obtained [65, 77, 62] in the case of approximating functions from Besov spaces using deep ReLU neural networks. 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 [76, 75]. 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 [65, 77, 62] can be appropriately modified to give the same rates for restricted approximation with a fixed suitably large value for  $M$  (independent of  $n$ ).*

## 10 Numerical Experiments

In this section, we present numerical experiments testing the loss functions described in §8 for the practical solution of elliptic PDEs. In our experiments, we solve the two-dimensional Poisson equation

$$-\Delta u = f \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega, \quad (10.1)$$

on  $\Omega = (0, 1)^2$  for various right-hand sides  $f$  and boundary values  $g$ . We take a uniform grid of collocation points on  $\Omega$ , where  $\tilde{m}$  is the total number of collocation points (where we penalize the residual of the equation), and  $\bar{m}$  is the number of collocation points which lie on the boundary (where we penalize the residual of the boundary values). We then test the following four loss functions, described in §8:

- The original (unweighted) PINNs loss:  $\mathcal{L}_{sq,1}$ ;
- A properly weighted PINNs loss:  $\mathcal{L}_{sq,\lambda(\bar{m})}$ , see (8.5);
- The loss  $\mathcal{L}_{sq,\gamma}^*$  from (7.2) (since  $d = 2$  we choose here  $\gamma = 1.1$ );
- The loss  $\mathcal{L}_{sq,2}^*$  (the loss function (7.2) with  $L_2$  on the domain).

Since we are in the case  $d = 2$ , the latter two loss functions involve a logarithm which is ignored in the numerical experiments. For all losses, we use the same neural network and training algorithm. We report both the final loss values achieved and the final relative error in the  $H^1(\Omega)$ -norm (estimated using a large number of collocation points).

Let us remark on the relative cost of implementing each of these four loss functions. In all our experiments, the efficiency of optimizing each of them was roughly the same. The reason for this is that although the discretization of the  $H^{1/2}$ -norm on the boundary used in the losses  $\mathcal{L}_{sq,\gamma}^*$  and  $\mathcal{L}_{sq,2}^*$  is more expensive to compute and differentiate, the training process is dominated by the cost of evaluating the network and its derivatives at all of the collocation points rather than evaluating the loss based upon these values. For this reason, using losses which discretize the  $H^{1/2}$ -norm doesn't significantly slow down the network training.

### 10.1 Network Architecture and Training Algorithm

When training PINNs, it is known that the optimization of the neural network can be particularly sensitive and tricky [60, 46, 71, 72, 18], and much research has been devoted to the development of improved training algorithms for PINNs [53, 17, 44, 57, 79, 70]. In this section, we describe our network architecture, initialization, and training algorithm in detail.

In our experiments, we use a deep neural network with  $L$  hidden layers of width  $W$ . The first layer uses a tanh activation function, while the remaining layers are residual layers [40] with the ReLU<sup>3</sup> 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<sup>3</sup> 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 [33, 39], 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 this network applying the energy natural gradient descent algorithm described in [57], using a line search to find the step size in each iteration. The full code and all experimental details can be found at <https://github.com/jwsiegel2510/consistent-PINNs>.

## 10.2 Results

**Experiment 1:** In our first experiment, we choose  $f$  and  $g$  so that the exact solution is given by

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

The function is harmonic, i.e  $f \equiv 0$ . The function  $u$  is very smooth and so for this problem it suffices to use a relatively small number of collocation points and a relatively small network. We test a grid of collocation points with 5, 10, 15 and 20 points in each direction. We use a small network ( $L = 3$ ,  $W = 5$ ) and iterate for 500 steps. We present the results in Table 1.

$\tilde{m}$	$\overline{m}$	$\mathcal{L}_{sq,1}$ Rel Error	$\mathcal{L}_{sq,1}$ Loss	$\mathcal{L}_{sq,\lambda(\overline{m})}$ Rel error	$\mathcal{L}_{sq,\lambda(\overline{m})}$ Loss	$\mathcal{L}_{sq,\gamma}^*$ Rel error	$\mathcal{L}_{sq,\gamma}^*$ Loss	$\mathcal{L}_{sq,2}^*$ Rel error	$\mathcal{L}_{sq,2}^*$ Loss
25	16	0.80%	$2.8 \cdot 10^{-5}$	0.68%	$1.1 \cdot 10^{-4}$	0.33%	$2.5 \cdot 10^{-5}$	0.30%	$3.1 \cdot 10^{-5}$
100	36	0.43%	$1.7 \cdot 10^{-5}$	0.21%	$2.3 \cdot 10^{-5}$	0.21%	$2.0 \cdot 10^{-5}$	0.25%	$2.2 \cdot 10^{-5}$
225	56	0.41%	$1.5 \cdot 10^{-5}$	0.15%	$2.1 \cdot 10^{-5}$	0.23%	$1.7 \cdot 10^{-5}$	0.27%	$2.0 \cdot 10^{-5}$
400	76	0.41%	$1.7 \cdot 10^{-5}$	0.086%	$7.18 \cdot 10^{-5}$	0.22%	$1.9 \cdot 10^{-5}$	0.33%	$3.6 \cdot 10^{-5}$

Table 1: Experimental results on the harmonic problem. All relative errors are in the  $H^1(\Omega)$ -norm, computed using a grid with 500 points in each direction. When the number of grid points is large, the properly weighted PINNs loss  $\mathcal{L}_{sq,\lambda(\overline{m})}$  results in significantly lower error than the other loss functions.

**Experiment 2:** In our second experiment, we choose  $f$  and  $g$  so that the exact solution is given by

$$u(x, y) = \frac{e^{2(x+y)} \cos(2\pi(y-x))}{1 + 8x^2 + y^2}, \quad (x, y) \in \Omega = (0, 1)^2.$$

In this case, both  $f$  and  $g$  are non-zero and the solution is significantly less smooth. To capture this, we use a slightly larger network ( $L = 3$  and  $W = 10$ ). All other parameters are the same as in the first experiment. The results are presented in Table 2.

$\tilde{m}$	$\overline{m}$	$\mathcal{L}_{sq,1}$ Rel Error	$\mathcal{L}_{sq,1}$ Loss	$\mathcal{L}_{sq,\lambda(\overline{m})}$ Rel error	$\mathcal{L}_{sq,\lambda(\overline{m})}$ Loss	$\mathcal{L}_{sq,\gamma}^*$ Rel error	$\mathcal{L}_{sq,\gamma}^*$ Loss	$\mathcal{L}_{sq,2}^*$ Rel error	$\mathcal{L}_{sq,2}^*$ Loss
25	16	26%	$1.2 \cdot 10^{-7}$	24%	$2.2 \cdot 10^{-6}$	24%	$3.3 \cdot 10^{-7}$	22%	$1.6 \cdot 10^{-7}$
100	36	1.5%	$3.7 \cdot 10^{-3}$	2.9%	$3.7 \cdot 10^0$	1.1%	$1.6 \cdot 10^{-2}$	0.69%	$2.0 \cdot 10^{-2}$
225	56	1.4%	$5.5 \cdot 10^{-3}$	1.1%	$4.1 \cdot 10^{-2}$	1.6%	$1.1 \cdot 10^{-1}$	0.90%	$1.1 \cdot 10^{-2}$
400	76	2.3%	$1.7 \cdot 10^{-2}$	0.88%	$3.2 \cdot 10^{-2}$	1.5%	$4.7 \cdot 10^{-2}$	0.75%	$2.2 \cdot 10^{-2}$

Table 2: Experimental results on the smooth problem. All relative errors are in the  $H^1(\Omega)$ -norm, computed using a grid with 500 points in each direction. Aside from large errors when the number of collocation points is relatively small, the consistent loss with  $L_2$ -penalty on the domain,  $\mathcal{L}_{sq,2}^*$ , performs best on this experiment.

**Experiment 3:** Finally, for our last experiment, we choose  $f$  and  $g$  so that the exact solution is given by

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

The solution  $u$  vanishes on the boundary (i.e.  $g = 0$ ) and is much less smooth than the preceding two experiments. For this reason, we use a significantly larger network ( $L = 3$ ,  $W = 15$ ), more collocation points (10, 20, 30 and 40 in each direction), and more training iterations (1000 steps). The results are presented in Table 3.

Our numerical experiments seem to imply that the optimization method plays a critical role in the final accuracy achieved when using any of these loss functions. Indeed, there is a fairly wide variation in the

$\tilde{m}$	$\overline{m}$	$\mathcal{L}_{sq,1}$ Rel Error	$\mathcal{L}_{sq,1}$ Loss	$\mathcal{L}_{sq,\lambda(\overline{m})}$ Rel error	$\mathcal{L}_{sq,\lambda(\overline{m})}$ Loss	$\mathcal{L}_{sq,\gamma}^*$ Rel error	$\mathcal{L}_{sq,\gamma}^*$ Loss	$\mathcal{L}_{sq,2}^*$ Rel error	$\mathcal{L}_{sq,2}^*$ Loss
100	36	8.4%	$9.0 \cdot 10^{-4}$	3.8%	$3.3 \cdot 10^{-3}$	10.4%	$1.0 \cdot 10^{-2}$	18.9%	$6.1 \cdot 10^{-3}$
400	76	18.8%	$5.2 \cdot 10^{-2}$	1.8%	$4.4 \cdot 10^{-2}$	5.1%	$2.2 \cdot 10^{-2}$	5.3%	$2.9 \cdot 10^{-2}$
900	116	10.9%	$1.1 \cdot 10^{-2}$	2.4%	$3.9 \cdot 10^{-2}$	3.1%	$1.9 \cdot 10^{-2}$	2.5%	$2.1 \cdot 10^{-2}$
1600	156	6.1%	$6.3 \cdot 10^{-3}$	1.2%	$7.5 \cdot 10^{-3}$	11.1%	$6.4 \cdot 10^{-2}$	1.4%	$1.1 \cdot 10^{-2}$

Table 3: Experimental results on the non-smooth problem. All relative errors are in the  $H^1(\Omega)$ -norm, computed using a grid with 500 points in each direction. We see that the properly weighted PINNs loss,  $\mathcal{L}_{sq,\lambda(\overline{m})}$ , and the consistent PINNs with  $L_2$ -penalty on the domain,  $\mathcal{L}_{sq,2}^*$ , perform significantly better than both other loss functions. Interestingly, the performance of the consistent loss  $\mathcal{L}_{sq,\gamma}^*$  deteriorates significantly with the largest number of collocation points.

computed final loss values. Accounting for this, in the first two tests, it seems that the correctly weighted PINNs loss  $\mathcal{L}_{sq,\lambda(\overline{m})}$  performs the best. The loss (7.2) with  $L_2$ -penalty on the residual,  $\mathcal{L}_{sq,2}^*$ , is performing similarly well. However, more experiments are needed to confirm our observations.

The results of Experiment 3 are more conclusive. The solution in this case is much less smooth and more difficult to fit using neural networks than in the first two experiments. In this experiment, we observe that optimizing the correctly weighted PINNs loss  $\mathcal{L}_{sq,\lambda(\overline{m})}$  and the loss (7.2) with  $L_2$ -penalty on the domain,  $\mathcal{L}_{sq,2}^*$ , result in significantly lower solution error than the other two loss functions. Interestingly, the loss  $\mathcal{L}_{sq,\gamma}^*$  performs nearly as well up until the final experiment where the collocation grid has 40 points in each direction.

Our interpretation of these results is the following. The (unweighted) original PINNs loss does not adequately fit the boundary values, while the loss (7.2) results in a less smooth numerical solution due to the optimization of the non-smooth  $L_\gamma$ -norm. Both the losses  $\mathcal{L}_{sq,\lambda(\overline{m})}$  and  $\mathcal{L}_{sq,2}^*$  overcome these issues by correctly weighting the boundary error and also resulting in neural network solutions which are smooth. This is corroborated by the solution plots in each case, which are shown in Figures 10.2 and 10.2. Observe that the solutions minimizing both of the losses  $\mathcal{L}_{sq,1}$  and  $\mathcal{L}_{sq,\gamma}^*$  fail to accurately fit the boundary values. However, the solution for  $\mathcal{L}_{sq,1}$  is relatively smooth, while the solution for  $\mathcal{L}_{sq,\gamma}^*$  is significantly less smooth (especially near the boundary). The solutions for both  $\mathcal{L}_{sq,\lambda(\overline{m})}$  and  $\mathcal{L}_{sq,2}^*$  are fairly smooth and fit the boundary values well.

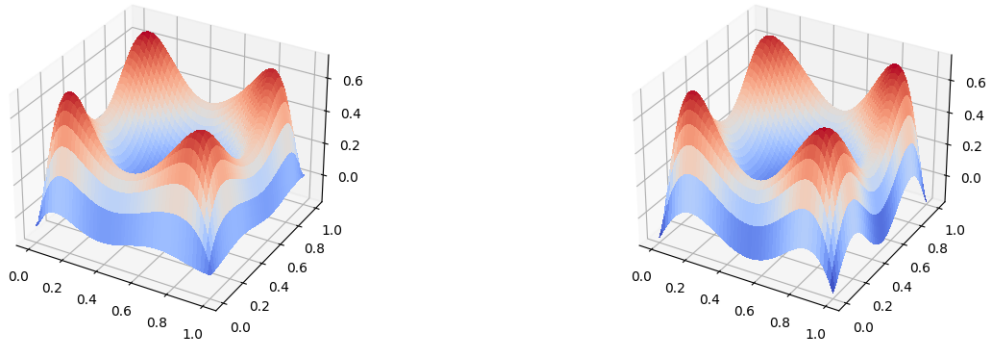


Figure 10.2: Experiment 3: Solution plots for the (unweighted) original PINNs loss  $\mathcal{L}_{sq,1}$  (left) and the consistent PINNs loss  $\mathcal{L}_{sq,\gamma}^*$  (right). We see that both fail to accurately fit the boundary values. Note that the consistent PINNs solution is much less smooth near the boundary.



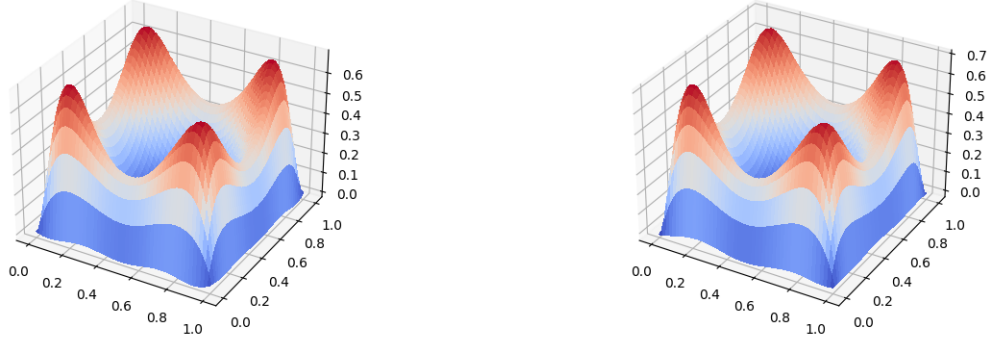


Figure 10.3: Experiment 3: Solution plots for the correctly weighted PINNs loss  $\mathcal{L}_{sq,\lambda(\overline{m})}$  (left) and the consistent PINNs loss with  $L_2$ -penalty on the domain  $\mathcal{L}_{sq,2}^*$  (right). We see that both accurately fit the boundary values and remain fairly smooth.

Based upon our numerical tests, we recommend using either the correctly weighted PINNs loss  $\mathcal{L}_{sq,\lambda(\overline{m})}$ , see (8.5), or the loss (7.2) with  $L_2$ -penalty on the domain,  $\mathcal{L}_{sq,2}^*$ , especially for difficult problems with limited regularity. However, our observations are naturally quite limited and significant additional experiments are needed to draw stronger conclusions.

## 11 Concluding Remarks

We have investigated collocation methods for solving elliptic PDEs of the form (1.1). 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. Our theoretical analysis was performed for  $\Omega = (0,1)^d$ , but it is well-known that the results carry over to any Lipschitz domain (see [58]).

We then turned to the study of theoretical 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. While the typical loss function used in PINNs does not provide an a posteriori bound on the error, we have introduced a variety of modified loss functions which do provide such a bound. We call the use of any one of these theoretically justified loss functions consistent PINNs. Finally, we perform several numerical experiments which compare these different loss functions and show their favorable performance in comparison to the typical loss used in PINNs.

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 suggested losses over  $\Sigma_n(M)$ . This is the same objection that can be made for almost all learning problems and certainly for PINNs. The usual approach in PINNs 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 enter into a discussion of the art of using gradient descent in optimization. However, we want to address two issues that are relevant to this paper. The first is the fact that we must optimize the loss over  $\Sigma_n(M)$  rather than  $\Sigma_n$  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 by adding a penalty to the consistent losses to guarantee that at each step of the optimization one remains in the constrained set  $\Sigma_n(M)$  (see [5]) but this does not ease the computational burden and one is still faced with a lack of convergence of the numerical optimization.

The second issue is understood by practitioners but not usually pointed out in the theory. Neural network spaces are a very unstable manifold in the sense 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 [9, 36, 35, 37]. However, 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,

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

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

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

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  $\ell_I$ , we have

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

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  $P \in \mathcal{P}_r$ .

We also will 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

$$\tilde{w}_r(f, t)_{L_p(I)}^p := t^{-d} \int_{\mathbf{h} \in [-t, t]^d} \int_{I_{r\mathbf{h}}} |\Delta_{\mathbf{h}}^r(f, \mathbf{x})|^p d\mathbf{x} d\mathbf{h}, \quad (12.3)$$

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  $\tilde{w}_r$  (see §2 of [23]),

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

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  $\omega_r$  replaced by  $\tilde{w}_r$ .

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

Before proceeding further, let us remark on why we introduce the averaged modulus of smoothness  $\tilde{w}_r$ . The advantage of  $\tilde{w}_r$  over  $\omega_r$  is that  $\tilde{w}_r^p$  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  $\Omega$  which form a partition of  $\Omega$ . Then, from the definition of  $\tilde{w}_r$ , we have

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

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

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

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

It is shown in Lemma 3.2 of [21] that if  $P_I \in \mathcal{P}_r$  is near best in  $L_p(I)$  with constant  $\lambda$ , then it is also near best in  $L_{\bar{p}}(I)$  whenever  $\bar{p} \geq p$ , i.e.,

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

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  $\lambda$  is near best on any larger cube  $J$  which contains  $I$  in the sense that

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

where now  $C$  depends additionally on  $|J|/|I|$ , see Lemma 3.3 in [21]. Note that even though Lemma 3.2 and Lemma 3.3 in [21] 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  $\bar{p} \geq p$ . 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 [23] or Lemma 3.1 and Lemma 3.2 in [22]. For any cube  $I$  in  $\mathbb{R}^d$  and any function  $g \in L_p(I)$ , we introduce the normalized  $L_p$  norm

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

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

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

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

$$|P|_{B_q^s(L_p(I))} \leq C\ell_I^{-s} \|P\|_{L_p(I)}, \quad (12.12)$$

with a constant independent of  $I$ . This is proved by dilation (see e.g. Corollary 5.2 of [21]).

### 12.3 Besov spaces and piecewise polynomial approximation

We now recall how membership in  $B_p^s(\Omega)$  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.

Recall that for  $k \geq 0$ , we denoted by  $\mathcal{D}_k$  the partition of  $\Omega$  into dyadic cubes  $I$  of side length  $2^{-k}$ , where the dyadic cube  $I \in \mathcal{D}_k$  is 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 and only if 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

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

where  $\chi_I$  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  $\bar{\Omega}$  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  $f \in B_p^s = B_\infty^s(L_p(\Omega))$ , with  $\Omega = (0, 1)^d$ , then*

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

with  $C$  depending only on  $p$  and  $s$ .

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

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

where  $\tilde{w}_r$  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

$$\|f - S_k\|_{L_p(\Omega)}^p \lesssim \sum_{I \in \mathcal{D}_k} \tilde{w}_r(f, 2^{-k})_{L_p(I)}^p \lesssim \tilde{w}_r(f, 2^{-k})_{L_p(\Omega)}^p \lesssim w_r(f, 2^{-k})_{L_p(\Omega)}^p \lesssim |f|_{B_p^s(\Omega)}^p 2^{-ksp}, \quad (12.16)$$

where we used the subadditivity (12.6) and the fact that  $\mathcal{D}_k$  is a partition of  $\Omega$ . Here we have also used the equivalence (12.4) of  $\tilde{w}_r$  and  $w_r$  and (2.8). This proves (12.14) in the case  $p < \infty$ . When  $p = \infty$ , these inequalities follow directly from (12.15) and the fact that  $\|\cdot\|_{L_\infty(I)} \leq \|\cdot\|_{L_\infty(\Omega)}$  for each  $I \subset \Omega$ .  $\square$

### 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  $p \leq \tau \leq \infty$ , then we have*

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

*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

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

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

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

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

$$\|R_k\|_{L_p(\Omega)} \leq \|f - S_k\|_{L_p(\Omega)} + \|f - S_{k-1}\|_{L_p(\Omega)} \leq C|f|_{B_p^s(\Omega)} 2^{-ks}, \quad k \geq 1, \quad (12.19)$$

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

$$f = \sum_{k=0}^{\infty} R_k \quad (12.20)$$

with the series converging in  $L_p(\Omega)$ . We now consider the following cases for  $\tau$ .

**Case 1:**  $\tau = \infty$ . On each dyadic cube  $I \in \mathcal{D}_k$ , we have

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

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

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

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

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

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

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

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

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

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

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

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

$$\|R_k\|_{L_\tau(\Omega)} \leq C2^{kd(1/p-1/\tau)} |f|_{B_p^s(\Omega)} 2^{-ks}, \quad k \geq 1. \quad (12.24)$$

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

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

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

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

which completes the proof in this case.  $\square$

**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

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

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

### 12.4.1 Proof of Theorem 2.3

For any  $f \in B_p^s(\Omega)$ , consider the corresponding  $S_k(f)$ , defined in (12.13) and the function  $\tilde{f}$ , see (12.23). As shown in Theorem 12.2,  $f = \tilde{f}$  a.e.. We want to observe that the function  $\tilde{f}$  is continuous on  $\Omega$  and moreover, (2.21) holds. Indeed, it is enough to consider  $0 < t \leq 1$ . Let  $\mathbf{x} \in \Omega$  and  $|\mathbf{h}| \leq t$  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  $\ell_J \leq rt$ . We argue as in the proof of (12.22) and (12.23), replacing  $\Omega$  by  $J$  to show that there is a polynomial  $P_J \in \mathcal{P}_r$  such that

$$\sup_{\mathbf{x} \in J} |\tilde{f}(\mathbf{x}) - P_J(\mathbf{x})| \leq C |f|_{B_p^s(J)} \ell_J^{s-d/p} \leq C |f|_{B_p^s(\Omega)} t^{s-d/p}.$$

It follows that

$$|\Delta_{r\mathbf{h}}(\tilde{f}, \mathbf{x})| \leq C |f|_{B_p^s(\Omega)} t^{s-d/p},$$

uniformly for  $\mathbf{x} \in \Omega$ . This proves that  $\tilde{f}$  is continuous and (2.21) holds.  $\square$

## 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

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

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

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

and therefore

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

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

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

which completes the proof of the theorem 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

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

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

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

which is the improved bound we want.

We now consider two cases.

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

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

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

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

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

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

We now integrate both sides of this inequality and find that

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

as desired.

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

$$\|f - P_I\|_{L_\infty(I)} \leq C \left\{ \sum_{j \geq k} 2^{-j\beta p'} \right\}^{1/p'} \left\{ \sum_{j \geq k} [2^{j(d/p+\beta)} \tilde{w}_r(f, 2^{-j})_{L_p(I)}^p] \right\}^{1/p}, \quad (12.35)$$

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

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

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.  $\square$

### 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

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

We define

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

It follows from (12.37) that

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

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

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

Since the function  $R_k^*(f)$  is continuous on  $\Omega$ , we have (see Theorem 2.1.2 in [8])

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

This gives

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

Writing,

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

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

## 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 [58], but for completeness, we give the following proof using the results of §2. Recall our notation  $B_p^s(\Omega) := B_\infty^s(L_p(\Omega))$  and the fact that all other Besov spaces  $B_q^s(L_p(\Omega))$ , with the same values for  $p$  and  $s$ , are contained in  $B_p^s(\Omega)$ . Therefore, it is enough to prove the upper bounds when  $\mathcal{B} = B_p^s(\Omega)$ . To prove these upper bounds, we need to exhibit a set of  $\tilde{m}$  data sites at which we achieve the claimed bound. We take these data sites as the tensor product grid points  $G_{k,r}$  of  $\bar{\Omega}$  where  $r > \max(s, 1)$ . This will yield the upper bounds when  $\tilde{m} = (2^k r)^d$ . The upper bounds for the other values of  $\tilde{m}$  follow from the fact that  $R_{\tilde{m}}^*$  is monotonically nonincreasing as  $\tilde{m} \rightarrow \infty$ .

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

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



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 (2.19)) that if  $\tau \geq p$  then

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

This bound also holds for  $\tau < p$  since then  $B_p^s(\Omega)$  is continuously embedded in  $B_\tau^s(\Omega)$ . The function  $S$  only depends on the data  $(f_i)_{i=1}^{\tilde{m}}$ . This shows that every function  $\tilde{f} \in \mathcal{F}_{\text{data}}$  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 (3.11). Then, the space  $L_\gamma(\Omega)$  is embedded into  $H^{-1}(\Omega)$  and we have*

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

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

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

where  $C$  is an absolute constant.

*Proof.* Consider first the case  $d \geq 3$ , in which case  $\gamma(d) > 1$ . Let  $\gamma'$  be the dual index of  $\gamma$ , that is,  $\frac{1}{\gamma'} = \frac{1}{2} - \frac{1}{d}$ , so that  $\gamma' < \infty$ . If  $v \in U(H_0^1(\Omega))$ , then by the Sobolev embedding theorem we have

$$\|v\|_{L_{\gamma'}(\Omega)} \lesssim \|v\|_{H^1(\Omega)} \lesssim |v|_{H_0^1(\Omega)} := \|\nabla v\|_{L_2(\Omega)}, \quad v \in H_0^1(\Omega), \quad (12.45)$$

where the last inequality is Poincaré's inequality. From Hölder's inequality, it follows that for any  $v \in H_0^1(\Omega)$  and any  $\tilde{f} \in L_\gamma(\Omega)$ , we have

$$\int_{\Omega} \tilde{f} v \leq \|\tilde{f}\|_{L_\gamma(\Omega)} \|v\|_{L_{\gamma'}(\Omega)} \lesssim \|\tilde{f}\|_{L_\gamma(\Omega)} |v|_{H_0^1(\Omega)}. \quad (12.46)$$

This gives

$$\|\tilde{f}\|_{H^{-1}(\Omega)} := \sup_{v \in U(H_0^1(\Omega))} \int_{\Omega} \tilde{f} v \lesssim \|\tilde{f}\|_{L_\gamma(\Omega)}, \quad \tilde{f} \in L_\gamma(\Omega), \quad (12.47)$$

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  $\tau'$  is the dual index to  $\tau$ . In this case we have a compact Sobolev embedding

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

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  $f$

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

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

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

Taking  $\tau'$ -th roots, we then get using that the  $\ell_{\tau'}$ -norm is bounded by the  $\ell_2$ -norm since  $\tau' \geq 2$  that

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

where  $C$  is independent of  $\tau'$ . 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  $\tau'$ . Utilizing (12.48), the same duality argument used for  $d \geq 3$  now completes the proof when  $d = 2$ .  $\square$

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  $\alpha_{-1}$  does not change over the  $p \geq \gamma$ , it is enough to consider the case when  $p \leq \gamma$ . Indeed, if  $p > \gamma$  then  $U(B_q^s(L_p(\Omega)))$  is contained in  $U(B_\infty^s(L_\gamma(\Omega))) =: U(B_\gamma^s(\Omega))$ . Therefore the upper bound follows once we have established the case  $\mathcal{F} = U(B_q^s(L_p(\Omega)))$  with  $p \leq \gamma$ . Similarly, when  $p \leq \gamma$ , then  $U(B_q^s(L_p(\Omega))) \subset U(B_\infty^s(L_p(\Omega)))$ . Accordingly, in going further we only need to consider the case when  $\mathcal{F} = U(B_\infty^s(L_p(\Omega))) =: U(B_p^s(\Omega))$  with  $p \leq \gamma$ .

Let  $f$  be any function in  $\mathcal{F} := U(B_\infty^s(L_p(\Omega)))$  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 (2.19), we have

$$\|f - S\|_{L_\gamma(\Omega)} \preceq |f|_{B_p^s(I)} 2^{-kd\alpha_{-1}}. \quad (12.50)$$

We now use Lemma 12.4 to obtain

$$\|f - S\|_{H^{-1}\Omega} \leq C \|f - S\|_{L_\gamma(\Omega)} \leq C m^{-\alpha_{-1}} |f|_{B_p^s(\Omega)}. \quad (12.51)$$

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  $C m^{-\alpha_{-1}}$  of  $S$  and hence proves that the Chebyshev radius of this set does not exceed  $C m^{-\alpha_{-1}}$ . 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

$$\|f - S\|_{L_p(\Omega)} \preceq |f|_{B_p^s(I)} 2^{-kd\alpha_{-1}}, \quad (12.52)$$

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

$$\|f - S\|_{H^{-1}\Omega} \leq C_p \|f - S\|_{L_p(\Omega)} \leq C_p m^{-\alpha_{-1}} |f|_{B_p^s(\Omega)}. \quad (12.53)$$

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

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

Applying Lemma 12.4 we get that

$$\|f - S\|_{H^{-1}\Omega} \leq C_p \|f - S\|_{L_p(\Omega)} \leq \frac{C}{\tau - 1} m^{-\alpha-1} m^{1-1/\tau} |f|_{B_p^s(\Omega)}. \quad (12.55)$$

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

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

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

**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  $\mathbf{x}_i$ ,  $i = 1, \dots, m$ , are any  $m$  data sites. In order to prove a lower bound for  $R^*(\mathcal{F})_X$ , we construct a function  $\eta$  in  $\mathcal{F}$  so that

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

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

where  $\alpha$  is the appropriate index for  $X$ . Since both  $\eta$  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  $\eta$  for each of the choices of  $X$ . Let  $\varphi$  be a smooth non-negative function on  $\mathbb{R}^d$  which vanishes outside  $\Omega$  and additionally satisfies

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

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

We choose  $\varphi$  with these properties which has the smallest norm

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

So,  $\varphi$  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  $\xi_I$  be the smallest vertex of  $I$  and as usual  $\ell_I$  is its side length. We define the function

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

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

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

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  $X_m := \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \in \Omega$  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

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

It follows that each  $\eta_I$  is in  $\mathcal{F}$ . Also, each of the  $\eta_I$  vanishes at each of the data sites. We shall use the  $\eta_I$  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  $\eta_I$  defined by (12.61). We take  $I$  as any fixed cube from  $\mathcal{I}$  and define  $\eta := \eta_I$ . Then  $\eta$  vanishes at each of the data sites. Also, it follows from the definition of  $\eta$  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  $U(B_q^s(L_\infty(\Omega))) \subset U(B_q^s(L_p(\Omega)))$ , it is sufficient to prove the lower bound in the case  $p = \infty$ , i.e., when  $\mathcal{F} = U(B_q^s(L_\infty(\Omega)))$  with  $0 < q < \infty$  arbitrary but fixed. We define

$$\eta = \kappa \sum_{I \in \mathcal{I}} \eta_I, \quad (12.62)$$

with  $\kappa$  a fixed constant chosen so that  $\eta \in U(B_q^s(L_\infty(\Omega)))$ . Here,  $\eta_I$  is defined as in (12.61) with  $p = \infty$ . We next derive a lower bound for  $\kappa$ . The terms  $\eta_I$  in (12.62) each have  $L_\infty$  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  $\eta$  satisfies

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

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

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

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  $p \leq 2$ . We take  $\eta = \eta_I$  for any fixed  $I \in \mathcal{I}$ . We estimate

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

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  $p \leq \gamma$ , where  $\gamma$  is given by (3.11). In this case, we choose  $\eta = \eta_I$  where  $I$  is any cube in  $\mathcal{I}$ . The function  $\eta$  vanishes at all of the data sites. By construction, we have that  $\eta \in \mathcal{F}$  and the lower bound

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

where  $I_0 = [\xi_I - \ell_I/4, \xi_I + 1/4]^d$  is the subcube corresponding to  $\Omega_0$  in  $I$  (see (12.57)).

We want to bound the  $H^{-1}(\Omega)$  norm of  $\eta$  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

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

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  $\eta_I$ , we have

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

It follows that

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

Since  $|I_0| \geq cm^{-1}$  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  $R^*(\mathcal{F})_{H^{-1}(\Omega)} \geq cm^{-\alpha-1}$  in the case  $p \leq \gamma$  since  $1/p - 1/\gamma \geq 0$ .

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  $\eta$  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  $\eta$ ,

$$v(x) = c\varphi(x), \quad (12.70)$$

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

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

because  $\kappa \geq c$  and  $\eta_I \geq 2^{-ks}$  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  $\Omega_0$  is  $2^{-d}(4^d m) = 2^d m$ . Since  $\#\mathcal{I} \geq (4^d - 1)m$ , we see that

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

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

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

This implies that  $R^*(\mathcal{F})_{H^{-1}(\Omega)} \geq cm^{-\alpha-1}$  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.  $\square$

### 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

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

where the next to last inequality uses the estimate (2.20) and also the fact that  $m \asymp 2^{k(d-1)}$ . This proves (ii).

It follows that

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

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

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

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  $\mathcal{Z} := \{\mathbf{z}_1, \dots, \mathbf{z}_m\}$ ,  $m = 2^{k(d-1)} - 1$ , be any proposed set of data sites on  $\partial\Omega$ . Let  $F := \{\mathbf{x} \in \Omega : \mathbf{x} \cdot \mathbf{e}_1 = 0, \mathbf{e}_1 = (1, 0, \dots, 0) \in \mathbb{R}^d\}$ , 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  $\varphi_J$  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  $\xi'_J$  is the center of  $J$ , then the point  $\xi'_J - (2^{-k-1}, 0, \dots, 0)$  is the center of  $\overline{J}$ .

We now define

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

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

$$\eta := T_{\partial\Omega}v \quad (12.77)$$

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

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

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

with  $c$  not depending on  $m$ . This will prove the lower bound we seek. First consider the function  $\varphi$  defined in (12.57). Let  $\mathbf{e}_1 = (1, 0, \dots, 0) \in \mathbb{R}^d$ . Because of the tensor product structure of  $\varphi$  described after (12.57), the trace  $\eta_0$  of  $\varphi$  onto the hyperplane  $\mathbf{x} \cdot \mathbf{e}_1 = 1/2$  is the  $d-1$  version of  $\varphi$ . We define

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

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

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

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

## 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

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

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

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

and thus we have

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

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

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

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] Robert A Adams and John JF Fournier, *Sobolev spaces*, Elsevier, 2003.
- [2] D. Arnold and J. Saraven, *On the asymptotic convergence of spline-collocation methods for partial differential equations*, SIAM J. Num. Anal. **21** (1984), 459–472.
- [3] 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).
- [4] Jens Berg and Kaj Nyström, *A unified deep artificial neural network approach to partial differential equations in complex geometries*, Neurocomputing **317** (2018), 28–41.
- [5] Peter Binev, Andrea Bonito, Ronald DeVore, and Guergana Petrova, *Optimal learning*, Calcolo **61** (2024), no. 1, 15.
- [6] Andrea Bonito, Claudio Canuto, Ricardo H Nochetto, and Andreas Veiser, *Adaptive finite element methods*, Acta Numerica **33** (2024), 163–485.
- [7] 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.
- [8] Philippe G Ciarlet, *The finite element method for elliptic problems*, SIAM, 2002.
- [9] Albert Cohen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk, *Optimal stable nonlinear approximation*, Foundations of Computational Mathematics **22** (2022), 607–648.
- [10] ———, *Optimal stable nonlinear approximation*, Foundations of Computational Mathematics **22** (2022), no. 3, 607–648.
- [11] 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.
- [12] 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.
- [13] 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.

- [14] ———, *Optimal approximation of elliptic problems by linear and nonlinear mappings ii*, Journal of Complexity **22** (2006), no. 4, 549–603.
- [15] Feng Dai, E Kosov, and V Temlyakov, *Some improved bounds in sampling discretization of integral norms*, Journal of Functional Analysis **285** (2023), no. 4, 109951.
- [16] Feng Dai, Andriy Prymak, A Shadrin, V Temlyakov, and S Tikhonov, *Sampling discretization of integral norms*, Constructive Approximation **54** (2021), no. 3, 455–471.
- [17] Caio Davi and Ulisses Braga-Neto, *Pso-pinn: Physics-informed neural networks trained with particle swarm optimization*, arXiv preprint arXiv:2202.01943 (2022).
- [18] Tim De Ryck, Florent Bonnet, Siddhartha Mishra, and Emmanuel de Bézenac, *An operator preconditioning perspective on training in physics-informed machine learning*, arXiv preprint arXiv:2310.05801 (2023).
- [19] Tim De Ryck and Siddhartha Mishra, *Numerical analysis of physics-informed neural networks and related models in physics-informed machine learning*, arXiv preprint arXiv:2402.10926 (2024).
- [20] R. DeVore and G. Lorentz, *Constructive approximation*, Springer-Verlag, 1993.
- [21] R. DeVore and V. Popov, *Interpolation of Besov spaces*, Transactions AMS **305** (1988), no. 1, 397–414.
- [22] R. DeVore and R. Sharpley, *Maximal functions measuring smoothness*, Memoirs of AMS **293** (1984).
- [23] ———, *Besov spaces on domains in  $R^d$* , TAMS **335** (1993), no. 2, 843–364.
- [24] Ronald DeVore, Boris Hanin, and Guergana Petrova, *Neural network approximation*, Acta Numerica **30** (2021), 327–444.
- [25] 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.
- [26] 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.
- [27] Nathan Doumèche, Gérard Biau, and Claire Boyer, *Convergence and error analysis of pinns*, arXiv preprint arXiv:2305.01240 (2023).
- [28] 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.
- [29] Francisco Eiras, Adel Bibi, Rudy R Bunel, Krishnamurthy Dj Dvijotham, Philip Torr, and M Pawan Kumar, *Efficient error certification for physics-informed neural networks*, Forty-first International Conference on Machine Learning, 2024.
- [30] Alexandre Ern and Jean-Luc Guermond, *Finite elements i: Approximation and interpolation*, vol. 72, Springer Nature, 2021.
- [31] Dimitrios Gazoulis, Ioannis Gkanis, and Charalambos G Makridakis, *On the stability and convergence of physics informed neural networks*, arXiv preprint arXiv:2308.05423 (2023).
- [32] David Gilbarg and Neil S Trudinger, *Elliptic partial differential equations of second order*, vol. 224, Springer, 1977.
- [33] 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.



- [34] Pierre Grisvard, *Elliptic problems in nonsmooth domains*, SIAM, 2011.
- [35] Guergana and Przemyslaw Wojtaszczyk, *Limitations on approximation by deep and shallow neural networks*, Journal of Machine Learning Research **24** (2023), no. 353, 1–38.
- [36] ———, *Lipschitz widths*, Constructive Approximation **57** (2023), 759–805.
- [37] ———, *Neural networks: deep, shallow, or in between?*, arXiv:2310.07190 (2023).
- [38] 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.
- [39] 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.
- [40] ———, *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [41] 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.
- [42] David Jerison and Carlos E Kenig, *The inhomogeneous dirichlet problem in lipschitz domains*, Journal of functional analysis **130** (1995), no. 1, 161–219.
- [43] 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.
- [44] 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.
- [45] David Krieg, Erich Novak, and Mathias Sonnleitner, *Recovery of Sobolev functions restricted to iid sampling*, Mathematics of Computation **91** (2022), no. 338, 2715–2738.
- [46] Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney, *Characterizing possible failure modes in physics-informed neural networks*, Advances in neural information processing systems **34** (2021), 26548–26560.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] Yiping Lu, Haoxuan Chen, Jianfeng Lu, Lexing Ying, and Jose Blanchet, *Machine learning for elliptic pdes: Fast rate generalization bound, neural scaling law and minimax optimality*, arXiv preprint arXiv:2110.06897 (2021).
- [51] 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.

- [52] 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.
- [53] Levi D McClenney and Ulisses M Braga-Neto, *Self-adaptive physics-informed neural networks*, Journal of Computational Physics **474** (2023), 111722.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] ———, *Achieving high accuracy with PINNs via energy natural gradient descent*, International Conference on Machine Learning, PMLR, 2023, pp. 25471–25485.
- [58] Erich Novak and Hans Triebel, *Function spaces in lipschitz domains and optimal rates of convergence for sampling*, Constructive approximation **23** (2006), 325–350.
- [59] 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.
- [60] Pratik Rathore, Weimu Lei, Zachary Frangella, Lu Lu, and Madeleine Udell, *Challenges in training pinns: A loss landscape perspective*, Forty-first International Conference on Machine Learning, 2024.
- [61] Thomas Sauer and Yuan Xu, *A case study in multivariate Lagrange interpolation*, Approximation Theory, Wavelets and Applications **454** (1995), 443–452.
- [62] 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.
- [63] 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.
- [64] 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.
- [65] 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.
- [66] 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.
- [67] E. Stein, *Singular integrals and differentiability properties of functions*, Princeton University Press, 1970.
- [68] 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.
- [69] Jan Vybíral, *Sampling numbers and function spaces*, Journal of Complexity **23** (2007), no. 4-6, 773–792.
- [70] Sifan Wang, Shyam Sankaran, Hanwen Wang, and Paris Perdikaris, *An expert’s guide to training physics-informed neural networks*, arXiv preprint arXiv:2308.08468 (2023).

- [71] Sifan Wang, Yujun Teng, and Paris Perdikaris, *Understanding and mitigating gradient flow pathologies in physics-informed neural networks*, SIAM Journal on Scientific Computing **43** (2021), no. 5, A3055–A3081.
- [72] Sifan Wang, Xinling Yu, and Paris Perdikaris, *When and why pinns fail to train: A neural tangent kernel perspective*, Journal of Computational Physics **449** (2022), 110768.
- [73] 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.
- [74] Jinchao Xu, *Finite neuron method and convergence analysis*, Communications in Computational Physics **28** (2020), no. 5, 1707–1745.
- [75] 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).
- [76] 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).
- [77] Dmitry Yarotsky, *Optimal approximation of continuous functions by very deep ReLU networks*, Conference on learning theory, PMLR, 2018, pp. 639–649.
- [78] 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).
- [79] 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.