Optimal sampling and reduced modeling

Matthieu Dolbeault

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

SORBONNE UNIVERSITÉ
LJLL

École doctorale **École Doctorale Sciences Mathématiques de Paris Centre**

Unité de recherche **Laboratoire Jacques-Louis Lions**

Thèse présentée par **Matthieu DOLBEAULT**

Soutenue le **14 juin 2023**

En vue de l'obtention du grade de docteur de Sorbonne Université

Discipline **Mathématiques appliquées**

Spécialité **Analyse numérique**

Échantillonnage optimal et réduction de modèle

Optimal Sampling and Model Order Reduction

Thèse dirigée par Albert COHEN

Composition du jury

Rapporteurs	Virginie EHRLACHER	professeure à l'École des Ponts ParisTech
	Tino ULLRICH	professeur à l'Université de Technologie de Chemnitz
Examinateurs	Markus BACHMAYR	professeur à l'Université RWTH Aachen
	Bruno DESPRÉS	professeur à Sorbonne Université
	Tony LELIÈVRE	professeur à l'École des Ponts ParisTech
	Yvon MADAY	professeur à Sorbonne Université
	Anthony NOUY	professeur à l'École Centrale Nantes
Directeur de thèse	Albert COHEN	professeur à Sorbonne Université

Remerciements

Je voudrais tout d'abord remercier Albert Cohen, pour les sujets variés et stimulants sur lesquels il m'a proposé de travailler, pour l'attention qu'il m'a accordée tout au long de ma thèse, son soutien constant, sa sérénité à toute épreuve, sa curiosité mathématique qui m'a motivé à toujours chercher des solutions mathématiques plus satisfaisantes, sa vision d'ensemble sur un très large champ de recherche, et enfin pour ses conseils sur les versants les mieux enneigés des Alpes ! Grâce à toi, travailler sur les sujets de ma thèse fut un immense plaisir, dont j'espère pouvoir continuer à profiter à l'avenir.

Tino Ullrich, thank you a lot for your careful reading of my thesis, all your constructive comments, and before that for inviting me to join the IBC community, of which you and your team form a prominent component. Je remercie Virginie Ehrlacher d'avoir accepté de rapporter ma thèse, de m'avoir accompagné tant en conférences que dans l'enseignement aux Ponts, et d'avoir toujours eu des remarques bienveillantes à mon égard.

Many thanks to Markus Bachmayr for being part of my jury, and for all the projects we talked about for next year. Merci beaucoup à Anthony Nouy pour les invitations aux workshop et séminaire, et surtout pour l'intérêt qu'il a témoigné pour mes travaux au cours de nos nombreuses discussions. Merci aussi à Tony Lelièvre, Bruno Després et Yvon Maday de prendre le temps de participer à mon jury de thèse.

Merci à Olga Mula pour l'énorme effort qu'a dû représenter l'organisation du CEMRACS avec Virginie, pour les liens créés entre problèmes abstraits et données concrètes, et pour ton amabilité en toutes circonstances. Thank you Mario Ullrich and David Krieg for your open-mindedness, your dynamism, and your thoroughness in the redaction of papers. Thanks to Wolfgang Dahmen for the rich set of methods and viewpoints you brought in our discussions. Merci à Abdellah Chkifa, pour les idées originales et l'approche algorithmique qui apportent beaucoup à nos discussions théoriques.

Pour avoir assuré le bon fonctionnement du laboratoire, merci à Emmanuel Trélat, à Khashayar Dadras et au secrétariat, en particulier à Malika Larcher, Salima Lounici, Corentin Maday et Eryka Loyson qui ont toujours réglé mes soucis administratifs, malgré quelques ordres de mission de dernière minute. Merci à Isabelle Gallagher et Didier Smets d'avoir assuré mon comité de suivi, à François Murat, Frédéric Hecht, et Émile Parolin pour m'avoir fait avancer sur certains de mes problèmes mathématiques. Mes remerciements vont aussi à tous les permanents qui, par leur présence en salle café et dans les couloirs, contribuent à la convivialité et à l'atmosphère chaleureuse du laboratoire.

Parce qu'une thèse c'est aussi des institutions, merci à l'ENS pour les trois années de ~~crous~~ cours qu'elle m'a offertes, à l'École des Ponts pour son éloignement de Paris, idéal pour le cyclisme, à Sorbonne Université pour son système de paiement des vacations digne des douze travaux d'Astérix. Merci aussi à ChatGPT pour m'avoir écrit le premier paragraphe du manuscrit.

J'en arrive aux doctorants. À commencer par l'ancienne génération, Amaury, Lydie, Hanouk, Idriss, Gabriela, David, Cyril, Valentin, Élise, Alexandre P., merci de m'avoir accueilli dans le labo ! Merci à Alexandre R. pour les parties de frisbee, à Nicolas pour les baignades, merci aux gens du 15-16-301, bureau rarement vide (à ma connaissance du moins) : Olivier pour ton humour, Po-Yi pour ta paisibilité, Ludovic G.-C. pour ton style rhétorique et vestimentaire, Fatima pour ton ardeur au travail et tes conseils sportifs, Allen pour les blagues dans les mails du GTT, Gontran pour le bureau d'angle que tu m'as laissé, Ramón pour tes opinions

imperturbables, Dawei pour tes apparitions discrètes, Mehdi pour ta présence, Rui pour avoir supporté mon vélo à Arcachon, Nga, Mingyue et Eleanor pour vos discussions joyeuses et animées. Roxane, bon séjour au CEA, encore bravo Dr. Anatole, et Lucas E. bon courage pour la fin !

Merci à Juliette et Guillaume pour votre implication dans le comité environnement, à Yipeng, Noemi, Maria, Charles, Guillaume, Lucas, Ludovic et Zhe d'avoir pris la relève du GTT, à David, Yoan, Farf d'avoir organisé celui du LPSM, à Antoine, Eva et Antonio pour la coordination des doctorants, à Jules P. de nous avoir apporté ambiance, organisation et rémunération.

Merci au groupe des Italiens pour les notes festives, Emma, Giorgia, Noemi, Eugenio, Elena, y gracias a los hispano-hablantes Jesús, Ramón, Emilio, Nicolás, Cristobál, con quien es un placer charlar. Merci à Sylvain de nous avoir donné les nouvelles de Paris 7, aux anciens doctorants des Ponts, Gaspard, Rémi G., Inas, Rutger, Laurent, Jean, Louis-Pierre pour les déjeuners là-bas, à Tianrui, Jules G., Ioanna-Maria, Toai, Assane, Fabrice, pour les moments de détente en salle café, aux nouveaux, Kala, Alessandro, Ruikang, Jean-Guillaume, aux stagiaires Siguang et Marie, à tous les autres aussi ! Merci à Nicolaï, Lyangying, Alexiane, Pauline, pour les bons moments passés à discuter dans vos bureaux sans aucune envie de travailler, à Chourouk pour toutes les pistaches, fruits sec et biscuits dont tu m'as régalé, à Agustín pour ta perpétuelle bonne humeur, tes mimiques hilarantes, et ta motivation à faire avancer les projets.

Merci à Lise pour tous tes conseils, à Robin pour ton chill omniprésent, à Thomas pour cette incroyable croisière en bateau, puisses-tu faire le tour du monde à la voile ! Merci à Barbo pour le choix des chalets, à Julien et Yohan pour l'accueil à Grenoble, à Alfred malgré mes défaites à Smallworld, à Rémi pour l'ambition des sommets, et à Lucas J. pour l'élégance des descentes.

Merci au C4, à la coloc à 7 puis à 5 puis à 3 puis à 2 pour avoir concentré chez les derniers occupants une incroyable collection de tupperwares, à Malachi pour tes mail-feuilletons, à Béranger pour ta liberté de vie, à Lucas et Pauline pour votre pragmatisme face aux proprios et aux politiques irrationnels, à Robin et Lola pour les soirées chez vous, à Raphaël pour les maths, les nouilles faites maison et les animés sur le canapé, à Assil pour avoir rendu la coloc vivante, ludique et apaisante, à Adonaï pour ton optimisme face à toutes les épreuves.

Merci aux copains de terminale, Cécile et Christian pour les séjours à Nice et Noyers, Antoine et Sophie pour votre persévérance à faire des repas vegan, Samy et Moïse pour nos discussions tardives, les traquenards festifs et mathématiques, Thomas, François et Inès pour vos projets de vie si différents de ma bulle académique, et Pierre pour ton anticipation qui surpasse de loin ma procrastination.

Pour terminer, merci à toute la famille, à Jeff, Agnès, Dom, François, Bulle, Valérie, Guillaume, Marie, Matthieu, Louis-Alexandre, Manon et Maya, pour votre soutien sans faille, et pour l'admiration mêlée de scepticisme que vous témoignez vis-à-vis des maths. Merci à Françoise pour ton accueil à chacun de mes passages dans le sud, à Rémy pour tes expérimentations détonantes, à Paps et Mams pour l'énergie frénétique et le plaisir de travailler comme de se reposer que vous m'avez insufflés.

Et e e it ot e a in est ou oi, a ou i ou : e i ou es a an es, à o iou e et au é ec, a e ia, an, a ille et y ie ; e i ou ou e emps a é en em e, ou ou e on ai e et a on é, a e e e ai e a io é ent !

Échantillonnage optimal et réduction de modèle
Optimal Sampling and Model Order Reduction

Résumé

Cette thèse porte d'une part sur la conception de *modèles réduits* qui approchent optimalement des classes complexes de fonctions, et d'autre part sur l'utilisation de ces modèles réduits pour reconstruire des fonctions à partir d'un nombre limité de mesures, en particulier d'évaluations ponctuelles.
La partie I de la thèse traite de deux thématiques en réduction de modèle linéaire et non linéaire. Dans le chapitre 2, nous construisons des modèles réduits linéaires pour des EDP elliptiques paramétriques avec diffusion à fort contraste. Nous prouvons qu'un taux de décroissance exponentiel peut être obtenu dans ce cadre avec un modèle réduit linéaire, à la fois pour la simulation directe et pour les problèmes inverses, et ce malgré la dégénérescence des coefficients. Dans le chapitre 3, nous introduisons un cadre général pour la résolution de problèmes inverses avec des modèles réduits non linéaires, avec en particulier une application à la reconstruction de fonctions régulières par morceaux à partir de valeurs moyennes sur chaque maille.
La partie II de la thèse aborde le problème fondamental de l'approximation d'une fonction à partir de ses valeurs ponctuelles en des positions prédéfinies, en mettant l'accent sur les stratégies aléatoires et déterministes de sélection optimales de ces points. Le chapitre 4 étudie les problèmes numériques posés par la manipulation de la densité d'échantillonnage optimale, et propose des méthodes multi-niveaux de complexité algorithmique réduite, avec une analyse approfondie du cas de l'approximation par des polynômes multivariés dans des domaines généraux. Dans le chapitre 5, nous améliorons la stratégie d'échantillonnage aléatoire en ramenant la taille de l'échantillon au même ordre que la dimension du modèle réduit. Le chapitre 6 étudie un cadre déterministe, en supposant que la classe de fonctions est incluse dans la boule unité d'un espace de Hilbert à noyau reproduisant. Enfin, le chapitre 7 fait de nouvelles avancées dans le contexte aléatoire, en atteignant un ratio de sur-échantillonnage minimal, ce qui aboutit à de nouvelles estimées d'interpolation.
Les chapitres 2, 3, 4, 5 et 6 sont issus des articles [a], [b], [c], [d] et [e] respectivement, tandis que le dernier chapitre est un travail en cours.

Mots clés : réduction de modèle, échantillonnage aléatoire, problèmes inverses, moindres carrés à poids

Abstract

This thesis is concerned, on the one hand, with the design of *reduced order models* that optimally approximate complex classes of functions, and on the other hand with the use of such reduced models to recover functions from a limited amount of measurements, in particular point evaluations.
Part I of the thesis deals with two topics in linear and nonlinear reduced modeling. In Chapter 2, we construct linear reduced models for parametric elliptic PDEs with high contrast diffusion. We prove that exponential decay rates can be obtained in that setting with a linear reduced model, both for forward simulation and inverse problems, despite the degeneracy of the coefficients. In Chapter 3, we introduce a general framework for solving inverse problems with nonlinear reduced models, with a particular application to the reconstruction of piecewise smooth functions from cell-average data.
Part II of the thesis adresses the ubiquitous problem of approximating a function from its values at some predefined sample points, with a focus on optimal random and deterministic selection strategies of such points. Chapter 4 investigates the numerical issues arising in handling the optimal sampling density, and proposes multistep algorithms to control the computational cost of the method, with a thorough analysis of the case of approximation by multivariate polynomials on general domains. In Chapter 5, we improve the randomized sampling strategy by reducing the sample size to the same order as the reduced model dimension. Chapter 6 studies a deterministic setting, through the assumption that the class of functions is included in the unit ball of a reproducing kernel Hilbert space. Finally, Chapter 7 progresses further in the randomized context, achieving a minimal oversampling ratio, which culminates in novel interpolation estimates.
Chapters 2, 3, 4, 5 and 6 are based on the articles [a], [b], [c], [d] and [e] respectively, whereas the last chapter is an ongoing work.

Keywords: reduced modeling, randomized sampling, inverse problems, weighted least-squares

Laboratoire Jacques-Louis Lions
Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France

Table des matières

Chapter 1

Introduction

In many fields of science and engineering, simulations are used to study complex phenomena for which practical experiments are either impossible or prohibitively expensive. However, these models often require significant computational resources, making them too slow to run a large number of times. *Reduced modeling* [28, 56] refers to a variety of numerical techniques that aim at simplifying these complex models while retaining their essential features. The resulting reduced order models should be computationally efficient, in order to fully explore the behavior of the system when its physical parameters vary. Reduced models are thus typically used in order to accelerate forward numerical simulation when the solution should be queried for many parameter values. They are also of important use in inverse problems, when the parameters and the state are unknown, and one can access the solution only through a few measurements.

In general mathematical terms, one considers an unknown element u of some Banach space V. A priori knowledge on u is expressed through its membership to a compact class $\mathcal{K} \subset V$, which could account for the regularity of u, the physical laws it obeys, and in a broader sense the equations and bounds that u is expected to satisfy. We are interested in representing u in a simplified manner, using a small number n of real coefficients, from which an approximation $\tilde{u}$ will be calculated. In the next section, we introduce various tools for gauging the trade-off between simplicity and accuracy of this representation. Their study is at the core of the results presented in this thesis.

1.1 Complexity measures

The procedure for deriving $\tilde{u}$ can be decomposed as encoding and then decoding u with some continuous maps $E : V \to \mathbb{R}^n$ and $D : \mathbb{R}^n \to V$. Here E expresses the n coefficients of the reduced model as information queried on u, while D constructs a surrogate $\tilde{u}$ based on $E(u)$. The resulting approximation $\tilde{u} = D(E(u))$ should remain close to u. The accuracy of such a method is assessed in the norm $\|\cdot\|_V$ associated to V, through a uniform error bound

$$\max_{u \in \mathcal{K}} \|u - D(E(u))\|_V. \tag{1.1}$$

Note that the maximum is indeed attained, due to the compactness assumption on $\mathcal{K}$.

Defining the *reduced order model* as $V_n = D(\mathbb{R}^n)$, the search for maps E and D making (1.1) small can be reformulated as looking for a set $V_n \subset V$ and a continuous map $P : V \to V_n$ such that $\max_{u \in \mathcal{K}} \|u - Pu\|_V$ is as small as possible. If we assume that $D(E(u)) = u$ when $u \in V_n$, we may view $P = D \circ E$ as a (generally nonlinear) projection onto V_n.

As computations should be much faster in the reduced model V_n than in the whole space V, one is frequently led to impose some constraints on V_n, and therefore on D. For example, only linear maps D could be considered, resulting in linear spaces V_n, as illustrated in Figure 1.1. Even for nonlinear decoders, D is commonly assumed to be smooth, making V_n an n-dimensional differentiable manifold, see Figure 1.2. On the other hand, the choice of the encoder E depends on the measurements that can be performed on u. One might have access only to linear forms evaluated on u, or to a more restrictive class such as point evaluations when u is a function.

Optimizing over D and E with the various above-mentioned constraints defines so-called *n-widths* of the

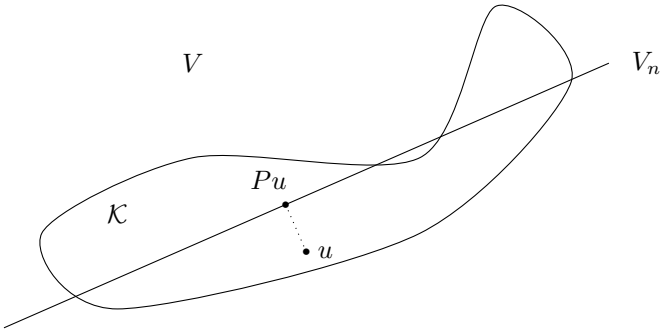


Figure 1.1 – Approximation by a linear reduced model V_n

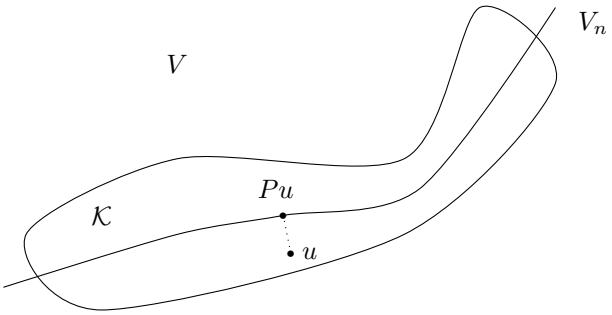


Figure 1.2 – Approximation by a nonlinear reduced model V_n. The increase in computational complexity caused by the nonlinear setting is expected to be compensated by an improved accuracy

class $\mathcal{K}$ (see [150], or [147] and Chapter 5 of [151] for equivalent definitions in operator theory), of the form

$$\inf_{E:V\to\mathbb{R}^n} \inf_{D:\mathbb{R}^n\to V} \max_{u\in\mathcal{K}} \|u - D(E(u))\|_V,$$

as summarized in the following Table 1.1.

Decoder / Encoder (E / D)	linear	nonlinear
nonlinear	d_n Kolmogorov widths	δ_n nonlinear widths
linear	a_n approximation numbers	s_n sensing numbers
point values	ρ_n sampling numbers	$\tilde{\rho}_n$ nonlinear sampling numbers

Table 1.1 – Different n-widths can be defined, depending on the constraints on the encoder E and decoder D. Observe that $d_n \leqslant a_n \leqslant \rho_n$, that all linear widths are larger than their nonlinear counterparts on the right column, and that all these numbers decrease as n increases

Probably the most popular are the *Kolmogorov n-widths* $d_n(\mathcal{K})_V$, which require D to be linear, and can be seen to satisfy

$$d_n(\mathcal{K})_V = \inf_{\substack{V_n \text{ linear} \\ \dim(V_n)=n}} \max_{u\in\mathcal{K}} \min_{v\in V_n} \|u - v\|_V \tag{1.2}$$

by taking the best approximation $P_n u \in \arg\min_{v\in V_n} \|u-v\|_V$. Note that, as $\mathcal{K}$ is compact, it can be covered by

balls of any radius $\varepsilon > 0$ centered at points $u_1, \dots, u_n \in \mathcal{K}$ for $n = n(\varepsilon)$ large enough. Taking V_n that contains $u_1, \dots, u_n$ shows that $d_n(\mathcal{K})_V \leqslant \varepsilon$, and therefore

$$d_n(\mathcal{K})_V \underset{n\to\infty}{\longrightarrow} 0.$$

One alternative to the Kolmogorov widths consists in imposing that E is also linear, which leads to the *approximation numbers*

$$a_n(\mathcal{K})_V = \inf_{\substack{P\in\mathcal{L}(V,V)\\ \operatorname{rank}(P)\leqslant n}} \max_{u\in\mathcal{K}} \|u - Pu\|_V,$$

where $\mathcal{L}(V, W)$ denotes the set of continuous linear operators from V to W. If V is a Hilbert space, the optimal projection P_n is linear, so $a_n(\mathcal{K})_V = d_n(\mathcal{K})_V$.

An important aspect for quantifying the effectiveness of linear reduced modeling is to determine how fast the Kolmogorov widths of $\mathcal{K}$ decay. Such decay properties are well understood for standard regularity classes [67, 167, 171, 173], and have more recently been studied for classes of solutions to parametric elliptic Partial Differential Equations (PDEs), under the systematic assumption that ellipticity holds uniformly over the parameter domain [20–22, 84, 174]. In Chapter 2, we address this question when the uniform ellipticity assumption is no more valid, namely for a stationary diffusion equation with degenerate coefficients, that correspond to the so-called high-contrast regime.

In some situations, the Kolmogorov widths may decrease too slowly, and a natural approach to tackle this issue consists in using nonlinear reduced models. This leads to the notion of *nonlinear widths* (or manifold widths) [169]

$$\delta_n(\mathcal{K})_V = \inf_{E\in C^0(V,\mathbb{R}^n)} \inf_{D\in C^0(\mathbb{R}^n,V)} \max_{u\in\mathcal{K}} \|u - D(E(u))\|_V,$$

Note that $V_n = D(\mathbb{R}^n)$ is now a nonlinear space. One important variant, called *stable nonlinear widths* [57] and denoted $\delta_n^L(\mathcal{K})_V$, imposes in addition that E and D are both L-Lipschitz continuous for a certain choice of norm on $\mathbb{R}^n$.

An important chapter of nonlinear approximation, that was highlighted by the theory of Compressed Sensing [42, 72], consists in imposing that encoding is made by linear measurements while decoding could be nonlinear. This brings us to introduce the *sensing numbers*

$$s_n(\mathcal{K})_V = \inf_{E\in\mathcal{L}(V,\mathbb{R}^n)} \inf_{D:\mathbb{R}^n\to V} \max_{u\in\mathcal{K}} \|u - D(E(u))\|_V. \tag{1.3}$$

There is a close connection between s_n and the *Gelfand widths*, defined as

$$g_n(\mathcal{K})_V := \inf_{E\in\mathcal{L}(V,\mathbb{R}^n)} \max_{\substack{u\in\mathcal{K}\\ E(u)=0}} \|u\|_V.$$

One can observe, as in [52], that

$$s_n(\mathcal{K})_V \leqslant g_n(\mathcal{K} - \mathcal{K})_V \leqslant 2\, s_n(\mathcal{K})_V.$$

In particular, when $\mathcal{K}$ is convex and centrally symmetric, a Hahn-Banach extension of $\mathrm{id}_{\operatorname{Ker} E}$ to V in the norm $\|v\|_{\mathcal{K}} := \min\{\lambda \in [0,\infty] \ : \ v/\lambda \in \mathcal{K}\}$ shows that $s_n(\mathcal{K})_V = g_n(\mathcal{K})_V$.

Remark 1.1. In the definition of sensing numbers, we can omit the assumption that D is continuous, without changing the value of s_n. Indeed, for any $\varepsilon > 0$, covering $\mathcal{K}$ by a finite union of sets $U_z = \{u + v \ : \ u \in \mathcal{K}, E(u) = z \text{ and } \|v\|_V \leqslant \varepsilon\}$, and applying a partition of unity on the covering of $E(\mathcal{K})$ by the $E(U_{z_i})$, one can define a continuous decoder D_ε interpolating any decoder D at points z_i, with an additional error at most ε. This will be useful in Chapter 3, where we prove bounds on s_n by the construction of a possibly discontinuous decoder.

In Chapter 3, we discuss how nonlinear reduced models can be used in the context of inverse problems, and establish theoretical bounds on the achievable accuracy for a given set of linear or nonlinear measurements. Indeed, in many applications, not all measurements of u are accessible, thus optimal encoders E are out of reach [131]. As a particular application, we discuss the performance of nonlinear reconstructions for classes $\mathcal{K}$ of piecewise smooth functions from their encoding by cell averages, and show that it decays faster than the

Kolmogorov widths of $\mathcal{K}$. This means that s_n and δ_n decay significantly faster than a_n and d_n for such $\mathcal{K}$.

Finally, a typical setting, in the case where u is a real or complex valued function defined on some domain Ω, is when encoding of u is perfomed by point evaluations only [83]. This naturally leads to the notions of *linear sampling numbers* that are defined as

$$\rho_n(\mathcal{K})_V = \inf_{x^1,\dots,x^n\in\Omega} \inf_{D\in\mathcal{L}(\mathbb{R}^n,V)} \max_{u\in\mathcal{K}} \|u - D(u(x^1),\dots,u(x^n))\|_V, \tag{1.4}$$

Once again we may relax the assumption that D is linear, leading to the *nonlinear sampling numbers* $\tilde{\rho}_n(\mathcal{K})_V$. Another important variant that will be defined further are the randomized sampling numbers, that correspond to the case where the x^i are picked at random and error is measured in an expectation sense.

This practical situation of point evaluations is the main focus of the rest of this thesis. Part II, composed of Chapters 4, 5, 6 and 7, explores new bounds on sampling numbers. In these contributions, decoding/reconstruction strategies are linear and very classical, such as interpolation or least-squares. The main contribution lies in the careful selection of the points where u should be evaluated, that is in the encoding/sampling strategy, a topic which has been the object of intensive research in recent years [143, 144, 167, 184]. In particular, we provide in various contexts estimates for the sampling numbers $\rho_n(\mathcal{K})_V$ that compare favorably to the Kolmogorov widths, when V is either $L^2(D)$ or $L^\infty(D)$.

Let us stress that there also exist many nonlinear reconstruction methods based on point sampling, and therefore falling in the category described by the nonlinear sampling numbers $\tilde{\rho}_n$ [162, 189]. We may single out the very active area of neural network approximation, with recent progress on learning Banach subsets, see [4] and the references therein. However, nonlinear sampling numbers $\tilde{\rho}_n$ do not compare as favorably to nonlinear widths δ_n as in the linear case, due to a theory-to-practice gap in deep learning [82]. As the underlying mathematical concepts are also quite different from the ones studied here, these methods are left outside the scope of this thesis.

Remark 1.2. Here we emphasized the role of reduced order models based on continuous coefficients $c \in \mathbb{R}^n$. Another approach consists in encoding information on u in a discrete way, by replacing $\mathbb{R}^n$ with a finite set $\{1,\dots,N\}$. Appraising the accuracy of this technique is the goal of covering numbers, packing numbers, and the related entropy numbers, that have natural connections with the complexity measures described above [45].

In the next sections, we go into further detail on each of the previously mentioned themes, and summarize the content of each chapter.

1.2 Reduced modeling for parametric elliptic PDEs

Part I of the thesis begins with Chapter 2, based on our article [a], which deals with linear reduced modeling for parametric elliptic PDEs with high contrast diffusion coefficients.

When simulating physical phenomena, u often corresponds to an intensive quantity, such as a temperature, pressure, concentration, or a velocity field. Besides the spatial coordinates $x \in \Omega$, u may depend on some parameters gathered in a vector $y \in Y \subset \mathbb{R}^d$, which could account both for physical properties of the materials, and for geometric variables describing for instance the shape of the domain Ω and its boundary conditions. Modeling this function $u(y)$ by a PDE

$$\mathcal{P}(u,y) = 0, \quad x \in \Omega,$$

creates a class of solutions

$$\mathcal{K}_Y = \{u(y) \ : \ y \in Y\} \tag{1.5}$$

in the Banach space V for which the PDE is well-posed.

One typically has access to a high fidelity numerical solver, which computes $u(y)$ for any given parameter vector $y \in Y$. Such solvers are computationally costful, which is particularly problematic in the "many query" context, that is, when the solution is needed for many different y. This motivates the offline search of a reduced model V_n for $\mathcal{K}$, that aims at collectively approximating all solutions as best as possible for the given dimension n. The two prominent approaches are the reduced basis method [84, 157, 160] and the principal orthogonal decomposition [46, 180, 190]. The main challenges arising in this context consist in

- estimating the Kolmogorov widths $d_n(\mathcal{K})_V$ from (1.2) that corresponds to the performance of optimal, and often out of reach, n-dimensional spaces;
- constructing a reduced space V_n such that the distance $\max_{u\in\mathcal{K}} \min_{v\in V_n} \|u-v\|_V$ from $\mathcal{K}$ to V_n compares favorably with d_n or satisfies similar decay estimates;
- once V_n is fixed, and given $y \in Y$, computing a surrogate $\tilde{u}(y) \in V_n$ of $u(y)$, with an error $\|u(y)-\tilde{u}(y)\|_V$ again satisfying a similar bound.

The online computation of the surrogate $\tilde{u}(y)$ is typically performed by the Galerkin method in the space V_n, that amounts to solving a system of moderate size n, resulting in substantial computational savings compared to the high fidelity solver.

Chapter 2 concentrates on an archetypal example of parametrized elliptic problem

$$\mathcal{P}(u,y) := f + \operatorname{div}(a(y)\nabla u) = 0, \quad x \in \Omega \tag{1.6}$$

modeling the stationary solution $u \in V := H^1_0(\Omega)$ of a diffusion equation in a heterogeneous domain, with source term $f \in V' = H^{-1}(\Omega)$, diffusion coefficient $a(y) \in L^\infty(\Omega)$, and homogeneous Dirichlet boundary conditions $u|_{\partial\Omega} = 0$.

Here, Ω is a fixed smooth spatial domain, and we assume that f is independent from the parameters y, because u depends linearly on f, hence a parametrized source term would pose no particular difficulty. On the contrary, the evolution of u when a varies is much more complex. We restrict ourselves to a piecewise constant geometry $a(y) = y_j$ on Ω_j, where $\{\Omega_1, \dots, \Omega_d\}$ is a fixed partition of Ω and $y \in Y \subset (0,\infty)^d$, as illustrated in Figure 1.3.

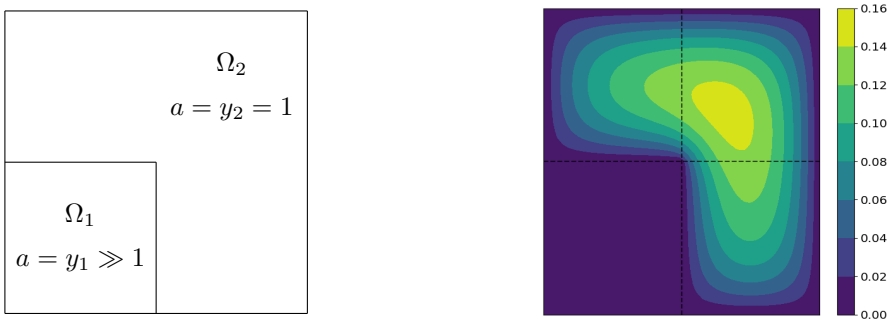


Figure 1.3 – Example of a partition of $\Omega = [-1,1]^2$ into $d = 2$ subdomains $\Omega_1 = [-1,0]^2$ and $\Omega_2 = \Omega \setminus \Omega_1$. If we take a source term $f = 1$ and a piecewise constant diffusion coefficient $a = y_j$ on Ω_j with $y_1 \gg 1$ and $y_2 = 1$, the solution $u(y)$, plotted on the right, is almost constant on Ω_1, with a value close to zero due to the homogeneous Dirichlet boundary conditions

This problem has been widely studied in recent years, see in particular [20, 22, 174], who all work under a uniform ellipticity assumption

$$Y \subset [a_{\min}, a_{\max}]^d, \quad 0 < a_{\min} < a_{\max} < \infty.$$

Even when the parameter dimension d goes to infinity, sparse approximations [3, 21, 56, 58] have been proved to achieve polynomial decay rates, for any affine parametrization of a as a function of y. For finite d, exponential rates can be obtained, by studying certain sparse polynomial expansions of $u(y)$ with respect to the parameter coordinates $(y_j)_{1\leqslant j\leqslant d}$. This is the strategy adopted in [20], yielding the following result:

Theorem 1.3. *If the diffusion coefficient a is piecewise constant with value y_j on Ω_j (or any affine function $a(y) = \overline{a} + \sum_{j=1}^d y_j\psi_j$) and $a_{\min} \leqslant a(x,y) \leqslant a_{\max}$ for all $x \in \Omega$ and $y \in Y$, then*

$$d_n(\mathcal{K}_Y)_V \leqslant C\exp(-cn^{1/d}),$$

where C and c are positive constants depending on d, $a_{\min}$, $a_{\max}$, and the geometry of the partition (or on the

affine basis functions $\overline{a}$ and $\psi_1, \dots, \psi_d$).

However, the constraint that a is bounded from above and below is quite stringent, as the constant C degrades proportionally to the level of contrast $a_{\max}/a_{\min}$. For heat diffusion in composite materials, metallic and cristalline parts may have conductivities hundreds or thousands times higher than organic compounds or air. An even worse gap occurs in modeling the diffusion of a contaminant in underground water flows, as the diffusivity is much higher in flooded cavities than in the surrounding porous rocks.

Many techniques have been proposed to avoid deterioration of the convergence of numerical solvers when the level of contrast increases. They range from preconditionners [9, 10, 75] to a posteriori error estimation [8, 30], not forgetting domain decomposition methods adapted to drastic changes in the coefficients [76, 77] or even to changes in the nature of the PDE [78]. The main novelty of Chapter 2 is to establish such robust estimates for model order reduction.

A first step consists in observing that, due the homogeneity property $u(y) = tu(ty)$, we only have to consider the case of large values $y \to \infty$. We thus assume, up to a rescaling, that $y_j \geqslant 1$ for all $1 \leqslant j \leqslant d$, and show a uniform reduced model error

$$\sup_{y \in [1,\infty)^d} \|\tilde{u}(y) - u(y)\|_V \leqslant \varepsilon.$$

This results in a similar uniform error bound, in the relative error sense

$$\sup_{y \in (0,\infty)^d} \frac{\|\tilde{u}(y) - u(y)\|_V}{\|u(y)\|_V} \leqslant \varepsilon,$$

over the full range of parameter $Y = (0,\infty)^d$, which is therefore robust to arbitrarily high contrast.

One key ingredient to our analysis is the fact that the solution u to the variational problem associated to (1.6),

$$\int_\Omega a(y) \nabla u \cdot \nabla v = \int_\Omega f v, \quad v \in V = H^1_0(\Omega),$$

converges, when some of y_j tend to infinity, to the solution $u_S \in V_S$ of a limit problem

$$\int_{\Omega^c_S} a(y) \nabla u_S \cdot \nabla v = \int_\Omega f v, \quad v \in V_S,$$

where S is the set of indices j such that $y_j = \infty$, $\Omega_S = \bigcup_{j \in S} \Omega_j$ is the part of Ω where $a(y)$ is infinite, and

$$V_S = \{v \in V \ : \ \nabla v = 0 \text{ on } \Omega_S\}.$$

The weak convergence of $u(y)$ towards u_S follows from a compactness argument, which can be found in Chapter 3 of [101] in the context of homogeneization. To obtain strong convergence, one additionally needs the convergence of the energy norm

$$\|v\|^2_y := \int_\Omega a(y) |\nabla v|^2.$$

A more detailed convergence analysis when there are only $d = 2$ subdomains can be found in [41], where asymptotic expansions up to any order are performed. One of our main contributions is to quantify more precisely the rate of this strong convergence.

The central idea for manufacturing a reduced model V_n of $\mathcal{K}_Y$ is then to use polynomial expansions adapted from [20] in subsets of the parametric domain $Y = [1,\infty]^d$ corresponding to moderate values of y, and to take advantage from the proximity between $u(y)$ and a limit solution when some of the y_j are large. Our main result is in turn the following.

Theorem 1.4. *For $Y = [1,\infty]^d$ and $a(y) = y_j$ on Ω_j, the class $\mathcal{K}_Y$ defined by (1.5) satisfies*

$$d_n(\mathcal{K})_V \leqslant C \exp(-c n^{1/2d})$$

for some constants $C, c > 0$ that depend on d and the geometry of the partition.

As the proof is constructive, it also provides a space V_n achieving the following bound, however at the

expense of computing derivatives of the solutions $u(y)$ with respect to y. Classical reduced models, selected as subspaces of the span of a limited number of solutions $u(y^1), \dots, u(y^N)$, are nevertheless proved to work at least as well [33, 40, 65, 179], and in practice reach much better rates than the pessimistic bound in $\exp(-cn^{1/2d})$.

Concerning the construction of the surrogate $\tilde{u}(y)$, the best approximation would be the H_0^1-orthogonal projection of $u(y)$ onto V_n, which is accessible only by first computing $u(y)$, which is exactly what we want to avoid. A more efficient approach is to compute the Galerkin projection, that is, the orthogonal projection of $u(y)$ onto V_n with respect to the energy norm $\|\cdot\|_a$. This only asks to solve the PDE in the reduced model, which comes down to solving an $n \times n$ linear system. We show estimates of the same kind as Theorem 1.4 for the error of approximation by Galerkin projection, despite the fact that the high-contrast regime does not allow a straightforward use of Cea's lemma.

Finally, as a consequence of the above results, we demonstrate that the linear Parametrized Background Data-Weak (PBDW) method [33, 124] allows to solve the inverse problem of recovering an unknown state u or the underlying parameters y, based on a few measurements of the solution.

1.3 Nonlinear reduced modeling and inverse problems

Chapter 3 is based on article [b], which deals with the use of nonlinear reduced models for solving inverse problems.

In full generality, we address the inverse problem of recovering an unknown function $u \in V$ from m linear measurements performed by given functionals $\ell_1, \dots, \ell_m : V \to \mathbb{R}$. The properties of u are again modeled by its membership in a class $\mathcal{K}$ that could be the solution manifold of a parametric PDE.

The Parametrized Background Data-Weak (PBDW) method consists in introducing a linear reduced model space V_n for $\mathcal{K}$, with $n \leqslant m$, and defining the recovery as the optimizer u^* in the pair

$$(u^*, \tilde{u}) \in \operatorname*{arg\,min}_{v^* \in V_z,\, \tilde{v} \in V_n} \|v^* - \tilde{v}\|_V,$$

where $V_z = \{v \in V \ : \ \ell(v) = z\}$ is the space of codimension m of all elements $v \in V$ that have the same measurements $z = \ell(u)$ as the unknown element u, with the notation $\ell = (\ell_1, \dots, \ell_m) : V \to \mathbb{R}^m$. The PBDW estimator u^* entirely trusts the data, expressed through V_z, and considers the parametrized space V_n as an inexact approximation, contrarily to the best fit estimator $\tilde{u}$ which heavily relies on the precision of V_n, while allowing noisier measurements.

If we assume that V is a Hilbert space and that $\ell \in \mathcal{L}(V, \mathbb{R}^m)$ is linear, one can take Riesz representers ω_i of the ℓ_i

$$\ell_i(v) = \langle \omega_i, v \rangle_V, \quad 1 \leqslant i \leqslant m, \quad v \in V,$$

and characterize the measurements as the orthogonal projection P_{W_m} onto the observation space

$$W_m = \operatorname{span}\{\omega_1, \dots, \omega_m\},$$

see Figure 1.4. Equivalently, we have $V_z = u + W_m^\perp$ and we could have defined

$$u^* = \operatorname*{arg\,min}_{v \in V_z} \|v - P_{V_n} v\|_V \quad \text{and} \quad \tilde{u} = \operatorname*{arg\,min}_{v \in V_n} \|P_{W_m}(u - v)\|_V.$$

If moreover V_n is a linear space, it was proved in [33, 124] that these estimators are near-optimal:

Theorem 1.5. *Given V_n and noiseless observations $P_{W_m} u$ of u, the estimators u^* and $\tilde{u}$ satisfy*

$$\max(\|u - u^*\|_V, \|u - \tilde{u}\|_V) \leqslant \mu_n^m \min_{v \in V_n} \|u - v\|_V,$$

where

$$\mu_n^m = \max_{v \in V_n} \frac{\|v\|}{\|P_{W_m} v\|_V}.$$

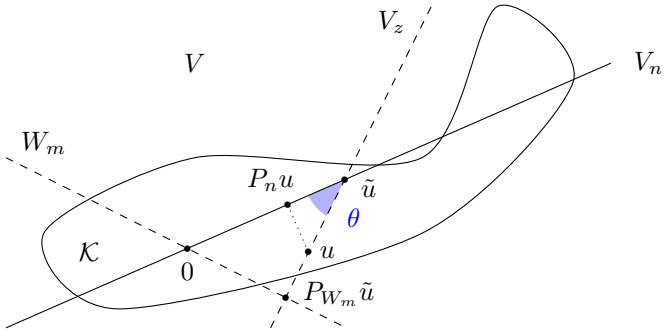


Figure 1.4 – Approximation of u in a linear reduced model V_n, based on measurements by projection on W_m.

The quantity μ_n^m may be thought of as the cosecant $1/\sin(\theta)$ of the minimal angle θ between $W_m^\perp$ (or the affine space V_z) and V_n. In other words, μ_n^m indicates how close V_n is from W_m. In particular, $\mu_n^m = 1$ if $V_n \subset W_m$, in which case the optimal approximation $\tilde{u} = P_{V_n} u$ can be computed from the observations $P_{W_m} u$. On the opposite, if there exists a non-zero $v \in V_n \cap W_m^\perp$, the corresponding coordinate $\langle u, v \rangle_V$ cannot be deduced from the measurements, and $\mu_n^m = \infty$.

Remark 1.6. In Figure 1.4, $\sin\theta$ corresponds both to the ratio between optimal error $\|u - P_n u\|_V$ and actual error $\|u - \tilde{u}\|_V$, and to the fraction $\|P_{W_m}\tilde{u}\|_V / \|\tilde{u}\|_V$. This motivates the definition of the stability constant μ_n^m in Theorem 1.5

In Chapter 3, we establish an extension of the above result to nonlinear spaces V_n in general Banach spaces V. For a linear measurement map $\ell : V \to \mathbb{R}^m$, Theorems 3.3, 3.5 and 3.8 imply:

Theorem 1.7. *Given V_n and noiseless linear observations $\ell(u)$ of u, the estimators u^* and $\tilde{u}$ satisfy*

$$\max(\|u - u^*\|_V, \|u - \tilde{u}\|_V) \leqslant (1 + 2\mu_n^W) \min_{v \in V_n} \|u - v\|_V$$

where

$$\mu_n^W = \sup_{v_1, v_2 \in V_n} \frac{\|v_1 - v_2\|_V}{\|\ell(v_1) - \ell(v_2)\|_W},$$

for the norm $\|z\|_W := \min_{v \in V_z} \|v\|_V$ on $\mathbb{R}^m$.

Note that μ_n^W is the inverse Lipschitz stability constant of ℓ in V_n. It coincides with the factor μ_n^m from the previous theorem under the corresponding assumptions, as illustrated on Figure 1.5.

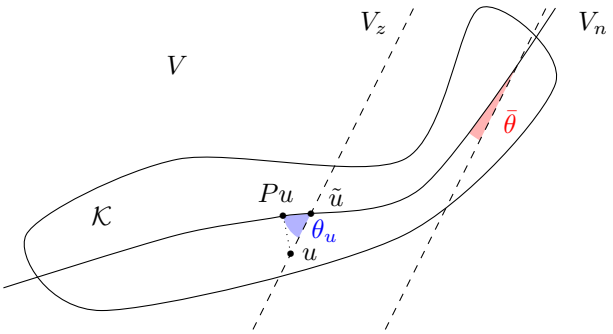


Figure 1.5 – Approximation of u in a nonlinear reduced model V_n, based on linear measurements expressed by V_z. The stability constant is again inversely proportional to $\sin\theta_u$, which we bound by considering the minimal angle $\bar{\theta} = \min_{u \in \mathcal{K}} \theta_u$. As V_n is nonlinear, these angles are in fact defined by considering differences between two elements of V_n, see Theorem 1.7

Furthermore, in the case of noisy observations $z = \ell(u) + \eta$, we obtain similar bounds, with an additional term proportional to $\|\eta\|_p$, the norm of the noise in $\ell^p(\mathbb{R}^m)$, for any $p \geqslant 1$. When the level of noise grows, a norm $\|\cdot\|_Z$ different from $\|\cdot\|_W$ should be chosen in Theorem 1.7, to preserve a balance between these two terms.

The above analysis is then carried on to a concrete example in Sections 3.4 and 3.5: one wishes to recover a piecewise smooth function $u : [0,1]^2 \to \mathbb{R}$ from its local averages $\fint_T u$ on cells of the form $T_{i,j} = [(i-1)h, ih] \times [(j-1)h, jh]$, where $h > 0$ is the cell size and $1 \leqslant i, j \leqslant 1/h$. This typically applies to cartoon images, for which one may construct a fine resolution image based on a pixelized version. Another domain where this example could be of interest is the numerical resolution of hyperbolic equations with shock singularities.

Assuming that u is in fact piecewise constant with values 0 and 1, the strategy consists in looking for the best approximation of u by the indicator of a half-plane, on each stencil of 3×3 cells. We show in Proposition 3.15 that $\mu_n^W = 3/2$, which implies the following bound, adapted from Theorem 3.18:

Theorem 1.8. *If u the indicator of a smooth domain in $[0,1]^2$, we can construct an approximation $\tilde{u}$ based on its average values on $n = h^{-2}$ cells such that*

$$\|u - \tilde{u}\|_{L^q}^q \leqslant \frac{C}{n}.$$

This is better than the rate $1/\sqrt{n}$ achieved by a linear method, this rate being optimal for $q = 2$. In other words, we proved that for the class $\mathcal{K}$ of indicators of smooth domains, the sensing numbers defined in (1.3) decrease as

$$s_n(\mathcal{K})_{L^2} \lesssim n^{-1/2},$$

whereas the Kolmogorov widths (or approximation numbers) only decay as $d_n(\mathcal{K})_{L^2} \approx n^{-1/4}$.

Finally, in Section 3.6, we show that for the space V_n of n-sparse vectors in $V = \mathbb{R}^N$, optimal accuracy of the reconstruction is equivalent, up to a factor 2 in the constants, to the null space property [52], which is known to certify the quality of the measurements in compressed sensing (it is in particular implied by the the restricted isometry property [43, 72]).

1.4 Weighted least-squares

In Part II of the thesis, we investigate more specifically linear recovery methods from point value data.

Thus the information on u consists of its evaluations at a few selected points $x^1, \ldots, x^m$, possibly affected by noise. The critical aspect of the setting we consider is that the user is allowed to pre-select the sample points. This situation happens:

— in the design of physical experiments, if u is a spatially-dependent physical quantity, and one can pick the positions $x^1, \ldots, x^m$ of sensors measuring $u(x^i)$,

— in the model reduction of a parametric PDE, whose solution u is seen as a function of the parameters y, since the user can pick the set of parameters y^i when lauching the fine numerical solver.

This *active learning* setting is in shear contrast to the regression problem in statistical learning, where the data points are drawn from an unknown underlying distribution. It puts in the forefront the problem of optimally selecting the sampling point, as already hinted in the definition (1.4) of the sampling numbers ρ_n.

As m represents a number of sensors or a number of expensive numerical solves, one is interested in establishing convergence estimates that compare favorably to the accuracy of the reduced model V_n, of given dimension n, that is used for the reconstruction, and hold for the smallest possible sampling budget m, if possible of the same order as n.

Before detailing the contents of each chapter of Part II, let us briefly present a simplified version of the main result of [59], see § 4.2.2, boosted by adding a conditioning step from [85], with a strategy of proof following the same lines as [d]. The proof is essentially summed up in Lemma 5.2, and the main result of this section is almost a copy of Lemma 5.4.

Let Ω be a multivariate domain equipped with a measure μ, and let

$$V := L^2(\Omega, \mu).$$

Let $V_n \subset V$ be a given linear space of dimension $n \in \mathbb{N}$. We seek to approximate an unknown function $u \in V$ by a surrogate $\tilde{u} \in V_n$, based on its point values. Take $(\varphi_1, \dots, \varphi_n)$ an orthonormal basis of V_n, and define

$$k_n(x) = \frac{1}{n} \sum_{j=1}^{n} |\varphi_j(x)|^2, \quad x \in \Omega, \tag{1.7}$$

the inverse Christoffel function, up to the scaling factor $1/n$. For $m \geqslant n$ to be determined later, we draw points $x^1, \dots, x^m \in \Omega$ i.i.d according to the sampling measure $k_n(x)\, d\mu(x)$. Moreover, we define weights $w_i = 1/k_n(x^i)$, and a discrete semi-norm

$$\|v\|_m^2 = \frac{1}{m} \sum_{i=1}^{m} w_i |v(x^i)|^2, \quad v \in V.$$

In the noise-free case, the weighted least-squares approximation

$$P_n^m u = \operatorname*{arg\,min}_{v \in V_n} \|u - v\|_m$$

is the orthogonal projection of u onto V_n with respect to the norm $\|\cdot\|_m$. This definition should be viewed in an almost sure sense, when applied to a function representative of $u \in L^2(\Omega, \mu)$.

We introduce the random Gram matrix

$$G_m := (\langle \varphi_j, \varphi_k \rangle_m)_{1 \leqslant j,k \leqslant n} = \frac{1}{m} \sum_{i=1}^{m} a_i a_i^\dagger, \tag{1.8}$$

where $a_i = \sqrt{w_i}\, \varphi(x^i) \in \mathbb{C}^n$. Following [85], we define a conditional estimator

$$\tilde{u} = \mathbb{E}\left(P_n^m u \,\middle|\, G_m \succcurlyeq \frac{1}{2} I \right),$$

where $A \succcurlyeq B$ means that $A - B$ is a positive semi-definite matrix. The estimator $\tilde{u}$ can be computed by redrawing samples $(x^1, \dots, x^m)$ until the event $G_m \succcurlyeq \frac{1}{2} I$ occurs.

Theorem 1.9. *For $m \geqslant 10\, n \ln(2n)$, the expected numbers of redraws is at most 2, and*

$$\mathbb{E}\left(\|u - \tilde{u}\|_{L^2}^2 \right) \leqslant 5 \min_{v \in V_n} \|u - v\|_{L^2}, \tag{1.9}$$

The main tool for proving Theorem 1.9 is the following matrix Chernoff inequality, originally proved by [7]. A review on matrix concentration inequalities can be found in [176], and the version given below is from [c, d], see Lemmas 4.1 and 5.3.

Proposition 1.10. *Let $a_1, \dots, a_m \in \mathbb{C}^n$ be i.i.d random vectors such that $\mathbb{E}(a_i a_i^\dagger) = I$ and $|a_i|^2 \leqslant n$ a.s. Then*

$$\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{m} \sum_{i=1}^{m} a_i a_i^\dagger \right) < \frac{1}{2} \right) \leqslant n \exp\left(-\frac{m}{10n} \right).$$

Proof of Theorem 1.9. The random vectors $a_i = \sqrt{w_i}\, \varphi(x^i)$ are i.i.d, bounded since $|a_i|^2 = n$ by definition of the w_i, and satisfy

$$\mathbb{E}(a_i a_i^\dagger) = \int_\Omega \frac{1}{k_n(x^i)} \varphi(x^i) \varphi(x^i)^\dagger k_n(x^i)\, d\mu(x^i) = \int_\Omega \varphi \varphi^\dagger d\mu = I.$$

Applying Proposition 1.10 with $m \geqslant 10 n \ln(2n)$, we see that the probability $p := \mathbb{P}(G_m \succcurlyeq \frac{1}{2} I)$ is at least $\frac{1}{2}$. The expected number of redraws is therefore $1/p \leqslant 2$.

For the estimate (1.9), recalling that $P_n u$ is the orthogonal projection of u onto V_n for the norm $\|\cdot\|_V = \|\cdot\|_{L^2}$,

$$\mathbb{E}\left(\|u-\tilde{u}\|_{L^2}^2\right) = \|u-P_n u\|_{L^2}^2 + \mathbb{E}\left(\|\tilde{u}-P_n u\|_{L^2}^2\right). \tag{1.10}$$

By definition of the estimator $\tilde{u}$,

$$\mathbb{E}\left(\|\tilde{u}-P_n u\|_{L^2}^2\right) = \mathbb{E}\left(\|P_n^m u - P_n u\|_{L^2}^2 \,\middle|\, G_m \succcurlyeq \frac{1}{2} I\right). \tag{1.11}$$

Decomposing $v = P_n^m u - P_n u \in V_n$ as $v = \sum_{j=1}^n \nu_j \varphi_j$, observe that

$$\|v\|_m^2 = \nu^* G_m \nu \geqslant \lambda_{\min}(G_m)\, \nu^*\nu \geqslant \frac{1}{2}\|v\|_{L^2}^2 \quad a.s.,$$

where we used the fact that $\lambda_{\min}(G_m) \geqslant \frac{1}{2}$ almost surely holds for the conditioned sample. Thus

$$\|P_n^m u - P_n u\|_{L^2}^2 \leqslant 2\,\|P_n^m u - P_n u\|_m^2 \leqslant 2\,\|u-P_n u\|_m^2 \quad a.s., \tag{1.12}$$

since $P_n^m u - P_n u = P_n^m(u - P_n u)$ is the orthogonal projection of $u - P_n u$ onto V_n for the discrete norm $\|\cdot\|_m$. Finally, for any random variable $X \geqslant 0$ and event E with $\mathbb{P}(E) > 0$, we have

$$\mathbb{E}(X|E) = \frac{\mathbb{E}(X\chi_E)}{\mathbb{P}(E)} \leqslant \frac{\mathbb{E}(X)}{\mathbb{P}(E)}.$$

Hence,

$$\mathbb{E}\left(\|u-P_n u\|_m^2 \,\middle|\, G_m \succcurlyeq \frac{1}{2} I\right) \leqslant \frac{1}{p}\,\mathbb{E}\left(\|u-P_n u\|_m^2\right) \leqslant 2\,\|u-P_n u\|_{L^2}^2. \tag{1.13}$$

We conclude by combining the above bounds (1.10), (1.11), (1.12) and (1.13). □

Theorem 1.9 can be rephrased in terms of *randomized sampling numbers* [185]

$$\rho_m^{\mathrm{rand}}(\mathcal{K})_V^2 = \inf_{\sigma_m \in \mathrm{Prob}(\Omega^m)} \inf_{D:\Omega^m \to \mathcal{L}(\mathbb{R}^m, V)} \max_{u\in\mathcal{K}} \mathbb{E}\left(\|u - D_{(x^1,\ldots,x^m)}(u(x^1),\ldots,u(x^m))\|_V^2\right), \tag{1.14}$$

where $\mathrm{Prob}(\Omega^m)$ is the set of probability measures on Ω^m, and the expectation is taken over the sample $(x^1,\ldots,x^m)$ of law σ_m. Note that these randomized numbers are smaller than the deterministic sampling numbers $\rho_m(\mathcal{K})_V$ from Table 1.1, defined in (1.4), since a deterministic sample may be thought of as a particular instance of random sample.

Corollary 1.11. *If $m \geqslant 10\, n \ln(2n)$, for any compact subset $\mathcal{K}$ of $V = L^2(\Omega,\mu)$,*

$$\rho_m^{\mathrm{rand}}(\mathcal{K})_{L^2}^2 \leqslant 5\, d_n(\mathcal{K})_{L^2}^2.$$

Remark 1.12. The main result in [59] proposes some improvements over the simplified Theorem 1.9: in particular, the conditioning step can be avoided if $u \in L^\infty$, by considering a truncated estimator. Moreover, the factor 5 is replaced by $1 + 8\,n/m$, and this new factor 8 could be reduced to any constant larger than 1, by increasing the ratio between m and $n\ln(n)$. This leads to a stability constant $1 + n/m + o(n/m)$ as $m \to \infty$ for a fixed value of n, which is optimal for i.i.d sampling methods.

Remark 1.13. In the definition (1.14) of $\rho_m^{\mathrm{rand}}(\mathcal{K})_V$, we did not enforce the sample points to be i.i.d or even independent, which would transcribe as $\sigma_m = \sigma^{\otimes m}$ for $\sigma \in \mathrm{Prob}(\Omega)$, or as $\sigma_m \in \mathrm{Prob}(\Omega)^m$, respectively. Indeed, the conditioning procedure does not preserve the independence of the points, and there always is a probability of failure for an algorithm based on i.i.d points.

1.5 Christoffel function on general domains

Chapter 4 is based on our article [c], and investigates the computational issues arising from the use of the Christoffel function on general domains, as well as multistep algorithms circumventing these difficulties.

One main computational challenge for applying the above least-squares strategy is to compute an orthonormal basis $(\varphi_1, \dots, \varphi_n)$ of V_n in $L^2(\Omega, \mu)$. This basis is essential for evaluating the Gram matrix G_m (1.8), and even more crucially for computing the inverse Christoffel function $k_n(x)$ (1.7), which is needed both as the optimal sampling density, and for assigning the correct weights $w_i = 1/k_n(x^i)$.

On general domains Ω, one might only have access to V_n through a non-orthonormal basis $(\phi_1, \dots, \phi_n)$, and the continuous norm $\|\cdot\|_{L^2}$ (and associated inner product $\langle\cdot,\cdot\rangle_{L^2}$) are usually not exactly computable. One is therefore constrained to use an approximately orthonormal basis $(\tilde{\varphi}_1, \dots, \tilde{\varphi}_n)$, and the related sampling density

$$\tilde{k}_n(x) = \frac{1}{n}\sum_{j=1}^{n} |\tilde{\varphi}_j(x)|^2.$$

An important remark is that, if only an approximation $\tilde{k}_n$ of k_n is known, it is sufficient to draw $\|Zk_n/\tilde{k}_n\|_{L^\infty}$ times more points according to the probability measure $\frac{1}{Z}\tilde{k}_n\, d\mu$, and to use weights $w_i = 1/\tilde{k}_n(x^i)$. Although the normalization constant

$$Z = \int_\Omega \tilde{k}_n(x)\, d\mu(x)$$

may not be known, classical sampling methods such as rejection sampling or Markov Chain Monte Carlo (MCMC) do not require its knowledge, and estimates on Z from above are sufficient to bound the factor $\|Zk_n/\tilde{k}_n\|_{L^\infty}$ in the sample size.

As a consequence, our algorithm is divided into two steps: the first one consists in drawing a large sample $y^1, \dots, y^M$ according to μ (or any known a priori sampling measure), to compute an inexact quadrature formula for $\|\cdot\|_{L^2}$ on V_n and obtain the density $\tilde{k}_n$, while the second draws a smaller sample $x^1, \dots, x^m \sim \tilde{k}_n\, d\mu$, evaluates u at these points, and returns the weighted least-squares approximation. This is the subject of Chapter 4, and in particular Section 4.3. Similar results have been obtained in [5, 6, 132, 133], with an emphasis on numerical error analysis in [132, 133], on frame discretizations in [6], and on the adaptivity to a nested sequence of approximation spaces $(V_{n_p})_{p\geqslant 1}$ in [5].

By the same arguments as for m, the number M of sample points in the first part has to grow at least like $M \geqslant \|k_n\|_{L^\infty} n\ln(n)$. When the domain Ω is very irregular, $\|k_n\|_{L^\infty}$ may increase too fast with n, which is problematic in terms of computational complexity of the offline stage. To get around this obstacle, we try to sample $y^1, \dots, y^M$ according to a better measure than μ. The optimal sampling density for the y^i is again the inverse Christoffel function k_n, which is precisely the quantity we would like to compute.

To avoid this loop, we use as in [5] a nested sequence of spaces

$$V_{n_1} \subset \cdots \subset V_{n_q} = V_n,$$

where $n_1 < \cdots < n_q$ and $\dim(V_{n_p}) = n_p$ for $1 \leqslant p \leqslant q$. These spaces are obtained by progressively adding basis elements, for instance by taking $V_{n_p} = \operatorname{span}\{\phi_1, \dots, \phi_{n_p}\}$. Our main result in this direction is Theorem 4.12, which states that near-optimal accuracy can be obtained with a quasilinear sample size m, using preliminary samples $y^{p,1}, \dots, y^{p,M_p}$ of size $M_p \sim n_p \ln(n_p)$ for $1 \leqslant p \leqslant q$, provided that there exists a constant κ such that

$$n_p\, k_{n_p} \leqslant n_{p+1}\, k_{n_{p+1}} \leqslant \kappa\, n_p\, k_{n_p}. \tag{1.15}$$

This multilevel strategy differs from the one in [87], in which it is assumed that the numerical solver can be queried to obtain values $u(x^i)$ with variable precision, depending on the computational effort invested. Moreover, it can be complemented with the variance reduction techniques discussed in [132]. Theoretical and empirical methods for choosing the intermediate sample sizes M_p are discussed in [34], in the connected setting of randomized weak greedy algorithms.

In order to guess the necessary sample sizes and to find κ satisfying (1.15), it remains to prove estimates

on the inverse Christoffel function. We address this question in Section 4.5 for the space V_n of multivariate total degree polynomials of a certain order, on a domain $\Omega \subset \mathbb{R}^d$, with μ the uniform probability measure on Ω. There is already a vast literature on the subject, see for instance [48, 66, 115, 153, 191]. Our main results in that matter are summarized below:

Theorem 1.14. *Let V_{n_p} be the space of total degree polynomials of order p on a domain $\Omega \subset \mathbb{R}^d$. If Ω has a $\mathcal{C}^2$ boundary,*

$$c_{\text{smooth}}\, n^{1/d} \leqslant \|k_n\|_{L^\infty} \leqslant C_{\text{smooth}}\, n^{1/d},$$

where c_{smooth} and C_{smooth} depend on Ω. If Ω has a Lipschitz boundary,

$$\|k_n\|_{L^\infty} \leqslant C_{\text{Lipschitz}}\, n, \quad \text{and} \quad k_n(x) \geqslant c_{\text{Lipschitz}}(x)\, n$$

for any outward corner $x \in \partial\Omega$. Finally, in the case of outward cusps, the reference domain $\Omega = \left\{x \in [-1,1]^d \,:\, \max_{1\leqslant j\leqslant d-1} |x_j|^{\alpha_i} \leqslant x_d\right\}$ satisfies

$$c_\alpha\, n^{\frac{1}{d}\left(2+\sum_{j=1}^{d-1} 2/\alpha_j\right)} \leqslant n\, k_n(0) \leqslant C_\alpha\, n^{\frac{1}{d}\left(2+\sum_{j=1}^{d-1} 2/\alpha_j\right)}$$

for any $\alpha_1, \ldots, \alpha_{d-1} \in (0, 2]$.

We also obtain a pointwise framing on k_n up to constants for piecewise smooth domains with outward corners, see Theorem 4.27, which implies (1.15) for such domains.

It is interesting to note that the sources of fast growth of k_n are the outward corners of the domain, in contrast to geometric singularities in the solutions to elliptic problems, which are caused by reentrant corners.

1.6 Reducing the sample size

In Chapter 5, based on our article [d], we obtain estimates similar to Theorem 1.9 and Corollary 1.11, with a sample size m' that scales linearly with n, therefore removing the logarithmic oversampling.

The main ingredient in this reduction of the sample size is the celebrated solution [128] of the Kadison-Singer problem. This problem was a conjecture posed by Kadison and Singer [103] in 1959, concerning extensions of C*-algebras in the formalization of quantum mechanics. It was later linked to a paving conjecture in [13], and brought back to finite dimension in [187], before being solved by Markus, Spielman and Srivastava, based on their earlier works on interlacing families of polynomials [127].

Although the focus of the authors of [128] was mainly to find so-called graph sparsifiers, their method helped advancing on the problem of frame discretization ([23, 73, 121, 139, 140], see also the surveys [38, 51]): given a vector-valued function $\varphi : \Omega \to \mathbb{C}^n$ such that

$$A_0 I \preccurlyeq \int_\Omega \varphi\varphi^*\, d\sigma \preccurlyeq B_0 I$$

for some probability measure σ and continuous frame bounds $A_0, B_0 > 0$, can one find a finite set $S \subset \Omega$ of controlled cardinality such that

$$AI \preccurlyeq \frac{1}{|S|} \sum_{x\in S} \varphi(x)\varphi(x)^* \preccurlyeq BI,$$

with discrete frame bounds $A, B > 0$?

As noticed in [170], if one takes $\varphi = (\varphi_1, \ldots, \varphi_n)$ an orthonormal basis of V_n in $V = L^2(\Omega, \mu)$, the first estimate is valid for $\sigma = \mu$ and $A = B = 1$, and the frame discretization yields a set of points S satisfying a Marcinkiewicz-Zygmund inequality:

$$A\, \|v\|_{L^2}^2 \leqslant \frac{1}{|S|} \sum_{x\in S} |v(x)|^2 \leqslant B\, \|v\|_{L^2}^2, \quad v \in V_n.$$

This inequality is in turn strongly related to approximation from point values, see [81], since it is equivalent to the eigenvalues of the Gram matrix G_m being comprised between A and B. In Chapter 5, we use this

subsampling theory to reduce the size of our sample $x^1, \dots, x^m$, by drawing points as in Section 1.4 and then randomly removing some of them. We detail the main steps below.

The main result from [128], see Corollary 1.5 there, can be stated as:

Theorem 1.15. *Let $r \in \mathbb{N}$ and $a_1, \dots, a_m \in \mathbb{C}^n$ such that $\frac{1}{m}\sum_{i=1}^m a_i a_i^\dagger = I$ and $|a_i|^2 \leqslant \delta$ for $1 \leqslant i \leqslant m$. Then there exists a partition $\{S_1, \dots, S_r\}$ of $\{1, \dots, m\}$ such that*

$$G_{S_s} := \frac{r}{m}\sum_{i \in S_s} a_i a_i^\dagger \preccurlyeq \left(1 + \sqrt{r\delta}\right)^2 I, \quad 1 \leqslant s \leqslant r.$$

To obtain a lower frame bound, it suffices to consider the case $r = 2$, and to notice that $G_{S_1} + G_{S_2} = I$, so $\lambda_{\min}(G_{S_1}) = 1 - \lambda_{\max}(G_{S_2})$. By a similarity transformation, one can send each G_{S_s} to the identity, and repeat the partitioning operation. This leads to a dyadic splitting of the initial sample $\{1, \dots, m\}$, satisfying the following duplicate of Lemma 5.8, which is mainly inspired from [140], Lemma 2, up to the slight difference that we keep track of all partition classes.

Lemma 1.16. *Let $a_1, \dots, a_m \in \mathbb{C}^n$ such that $\frac{1}{2}I \preccurlyeq \frac{1}{m}\sum_{i=1}^m a_i a_i^\dagger \preccurlyeq \frac{3}{2}I$ and $|a_i|^2 = \frac{n}{m}$ for $1 \leqslant i \leqslant m$. Then there exists $L \sim \ln(m/n)$ and a partition $\{J_1, \dots, J_{2^L}\}$ of $\{1, \dots, m\}$ such that $|J_s| \leqslant C_0 n$ and*

$$c_0 I \preccurlyeq \frac{1}{n}\sum_{i \in J_s} a_i a_i^\dagger \preccurlyeq C_0 I, \qquad 1 \leqslant s \leqslant 2^L,$$

for some universal constants c_0 and C_0.

With this, randomly selecting the partition class J_s with probability $|J_s|/m$, we are able to obtain an error bound for weighted least-squares that is optimal up to a constant, and has a linear sampling budget $|J_s| \leqslant C_0 n$. In Corollary 5.9, we indeed show that for any compact $\mathcal{K} \subset V = L^2(\Omega, \mu)$,

$$\rho^{\mathrm{rand}}_{C_0 n}(\mathcal{K})_{L^2} \leqslant \left(1 + 2\,\frac{C_0}{c_0}\right) d_n(\mathcal{K})_{L^2}.$$

This result therefore shows that, in the $\|\cdot\|_{L^2}$ norm, sampling numbers compare favorably to Kolmogorov widths, up to constant factors in the error bounds and the oversampling ratio $|J_s|/n$. Note however that the uniformity of $\rho^{\mathrm{rand}}_{C_0 n}(\mathcal{K})_{L^2}$ with respect to the element $u \in \mathcal{K}$ only holds with an expectation over a randomized sample inside the supremum $\sup_{u \in \mathcal{K}}$. Indeed, there can be no deterministic sample achieving a recovery error close to $d_n(\mathcal{K})_{L^2}$ for general classes $\mathcal{K}$.

On the other hand, if more regularity than having a finite n-width is assumed on $\mathcal{K}$, deterministic bounds can be attained. This topic has attracted much attention in recent years, see for instance [23, 104, 106, 113, 121, 137, 168], and is discussed in the next sections.

1.7 Reproducing Kernel Hilbert Spaces

Chapter 6 is based on our article [e], and studies deterministic sampling numbers $\rho_n(\mathcal{K})_{L^2}$ from (1.4), when $\mathcal{K}$ is the unit ball of a separable reproducing kernel Hilbert space (RKHS), or a compact class with sufficiently fast decay of its Kolmogorov widths.

The need for randomized error norms in the previous sections stems from the fact that point evaluations are not continuous linear forms in $L^2(\Omega, \mu)$. In contrast, a RKHS is a Hilbert space H of functions on some domain Ω, equipped with a kernel $K : \Omega \times \Omega \to \mathbb{C}$ such that

$$v(x) = \langle v, K(x, \cdot)\rangle_H, \quad v \in H,\ x \in \Omega.$$

We first consider the case where $\mathcal{K}$ is the unit ball of a RKHS H, itself compactly embedded into $L^2(\Omega, \mu)$. In

addition, we require the separability of H and the finite trace of its kernel:

$$\int_\Omega K(x,x)\,d\mu < \infty.$$

Under this assumption, our main result is the following.

Theorem 1.17. *There exists a universal constant $c \in \mathbb{N}$ such that*

$$\rho_{cn}(\mathcal{K})_{L^2}^2 \leqslant \frac{1}{n}\sum_{k\geqslant n} d_k(\mathcal{K})_{L^2}^2, \quad n \in \mathbb{N}. \tag{1.16}$$

Note that the separability assumption could be dropped by following [135], whereas the finite trace assumption is essential since it implies the ℓ^2 summability of the Kolmogorov widths $d_n(\mathcal{K})_V$.

This result is a replica of Theorem 6.1, which comes as the culmination of a series of earlier works [106, 113, 117, 137, 184] proving estimates of the same kind. It matches, up to a change in the constant c, the lower bounds on sampling numbers given in [95–97], and solves or partially answers open problems posed in [67, 95, 113, 144].

It notably implies the following bounds on decay rates of sampling numbers, see Corollary 6.2:

Corollary 1.18. *If $d_n(\mathcal{K})_{L^2} \leqslant c\,n^{-\alpha}\ln^{-\beta} n$ for some $\alpha > 0$, $\beta \in \mathbb{R}$, and $c > 0$, then there exists $C > 0$ such that*

$$\rho_m(\mathcal{K})_{L^2} \leqslant \begin{cases} C\,m^{-\alpha}\ln^{-\beta} m & \text{if } \alpha > 1/2, \\ C\,m^{-\alpha}\ln^{-\beta+1/2} m & \text{if } \alpha = 1/2 \text{ and } \beta > 1/2. \end{cases}$$

The main tool for the proof of Theorem 1.17 is an infinite-dimensional version of [128], see Proposition 6.17, which exhibits some similarities with the discretization of infinite-dimensional frames in [73].

Secondly, in Section 6.6, we consider more general classes $\mathcal{K}$, that are only assumed to be compactly embedded into L^2, separable, and with continuous point evaluation functionals. By constructing the appropriate RKHS containing $\mathcal{K}$ as in [114], we obtain bounds similar to (1.16), up to a logarithmic loss. This in particular yields the following decay rates, see Corollary 6.4:

Corollary 1.19. *If $d_n(\mathcal{K})_{L^2} \leqslant c\,n^{-\alpha}\ln^{-\beta} n$ for some $\alpha > 0$, $\beta \in \mathbb{R}$, and $c > 0$, then there exists $C > 0$ such that*

$$\rho_m(\mathcal{K})_{L^2} \leqslant \begin{cases} C\,m^{-\alpha}\ln^{-\beta} m & \text{if } \alpha > 1/2, \\ C\,m^{-\alpha}\ln^{-\beta+1} m & \text{if } \alpha = 1/2 \text{ and } \beta > 1, \\ C & \text{otherwise.} \end{cases}$$

Finally, we provide examples of classes $\mathcal{K}$ for which the rates of Corollaries 1.18 and 1.19 are sharp. In particular, this shows that sampling numbers may behave differently from Kolmogorov widths, and that this gap is amplified in a non-Hilbert setting.

1.8 Weighted least-squares with minimal oversampling

We conclude this introduction with the presentation of Chapter 7, which is based on an ongoing joint project with Abdellah Chkifa. This work improves significantly on Chapter 5, both in terms of computational complexity and sampling budget, by means of a randomized greedy algorithm for selecting the sampling points, inspired from [24] and [119].

The strategy is in fact simpler than in Chapter 5, since its sample points are drawn directly from continuous measures, following an idea from [61], instead of being subsampled from a larger initial sample.

Algorithm 4 is inspired by [118, 119], themselves randomized versions of [24, 164]. Our main theorem below shows that it achieves a randomized error bound as soon as $m \geqslant n$, see Theorem 7.1 for details.

Theorem 1.20. *For any $m \geqslant n$, one can compute a weighted least-squares approximation $\tilde{u} \in V_n$ using m evaluations of u such that*

$$\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2) \leqslant \left(1 + \frac{1}{(1-\sqrt{r})^2}\right) \min_{v \in V_n} \|u - v\|_{L^2}^2$$

where $r = (n-1)/m < 1$ is the oversampling ratio.

Interestingly, the same algorithm can also produce an estimator $\tilde{u}$ satisfying a deterministic error bound, when u and V_n are in L^∞. This setting has been developed in [62, 121, 168], with a structure of proof based on Lemma 5.12, see also [59, 81, 167] or Theorem 2.1 in [168].

Theorem 1.21. *For any $m \geqslant n$, one can compute a weighted least-squares approximation $\tilde{u} \in V_n$ using m evaluations of u such that*

$$\|u - \tilde{u}\|_{L^2}^2 \leqslant \left(1 + \frac{1}{(1-\sqrt{r})^2}\right) \min_{v \in V_n} \|u - v\|_{L^\infty}^2 \quad a.s.,$$

where $r = (n-1)/m < 1$ is the oversampling ratio.

As a consequence of Theorem 1.20, we obtain a randomized interpolation result, with a Lebesgue constant of order $\mathcal{O}(n^2)$. We repeat Corollary 7.2 below:

Corollary 1.22. *For $m = n$, one can compute an interpolation $\tilde{u} \in V_n$ of u such that*

$$\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2) \leqslant 4n^2 \min_{v \in V_n} \|u - v\|_{L^2}^2.$$

To our knowledge, there are no pre-existing results on interpolation of L^2 functions in expectation with respect to the L^2 norm. It is much more classical to study the stability of interpolation in the uniform norm. This can also be done here, by taking $m = n$ in Theorem 1.21, however requiring an additional inverse inequality between $L^2(\Omega, \mu)$ and L^∞ in V_n. Such an inequality involves the L^∞ norm of the inverse Christoffel function $\|nk_n\|_{L^\infty}^{1/2} = \max_{v \in V_n} \|v\|_{L^\infty} / \|v\|_{L^2}$, and gives our last result:

Corollary 1.23. *For $m = n$, one can compute an interpolation $\tilde{u} \in V_n$ of u such that*

$$\|u - \tilde{u}\|_{L^\infty} \leqslant \left(1 + 2n \|nk_n\|_{L^\infty}^{1/2}\right) \min_{v \in V_n} \|u - v\|_{L^\infty}^2.$$

Using Theorem 1.14 or the results from Section 4.5, we at once obtain strong stability estimates for polynomial interpolation on general domains. Moreover, in contrast to the Fekete points, the sample is computable in polynomial time, with Algorithm 4. In the monodimensional case of a finite union of intervals, one can see for instance that $\|k_n\|_{L^\infty} \sim \sqrt{n}$, resulting in a Lebesgue constant

$$\Lambda_n := \max_{v \in V_n} \frac{\|\tilde{v}\|_{L^\infty}}{\|v\|_{L^\infty}} = \mathcal{O}(n^{7/8}),$$

beating other greedy strategies based on Leja points [14, 47, 49] which only achieve $\Lambda_n = \mathcal{O}(n^{13/4})$ or $\Lambda_n = \mathcal{O}(n^{1+\log_2(3)})$. Of course, in this very elementary case, it is well known that Chebyshev points give the optimal rate $\Lambda_n = \mathcal{O}(\ln n)$ for the Lebesgue constant.

Corollary 1.23 also applies for different spaces V_n. For instance, on the d-dimensional torus $\Omega = \mathbb{T}^d$, and for any space V_n of trigonometric polynomials, it holds $k_n = 1$ over the whole domain, by translation invariance of V_n and of the Lebesgue measure μ. This yields the bound

$$\Lambda_n \leqslant 2n\sqrt{n}.$$

As a concluding remark, let us note that the measure μ with respect to which we integrate in the L^2 norm is not fixed by the L^∞ constraint. Therefore, one may multiply μ by any density before applying Corollary 1.23. For

instance, in the case of univariate polynomials on the interval $[-1, 1]$, taking the arcsine density $\frac{4\,dx}{\pi\sqrt{1-x^2}}$ makes the Chebyshev polynomials orthogonal, and as they are uniformly bounded, we once more obtain $\Lambda_n = \mathcal{O}(n^{3/2})$.

This leaves open the challenge of attaining a bound of the same order for any finite dimensional space V_n of functions on a general domain Ω.

Acknowledgements: The author would like to thank the reviewers of papers [a], [b], [c], [d] and [e] for their careful work and constructive comments. Together with the coauthors of [a], he also thanks François Murat and Jules Pertinand for useful discussions in the understanding of the convergence process towards limit solutions, and Hamza Maimoune for leading them to this work through his remarks during his master project. Together with the coauthors of [e], he thanks Albert Cohen, Daniel Freeman, Aicke Hinrichs and Erich Novak for useful discussions and detailed corrections. Moreover, David Krieg and Mario Ullrich want to acknowledge that he obtained Theorem 6.1 on his own and presented it at the *MASCOT-NUM Workshop on "Optimal Sampling for Approximation"* at the IHP in Paris on March 10, 2022. Chapter 6 is the product of subsequent collaboration.

Part I

Reduced order modeling for PDEs and inverse problems

Chapter 2

Reduced order modeling for elliptic problems with high contrast diffusion coefficients

Abstract. We consider a parametric elliptic PDE with a scalar piecewise constant diffusion coefficient taking arbitrary positive values on fixed subdomains. This problem is not uniformly elliptic, as the contrast can be arbitrarily high, contrarily to the Uniform Ellipticity Assumption (UEA) that is commonly made on parametric elliptic PDEs. We construct reduced model spaces that approximate uniformly well all solutions with estimates in relative error that are independent of the contrast level. These estimates are sub-exponential in the reduced model dimension, yet exhibiting the curse of dimensionality as the number of subdomains grows. Similar estimates are obtained for the Galerkin projection, as well as for the state estimation and parameter estimation inverse problems. A key ingredient in our construction and analysis is the study of the convergence towards limit solutions of stiff problems when diffusion tends to infinity in certain subdomains.

2.1 Introduction

2.1.1 Reduced models for parametric PDEs

Parametric PDEs are commonly used to describe complex physical phenomena. With $y = (y_1, \dots, y_d)$ denoting a parameter vector ranging in some domain $Y \subset \mathbb{R}^d$, and $u(y)$ the corresponding solution to the PDE of interest, assumed to be well defined in some Hilbert space V, we denote by

$$\mathcal{K}_Y := \{u(y) \ : \ y \in Y\} \tag{2.1}$$

the collection of all solutions, called the *solution manifold*.

There are two main ranges of problems associated to parametric PDEs:

1. Forward modeling: in applications where many queries of the parameter to solution map $y \mapsto u(y)$ are required, one needs numerical forward solvers that efficiently compute approximations $\tilde{u}(y)$ with a prescribed accuracy.
2. Inverse problems: when the exact value of the parameter y is unknown, one is interested in either recovering an approximation to $u(y)$ (state estimation) or to y (parameter estimation), from a limited number of observations $z_i = \ell_i(u(y))$, possibly corrupted by noise.

Reduced order modeling is widely used for tackling both problems. In its most common form, its aim is to construct linear spaces V_n of moderate dimension n that approximate all solutions $u(y)$ with best possible certified accuracy. The natural benchmark for measuring the performance of such linear reduced models is provided by the *Kolmogorov n-width* of the solution manifold

$$d_n(\mathcal{K}_Y)_V := \inf_{\dim(V_n)=n} \operatorname{dist}(\mathcal{K}_Y, V_n)_V \tag{2.2}$$

that describes the performance of an optimal space. Here

$$\operatorname{dist}(\mathcal{K}_Y, V_n)_V := \sup_{u\in\mathcal{K}_Y} \inf_{v\in V_n} \|u-v\|_V = \sup_{y\in Y} \|u(y) - P_n u(y)\|_V,$$

where P_n is the V-orthogonal projector onto V_n. We refer the reader to the Section 1.1, see also [150], for a general treatment of n-widths.

While an optimal space achieving the above infimum is usually out of reach, there exist two main approaches aiming to construct "sub-optimal yet good" spaces. The first one consists in building expansions of the parameter to solution map, for example by polynomials

$$\tilde{u}(y) := \sum_{\nu\in\Lambda_n} u_\nu y^\nu, \quad y^\nu := y_1^{\nu_1}\dots y_d^{\nu_d}, \tag{2.3}$$

where $\Lambda_n \subset \mathbb{N}_0^d$ is a set of cardinality n. The coefficients u_ν are elements of V and therefore, for all $y \in Y$ the approximation $\tilde{u}(y)$ is picked from the space

$$V_n := \operatorname{span}\{u_\nu \ : \ \nu \in \Lambda_n\}.$$

Notice that $\tilde{u}(y)$ is not the orthogonal projection $P_n u(y)$ in this case, but $\tilde{u}(y)$ is easy to compute for a given query y once the u_ν have been constructed (usually through a high fidelity finite element solver). We refer to [19, 21, 22, 26, 56, 58, 174] for instances of this approach.

The second approach is the reduced basis method [84, 157, 160], that consists in taking

$$V_n := \operatorname{span}\{u(y^1), \dots, u(y^n)\},$$

where the $u(y^i)$ are particular solution instances corresponding to a selection of parameter vectors $y^i \in Y$. A close variant is the proper orthogonal decomposition method [46, 180, 190], where the reduced spaces are obtained by principal component analysis applied to large training set of such instances. In the reduced basis method, the parameter vectors $y^1, \dots, y^n$ can be selected by a greedy algorithm, introduced in [179] and originally studied in [40]. For such a selection process, it is proved in [33, 65] that if $d_n(\mathcal{K}_Y)_V$ has a certain algebraic or exponential rate of decay with n, then a similar rate is achieved by $\operatorname{dist}(\mathcal{K}_Y, V_n)_V$ for the reduced basis spaces.

It follows that the reduced basis spaces constructed by the greedy algorithm are close to optimal. This is in contrast to the spaces V_n spanned by the polynomial coefficients u_ν for which the approximation rate is not guaranteed to be optimal. We refer to [20] for instances where reduced basis methods can be proved to converge with a strictly higher rate than polynomial approximations. On the other hand, the polynomial constructions (2.3) have certain numerical advantages. Namely, for several relevant classes of parametric PDEs, it can be shown that the parameter to solution mapping $y \mapsto u(y)$ has certain smoothness properties that can be used to obtain a-priori bounds on the $\|u_\nu\|_V$ without actually computing these norms. This allows for an a priori selection of an appropriate set Λ_n, yielding concrete approximation estimates for the error $\sup_{y\in Y} \|u(y)-\tilde{u}(y)\|_V$. These estimates in turn provide an upper bound for $d_n(\mathcal{K}_Y)_V$, and therefore for reduced basis approximations.

2.1.2 Parametric elliptic PDEs

One prototypical instance where the convergence analysis described above has been deeply studied is the parametric second order elliptic equation

$$-\operatorname{div}(a(y)\nabla u(y)) = f \quad \text{in } \Omega, \qquad u|_{\partial\Omega} = 0 \quad \text{on } \partial\Omega, \tag{2.4}$$

where Ω is a spatial domain, $f \in H^{-1}(\Omega)$ a source term, and $a(y)$ has the *affine* form

$$a(x,y) = \overline{a}(x) + \sum_{j=1}^{d} y_j \psi_j(x), \quad x \in \Omega, \tag{2.5}$$

with $\overline{a}$ and $\psi_1, \dots, \psi_d$ some fixed functions in $L^\infty(\Omega)$.

The corresponding solution $u(y) \in H_0^1(\Omega)$ is defined through the standard variational formulation in $H_0^1(\Omega)$ equipped with its usual norm. Up to renormalization, it is usually assumed that the y_j range in $[-1,1]$, or equivalently $Y = [-1,1]^d$. To ensure existence and uniqueness of solutions, one typically assumes that the so-called *Uniform Ellipticity Assumption* (UEA) holds: for some fixed $0 < a_{\min} \leqslant a_{\max} < \infty$,

$$a_{\min} \leqslant a(x,y) \leqslant a_{\max}, \qquad x \in \Omega, \quad y \in Y, \tag{2.6}$$

or in short $a_{\min} \leqslant a(y) \leqslant a_{\max}$ for all $y \in Y$. Under this assumption, Lax-Milgram theory ensures that the solution map $y \mapsto u(y)$ is well defined from Y into $H_0^1(\Omega)$, with the uniform bound

$$\|u(y)\|_{H_0^1} := \|\nabla u(y)\|_{L^2} \leqslant \frac{C_f}{a_{\min}}, \quad y \in Y.$$

Here and throughout this chapter

$$C_f := \|f\|_{H^{-1}}. \tag{2.7}$$

It was proved in [22, 174] that, under UEA, polynomial approximations (2.3) of given total degree converge sub-exponentially: for $\Lambda_n = \{\nu \in \mathbb{N}_0^d \,:\, |\nu| \leqslant k\}$, one has

$$\sup_{y\in Y} \|u(y) - \tilde{u}(y)\|_{H_0^1} \leqslant C' \exp(-cn^{1/d}) \quad \text{with} \quad n = \binom{k+d}{k}. \tag{2.8}$$

Such sub-exponential rates show that the spaces V_n based on polynomial expansions or reduced bases perform significantly better than standard finite element spaces, at least for a moderate number d of parameters. It is possible to maintain a rate of convergence as d grows, and even when $d = \infty$, when assuming some anisotropy in the variable y_j through the decay of the size of ψ_j as $j \to \infty$, see in particular [21, 56, 58] for results of this type.

2.1.3 High constrast problems

The Uniform Ellipticity Assumption (2.6) implies that there is a uniform control on the level of contrast in the diffusion function

$$\kappa(y) := \frac{\max_{x\in\Omega} a(x,y)}{\min_{x\in\Omega} a(x,y)} \leqslant \frac{a_{\max}}{a_{\min}}, \quad y \in Y. \tag{2.9}$$

This assumption also plays a key role in the derivation of the above approximation results, since it guarantees that the parameter to solution map has a holomorphic extension to a sufficiently large complex neighborhood of Y. In this case, a good polynomial approximation $\tilde{u}$ may be defined by simply truncating the power series $\sum_{\nu\in\mathbb{N}_0^d} u_\nu y^\nu$, leading to the estimate (2.8).

On the other hand, there exist various situations where one would like to avoid such a strong restriction on the level of contrast. Perhaps the most representative setting is when the domain Ω is partitioned into disjoint subdomains $\{\Omega_1, \ldots, \Omega_d\}$, each of them admitting a constant diffusivity level that could vary strongly between subdomains. This is typically the case when modeling diffusion in materials having multiple layers or inclusions that could have very different nature, for example air or liquid versus solid. This situation can be encountered in groundwater flow applications, where certain subdomains correspond to cavities, for which the diffusion function becomes nearly infinite, as opposed to subdomains containing sediments or other porous rocks.

In such a case, we do not want to limit the contrast level. To represent this setting, we let

$$a(y)|_{\Omega_j} = y_j, \quad y_j \in (0,\infty) \tag{2.10}$$

or equivalently $a(y) = \sum_{j=1}^d y_j \chi_{\Omega_j}$, which corresponds to the affine form (2.5) with $\overline{a} = 0$ and $\psi_j = \chi_{\Omega_j}$ for $1 \leqslant j \leqslant d$, now with

$$Y := (0,\infty)^d. \tag{2.11}$$

We take (2.11) as the definition of the parameter domain Y for the remainder of this chapter. The solution

$u(y)$ satisfies the variational formulation

$$\sum_{j=1}^{d} y_j \int_{\Omega_j} \nabla u(y) \cdot \nabla v \, dx = \langle f, v \rangle_{H^{-1}, H_0^1}, \quad v \in H_0^1(\Omega), \tag{2.12}$$

or equivalently $-y_j \Delta u(y) = f$ as elements of $H^{-1}(\Omega_j)$ on each Ω_j, with the standard jump conditions $[a(y)\partial_{\vec{n}} u(y)] = 0$ across the boundaries between subdomains.

Let us observe that in this setting, it is hopeless to find spaces V_n that approximate all solutions $u(y)$ uniformly well. Indeed, the following homogeneity property obviously holds: for any $y \in Y$ and $t > 0$, one has

$$u(ty) = t^{-1} u(y). \tag{2.13}$$

This property implies in particular that $\|u(y)\|_{H_0^1}$ tends to infinity as $y \to 0$, and so does $\|u(y) - P_n u(y)\|_{H_0^1}$ in general. In fact, this also shows that the solution manifold $\mathcal{K}_Y$ is *not* relatively compact and does not have finite n-widths.

In addition to this principal difficulty, let us remind that when using the spaces V_n in forward modeling, we typically use the Galerkin method, which delivers the orthogonal projection P_n^y onto V_n, however for the energy norm

$$\|v\|_y^2 := \sum_{j=1}^{d} y_j \int_{\Omega_j} |\nabla v|^2 \, dx. \tag{2.14}$$

This approximation is thus optimal in $H_0^1(\Omega)$, however up to the constant $\kappa(y)^{1/2}$, which deteriorates with high contrast.

The main contribution of this chapter is to treat these issues, and derive approximation estimates that are robust to high contrast, in the sense that they are independent of $y \in Y$.

Due to the main objection coming from the homogeneity property (2.13), it is natural to look for uniform approximation estimates in relative error, that is, estimates of the form

$$\|u(y) - P_n u(y)\|_{H_0^1} \leqslant \varepsilon_n \|u(y)\|_{H_0^1}, \quad y \in Y, \tag{2.15}$$

with $\lim_{n\to\infty} \varepsilon_n = 0$, and similarly for $P_n^y u(y)$. Our main results, Theorems 2.16 and 2.21, exhibit spaces V_n ensuring the validity of such uniform estimates with ε_n having sub-exponential decay with n, similar to the known results under UEA.

Remark 2.1. High contrast problems have been the object of intense investigation, in particular with the objective of developing techniques for multilevel or domain decomposition preconditioning [9, 10, 75] and a-posteriori error estimation [8, 30], that are provably robust with respect to the level of contrast. To our knowledge, the present work is the first in which this robustness is established for reduced modeling methods.

2.1.4 Outline

Throughout this chapter, we consider the parametric elliptic PDE (2.4) with $a(y)$ having piecewise constant form (2.10) over a fixed partition. In view of the homogeneity property (2.13), we are led to consider the subset

$$Y' := [1, \infty)^d \tag{2.16}$$

of parameters corresponding to the coercive regime. Any result on relative approximation error that is established for Y' extends automatically to all of Y because of the homogeneity property. Accordingly, we let

$$\mathcal{K}_{Y'} := \{u(y) \ : \ y \in Y'\}. \tag{2.17}$$

In Section 2.2, we start by proving that $\mathcal{K}_{Y'}$ is a precompact set of $H_0^1(\Omega)$. One crucial ingredient for this analysis are the *limit solutions* of the so-called *stiff problem*, obtained as $y_j \to \infty$ for certain $j \in \{1, \dots, d\}$.

In Section 2.3, we construct specific reduced model spaces for which the approximation estimate (2.15) holds with ε_n decaying sub-exponentially. Our construction is based on partitioning the parametric domain Y' into rectangular regions and using a different polynomial approximation on each region. This results in a global reduced model space V_n for which the accuracy bound remains sub-exponential, however in $\exp(-cn^{\frac{1}{2d-2}})$. A key ingredient for establishing these sub-exponential rates is the derivation of quantitative estimates on the convergence of $u(y)$ towards limit solutions defined in Section 2.2 as some y_j tend to infinity. These estimates are established under an additional geometric assumption on the partition, similar results for a general partition of Ω being an open problem.

In Section 2.4, we discuss the use of these reduced model spaces in forward modeling and inverse problems. Our main result relative to forward modeling is that the estimate (2.15) also holds for the Galerkin projection with the same exponential decay ε_n. We show that such a result is only possible if V_n includes functions that have constant values over some subdomains. For the state estimation problem, we follow the Parametrized Background Data Weak (PBDW) method [33, 124], and obtain recovery bounds that are uniform over $y \in Y$ in relative error. For the parameter estimation problem, we introduce an ad-hoc strategy that specifically exploits the piecewise constant structure of the diffusion coefficient and obtain similar recovery bounds for the inverse diffusivity.

We conclude in Section 2.5 by presenting some numerical illustrations revealing the effectiveness of the reduced model spaces even in the high-contrast regime, as expressed by the approximation results.

2.2 Uniform approximation in relative error

In this section we work under no particular geometric assumption on the partition $\{\Omega_1, \dots, \Omega_d\}$ of Ω, and consider the solution manifold $\mathcal{K}_Y$ defined by (2.1), where $u(y) \in H^1_0(\Omega)$ is solution to the elliptic boundary value problem with variational formulation (2.12). Our objective is to show the existence of spaces V_n that uniformly approximate $\mathcal{K}_Y$ in the relative error sense expressed by (2.15).

2.2.1 Limit solutions and the extended solution manifold

Our first observation is that this collection can be continuously extended when $y_j = \infty$ for some values of j, through limit solutions of stiff inclusions problems. Such limit solutions have for example been considered in the context homogeneization, see e.g. p.98 of [101].

For this purpose, to any $S \subset \{1, \dots, d\}$, we associate the space

$$V_S := \{v \in H^1_0(\Omega) \ : \ \nabla v|_{\Omega_j} = 0, \ \ j \in S\}. \tag{2.18}$$

In other words, V_S consists of the functions from $H^1_0(\Omega)$ that have constant values on the subdomains Ω_j for $j \in S$ (or on each of their connected components if these subdomains are not connected). It is a closed subspace of $H^1_0(\Omega)$. We decompose the parameter vector y according to

$$y = (y_S, y_{S^c}), \quad y_S := (y_j)_{j \in S} \quad \text{and} \quad y_{S^c} := (y_j)_{j \in S^c}. \tag{2.19}$$

For any finite and positive vector y_{S^c}, similar to the $\|\cdot\|_y$ norm (2.14), we may define

$$\|v\|^2_{y_{S^c}} := \sum_{j \in S^c} y_j \int_{\Omega_j} |\nabla v|^2 \, dx, \tag{2.20}$$

which is a semi-norm on $H^1_0(\Omega)$, and a full norm equivalent to the H^1_0-norm on V_S. Also note that when $y = (y_S, y_{S^c})$ is finite, one then has $\|v\|_{y_{S^c}} = \|v\|_y$ for any $v \in V_S$.

For any finite and positive vector y_{S^c}, we define the function $u_S(y_{S^c}) \in V_S$ solution to the following stiff inclusions problem:

$$\sum_{j \in S^c} y_j \int_{\Omega_j} \nabla u_S(y_{S^c}) \cdot \nabla v \, dx = \langle f, v \rangle_{H^{-1}, H^1_0}, \quad v \in V_S. \tag{2.21}$$

The following result shows that this solution is well defined and is the limit of $u(y)$, when y_{S^c} is fixed and

$y_j \to \infty$ for $j \in S$. Note that the weak convergence is established in [101] (p.98) and so we concentrate the proof on the strong convergence.

Lemma 2.2. *There exists a unique $u_S(y_{S^c}) \in V_S$ solution to (2.21), which is the limit in $H^1_0(\Omega)$ of the solution $u(y_S, y_{S^c})$ as $y_j \to \infty$ for all $j \in S$.*

Proof. Using the bilinear form $(u,v) \mapsto \sum_{j\in S^c} y_j \int_{\Omega_j} \nabla u \cdot \nabla v \, dx$ in the space V_S, Lax-Milgram theory implies the existence of a unique solution $u_S(y_{S^c}) \in V_S$ to (2.21).

Consider now a sequence $(y^n)_{n\geqslant 1} \in Y^{\mathbb{N}}$, with $y^n_{S^c} = y_{S^c}$ and $y^n_j \to \infty$ for all $j \in S$. Denoting $u_n = u(y^n)$, it is readily seen that $(u_n)_{n\geqslant 1}$ is uniformly bounded in H^1_0 norm by $C = C_f\, c^{-1}$, where $c := \min_{n\geqslant 1} \min_{1\leqslant j\leqslant d} y^n_j > 0$, and that any weak limit of a sequence extraction is solution to the variational equation (2.21). Therefore the whole sequence $(u_n)_{n\geqslant 1}$ weakly converges to $\bar{u} = u_S(y_{S^c})$.

We finally prove strong convergence by writing

$$\begin{aligned} c\,\|u_n - \bar{u}\|^2_{H^1_0} &\leqslant \int_\Omega a(y^n)|\nabla(u_n - \bar{u})|^2\, dx \\ &= \langle f, u_n\rangle_{H^{-1},H^1_0} - 2\langle \bar{u}, u_n\rangle_{y_{S^c}} + \|\bar{u}\|^2_{y_{S^c}} \\ &\underset{n\to\infty}{\longrightarrow} \langle f, \bar{u}\rangle_{H^{-1},H^1_0} - \|\bar{u}\|^2_{y_{S^c}} = 0. \end{aligned}$$

□

The above lemma allows us to readily extend the solution manifold by introducing $\widetilde{Y} := (0,\infty]^d$ and

$$\overline{\mathcal{K}_Y} := \{u(y) \ : \ y \in \widetilde{Y}\},$$

where we have formally set

$$u(y) := u_S(y_{S^c}),$$

when $y_j = \infty$ for $j \in S$ and $y_j < \infty$ for $j \in S^c$. Note that when $S = \{1,\dots,d\}$ the space V_S is trivial and one has

$$u(\infty,\dots,\infty) = 0.$$

Remark 2.3. Although we do not make explicit use of it, it can be checked that despite the fact that $y_j = 0$ is excluded in the definition of $\overline{\mathcal{K}_Y}$, it indeed coincides with the closure of $\mathcal{K}_Y$ in $H^1_0(\Omega)$ due to the fact that $\|u(y)\|_{H^1_0} \to \infty$ as $y \to 0$.

Remark 2.4. More precisely, when some y_j tend to zero, $u(y)$ converges to the solution of the so-called soft inclusions problem (see [101], chapter 3), outside the corresponding subdomains Ω_j. Here, due to the fact that the approximation estimates that we prove further are in relative error, these other limit solutions are of no use in our analysis.

2.2.2 A compactness result

As already observed in the introduction, the manifold $\overline{\mathcal{K}_Y}$ is not bounded in $H^1_0(\Omega)$ due to the homogeneity property (2.13) and therefore not compact.

In order to treat this defect, we consider $\widetilde{Y}' := [1,\infty]^d$, and the submanifold

$$\overline{\mathcal{K}_{Y'}} := \{u(y) \ : \ y \in \widetilde{Y}'\},$$

which is now bounded in $H^1_0(\Omega)$, from the standard a-priori estimate

$$\|u(y)\|_{H^1_0} \leqslant \frac{C_f}{\min_{1\leqslant j\leqslant d} y_j} \leqslant C_f,$$

that is obtained by taking $v = u(y)$ in the variational formulation (2.12), with $C_f = \|f\|_{H^{-1}}$ as in (2.7). This estimate trivially extends to $u_S(y_{S^c})$ when the y_j have infinite value for $j \in S$. In addition we have the following result.

Theorem 2.5. *The set $\overline{\mathcal{K}_{Y'}}$ is compact in $H^1_0(\Omega)$.*

Proof. Consider any sequence of vectors $y^n = (y^n_1, \dots, y^n_d) \in \widetilde{Y}'$ for $n \geqslant 1$. We need to prove that the corresponding sequence of solutions $(u(y^n))_{n\geqslant 1}$ admits a converging subsequence. For this purpose, we observe that there exists a subset $S \in \{1, \dots, d\}$ such that, up to subsequence extraction,

$$\lim_{n\to\infty} y^n_j = \infty, \quad j \in S,$$

and

$$\lim_{n\to\infty} y^n_j = y_j < \infty, \quad j \in S^c.$$

Note that S could be empty, for instance in the case where the y^n_j are uniformly bounded for all j.

Let $\varepsilon > 0$. Using the strong convergence result in Lemma 2.2, for all $n \geqslant 1$ there exists an auxiliary vector $\bar{y}^n$ such that $\bar{y}^n_j = y^n_j$ when $y^n_j < \infty$, $\bar{y}^n_j < \infty$ when $y^n_j = \infty$, such that by having picked $\bar{y}^n_j$ large enough in the second case

$$\|u(y^n) - u(\bar{y}^n)\|_{H^1_0} \leqslant \varepsilon/3.$$

In addition we may assume that $\bar{y}^n_j \to \infty$ for $j \in S$. Next we introduce the vector $\tilde{y}^n$ such that $\tilde{y}^n_j = \bar{y}^n_j$ when $j \in S$ and $\tilde{y}^n_j = y_j$ when $j \in S^c$. Applying again Lemma 2.2, we find that with $y_{S^c} = (y_j)_{j\in S^c}$, one has

$$\|u(\tilde{y}^n) - u_S(y_{S^c})\|_{H^1_0} \leqslant \varepsilon/3,$$

for n sufficiently large. Finally we argue that

$$\|u(\tilde{y}^n) - u(\bar{y}^n)\|_{H^1_0} \leqslant \varepsilon/3,$$

for n large enough. This is a consequence of the following variant of Strang first lemma (which proof is similar and left as an exercise to the reader) that says that for two diffusion functions $\bar{a}$ and $\tilde{a}$, the corresponding solution $u(\bar{a})$ and $u(\tilde{a})$ with the same data f satisfy

$$\|u(\bar{a}) - u(\tilde{a})\|_{H^1_0} \leqslant \frac{C_f\, \|\bar{a} - \tilde{a}\|_{L^\infty}}{\min\{\bar{a}_{\min}, \tilde{a}_{\min}\}^2}.$$

We then apply this to $\bar{a} := \bar{a}_n = a(\bar{y}^n)$ and $\tilde{a} := \tilde{a}_n = a(\tilde{y}^n)$, observing that $\|\bar{a} - \tilde{a}\|_{L^\infty} = \max_{j\in S^c} |\bar{y}^n_j - y_j| \to 0$ as $n \to \infty$. Therefore $\|u(y^n) - u_S(y_{S^c})\|_{H^1_0} \leqslant \varepsilon$ for n sufficiently large, which concludes the proof. □

We next observe that any $y \in Y$ can be rewritten as

$$y = t\tilde{y},$$

with $\tilde{y} \in Y'$ and normalization $\min \tilde{y}_j = 1$, for some $t > 0$, and from (2.13) one has $u(y) = t^{-1}u(\tilde{y})$. This motivates the study of the further reduced manifold

$$\overline{\mathcal{K}_{Y''}} := \{u(y) \ : \ y \in \widetilde{Y}' = [1, \infty]^d \ : \ \min_{1\leqslant j\leqslant d} y_j = 1\}, \tag{2.22}$$

which is a subset of $\overline{\mathcal{K}_{Y'}}$.

One important observation is that the solutions contained in $\overline{\mathcal{K}_{Y''}}$ are also uniformly bounded from below, under mild assumptions on the data f.

Lemma 2.6. *The set $\overline{\mathcal{K}_{Y''}}$ is compact in $H^1_0(\Omega)$. Moreover, one has the framing*

$$\min_{1\leqslant j\leqslant d} \|f\|_{H^{-1}(\Omega_j)} \leqslant \|u(y)\|_{H^1_0} \leqslant C_f, \tag{2.23}$$

for all $u(y) \in \overline{\mathcal{K}_{Y''}}$.

Proof. The compactness of $\overline{\mathcal{K}_{Y''}}$ follows from that of $\overline{\mathcal{K}_{Y'}}$, since $\overline{\mathcal{K}_{Y''}}$ is a closed subset of $\overline{\mathcal{K}_{Y'}}$. For the framing,

as $a(y) \geqslant 1$ on Ω,

$$\|u\|_{H_0^1}^2 \leqslant \sum_{j \in S^c} y_j \int_{\Omega_j} |\nabla u(y)|^2 \, dx = \langle f, u(y) \rangle_{H^{-1}, H_0^1} \leqslant C_f \|u(y)\|_{H_0^1},$$

so $\|u(y)\|_{H_0^1} \leqslant C_f$. Now take $j \in \{1, \ldots, d\}$ such that $y_j = 1$, and consider $\phi \in H_0^1(\Omega_j)$. Then

$$\langle f, \phi \rangle_{H^{-1}, H_0^1} = \int_{\Omega_j} \nabla u(y) \cdot \nabla \phi \, dx \leqslant \|u(y)\|_{H_0^1(\Omega)} \|\phi\|_{H_0^1(\Omega_j)},$$

which gives the result by optimizing over ϕ. □

In the sequel of this chapter, we always work under the condition that the lower bound in (2.23) is strictly positive

$$c_f := \min_{1 \leqslant j \leqslant d} \|f\|_{H^{-1}(\Omega_j)} > 0. \tag{2.24}$$

Let us observe that when f is a function in $L^2(\Omega)$, this is ensured as soon as f is not identically zero on one of the Ω_j. We thus have

$$0 < c_f \leqslant \|u(y)\|_{H_0^1} \leqslant C_f, \tag{2.25}$$

for all $u(y) \in \overline{\mathcal{K}_{Y''}}$.

Remark 2.7. The condition $c_f > 0$ is in general necessary for controlling $\|u(y)\|_{H_0^1}$ from below. Indeed assume $\|f\|_{H^{-1}(\Omega_j)} = 0$ for some j such that $\mathbb{R}^d \setminus \overline{\Omega}_j$ is connected. Then taking $y_k = \infty$ for $k \neq j$ and $y_j = 1$, we find that $u(y) \in V_S$ with $S = \{j\}^c$, which is equivalent to $u(y) \in H_0^1(\Omega_j)$ since it vanishes on the other sub-domains. As $\|f\|_{H^{-1}(\Omega_j)} = 0$, we obtain $u(y) = 0$.

Remark 2.8. One also has the uniform framing in the $\|\cdot\|_y$ norm since

$$0 < c_f \leqslant \|u(y)\|_{H_0^1} \leqslant \|u(y)\|_y = \sqrt{\langle f, u \rangle_{H^{-1}, H_0^1}} \leqslant C_f \tag{2.26}$$

when the y_j are finite, and by continuity for all $u(y) \in \overline{\mathcal{K}_{Y''}}$ in the norm $\|\cdot\|_{y_{S^c}}$.

The framing (2.25) has an implication on the existence of reduced model spaces that approximate uniformly well all solutions $u(y) \in \overline{\mathcal{K}_Y}$ in relative error.

Theorem 2.9. *There exists a sequence of linear spaces $(V_n)_{n \geqslant 1}$ such that $\dim(V_n) = n$, and a sequence $(\varepsilon_n)_{n \geqslant 1}$ that converges to zero such that*

$$\|u(y) - P_n u(y)\|_{H_0^1} \leqslant \varepsilon_n \|u(y)\|_{H_0^1} \tag{2.27}$$

for all $y \in \widetilde{Y}$, where P_n is the $H_0^1(\Omega)$-orthogonal projector onto V_n.

Proof. Since $\overline{\mathcal{K}_{Y''}}$ is compact, there exists a sequence of spaces $(V_n)_{n \geqslant 1}$ with $\dim(V_n) = n$ and a sequence $(\sigma_n)_{n \geqslant 1}$ that tends to 0, such that

$$\|v - P_n v\|_{H_0^1} \leqslant \sigma_n, \quad v \in \overline{\mathcal{K}_{Y''}}.$$

Now let $y \in \widetilde{Y}$ differing from $(\infty, \ldots, \infty)$, for which there is nothing to prove since $u(\infty, \ldots, \infty) = 0$, and let $t = 1/\min_{1 \leqslant j \leqslant d} y_j < \infty$. By homogeneity, $t^{-1} u(y) = u(ty) \in \overline{\mathcal{K}_{Y''}}$, and therefore

$$\|u(y) - P_n u(y)\|_{H_0^1} = t \|u(ty) - P_n u(ty)\|_{H_0^1(\Omega)} \leqslant t \sigma_n.$$

On the other hand, $\|u(y)\|_{H_0^1(\Omega)} = t \|u(ty)\|_{H_0^1(\Omega)} \geqslant t c_f$ by framing (2.23), which proves Theorem 2.9 with $\varepsilon_n = \sigma_n / c_f$. □

The above theorem tells us that we can achieve contrast-independent approximation in relative error. It is however still unsatisfactory from two perspectives:

1. It does not describe the rate of decay of ε_n as the reduced dimension n grows. In practice, one would like to construct reduced spaces V_n such that this decay is fast, similar to the exponential decay obtained under UEA.

2. The approximation property is expressed in terms of the orthogonal projection P_n. In applications to forward modeling, we approximate the solution $u(y)$ in the space V_n by the Galerkin projection $P_n^y u(y)$. We thus wish for uniform estimates also for such approximations.

These two problems are treated in Section 2.3 and Section 2.4 respectively.

2.3 Approximation rates

Our construction of efficient reduced model spaces is based on a certain partitioning of the parameter domain $\widetilde{Y}'$ associated to the manifold $\overline{\mathcal{K}_{Y'}}$. To any $\ell = (\ell_1, \dots, \ell_d) \in \mathbb{N}_0^d$ we associate the dyadic rectangle

$$R_\ell = [2^{\ell_1}, 2^{\ell_1+1}] \times \cdots \times [2^{\ell_d}, 2^{\ell_d+1}], \tag{2.28}$$

For a positive integer L to be fixed further, we modify the definition of R_ℓ by replacing the interval $[2^{\ell_j}, 2^{\ell_j+1}]$ by $[2^{\ell_j}, \infty]$ when $\ell_j = L$ for some j. This leads to the partition

$$\widetilde{Y}' = \bigcup_{\ell \in \{0,\dots,L\}^d} R_\ell. \tag{2.29}$$

This partition is best visualized in the inverse parameter domain

$$(y_1^{-1}, \dots, y_d^{-1}) \in [0,1]^d. \tag{2.30}$$

Then, the inverse rectangles R_ℓ^{-1} split the unit cube, as shown on Figure 2.1. In particular, the rectangles touching the axes correspond to rectangles R_ℓ of infinite size.

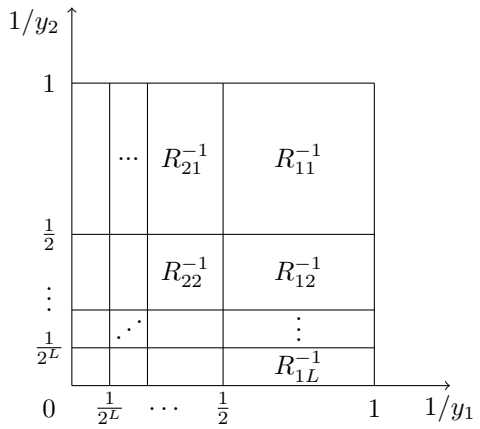


Figure 2.1 – Partition of $[0,1]^d$ by the inverse rectangles R_ℓ^{-1} in the case $d = 2$.

We build reduced model spaces through a piecewise polynomial approximation over this partition. In other words, for each $\ell \in \{0, \dots, L\}^d$, we use different polynomials

$$u_{\ell,k}(y) = \sum_{|\nu| \leqslant k} u_{\ell,\nu} y^\nu,$$

of total degree k for approximating $u(y)$ when $y \in R_\ell$, leading to a family of local reduced model spaces

$$V_{\ell,k} = \operatorname{span}\{u_{\ell,\nu} \ : \ |\nu| \leqslant k\}, \tag{2.31}$$

that can be either used individually when approximating $u(y)$ if the rectangle R_ℓ containing y is known, or summed up in order to obtain a global reduced model space.

In this section we show that this construction yields exponential convergence rates in (2.15), similar to those obtained under a Uniform Ellipticity Assumption. This requires a proper tuning between the total polynomial degree k and the integer L that determines the size of the partition. In the study of local polynomial approximation, we treat separately the inner rectangles for which $\ell \in \{0, \cdots, L-1\}^d$ and the infinite rectangles for which one or several ℓ_j are equal to L. The estimates obtained in the latter case rely on the additional assumption that the partition has a geometry of disjoint inclusions.

2.3.1 Polynomial approximation on inner rectangles

Inner rectangles R_ℓ are particular cases of rectangles of the form

$$R = [a_1, 2a_1] \times \cdots \times [a_d, 2a_d], \tag{2.32}$$

for some $a_j \geqslant 1$. The following lemma, adapted from [20], shows that one can approximate the parameter to solution map in the $\|\cdot\|_y$ and $\|\cdot\|_{H^1_0}$ norms on such rectangles, with a rate that decreases exponentially in the total polynomial degree.

Lemma 2.10. *Let R be any rectangle of the form (2.32). Then, for $k \geqslant 0$, there exists functions $u_\nu \in H^1_0(\Omega)$, $|\nu| \leqslant k$, such that*

$$\Big\| u(y) - \sum_{|\nu| \leqslant k} u_\nu y^\nu \Big\|_y \leqslant C 3^{-k}, \quad y \in R, \tag{2.33}$$

where $C := \frac{1}{\sqrt{3}} C_f$, and

$$\Big\| u(y) - \sum_{|\nu| \leqslant k} u_\nu y^\nu \Big\|_{H^1_0} \leqslant C 3^{-k}, \quad y \in R, \tag{2.34}$$

where $C := \frac{1}{\sqrt{6}} C_f$.

Proof. The exponential rate is established in [20] for a single parameter domain with uniform ellipticity assumption. Here the difficulty lies in the fact that we want the same estimate for all parametric rectangles R and thus without control on the uniform ellipticity. Still the technique of proof, based on power series, is similar.

The elliptic equation $-\mathrm{div}(a(y)u(y)) = f$ may be written in operator form

$$A_y u(y) = f,$$

where the invertible operator $A_y : H^1_0(\Omega) \to H^{-1}(\Omega)$ is defined by

$$\langle A_y v, w \rangle_{H^{-1}, H^1_0} := \int_\Omega a(y) \nabla v \cdot \nabla w \, dx = \langle v, w \rangle_y.$$

We introduce

$$\overline{y} := \frac{3}{2}(a_1, \ldots, a_d),$$

the center of the rectangle, and write any $y \in R$ as

$$y = \overline{y} + \tilde{y},$$

where the components $\tilde{y}_j$ of $\tilde{y}$ vary in $[-a_j/2, a_j/2]$. We may write $A_y = A_{\overline{y}} + \sum_{j=1}^d \tilde{y}_j A_j$, where the operators $A_j : H^1_0(\Omega) \to H^{-1}(\Omega)$ are defined by

$$\langle A_j v, w \rangle_{H^{-1}, H^1_0} := \int_{\Omega_j} \nabla v \cdot \nabla w \, dx.$$

This allows us to rewrite the equation as

$$(I + B(\tilde{y}))u(y) = g,$$

where $g := A_{\overline{y}}^{-1} f \in H_0^1(\Omega)$ and $B(\tilde{y}) = \sum_{j=1}^d \tilde{y}_j A_{\overline{y}}^{-1} A_j$ acts in $H_0^1(\Omega)$. We then observe that

$$\langle B(\tilde{y})v, w\rangle_{\overline{y}} = \langle A_{\overline{y}} B(\tilde{y})v, w\rangle_{H^{-1},H_0^1} = \sum_{j=1}^d \tilde{y}_j \langle A_j v, w\rangle_{H^{-1},H_0^1} = \sum_{j=1}^d \tilde{y}_j \int_{\Omega_j} \nabla v \cdot \nabla w \, dx,$$

and therefore, since $|\tilde{y}_j| \leqslant \frac{1}{3}\overline{y}_j$,

$$|\langle B(\tilde{y})v, w\rangle_{\overline{y}}| \leqslant \frac{1}{3}\sum_{j=1}^d \overline{y}_j \Big| \int_{\Omega_j} \nabla v \cdot \nabla w \, dx\Big| \leqslant \frac{1}{3}\|v\|_{\overline{y}}\|w\|_{\overline{y}},$$

which shows that $\|B(\tilde{y})\|_{\overline{y}\to\overline{y}} \leqslant \frac{1}{3}$. We may thus approximate $(I + B(\tilde{y}))^{-1}$ by the partial Neumann series

$$\sum_{l=0}^k (-1)^l B(\tilde{y})^l,$$

which is a polynomial in $\tilde{y}$ of total degree k. The corresponding polynomial approximation to $u(y)$ is given by

$$N_k u(y) = \sum_{l=0}^k (-1)^l B(\tilde{y})^l g = \sum_{l=0}^k (-1)^l \left(\sum_{j=1}^d \tilde{y}_j A_{\overline{y}}^{-1} A_j\right)^l g = \sum_{|\nu|\leqslant k} v_\nu \tilde{y}^\nu,$$

and coincides with the truncated power series of $\tilde{u}(\tilde{y}) := u(\overline{y} + \tilde{y})$ at $\tilde{y} = 0$, that is,

$$v_\nu := \frac{1}{\nu!}\partial^\nu u(\overline{y}), \qquad \nu! := \prod \nu_j!.$$

It can be rewritten in the form

$$N_k u(y) = \sum_{|\nu|\leqslant k} u_\nu y^\nu.$$

One has

$$\|u(y) - N_k u(y)\|_{\overline{y}} \leqslant \sum_{l>k} \|B(\tilde{y})^l g\|_{\overline{y}} \leqslant \Big(\sum_{l>k} 3^{-l}\Big)\|A_{\overline{y}}^{-1} f\|_{\overline{y}} = \frac{3^{-k}}{2}\|A_{\overline{y}}^{-1} f\|_{\overline{y}},$$

and

$$\|A_{\overline{y}}^{-1} f\|_{\overline{y}}^2 = \langle A_{\overline{y}} A_{\overline{y}}^{-1} f, A_{\overline{y}}^{-1} f\rangle_{H^{-1},H_0^1} = \langle f, u(\overline{y})\rangle_{H^{-1},H_0^1} \leqslant C_f \|u(\overline{y})\|_{H_0^1} \leqslant C_f^2,$$

where the last inequality follows from Lax-Milgram estimate since $a(\overline{y}) \geqslant 1$. This proves the estimate

$$\Big\|u(y) - \sum_{|\nu|\leqslant k} u_\nu y^\nu\Big\|_{\overline{y}} \leqslant C3^{-k}, \quad y \in R, \tag{2.35}$$

with $C := \frac{1}{2}C_f$. Using the inequalities

$$\|v\|_y^2 \leqslant \frac{4}{3}\|v\|_{\overline{y}}^2, \quad v \in H_0^1(\Omega),\ y \in R,$$

and

$$\|v\|_{H_0^1}^2 \leqslant \frac{2}{3}\|v\|_{\overline{y}}^2, \quad v \in H_0^1(\Omega),$$

we obtain the estimate (2.33) and (2.34) with the modified multiplicative constants. □

Remark 2.11. The above lemma shows that the set $\mathcal{K}_R := \{u(y) \ : \ y \in R\}$ can be approximated with accuracy $C3^{-k}$ by the space

$$V_R := \text{span}\{u_\nu \ : \ |\nu| \leqslant k\}. \tag{2.36}$$

The dimension of V_R is at most $\binom{k+d}{d}$, however, as noticed in [20], it can in fact be seen that

$$\dim(V_R) \leqslant \binom{k+d-1}{d-1}. \tag{2.37}$$

This stems from the fact that the operators defined in the above proof satisfy the dependency relation

$$A_{\overline{y}} = \sum_{j=1}^{d} \overline{y}_j A_j,$$

and therefore, one can rewrite A_y as

$$A_y := (1 + \tilde{y}_d/\overline{y}_d)A_{\overline{y}} + \sum_{j=1}^{d-1} (\tilde{y}_j - \tilde{y}_d \overline{y}_j/\overline{y}_d)A_j.$$

Using this form, the partial Neumann sum $N_k u(y)$ has at most $\binom{k+d-1}{d-1}$ independent terms.

We shall also make use of the following adaptation of the above lemma to the approximation of the limit solution map $y_{S^c} \mapsto u_S(y_{S^c})$, defined by (2.21). Its proof is an immediate adaptation of the previous one and is therefore omitted.

Lemma 2.12. *Let $S \subset \{1, \dots, d\}$, and for some $a_j \geqslant 1$, let R be a rectangle of the form*

$$R = \prod_{j \in S^c} [a_j, 2a_j]. \tag{2.38}$$

Then, there exists functions $u_\nu \in V_S$ such that

$$\Big\| u_S(y_{S^c}) - \sum_{|\nu| \leqslant k} u_\nu y_{S^c}^\nu \Big\|_{y_{S^c}} \leqslant C 3^{-k}, \quad y_{S^c} \in R, \tag{2.39}$$

where $C := \frac{1}{\sqrt{3}} C_f$, and

$$\Big\| u_S(y_{S^c}) - \sum_{|\nu| \leqslant k} u_\nu y_{S^c} \Big\|_{H_0^1} \leqslant C 3^{-k}, \quad y_{S^c} \in R, \tag{2.40}$$

where $C := \frac{1}{\sqrt{6}} C_f$.

2.3.2 Polynomial approximation on infinite rectangles

We now consider the infinite rectangles R_ℓ, corresponding to the ℓ such that some of the ℓ_j equal L. We define

$$S := \{j \ : \ \ell_j = L\}, \tag{2.41}$$

the set of such indices. When $y \in R_\ell$, we thus have

$$y_j \geqslant 2^L, \quad j \in S,$$

and so $u(y)$ should be close to $u_S(y_{S^c})$ as L is large. On the other hand y_{S^c} belongs to a rectangle of the form

$$R_{\ell_{S^c}} = \prod_{j \in S^c} [2^{\ell_j}, 2^{\ell_j + 1}].$$

Therefore, by Lemma 2.12, we can approximate $u_S(y_{S^c})$ by a polynomial of total degree k in these restricted variables.

In order to conclude that this polynomial is a good approximation to $u(y)$ on R_ℓ, we need a quantitative

estimate on the convergence of $u(y)$ towards $u_S(y_{S^c})$. Let us observe that since

$$\sum_{j=1}^{d} y_j \int_{\Omega_j} \nabla u(y) \cdot \nabla v \, dx = \sum_{j \in S^c} y_j \int_{\Omega_j} \nabla u_S(y_{S^c}) \cdot \nabla v \, dx = \langle f, v \rangle_{H^{-1}, H_0^1}, \quad v \in V_S,$$

the function $u_S(y_{S^c})$ coincides with the orthogonal projection of $u(y)$ onto V_S for the y-norm, as well as for the y_{S^c}-norm:

$$u_S(y_{S^c}) = P_{V_S}^{y} u(y) = P_{V_S}^{y_{S^c}} u(y). \tag{2.42}$$

In addition, with

$$\Omega_S := \bigcup_{j \in S} \Omega_j, \tag{2.43}$$

we have

$$2^L \|\nabla u(y)\|^2_{L^2(\Omega_S)} \leqslant \sum_{j \in S} y_j \int_{\Omega_j} |\nabla u(y)|^2 \, dx \leqslant \langle f, u(y) \rangle_{H^{-1}, H_0^1} \leqslant C_f^2,$$

since $\|u(y)\|_{H_0^1} \leqslant C_f$, and therefore, since $\nabla u_S(y_{S^c}) = 0$ on Ω_S, we find that

$$\|\nabla u(y) - \nabla u_S(y_{S^c})\|_{L^2(\Omega_S)} \leqslant C_f 2^{-L/2}. \tag{2.44}$$

Our objective is to obtain a similar error bound on the remaining domains Ω_j for $j \in S^c$. This turns out to be feasible, with an even better rate 2^{-L}, when making certain geometric assumptions on the partition of the domain Ω.

Definition 2.13. We say that $\{\Omega_1, \ldots, \Omega_d\}$ is a *Lipschitz partition* if and only if for any subset $T \subset \{1, \ldots, d\}$, the domain $\Omega_T = \bigcup_{j \in T} \Omega_j$ has Lipschitz boundaries.

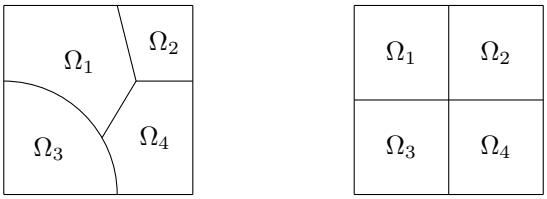


Figure 2.2 – A Lipschitz partition of Ω (left) and a counter-example (right) since $\Omega_1 \cup \Omega_4$ is not Lipschitz.

Note that such a property is stronger than just saying that each domain is Lipschitz, see Figure 2.2 (right) for a counter-example. In a Lipschitz partition, all subdomains Ω_j are Lipschitz, and the common boundary between two subdomains is either empty or a hypersurface of dimension $\dim(\Omega) - 1$, as illustrated on Figure 2.2 (left). In particular, it is easily checked that partitions consisting of a background domain and well separated subdomains that have Lipschitz boundaries fall in this category. Similar to the Ω_T, the individual Ω_j could have several connected components, that should then be well separated. Here by "well separated", we mean that δ-neighbourhoods of the subdomains remain disjoints for some $\delta > 0$.

For the inner domains Ω_T such that $\partial\Omega_T \cap \partial\Omega = \emptyset$, the classical Stein's extension theorem [165] guarantees the existence of continuous extension operators

$$E_T : H^1(\Omega_T) \to H^1(\Omega),$$

that satisfy $(E_T v)|_{\Omega_T} = v$ for all $v \in H^1(\Omega_T)$. We refer to chapter 5 of [1] for a relatively simple construction of the extension operator E_j by local reflection after using a partitioning of unity along the boundary of Ω_T and local transformations mapping the boundary to the hyperplane $\mathbb{R}^{n-1}$.

For the domains Ω_T touching the boundary $\partial\Omega$, these operators are modified in order to take into account the homogeneous boundary condition, and we refer to [192] for such adaptations. Here, the relevant space is

$$\tilde{H}^1(\Omega_T) := R_T(H_0^1(\Omega)), \tag{2.45}$$

where R_T is the restriction to Ω_T, over which $v \mapsto \|\nabla v\|_{L^2(\Omega_T)}$ is equivalent to the H^1 norm by Poincaré inequality. Then, there exists a continuous extension operator

$$E_T : \tilde{H}^1(\Omega_T) \to H^1_0(\Omega).$$

Note that the norm of all these operators depends on the geometry of the partition. These operators are instrumental in proving the following convergence estimate.

Lemma 2.14. *Assume that $\{\Omega_1, \dots, \Omega_d\}$ is a Lipschitz partition of Ω. Then there exists a constant C_0 that only depends on the geometry of the partition such that for any $S \subset \{1, \dots, d\}$ and $y = (y_S, y_{S^c}) \in Y'$, one has*

$$\|u(y) - u_S(y_{S^c})\|_{H^1_0} \leqslant C_0 C_f \max_{j \in S} y_j^{-1}. \tag{2.46}$$

In particular, for the infinite rectangle R_ℓ,

$$\|u(y) - u_S(y_{S^c})\|_{H^1_0} \leqslant C_0 C_f 2^{-L}, \quad y \in R_\ell, \tag{2.47}$$

with S defined by (2.41).

Proof. We first note that it suffices to prove (2.46) in the particular case where the largest y_j are those for which $j \in S$. Indeed, if this is not the case, we use the decomposition

$$u(y) - u_S(y_{S^c}) = (u(y) - u_{S'}(y_{S'^c})) - (u(y') - u_{S'}(y_{S'^c})) + (u(y') - u_S(y_{S^c})),$$

with $S' = \{i \ : \ y_i \geqslant \min_{j \in S} y_j\}$ and y' defined by $y'_j = \max_{i=1,\dots,d} y_i$ if $j \in S$, $y'_j = y_j$ otherwise, so that each term falls in this particular case and will be bounded in H^1_0 norm by $C_0 C_f \max_{j \in S} y_j^{-1}$. This leads to the same estimate (2.46) up to a factor 3 in constant C_0. In addition, up to reordering the subdomains Ω_j, we may assume $y_1 \geqslant \cdots \geqslant y_d$ and therefore $S = \{1, \dots, |S|\}$.

Fix $j \geqslant |S|$, and denote $u = u(y)$ and $u_S = u_S(y_{S^c})$ for simplicity. We define the Lipschitz domain $\Omega^j = \overline{\Omega}_1 \cup \cdots \cup \overline{\Omega}_j$, remarking that

$$\Omega_S = \bigcup_{j \in S} \Omega_j = \Omega^{|S|}.$$

Poincaré's inequality ensures that there exists a function c on Ω^j, constant on any connected component of Ω^j, and null on $\partial\Omega \cap \Omega^j$, such that

$$\|u - u_S - c\|_{H^1(\Omega^j)} \leqslant C_P \|\nabla(u - u_S)\|_{L^2(\Omega^j)},$$

with C_P the maximal Poincaré constant of all unions of subdomains from the partition. Moreover, there is an extension $v \in H^1_0(\Omega)$ of $u - u_S - c \in \tilde{H}^1(\Omega^j)$ such that

$$\|v\|_{H^1_0(\Omega)} \leqslant C_E \|u - u_S - c\|_{H^1(\Omega^j)} \leqslant C_E C_P \|\nabla(u - u_S)\|_{L^2(\Omega^j)},$$

with C_E the maximal norm of all extension operators E_T, $T \subset \{1, \dots, d\}$.

As $u - u_S - v = c$ on $\Omega_S \subset \Omega^j$, the function $u - u_S - v$ is in V_S, and therefore orthogonal to $u - u_S = u - P^y_{V_S} u$ for the $\|\cdot\|_y$ norm:

$$\begin{aligned}
0 &= \langle u - u_S, u - u_S - v \rangle_y \\
&= \sum_{i=1}^d y_i \int_{\Omega_i} |\nabla(u - u_S)|^2 - \sum_{i=1}^d y_i \int_{\Omega_i} \nabla(u - u_S) \cdot \nabla v \\
&= \sum_{i>j} y_i \int_{\Omega_i} |\nabla(u - u_S)|^2 - \sum_{i>j} y_i \int_{\Omega_i} \nabla(u - u_S) \cdot \nabla v
\end{aligned}$$

since $\nabla v = \nabla(u - u_S)$ on Ω^j. In particular, we obtain

$$\begin{aligned} y_{j+1}\|\nabla(u-u_S)\|^2_{L^2(\Omega_{j+1})} &\leqslant \sum_{i>j} y_i \int_{\Omega_i} |\nabla(u-u_S)|^2 \\ &\leqslant y_{j+1} \int_{\Omega\setminus\Omega^j} |\nabla(u-u_S)\cdot\nabla v| \\ &\leqslant y_{j+1}\|u-u_S\|_{H^1_0(\Omega)}\|v\|_{H^1_0(\Omega)} \\ &\leqslant y_{j+1}\|u-u_S\|_{H^1_0(\Omega)} C_P C_E \|\nabla(u-u_S)\|_{L^2(\Omega^j)}, \end{aligned}$$

and therefore

$$\|\nabla(u-u_S)\|^2_{L^2(\Omega^{j+1})} \leqslant (1+C_PC_E)\|\nabla(u-u_S)\|_{L^2(\Omega)}\|\nabla(u-u_S)\|_{L^2(\Omega^j)}.$$

Applying this inequality inductively for $j = d-1,\dots,d-r$, we get

$$\|\nabla(u-u_S)\|_{L^2(\Omega)} \leqslant (1+C_PC_E)^{2^r-1}\|\nabla(u-u_S)\|_{L^2(\Omega^{d-r})},$$

for any $r = 1,\dots,d-|S|$. For $r = d-|S|$, this results in the bound

$$\|\nabla(u-u_S)\|^2_{L^2(\Omega)} \leqslant C_0\|\nabla(u-u_S)\|^2_{L^2(\Omega_S)} = C_0\|\nabla u\|^2_{L^2(\Omega_S)}, \tag{2.48}$$

for any non-empty S, with $C_0 = (1+C_PC_E)^{2^{d-1}}$.

We now write

$$\begin{aligned} (\min_{i\in S} y_i)\|\nabla(u-u_S)\|^2_{L^2(\Omega_S)} &\leqslant \|u-u_S\|^2_y = \langle u, u-2u_S\rangle_y + \langle u_S, u_S\rangle_{y_{S^c}} \\ &= \langle f, u-u_S\rangle_{H^{-1},H^1_0} \leqslant C_f\|\nabla(u-u_S)\|_{L^2(\Omega)}, \end{aligned}$$

which, combined to the previous estimate, gives

$$\|u-u_S\|_{H^1_0} = \|\nabla(u-u_S)\|_{L^2(\Omega)} \leqslant C_0C_f \max_{i\in S} y_i^{-1},$$

therefore proving (2.46). For (2.47), we simply notice that $\max_{j\in S} y_j^{-1} \leqslant 2^{-L}$ for $y \in Y' \cap R_\ell$, and use a continuity argument when y takes infinite values. □

Combining the estimate (2.47) from the above lemma with (2.40) from Lemma 2.12, we obtain the following estimate for polynomial approximation on an infinite rectangle R_ℓ:

$$\Big\|u(y) - \sum_{|\nu|\leqslant k} u_\nu y^\nu_{S^c}\Big\|_{H^1_0} \leqslant \frac{C_f}{\sqrt{6}}\,3^{-k} + C_0C_f2^{-L}, \quad y \in R_\ell, \tag{2.49}$$

where C_0 is the constant in (2.47). This estimate hints how the level L in the partition should be tuned to the total polynomial degree k, so that the two contributions in the above estimate are of the same order.

Remark 2.15. Note that the constant $C_0 = (1+C_PC_E)^{2^{d-1}}$ becomes prohibitive even for moderate values of d. However, under more restrictive geometric assumptions, for instance if the subdomains $\overline{\Omega}_2,\dots,\overline{\Omega}_d$ are disjoint inclusions in a background Ω_1, better bounds can be obtained, with a constant C_0 that does not suffer a similar curse of dimensionality, by replacing the induction in the proof by a two-step procedure, consisting of extensions first from the high-diffusivity inclusions to the background, and then to the whole domain Ω.

2.3.3 Approximation rates and n-widths

We are now in position to establish an approximation result for the reduced model spaces. For this purpose, we fix the smallest level $L = L_k \geqslant 1$ such that

$$C_0C_f2^{-L} \leqslant \frac{C_f}{\sqrt{3}}\,3^{-k}.$$

In particular L scales linearly with k, with the bound $\alpha k + \beta \leqslant L_k \leqslant \alpha k + \gamma$, where

$$\alpha := \frac{\ln 3}{\ln 2}, \quad \beta := \frac{\ln(\sqrt{3}C_0)}{\ln 2}, \quad \gamma := \frac{\ln(2\sqrt{3}C_0)}{\ln 2}. \tag{2.50}$$

Then, the polynomial approximation estimates (2.34) and (2.49) show that for each $\ell \in \{0, \cdots, L_k\}^d$, there exist functions $u_{\ell,\nu} \in H^1_0(\Omega)$ such that

$$\Big\| u(y) - \sum_{|\nu| \leqslant k} u_{\ell,\nu} y^\nu \Big\|_{H^1_0} \leqslant \left(\frac{C_f}{\sqrt{6}} + \frac{C_f}{\sqrt{3}} \right) 3^{-k} \leqslant C_f 3^{-k}, \quad y \in R_\ell.$$

Note that in the case of an infinite rectangle R_ℓ, the $u_{\ell,\nu}$ are non trivial only for monomials of the form $y^\nu_{S^c}$ and they belong to V_S, where $S := \{j \ : \ \ell_j = L_k\}$.

Thus the solutions $u(y)$ for $y \in R_\ell$ are approximated with accuracy $C_f 3^{-k}$ in the space

$$V_{\ell,k} := \operatorname{span}\{u_{\ell,\nu} \ : \ |\nu| \leqslant k\},$$

which in view of Remark 2.11 has dimension at most $\binom{k+d-1}{d-1}$.

Note also that approximating the reduced manifold $\overline{\mathcal{K}_{Y''}}$ defined in (2.22) requires a smaller subset of rectangles, since

$$\{y \in \widetilde{Y}' \ : \ \min y_j = 1\} \subset \bigcup_{\ell \in E_k} R_\ell, \qquad E_k := \{0, \cdots, L_k\}^d \setminus \{1, \cdots, L_k\}^d.$$

We thus introduce the reduced model space

$$V_n := \bigoplus_{\ell \in E_k} V_{\ell,k}, \quad n = \dim(V_n) \leqslant \#(E_k) \binom{k+d-1}{d-1}, \tag{2.51}$$

and find that

$$\|u(y) - P_n u(y)\|_{H^1_0} \leqslant C_f 3^{-k}, \tag{2.52}$$

for all $y \in \widetilde{Y}'$ such that $\min y_j = 1$. In view of (2.50), there exists a constant C that depends on d and C_0, such that

$$n \leqslant ((L_k+1)^d - L_k^d) \binom{k+d-1}{d-1} \leqslant C(k+1)^{2d-2}. \tag{2.53}$$

This leads to the following approximation theorem.

Theorem 2.16. *Assume that the partition has the geometry of disjoint inclusions. The reduced basis space V_n defined in* (2.51) *then satisfies*

$$\|u(y) - P_n u(y)\|_{H^1_0} \leqslant C \exp\left(-cn^{\frac{1}{2d-2}} \right), \tag{2.54}$$

for all $y \in \widetilde{Y}' = [1, \infty]^d$ such that $\min y_j = 1$. The Kolmogorov n-width (2.2) *of the reduced manifold $\overline{\mathcal{K}_{Y''}}$ satisfies*

$$d_n(\overline{\mathcal{K}_{Y''}})_{H^1_0} \leqslant C \exp\left(-cn^{\frac{1}{2d-2}} \right). \tag{2.55}$$

Over the full manifold $\overline{\mathcal{K}_Y}$, one has the estimate in relative error

$$\|u(y) - P_n u(y)\|_{H^1_0} \leqslant C \exp\left(-cn^{\frac{1}{2d-2}} \right) \|u(y)\|_{H^1_0}, \tag{2.56}$$

for all $y \in \widetilde{Y} = (0, \infty]^d$. The positive constants c and C only depend on d, C_f, c_f, and on the geometry of the partition through the constant C_0.

Proof. The estimate (2.54) follows directly by combining (2.52) and (2.53), and (2.55) is an immediate consequence. We then derive (2.56) by using the homogeneity property (2.13) and the lower inequality in (2.25), similar to the proof of (2.27) in Theorem 2.9. □

Remark 2.17. In the above construction of V_n, the dimension n only takes the values $n_k := \#(E_k)\binom{k+d-1}{d-1}$ for $k \geqslant 0$. However it is easily seen that if we set $V_n = V_{n_k}$ for $n_k \leqslant n < n_{k+1}$, then all the estimates in the above theorem remain valid up to a change in the constants C and c.

Remark 2.18. Note that the union of the $V_{\ell,k}$ for $\ell \in E_k$ would suffice to approximate $\overline{\mathcal{K}_{Y''}}$ with uniform accuracy $C_f 3^{-k}$, their sum V_n is an overkill. When y is known, for example in forward modeling, it is therefore possible to first identify the proper space $V_{\ell,k}$ associated to the rectangle R_ℓ that contains y, and build the approximation to $u(y)$ from this space. This nonlinear reduced modeling strategy has been studied in [36] with similar local polynomial approximation under UEA, and in [68, 125, 194] with local reduced basis. The natural benchmark is given by the notion of library width introduced in [169], that is defined for any compact set $\mathcal{K}$ in a Banach space V as

$$d_{n,N}(\mathcal{K})_V := \inf_{\#(\mathcal{L}_n)\leqslant N} \sup_{u\in\mathcal{K}} \min_{V_n\in\mathcal{L}_n} \min_{v\in V_n} \|u-v\|_V, \tag{2.57}$$

where the first infimum is taken over all libraries $\mathcal{L}_n$ of n-dimensional spaces with cardinality at most N. Our results thus show that

$$d_{n,N}(\overline{\mathcal{K}_{Y''}})_{H_0^1} \leqslant C_f 3^{-k} \sim C\exp(-cn^{\frac{1}{d}}), \quad n := \binom{k+d-1}{d-1},\ N = (L_k+1)^d - L_k^d.$$

Note that the above sub-exponential rate can be misleading due to fact that the constant c has a hidden dependence in d. As an example, up to the constant C_f, we find that taking $k = 4, 7, 9$ leads to error bounds 3^{-k} of order $10^{-2}, 10^{-3}, 10^{-4}$, with $n = 15, 36, 55$ for $d = 3$, and $n = 35, 120, 220$ for $d = 4$, which is far better than the value of $\exp(-n^{\frac{1}{d}})$.

Remark 2.19. In view of the results from [33] and [65], we are ensured that a proper selection of reduced basis elements in the manifold $\overline{\mathcal{K}_{Y''}}$ should generate spaces V_n that perform at least with the same exponential rates as those achieved by the spaces V_n in Theorem 2.16. As explained in the introduction, reduced basis spaces may perform significantly better than reduced model spaces based on polynomial or piecewise polynomial approximation. This occurs in particular when the polynomial coefficients have certain linear dependency, as established in [20] for the elliptic problem with piecewise constant coefficients in the low contrast regime, and recalled in Remark 3.2. There, it is shown that the rate $\mathcal{O}(\exp(-cn^{\frac{1}{d}}))$ is at least improved to $\mathcal{O}(\exp(-cn^{\frac{1}{d-1}}))$ and that further improvements in the rate may result from certain symmetry properties of the domain partition, however not circumventing the curse of dimensionality. While we do not pursue this analysis in the present high contrast setting, we expect similar results to hold.

2.4 Forward modeling and inverse problems

2.4.1 Galerkin projection

In the context of forward modeling, the reduced model space V_n is used to approximate the parameter to solution map, by a map

$$y \mapsto \tilde{u}(y) \in V_n,$$

computed through the Galerkin method: $\tilde{u}(y) \in V_n$ is such that

$$\sum_{j=1}^{d} y_j \int_{\Omega_j} \nabla\tilde{u}(y)\cdot\nabla v\,dx = \langle f, v\rangle_{H^{-1},H_0^1}, \quad v \in V_n.$$

Therefore $\langle\tilde{u}(y), v\rangle_y = \langle u(y), v\rangle_y$, that is

$$\tilde{u}(y) = P_{V_n}^y u(y),$$

where $P_{V_n}^y$ is the projection onto V_n with respect to norm $\|\cdot\|_y$.

Hence, one would like to derive estimates on $\|u(y) - P_{V_n}^y u(y)\|_{H_0^1}$ in place of the estimates on $\|u(y) - P_n u(y)\|_{H_0^1}$ that we have obtained so far, since $P_n u(y)$ is not practically accessible. As explained in the intro-

duction, we cannot be satisfied with combining the latter estimates with the bound

$$\|u(y) - P_n^y u(y)\|_{H_0^1} \leqslant \kappa(y)^{1/2} \|u(y) - P_n u(y)\|_{H_0^1}$$

derived from Cea's lemma, since the multiplicative constant $\kappa(y)$ from (2.9) is not uniformly bounded over the manifolds $\mathcal{K}_Y$, $\mathcal{K}_{Y'}$ or $\overline{\mathcal{K}_{Y''}}$. Here, we shall employ another approach to derive the same rates of convergence for $\|u(y) - P_{V_n}^y u(y)\|_{H_0^1}$.

One first observation is that, in order for Galerkin projection P_n^y onto a reduced model space V_n to satisfy a convergence bound in relative error, it is critical that this space contains some functions from the limit spaces V_S. This is expressed by the following result.

Proposition 2.20. *Assume that there exists $S \subsetneq \{1, \ldots, d\}$ such that $V_n \cap V_S = \{0\}$. Then for any $C \in (0,1)$, there exists $y \in Y' = [1,\infty)^d$ such that*

$$\|u(y) - P_{V_n}^y u(y)\|_{H_0^1} \geqslant C \|u(y)\|_{H_0^1}. \tag{2.58}$$

Proof. Since $V_n \cap V_S = \{0\}$, the quantity $\|\nabla v\|_{L^2(\Omega_S)}$ is a norm on V_n and one can define

$$\alpha = \min_{v \in V_n} \frac{\|\nabla v\|_{L^2(\Omega_S)}}{\|v\|_{H_0^1}} > 0.$$

For any $\varepsilon > 0$, take $y_j = \varepsilon^{-2}$ for $j \in S$ and $y_j = 1$ for $j \in S^c$. Then, for $v = P_{V_n}^y u(y)$,

$$\frac{\alpha}{\varepsilon}\|v\|_{H_0^1} \leqslant \frac{1}{\varepsilon}\|\nabla v\|_{L^2(\Omega_S)} \leqslant \|v\|_y \leqslant \|u(y)\|_y \leqslant C_f \leqslant \frac{C_f}{c_f}\|u(y)\|_{H_0^1},$$

where we have used the framings (2.25) and (2.26). Therefore, taking $\varepsilon = \frac{c_f}{C_f}\alpha(1-C)$ implies $\|v\|_{H_0^1} \leqslant (1-C)\|u(y)\|_{H_0^1}$, and (2.58) follows. □

However, in the construction of V_n in Section 2.3, each space $V_{\ell,k}$ is a subset of V_S for $S = \{j \,:\, \ell_j = L_k\}$. This prevents the phenomenon described in the previous proposition from occurring. Instead, we obtain similar convergence bounds as those obtained for P_n, as expressed in the following result.

Theorem 2.21. *Assume that the partition of Ω has the geometry of disjoint inclusions. On the rectangles R_ℓ for $\ell \in \{0, \ldots, L\}^d$, the following uniform convergence estimates hold:*

$$\|u(y) - P_{V_{\ell,k}}^y u(y)\|_{H_0^1} \leqslant \frac{C_f}{\sqrt{3}} 3^{-k}, \quad y \in R_\ell, \tag{2.59}$$

if $\|\ell\|_\infty < L$, and

$$\|u(y) - P_{V_{\ell,k}}^y u(y)\|_{H_0^1} \leqslant \frac{C_f}{\sqrt{3}} 3^{-k} + C_0 C_f 2^{-L}, \quad y \in R_\ell, \tag{2.60}$$

if $\|\ell\|_\infty = L$. As a consequence, with $L = L_k$ and V_n defined as in § 2.3.3, one has the estimates

$$\|u(y) - P_n^y u(y)\|_{H_0^1} \leqslant C \exp\left(-cn^{\frac{1}{2d-2}}\right), \tag{2.61}$$

for all $y \in \widetilde{Y}'$ such that $\min y_j = 1$, and

$$\|u(y) - P_{V_{\ell,k}}^y u(y)\|_{H_0^1} \leqslant C \exp\left(-cn^{1/(2d-2)}\right) \|u(y)\|_{H_0^1}, \tag{2.62}$$

for all $y \in \widetilde{Y}$, with constants c and C that only depend on d, C_f, c_f, and on the geometry of the partition through the constant C_0.

Proof. For bounded rectangles R_ℓ with $\|\ell\|_\infty < L$, we know from Lemma 2.10, and more precisely from (2.33),

that

$$\|u(y) - P^y_{V_{\ell,k}} u(y)\|_y = \min_{v \in V_{\ell,k}} \|u(y) - v\|_y \leqslant \Big\|u(y) - \sum_{|\nu| \leqslant k} u_\nu y^\nu\Big\|_y \leqslant \frac{C_f}{\sqrt{3}} 3^{-k}$$

for any $y \in R_\ell$. Since all the y_j are greater or equal to 1, one has $\|v\|_{H^1_0} \leqslant \|v\|_y$ for all v and therefore (2.59) follows.

For infinite rectangles R_ℓ such that $\|\ell\|_\infty = L$, we again introduce $S = \{j \,:\, \ell_j = L\}$. Then, using (2.47),

$$\begin{aligned}\|u(y) - P^y_{V_{\ell,k}} u(y)\|_{H^1_0} &\leqslant \|u(y) - u_S(y_{S^c})\|_{H^1_0} + \|u_S(y_{S^c}) - P^y_{V_{\ell,k}} u(y)\|_{H^1_0} \\ &\leqslant C_0 C_f 2^{-L} + \|u_S(y_{S^c}) - P^y_{V_{\ell,k}} u(y)\|_{H^1_0}.\end{aligned}$$

Since $V_{\ell,k} \subset V_S$, we have

$$P^y_{V_{\ell,k}} u(y) = P^y_{V_{\ell,k}} P^y_{V_S} u(y) = P^y_{V_{\ell,k}} u_S(y_{S^c}) = P^{y_{S^c}}_{V_{\ell,k}} u_S(y_{S^c}),$$

Similarly to the previous case, we apply (2.39) from Lemma 2.12:

$$\|u_S(y_{S^c}) - P^y_{V_{\ell,k}} u_S(y_{S^c})\|_{H^1_0} \leqslant \|u_S(y_{S^c}) - P^y_{V_{\ell,k}} u_S(y_{S^c})\|_y \leqslant \frac{C_f}{\sqrt{3}} 3^{-k},$$

and we thus obtain (2.60).

After taking $L = L_k$ and defining V_n as the sum of the $V_{\ell,k}$ for $\ell \in E_k$, the derivation of (2.61) and (2.62) is exactly the same as for (2.54) and (2.56). □

Remark 2.22. As in Remark 2.19, it is expected that the same rate of convergence is attained if V_n is a reduced basis space generated by solutions $u(y^i)$, $i = 1, \ldots, n$, as long as there are $O\big(\binom{k+d-1}{d-1}\big)$ samples y^i in each rectangle, however with samples forced to be of the form $u_S(y^i_{S^c}) \in V_S$ in the case of infinite rectangles.

2.4.2 State and parameter estimation

The state estimation problem consists in retrieving the solution $u = u(y)$ when the parameter y is unknown, and one observes m linear measurements

$$z_i = \ell_i(u), \quad i = 1, \ldots, m,$$

where the ℓ_i are continuous linear functional on the Hilbert space V that contains the solution manifold. These linear functionals may thus be written in terms of Riesz representers

$$\ell_i(v) = \langle \omega_i, v \rangle_V.$$

The Parametrized Background Data Weak (PBDW) method, introduced in [124] and further studied in [33], exploits the fact that all potential solutions are well approximated by reduced model spaces V_n. It is based on a simple recovery algorithm that consists in solving the problem

$$\min_{v^* \in V_z} \min_{\tilde{v} \in V_n} \|v^* - \tilde{v}\|_V, \tag{2.63}$$

where, for $z = (z_1, \ldots, z_m) \in \mathbb{R}^m$,

$$V_z := \{v \in V \,:\, \ell_i(v) = z_i,\ i = 1, \ldots, m\},$$

is the affine space of functions that agree with the measurements.

The analysis of this problem is governed by the quantity

$$\mu_n^m = \mu(V_n, W_m) := \sup_{v \in V_n} \frac{\|v\|_V}{\|P_{W_m} v\|_V}, \tag{2.64}$$

where $W_m := \operatorname{span}\{\omega_1, \dots, \omega_m\}$, which is finite if and only if $V_n \cap W_m^\perp = \{0\}$. Then, there exists a unique minimizing pair

$$(u^*, \tilde{u}) = (u^*(z), \tilde{u}(z)) \in V_z \times V_n$$

to (2.63), which satisfies the estimates

$$\|u - \tilde{u}\|_V \leqslant \mu_n^m \min_{v \in V_n} \|u - v\|_V, \tag{2.65}$$

and

$$\|u - u^*\|_V \leqslant \mu_n^m \min_{v \in V_n + (W_m \cap V_n^\perp)} \|u - v\|_V. \tag{2.66}$$

The computation of $(u^*, \tilde{u})$ amounts to solving finite linear systems, and both solutions depend linearly on z.

Turning to our specific elliptic problem, and assuming that the ℓ_i belong to $H^{-1}(\Omega) = V'$ for $V = H_0^1(\Omega)$, we may apply the above PBDW method using the reduced basis spaces V_n introduced in Section 2.3. As an immediate consequence of Theorem (2.16), we obtain a recovery estimate in relative error.

Proposition 2.23. *Let $y \in \widetilde{Y}$ and $u = u(y)$. Then both estimators $\tilde{u} \in V_n$ and $u^* \in V_z$ satisfy*

$$\max(\|u - \tilde{u}\|_{H_0^1}, \|u - u^*\|_{H_0^1}) \leqslant C\mu_n^m \exp\left(-cn^{\frac{1}{2d-2}}\right) \|u\|_{H_0^1}. \tag{2.67}$$

The constants $C, c > 0$ only depend on d, C_f, c_f, and on the geometry of the partition through the constant C_0.

Proof. It follows readily by combining (2.56) with the recovery estimates (2.65) and (2.66). □

We next turn to the problem of parameter estimation, namely recovering an approximation y^* to y from the measurements z. In contrast to state estimation, this is a nonlinear inverse problem since the first mapping in

$$y \mapsto u \mapsto z$$

is typically nonlinear. One way of relaxing this problem into a linear one is by first using a recovery u^* of the state u, for example obtained by the PBDW method. One then defines y^* as the minimizer over $\widetilde{Y}$ of the residual

$$R(y) := \|\operatorname{div}(a(y)\nabla u^*) + f\|_{H^{-1}}.$$

This is a quadratic problem when $a(y)$ has an affine dependence in y, that can be solved by standard quadratic optimization methods. The rationale for this approach is the fact that

$$R(y) = \|A_y u^* - A_y u(y)\|_{H^{-1}} \sim \|u^* - u(y)\|_{H_0^1},$$

and therefore we should be close to finding the parameter y that best explains the approximation u^*. Unfortunately, this approach is not much viable in the high-contrast regime since the equivalence $\|A_y v\|_{H^{-1}} \sim \|v\|_{H_0^1}$ has constants that are not uniform in y and deteriorate with the level of contrast $\kappa(y)$.

Instead, we propose a more specific approach that exploits the piecewise constant structure of $a(y)$, assuming that V_n is a reduced space of the form

$$V_n = \operatorname{span}\{u(y^1), \dots, u(y^n)\},$$

for some properly selected parameter vectors

$$y^i = (y_1^i, \dots, y_d^i), \quad i = 1, \dots, n.$$

As mentioned, see Remark (2.19), these spaces satisfy the same exponential convergence bounds as the spaces constructed in Section 2.3.

The PBDW estimator $u^* = u^*(z) \in V_n$ thus has the form

$$u^* = \sum_{i=1}^n c_i u(y^i) \in V_n$$

and satisfies a similar bound (2.67) as in the above proposition. Then, on the particular domain Ω_j, one has

$$\frac{f}{y_j} = -\Delta u|_{\Omega_j} \approx -\sum_{i=1}^{n} c_i \Delta u(y^i) = \sum_{i=1}^{n} c_i \frac{f}{y_j^i},$$

and therefore, a natural candidate for the parameter estimate is $y^* = (y_1^*, \dots, y_d^*)$ with

$$y_j^* := \left(\sum_{i=1}^{n} \frac{c_i}{y_j^i} \right)^{-1}. \tag{2.68}$$

The following result gives a recovery bound in relative error for the inverse diffusivity.

Proposition 2.24. *With the notation $1/y = (1/y_1, \dots, 1/y_d)$, the estimator y^* defined by (2.68) satisfies the bound*

$$\left\| \frac{1}{y^*} - \frac{1}{y} \right\|_\infty \leqslant \frac{C_f}{c_f} C \mu_n^m \exp\left(-cn^{\frac{1}{2d-2}} \right) \left\| \frac{1}{y} \right\|_\infty, \tag{2.69}$$

where C_f and c_f are as in (2.25), and the other constants as in (2.67).

Proof. For $1 \leqslant j \leqslant d$, take $\phi \in H_0^1(\Omega_j)$, then

$$\begin{aligned}
\left| \frac{1}{y_j^*} - \frac{1}{y_j} \right| |\langle f, \phi \rangle_{H^{-1}, H_0^1}| &= \left| \sum_{i=1}^{n} \frac{c_i}{y_j^i} \int_{\Omega_j} y_j^i \nabla u(y^i) \cdot \nabla \phi \, dx - \frac{1}{y_j} \int_{\Omega_j} y_j \nabla u \cdot \nabla \phi \, dx \right| \\
&= \left| \int_{\Omega_j} \nabla (v^* - u) \cdot \nabla \phi \, dx \right| \\
&\leqslant \|v^* - u\|_{H_0^1(\Omega)} \|\phi\|_{H_0^1(\Omega_j)}.
\end{aligned}$$

Optimizing over ϕ gives

$$\left\| \frac{1}{y^*} - \frac{1}{y} \right\|_\infty \leqslant c_f^{-1} \|v^* - u\|_{H_0^1},$$

which combined with (2.67) gives

$$\left\| \frac{1}{y^*} - \frac{1}{y} \right\|_\infty \leqslant c_f^{-1} C \mu_n^m \exp\left(-cn^{\frac{1}{2d-2}} \right) \|u\|_{H_0^1}.$$

Using the Lax-Milgram estimate

$$\|u\|_{H_0^1} \leqslant C_f \left\| \frac{1}{y} \right\|_\infty,$$

we reach (2.69). □

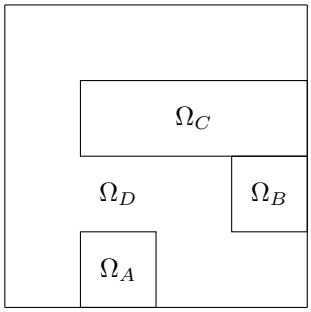
Remark 2.25. The bound (2.69) is not entirely satisfactory since the approximation error on y_j remains high when $y \in \overline{\mathcal{K}_{Y''}}$ with $y_j \gg 1$. We do not know if a bound of the form

$$\left| \frac{1}{y_j^*} - \frac{1}{y_j} \right| \leqslant \frac{\varepsilon_n}{y_j}, \quad 1 \leqslant j \leqslant d,$$

which would imply $|y_j^* - y_j| \leqslant \varepsilon_n / (1 - \varepsilon_n)\, y_j$, holds uniformly over $\overline{\mathcal{K}_{Y''}}$ with $\varepsilon_n \underset{n \to \infty}{\longrightarrow} 0$.

2.5 Numerical illustration

The base model that will be used all along the numerical illustrations is the diffusion equation (2.4) with data $f = 1$ set on the two-dimensional square $\Omega = [-1, 1]^2$ with homogeneous Dirichlet boundary conditions.



(a) Lipschitz partition of Ω.

Ω_D	Ω_B	Ω_C	Ω_A
Ω_B	Ω_D	Ω_A	Ω_C
Ω_C	Ω_A	Ω_D	Ω_B
Ω_A	Ω_C	Ω_B	Ω_D

(b) Non-lipschitz partition of Ω.

Figure 2.3 – Different partitions of $\Omega = [-1, 1]^2$ into four subdomains, considered in the numerical tests

We consider a piece-wise constant diffusion coefficient

$$a|_{\Omega_j} = y_j, \quad 1 \leqslant j \leqslant d,$$

on a partition of Ω into 16 squares of quarter side-length.

As such this partition does not satisfy the geometric assumption of "Lipschitz partition" that was critical in our analysis for the application of Lemma 2.14. Therefore we consider sub-partitions that comply to the assumptions, such as illustrated on Figure 2.3a, which amounts to equate the parameters y_j of squares belonging to the same sub-domain. This way we can consider that $y = (y_A, y_B, y_C, y_D)$ consists of four parameters, one for each subdomain.

The numerical results that we present next aim at illustrating the robustness to high contrast of the reduced basis method, and discuss in addition the effect of parameter selection, higher parametric dimensions, and inclusions that are not satisfying the geometric assumption, as exemplified on Figure 2.3b.

We construct different reduced bases $\{u(y^1), \dots, u(y^n)\}$ of moderate dimension $1 \leqslant n \leqslant 15$, for certain parameter selections $y^1, \dots, y^n$. Each reduced basis element $u(y^i)$ is numerically computed by the Galerkin method in a background finite element space V_h of dimension 6241.

The reduced basis spaces are thus subspaces of V_h, thus strictly speaking $V_n = V_{n,h}$ depends both on n and on the meshsize h. In our numerical computation, we always assess the error

$$P^y_{V_h} u(y) - P^y_{V_{n,h}} u(y).$$

We noticed that for the considered values of $n = 1, \dots, 15$ the error curves do not vary much when further reducing the mesh size h. In fact they are essentially identical when the dimension of V_h is four times smaller. Therefore, for simplicity of the presentation, we still write

$$u(y) - P^y_{V_n} u(y),$$

bearing in mind that the additional finite element error $u(y) - P^y_{V_h} u(y)$ depends on h (with algebraic decay in the finite element dimension).

All the tests were done using Python 3.8. For more information and experiments not presented here we invite the reader to look into the github repository `https://github.com/agussomacal/ROMHighContrast`.

2.5.1 Parameter selection

We first study the case of a one-parameter family: the diffusion coefficient y_A of Ω_A in Figure 2.3a varies from 1 to ∞, while the other subdomains are considered as background with all coefficents equal to 1. Thus the y^i are of the form $y^i = (y^i_A, 1, 1, 1)$.

In reduced basis constructions, the approach for selecting parameter sets is usually either random or greedy. Random selection usually performs well enough in many situations, however we shall see that it fails in the high contrast regime. This is in particular due to the fact that it does not capture the limit solutions, while we have observed in Section 2.4 that robust convergence of the Galerkin method in the high-contrast regime critically

requires to include limit solutions in the space V_n. Here, there is only one limit solution $u_\infty = u(y_\infty)$ where $y_\infty = (\infty, 1, 1, 1)$, and this element is picked by the greedy method if initialized at any other point.

More precisely, we compare four strategies for selecting the $y_A^i \in [1, \infty]$:

— Random: The y_A^i are drawn independently according to the uniform law for $1/y_A \in [0, 1]$.

— Random-∞: First the limit solution, corresponding to $y_A = \infty$, is inserted in the basis. The rest of the elements are randomly picked as in the previous case.

— Greedy H_0^1: The y^i are picked incrementally, y^{i+1} maximizing the relative H_0^1 projection error

$$\frac{\|u(y) - P_{V_i} u(y)\|_{H_0^1}}{\|u(y)\|_{H_0^1}}.$$

— Greedy Galerkin: The y^i are picked incrementally, y^{i+1} maximizing the relative H_0^1 error of the Galerkin projection

$$\frac{\|u(y) - P_{V_i}^y u(y)\|_{H_0^1}}{\|u(y)\|_{H_0^1}}.$$

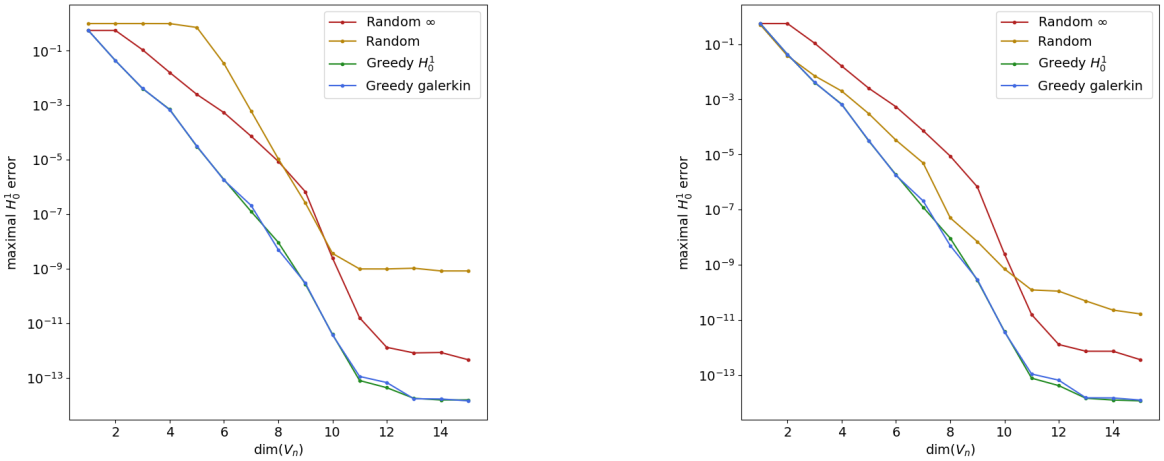


Figure 2.4 – Galerkin (left) and H_0^1 (right) projection error, both measured in H_0^1 relative error, maximized over the parameter domain, for different reduced bases, case $d = 1$.

Figure 2.4 displays on the left the evolution of the maximal relative error of the Galerkin projection

$$\sup_{y_A \in [1,\infty]} \frac{\|u(y) - P_n^y u(y)\|_{H_0^1}}{\|u(y)\|_{H_0^1}},$$

as a function of $n = \dim(V_n)$ for these various selection strategies. It reveals the superiority of the greedy selection that reaches machine precision after picking $n = 11$ reduced basis elements, and the gain in including the limit solution in the case of a random selection. As a comparison, we display on the right the decay of the relative H_0^1-orthogonal projection error

$$\sup_{y_A \in [1,\infty]} \frac{\|u(y) - P_n u(y)\|_{H_0^1}}{\|u(y)\|_{H_0^1}}$$

for the same parameter selection strategies. Here, we notice that the inclusion of the limit solution u_∞ is not anymore critical for reaching good accuracy. Nevertheless, these errors still decay faster for the greedy strategies.

Remark 2.26. As the diffusion coefficient is piecewise constant on the partition $\Omega_A \cup \Omega_A^c$, the parameter space dimension is $d = 2$ in this numerical example. The theoretical results thus provide a bound on the error of order

$\exp(-c\sqrt{n})$. However, this bound is obtained with local reduced spaces $V_{\ell,k}$ on dyadic intervals, which does not perform as well as $V_n = \bigoplus_{\ell \in E_k} V_{\ell,k}$, for which one might expect a rate closer to $\exp(-cn)$. In Figure 2.4 for $n \leqslant 11$, that is, until numerical precision issues arise, we even observe a faster than exponential convergence, that could be due to the superiority of reduced bases over polynomial approximations.

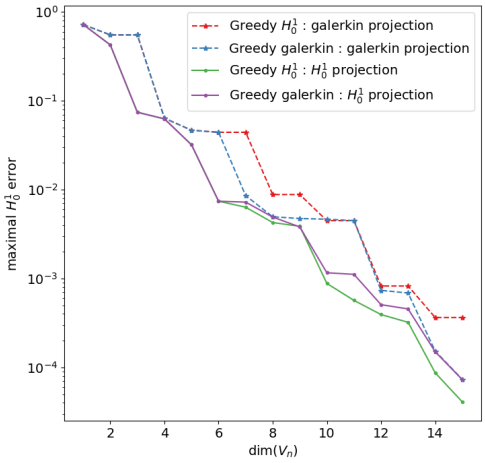


Figure 2.5 – Galerkin and H_0^1 projection error (both measured in H_0^1 relative error maximized over the parameter domain) for different reduced bases, case $d = 2$.

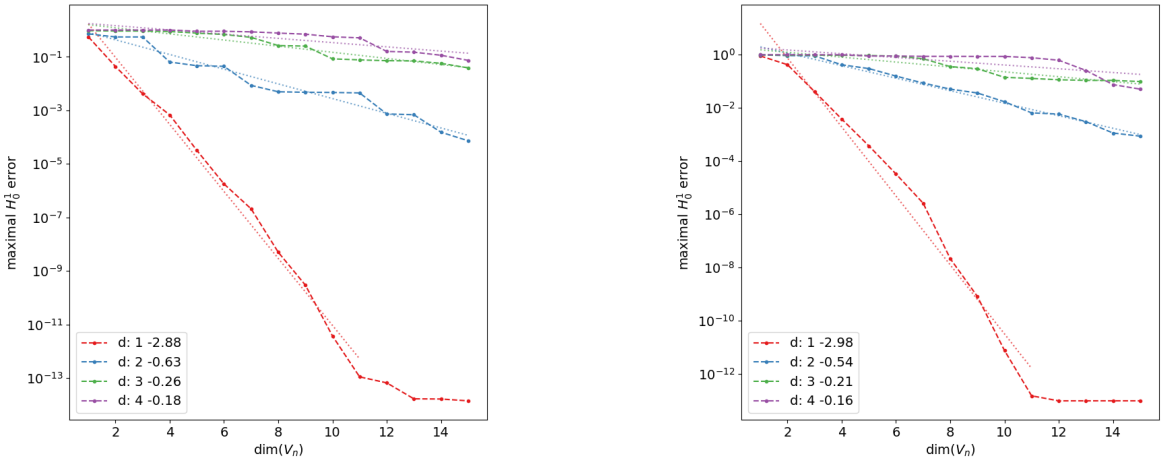


Figure 2.6 – The Galerkin projection of Greedy Galerkin method for increasing dimensionality in geometries satisfying (left) or not (right) the assumptions.

Remark 2.27. It is well known that the reduced basis can be very ill-conditioned, since $u(y^n)$ becomes extremely close to $V_{n-1} = \text{span}\{u(y^1), \ldots, u(y^{n-1})\}$ as n gets moderately large. In order to avoid numerical instabilities, prior to the computation of the Galerkin or H_0^1 projection onto V_n, we need to perform a change of basis, typically by some orthonormalization process. In our numerical test, we perform this orthonormalization with respect to the discrete ℓ^2 inner product for the nodal values in the background finite element representation, using the QR decomposition, and obtain a satisfactory stable numerical behavior. However, this process

is not invariant under permutations, and we observe that it behaves better in terms of numerical stability when sorting the reduced basis elements from higher contrast to lower contrast.

In this one parameter scenario, both greedy strategies behaved equally well. However, as we increase the dimensionality of the problem $d > 1$, Greedy Galerkin appears to be the best selection procedure, as could be expected since it optimizes the error based on the approximation which is effectively computed in forward modeling. Figure 2.5 shows this effect when $d = 2$, where y_A and y_B are allowed to vary independently while y_C and y_D are taken as background always equal to 1.

2.5.2 Influence of dimensionality and geometry

In order to study the impact of dimensionality on the approximation rates, we compare the behavior of the Greedy Galerkin selection method, as we increase the number of freely varying parameters. As before, we will have for $y = (y_A, 1, 1, 1)$ when $d = 1$, then $y = (y_A, y_B, 1, 1)$ when $d = 2$, until having all four subdomains freely varying between 1 and ∞.

In Figure 2.6 the degradation with respect to dimension is clearly observed as the approximation capabilities strongly decrease. Even thought the exponential decay rate is still conserved, the decay parameter shrinks from almost 3 down to 0.22 when $d = 4$.

Secondly, we study the case where the geometric assumptions are not satisfied. We follow the same incremental subdomains unfreezing as in the previous case but using the geometry stated in Figure 2.3b. We observe that the reduced basis approach still achieves exponential approximation rates, actually higher than in the previous example. This hints that the geometric assumptions which are needed in our proofs could be artificial, and leaves open the question of achieving such results without relying on these assumptions.

Chapter 3

Nonlinear approximation spaces for inverse problems

Abstract. This chapter is concerned with the ubiquitous inverse problem of recovering an unknown function u from finitely many measurements, possibly affected by noise. In recent years, inversion methods based on *linear* approximation spaces were introduced in [33, 124] with certified recovery bounds. It is however known that linear spaces become ineffective for approximating simple and relevant families of functions, such as piecewise smooth functions that typically occur in hyperbolic PDEs (shocks) or images (edges). For such families, *nonlinear* spaces [63] are known to significantly improve the approximation performance. The first contribution of this chapter is to provide certified recovery bounds for inversion procedures based on nonlinear approximation spaces. The second contribution is the application of this framework to the recovery of general bidimensional shapes from cell-average data. We also discuss how the application of our results to n-term approximation relates to classical results in compressed sensing.

3.1 Introduction

3.1.1 The recovery problem

In this chapter, we treat the following state estimation problem in a general Banach space V. We want to recover an approximation to an unknown function $u \in V$ from data given by m observations

$$z_i \coloneqq \ell_i(u) + \eta_i, \quad i = 1, \dots, m, \tag{3.1}$$

where $\ell_i : V \mapsto \mathbb{R}$ are known measurement functionals, and η_i is additive noise. The functionals ℓ_i often correspond to the response of a physical measurement device but they can have a different interpretation depending on the application. Their behavior can be linear (in which case the ℓ_i are linear functionals from V', the dual of V) or nonlinear. This type of recovery problem is clearly ill-posed when the dimension of V exceeds m. It nevertheless prevails in sampling and inverse problem applications where V is infinite dimensional (to name a few, see [2, 16, 74, 99]).

One natural strategy to address this difficulty is to search for a recovery of u by an element of a low-dimensional reconstruction space $V_n \subset V$. The space V_n could be either an n-dimensional linear subspace, or more generally a nonlinear approximation space parametrized by n degrees of freedom, with $n \leqslant m$.

In order to obtain quantitative results for such recovery procedures, it is necessary to possess additional information about u, usually as an assumption that u belongs to a certain model class $\mathcal{K}$ contained in V. The approximation space V_n is chosen in order to collectively approximate the elements of $\mathcal{K}$ as well as possible, in the sense that

$$\operatorname{dist}(\mathcal{K}, V_n)_V \coloneqq \max_{u \in \mathcal{K}} \min_{v \in V_n} ||u - v||_V$$

is as small as possible for moderate values of n.

Multiple theoretical results and numerical algorithms have been proposed in various fields to study and solve

the above recovery problem (we recall some relevant results below). However, to the best of our knowledge, they all involve at least one or several of the following assumptions:

— The ℓ_i are *linear* functionals,
— V_n is a *linear* (or affine) subspace of V,
— V is a *Hilbert* space,
— The model class $\mathcal{K}$ is a *ball in a smoothness space*, e.g., a unit ball in Lipschitz, Sobolev, or Besov spaces. Results involving this type of model classes have been intensively studied in the field of optimal recovery (see [35, 131, 142]).

The goal of this chapter is to develop and analyze inversion procedures that do not require any of the above assumptions. Our analysis and numerical algorithms can thus be applied to virtually any recovery problem. The starting point of our development is based on algorithms introduced for inverse state estimation using reduced order models of parametric Partial Differential Equations (PDEs). We next recall that specific framework. The presentation will also serve to explain more in depth the motivations leading to propose the present generalization.

3.1.2 State estimation with reduced models for parametric PDEs

A relevant scenario in inverse state estimation is when the model class $\mathcal{K}$ is given by the set of solutions to some parameter-dependent PDE of the general form

$$\mathcal{P}(u, y) = 0, \tag{3.2}$$

where $\mathcal{P}$ is a differential operator, y a vector of parameters ranging in some domain Y in $\mathbb{R}^d$, and u is the solution. If well-posedness holds in some Banach space V for each $y \in Y$, we denote by $u(y) \in V$ the corresponding solution for the given parameter value y and by

$$\mathcal{K}_Y \coloneqq \{u(y) \ : \ y \in Y\},$$

the *solution manifold*.

In inverse state estimation, we take $\mathcal{K} = \mathcal{K}_Y$ for the model class, so the unknown u to recover belongs to $\mathcal{K}$. However, the parameter y that satisfies $u = u(y)$ is unknown, so we cannot solve the forward problem (3.2) to approximate u. Instead, we must approximate u from the partial observational data (3.1), and the knowledge of the model class $\mathcal{K}_Y$.

For the manifold $\mathcal{K}_Y$, efficient approximation spaces V_n are usually obtained by reduced modeling techniques. In their most simple format, reduced models consist into linear spaces $(V_n)_{n\geqslant 0}$ with $\dim(V_n) = n$. The ideal benchmark in this linear approximation setting is provided by the Kolmogorov n-width

$$d_n(\mathcal{K}_Y)_V \coloneqq \inf_{\dim(V_n)\leqslant n} \operatorname{dist}(\mathcal{K}_Y, V_n)_V,$$

which describes the optimal approximation performance achievable by an n-dimensional space over the set $\mathcal{K}_Y$.

Apart from very simplified cases, the space V_n achieving the above infimum is usually out of reach. Practical model reduction techniques such as polynomial approximation in the parametrized domain [56, 58, 174] or reduced bases [68, 92, 125, 157, 194] construct spaces V_n that are "suboptimal yet good". In particular, the reduced basis method, which generates V_n by a specific selection of particular solution instances $u(y^1), \ldots, u(y^n) \in \mathcal{K}_Y$, has been proved to have approximation error $\operatorname{dist}(\mathcal{K}_Y, V_n)_V$ that decays with the same polynomial or exponential rates as $d_n(\mathcal{K}_Y)_V$, and in that sense are close to optimal [65].

3.1.3 The PBDW method

We take the *Parametrized Background Data Weak* (PBDW) method as a starting point for our analysis. The PBDW method, first introduced in [124], as well as several extensions, has been the object of a series of works [32, 33, 53, 54] on its optimality properties as a recovery algorithm. It has also been used for different practical applications, see [16, 74, 88]. We refer to [136] for an overview of the state of the art on this approach, and its connections with different fields. For our current purposes, it will suffice to recall the first version of the algorithm, which is the goal of this section.

The PBDW method uses a linear approximation space V_n of dimension $n \leqslant m$. Usually this space is a reduced model in applications. It is assumed that the ℓ_i are continuous linear functionals, that is $\ell_i \in V'$, and that V is a Hilbert space. Then, introducing the Riesz representers $\omega_i \in V$ such that $\ell_i(v) = \langle \omega_i, v \rangle_V$, the data of the noise-free observation

$$z = \ell(u) \coloneqq (\ell_1(u), \dots, \ell_m(u)),$$

is equivalent to that of the orthogonal projection $w = P_{W_m} u$ on the *Riesz measurement space*

$$W_m \coloneqq \operatorname{span}\{\omega_1, \dots, \omega_m\}.$$

Assuming linear independence of the ℓ_i, this space has dimension m. A critical quantity is the number

$$\mu_n^m = \mu(V_n, W_m) \coloneqq \max_{v \in V_n} \frac{\|v\|_V}{\|P_{W_m} v\|_V}, \tag{3.3}$$

that describes the "stability" of the description of an element of V_n by its projection onto W_m, and may be thought of as the inverse cosine of the angle between W_m and V_n. In particular, this quantity is finite only when $n \leqslant m$. It can be explicitly computed as the inverse of the smallest singular value of a cross-grammian matrix between orthonormal bases of V_n and W_m (see [33, 136]).

The PBDW method consists in solving the minimization problem

$$\min_{v^* \in V_w} \min_{\tilde{v} \in V_n} \|v^* - \tilde{v}\|_V,$$

where $V_w \coloneqq w + W_m^\perp$ is the set of all states v such that $P_{W_m} v = w$. We denote by $(u^*, \tilde{u}) \in V_w \times V_n$ the minimizing pair, which is unique when $\mu_n^m < \infty$, and can be computed by solving an $n \times n$ linear system. The function $\tilde{u}$ may be seen as a particular best fit estimator of u on V_n, since it is also defined by

$$\tilde{u} \coloneqq \arg\min_{v \in V_n} \|P_{W_m} v - w\|_V.$$

The function u^* can be derived from $\tilde{u}$ by the correction procedure

$$u^* \coloneqq \tilde{u} + (w - P_{W_m} \tilde{u}),$$

which shows that $u^* \in V_n + W_m$. It may be thought of as a generalized interpolation estimator, since it agrees with the observed data ($P_{W_m} u^* = P_{W_m} u$). In the case of noise-free data, it is proved in [33, 124] that these estimators satisfy the recovery bounds

$$\|u - \tilde{u}\|_V \leqslant \mu_n^m \min_{v \in V_n} \|u - v\|_V \quad \text{and} \quad \|u - u^*\|_V \leqslant \mu_n^m \min_{v \in V_n \oplus (W_m \cap V_n^\perp)} \|u - v\|_V.$$

These bounds reflect a typical trade-off in the choice of the reduced basis space, since making n larger has both effects of decreasing the approximation error $\min_{v \in V_n} \|u - v\|_V$ and increasing the stability constant $\mu_n^m = \mu(V_n, W_m)$.

When the PBDW method is applied to noisy data, amounting in observing a perturbed version $\overline{w}$ of $w = P_{W_m} u$, the recovery bounds remain valid up to the additional term $\mu_n^m \|w - \overline{w}\|_V$. In summary, one has for both estimators

$$\max\{\|u - \tilde{u}\|_V, \|u - u^*\|_V\} \leqslant C \mu_n^m (e_n(u) + \|w - \overline{w}\|_V), \tag{3.4}$$

where

$$e_n(u) \coloneqq \min_{v \in V_n} \|u - v\|_V$$

is the reduced model approximation error, $C > 0$, and $\|w - \overline{w}\|_V$ is the noise error measured in the space W_m. Note that since the additive perturbations η_i are applied to the data $\ell_i(u)$, a natural model for the measurement noise is to assume a bound of the norm $\|\eta\|_p \leqslant \varepsilon$, for the vector $\eta = (\eta_1, \dots, \eta_m)$, typically in the max norm

$p = \infty$ or euclidean norm $p = 2$. Therefore, one has $\|w - \overline{w}\|_V \leqslant \beta_m^p \varepsilon$, where

$$\beta_m^p \coloneqq \max_{v \in W_m} \frac{\|v\|_V}{\|\ell(v)\|_p},$$

resulting in a bound of the form $C\mu_n^m(e_n(u) + \beta_m^p \varepsilon)$ for both estimators.

3.1.4 Towards nonlinear approximation spaces

The simplicity of the PBDW method and its variants comes together with a fundamental limitation on its performance: it is by essence a linear reconstruction method with recovery bounds tied to the approximation error $e_n(u)$. When the only prior information is that the unknown function u belongs to a class $\mathcal{K}$, with $\mathcal{K} = \mathcal{K}_Y$ the solution manifold in the case of parametric PDEs, its best performance over $\mathcal{K}$ is thus limited by the n-width $d_n(\mathcal{K})_V$ and in turn by $d_m(\mathcal{K})_V$ since $n \leqslant m$.

In several simple yet relevant settings, it is known that n-widths have poor decay with n. One instance is when the class $\mathcal{K}$ contains piecewise smooth states, with a state-dependent location of jump discontinuities. As an elementary example, one can easily check that if $V = L^2([0,1])$ and $\mathcal{K}$ is the set of all indicator functions $u = \chi_{[a,b]}$ with $a, b \in [0,1]$, one has $d_n(\mathcal{K})_V \sim n^{-1/2}$. This decay is of course even slower for more general classes of piecewise smooth functions in higher dimension, see in particular [28, Chapter 3, equation (3.76)]. Such functions are typical in parametric hyperbolic PDEs, due to the presence of shocks with positions that differ when parameters entering in the velocity field vary. We refer to [25, 32, 69, 80, 145, 188] for other examples of parametric PDEs whose solution manifold has slow Kolmogorov n-width decay.

For such classes of functions, nonlinear approximation methods are well known to perform significantly better than their linear counterparts. Typical representatives of such methods include approximation by rational fractions, free knot splines or adaptive finite elements, best n-term approximation in a basis or dictionary, neural network or various tensor formats. In all these instances the space V_n still depends on n or $\mathcal{O}(n)$ parameters but is not anymore a linear space. We refer to [63] for a general introduction on the topic of nonlinear approximation.

3.1.5 Objective and outline

The objective of this chapter is to study the natural extensions of the PBDW method to such nonlinear approximation spaces and identify the basic structural properties that lead to near optimal recovery estimates similar to (3.4).

We begin in Section 3.2 by considering the most general setting where V is a Banach space, V_n a nonlinear approximation family, and the ℓ_i are functionals defined on V that are not necessarily linear, but Lipschitz continuous, that is

$$\|\ell(v) - \ell(\tilde{v})\|_Z \leqslant \alpha_Z \|v - \tilde{v}\|_V, \quad v, \tilde{v} \in V. \tag{3.5}$$

Here $\|\cdot\|_Z$ can be any given norm defined over $\mathbb{R}^m$ with the constant α_Z depending on this choice of norm. In this framework, we discuss the best fit estimation procedure that consists in minimizing the distance to the observed data in a given norm $\|\cdot\|_Z$.

Our main structural assumption on V_n is the following *inverse stability property*: the reduced model is stable with respect to the measurement functionals if there exists a finite constant μ_n^Z such that

$$\|v - \tilde{v}\|_V \leqslant \mu_n^Z \|\ell(v) - \ell(\tilde{v})\|_Z, \quad v, \tilde{v} \in V_n. \tag{3.6}$$

The stability constant μ_n^Z depends on the Z norm and plays a role similar to that of μ in the linear case. In particular, we show that this constant is finite only if $n \leqslant m$. The resulting estimator $\tilde{u}$ is then proved to satisfy a general recovery bound of the form

$$\|u - \tilde{u}\|_V \leqslant C_1 e_n(u) + C_2 \|\eta\|_p,$$

where $e_n(u) \coloneqq \min_{v \in V_n} \|u - v\|_V$ is the nonlinear reduced model approximation error, $\|\eta\|_p$ the level of measurement noise in ℓ^p norm, and the constants C_1 and C_2 depend on α_Z and μ_n^Z.

In Section 3.3, we consider the more particular setting where the ℓ_i are linear functionals. Then, we show that constants C_1 and C_2 are each minimized by a different choice of norm $\|\cdot\|_Z$, resulting in two different

best fit estimators $\tilde{u}$, as already observed in [29] in the case of linear reduced models. This particular setting also allows us to introduce a generalized interpolation estimator u^* and establish similar recovery estimates for $\|u - u^*\|_V$.

We next apply our framework to the inverse problem that consists in recovering a general shape $\mathcal{D}$, identified to its characteristic function $\chi_{\mathcal{D}}$, based on cell average data

$$a_T(\mathcal{D}) \coloneqq \frac{1}{|T|}\int_T \chi_{\mathcal{D}}, \quad T \in \mathcal{T},$$

where $\mathcal{T}$ is a fixed cartesian mesh. One motivation for this problem is the design of finite volume schemes for the computation of solutions to transport PDEs on such meshes.

We first discuss in Section 3.4 the best estimation rate in terms of the mesh size h that can be achieved by standard linear reconstructions, and which is essentially that of piecewise constant approximations, that is $\mathcal{O}(h^{1/q})$ regardless of the smoothness of the boundary $\partial\mathcal{D}$. This intrinsic limitation is due to the presence of the jump discontinuity that is not well resolved by the mesh.

We then discuss in Section 3.5 a local recovery strategy based on a nonlinear approximation space V_n that consists of characteristic functions of half-planes which can fit the boundary of $\mathcal{D}$ at a subcell resolution level, as already proposed in [15, 148, 149, 154]. One main result, whose proof is given in an appendix, is that this approximation space is stable in the sense of (3.6) with respect to cell average measurements on a stencil of 3×3 squares. In turn, if $\mathcal{D}$ has a C^2 boundary, the recovered shape $\widetilde{\mathcal{D}}$ is proved to satisfy an estimate of the form

$$\|\chi_{\mathcal{D}} - \chi_{\widetilde{\mathcal{D}}}\|_{L^q} \leqslant Ch^{2/q},$$

where h is the mesh size, which cannot be achieved by any linear reconstruction. This paves the way to higher order reconstruction methods for smoother boundaries by using local nonlinear approximation spaces with curved boundaries and larger stencils.

Finally, we discuss in Section 3.6 the application of our results to the recovery of large vectors of size N from $m < N$ linear measurements, up to the error of best n-term approximation. This problem is well-known in compressed sensing [42, 72], and was in particular studied in [52] which discusses the importance of the recovery norm $\|\cdot\|_V$ to understand if near-optimal recovery bounds can be achieved with m not much larger than n. We show that the structural assumptions identified in our general setting are naturally related to the so-called *null space property* introduced in [52].

3.2 Nonlinear reduction of inverse problems

3.2.1 A general framework

In full generality we are interested in recovering functions u in a general Banach space V with norm $\|\cdot\|_V$, from the measurement vector $z = (z_1, \ldots, z_m) \in \mathbb{R}^m$ given by (3.1). A recovery (or inversion) map

$$z \to R(z),$$

takes this vector to an approximation $R(z)$ of u. We are interested in controlling the recovery error $\|u - R(z)\|_V$.

To build the recovery map R, we use a nonlinear approximation space of dimension n, that is, a family of functions that can be described by n parameters. Loosely speaking, this means that there exists a set $S \subset \mathbb{R}^n$ and a continuous map $D : S \to V$ such that

$$V_n \coloneqq \{D(x) \ : \ x \in S\}.$$

Note that this definition covers the case of an n dimensional linear subspace since we can choose $S = \mathbb{R}^n$ and D a linear map.

Our main assumptions are the Lipschitz stability of the functionals ℓ_i over the whole space V and their inverse Lipschitz stability over the nonlinear approximation space V_n, expressed by (3.5) and (3.6), respectively. Note that since $\mathbb{R}^m$ is finite dimensional, the norm $\|\cdot\|_Z$ that is chosen in $\mathbb{R}^m$ to express these properties could be arbitrary up to a modification of the stability constants α_Z, μ_n^Z. These constants can be optimally defined

as

$$\alpha_Z = \sup_{v_1,v_2\in V} \frac{\|\ell(v_1)-\ell(v_2)\|_Z}{\|v_1-v_2\|_V},$$

and

$$\mu_n^Z = \sup_{v_1,v_2\in V_n} \frac{\|v_1-v_2\|_V}{\|\ell(v_1)-\ell(v_2)\|_Z}.$$

Note that one always has $\alpha_Z\mu_n^Z \geqslant 1$.

Remark 3.1. Note that when V_n is an n-dimensional space and the ℓ_i are linear functionals, the quantity μ_n^Z may be rewritten as

$$\mu_n^Z = \max_{v\in V_n} \frac{\|v\|_V}{\|\ell(v)\|_Z}.$$

As discussed further, the quantity μ_n^m defined in (3.3) for the analysis of the PBDW method is an instance of μ_n^Z corresponding to a particular choice of norm $\|\cdot\|_Z$. Assuming the ℓ_i are independent functionals, one easily checks that finiteness of this quantity imposes that $n \leqslant m$. Indeed, if $n > m$, there exists a non-trivial $v \in V_n \cap \mathcal{N}$, where

$$\mathcal{N} \coloneqq \{v \in V \ : \ \ell(v) = 0\}$$

is the null space of the measurement map that has codimension m, and therefore μ_n^Z is infinite.

Remark 3.2. The restriction $n \leqslant m$ is also needed for nonlinear spaces V_n and measurement map ℓ, under assumptions expressing that m and n are local dimensions. More precisely, assume that the map D defining V_n is differentiable at some c_0 in the interior of S, that ℓ is differentiable at $v_0 = D(c_0)$, and that both tangent maps have full rank at these points, that is,

$$\dim(dD_{c_0}(\mathbb{R}^n)) = n \quad \text{and} \quad \dim(d\ell_{v_0}(V)) = m.$$

Then, by taking $v_1 = v_0$ and $v_2 = D(c_0 + tc)$ in the quotient that defines μ_n^Z, and letting $t \to 0$ for arbitrary $c \in \mathbb{R}^n$, one finds that

$$\mu_n^Z \geqslant \max_{v\in dD_{c_0}(\mathbb{R}^n)} \frac{\|v\|_V}{\|d\ell_{v_0}(v)\|_Z},$$

and therefore it is infinite if $n > m$, by the same argument as in the previous remark.

3.2.2 The best fit estimator

We define a first recovery map $z \mapsto \tilde{u} = R(z)$ as the best fit estimator in the $\|\cdot\|_Z$ norm

$$\tilde{u} \coloneqq \operatorname*{arg\,min}_{v\in V_n} \|z-\ell(v)\|_Z. \tag{3.7}$$

The existence of such a minimizer is trivial if the space V_n and the measurement map ℓ are linear. It can also be ensured in the nonlinear case under additional assumptions, for example compactness of the set S defining the nonlinear space V_n, which will be the case in the application to shape recovery discussed in Section 3.5. If the minimizer does not exist, we may consider a near minimizer, that is $\tilde{u} \in V_n$ satisfying

$$\|z-\ell(\tilde{u})\|_Z \leqslant C\|z-\ell(v)\|_Z, \quad v \in V_n,$$

for some fixed $C > 1$. Inspection of the proofs of our main results below reveals that similar recovery bounds can be obtained for such a near minimizer, up to the multiplicative constant C.

Recall that our assumption on the noise model is a control on $\|\eta\|_p$ for some $1 \leqslant p \leqslant \infty$. For this value of p, we introduce the quantity

$$\beta_Z^p \coloneqq \max_{z\in\mathbb{R}^m} \frac{\|z\|_Z}{\|z\|_p}$$

We are now in position to state a recovery bound in this general framework.

Theorem 3.3. *The best fit estimator $\tilde{u}$ from* (3.7) *satisfies the estimate*

$$\|u - \tilde{u}\|_V \leqslant C_1 e_n(u) + C_2 \|\eta\|_p. \tag{3.8}$$

where $C_1 \coloneqq 1 + 2\alpha_Z \mu_n^Z$ and $C_2 \coloneqq 2\beta_Z^p \mu_n^Z$.

Proof. Consider any $v \in V_n$ and write

$$\|u - \tilde{u}\|_V \leqslant \|u - v\|_V + \|v - \tilde{u}\|_V \leqslant \|u - v\|_V + \mu_n^Z \|\ell(v) - \ell(\tilde{u})\|_Z,$$

where we have used (3.6). On the other hand, the minimizing property of $\tilde{u}$ ensures that

$$\|\ell(v) - \ell(\tilde{u})\|_Z \leqslant \|z - \ell(v)\|_Z + \|z - \ell(\tilde{u})\|_Z \leqslant 2\|z - \ell(v)\|_Z.$$

Furthermore, using the stability (3.5) of ℓ and the definition of β_Z^p, we have

$$\|z - \ell(v)\|_Z \leqslant \|\ell(u) - \ell(v)\|_Z + \|\eta\|_Z \leqslant \alpha_Z \|u - v\| + \beta_Z^p \|\eta\|_p.$$

Combining the three estimates, we reach

$$\|u - \tilde{u}\|_V \leqslant \big(1 + 2\alpha_Z \mu_n^Z\big)\|u - v\|_V + 2\beta_Z^p \mu_n^Z \|\eta\|_p,$$

which gives (3.8) by optimizing over $v \in V_n$. □

The constants C_1 and C_2 in the above recovery estimate depend on the choice of norm $\|\cdot\|_Z$. Note that they are invariant when this norm is scaled by a factor $t > 0$, since this has the effect of multiplying α_Z and β_Z^p by t and dividing μ_n^Z by t, which is consistant with the fact that the resulting estimator $\tilde{u}$ is left unchanged by such a scaling. In the next section we show, in the particular setting of linear measurements, that specific choices of $\|\cdot\|_Z$ can be used to minimize C_1 or C_2. This setting also allows us to introduce and study a generalized interpolation estimator, which is not relevant to the present section since the nonlinear measurement map ℓ is not assumed to be surjective: in the presence of noise, there might exist no $v \in V$ that agrees with the data, in the sense that $z = \ell(u) + \eta$ does not belong to the range of ℓ.

3.3 Linear observations

In this section, we assume that the $\ell_i \in V'$ are independent linear functionals, still allowing V_n to be a general nonlinear space. In this framework, which contains the example of shape recovery discussed in Section 3.5, one has

$$\alpha_Z = \max_{v \in V} \frac{\|\ell(v)\|_Z}{\|v\|_V}$$

and

$$\mu_n^Z = \max_{v \in V_n^{\mathrm{diff}}} \frac{\|v\|_V}{\|\ell(v)\|_Z},$$

where

$$V_n^{\mathrm{diff}} = V_n - V_n \coloneqq \{v_1 - v_2 \ : \ v_1, v_2 \in V_n\}.$$

In this particular setting, we can identify the norms $\|\cdot\|_Z$ that minimize the constants $C_1 \coloneqq 1 + 2\alpha_Z \mu_n^Z$ and $C_2 \coloneqq 2\beta_Z^p \mu_n^Z$, respectively.

3.3.1 Optimal norms

As $\ell : V \to \mathbb{R}^m$ is linear continuous and surjective, we can define a norm on $\mathbb{R}^m$ through

$$\|z\|_W = \min\{\|v\|_V \ : \ \ell(v) = z\}. \tag{3.9}$$

Remark 3.4. If V is a Hilbert space, the minimizer is unique by strict convexity of $\|\cdot\|_V$, and the m-dimensional space

$$W := \left\{ \operatorname*{arg\,min}_{\ell(v)=z} \|v\|_V \ : \ z \in \mathbb{R}^m \right\}$$

is exactly the span of the Riesz representers of the observation functionals $\ell_i \in V'$. Moreover, denoting P_W the orthogonal projection on W, we have

$$\|\ell(v)\|_W = \|P_W v\|_V, \quad v \in V.$$

For this reason, we sometimes refer to $\|\cdot\|_W$ as the *Riesz norm* even in the case of a more general Banach space.

The following result shows that the choice $\|\cdot\|_Z := \|\cdot\|_W$ is the one that minimizes the constant C_1, while C_2 is minimized by simply taking the ℓ^p norm $\|\cdot\|_Z = \|\cdot\|_p$.

Theorem 3.5. *For any norm $\|\cdot\|_Z$, one has*

$$\alpha_W \mu_n^W = \mu_n^W \leqslant \alpha_Z \mu_n^Z,$$

and

$$\beta_p^p \mu_n^p = \mu_n^p \leqslant \beta_Z^p \mu_n^Z,$$

where $(\alpha_W, \beta_W^p, \mu_n^W)$ and $(\alpha_p, \beta_p^p, \mu_n^p)$ are the triplets $(\alpha_Z, \beta_Z^p, \mu_n^Z)$ when $\|\cdot\|_Z := \|\cdot\|_W$ and $\|\cdot\|_Z = \|\cdot\|_p$, respectively.

Proof. One has

$$\alpha_W = \max_{v \in V} \frac{\|\ell(v)\|_W}{\|v\|_V} = \max_{z \in \mathbb{R}^m} \max_{\ell(v)=z} \frac{\|z\|_W}{\|v\|_V} = 1,$$

and so

$$\alpha_W \mu_n^W = \mu_n^W = \max_{v \in V_n^{\mathrm{diff}}} \frac{\|v\|_V}{\|\ell(v)\|_W} \leqslant \max_{v \in V_n^{\mathrm{diff}}} \frac{\|\ell(v)\|_Z}{\|\ell(v)\|_W} \max_{v \in V_n^{\mathrm{diff}}} \frac{\|v\|_V}{\|\ell(v)\|_Z} = \max_{v \in V_n^{\mathrm{diff}}} \frac{\|\ell(v)\|_Z}{\|\ell(v)\|_W} \mu_n^Z.$$

We now observe that from the definition of $\|\cdot\|_W$, one has

$$\max_{v \in V_n^{\mathrm{diff}}} \frac{\|\ell(v)\|_Z}{\|\ell(v)\|_W} \leqslant \max_{z \in \mathbb{R}^m} \frac{\|z\|_Z}{\|z\|_W} = \max_{z \in \mathbb{R}^m} \max_{\ell(v)=z} \frac{\|z\|_Z}{\|v\|_V} = \alpha_Z.$$

We have thus obtained the first claim $\alpha_W \mu_n^W = \mu_n^W \leqslant \alpha_Z \mu_n^Z$. For the second claim, note that we trivially have $\beta_p^p = 1$, and so

$$\beta_p^p \mu_n^p = \mu_n^p = \max_{v \in V_n^{\mathrm{diff}}} \frac{\|v\|_V}{\|\ell(v)\|_p} \leqslant \max_{v \in V_n^{\mathrm{diff}}} \frac{\|\ell(v)\|_Z}{\|\ell(v)\|_p} \max_{v \in V_n^{\mathrm{diff}}} \frac{\|v\|_V}{\|\ell(v)\|_Z} \leqslant \beta_Z^p \mu_n^Z.$$

□

Remark 3.6. In the particular case where V is a Hilbert space, V_n a linear subspace and $p = 2$, it was already observed in [29] that the reconstruction operators based on the choice $\|\cdot\|_Z = \|\cdot\|_W$ or $\|\cdot\|_Z = \|\cdot\|_2$ are the most stable with respect to the approximation error and the noise error, respectively. The above result may thus be seen as a generalization of this state of affairs to the case of nonlinear subspaces of Banach spaces, and ℓ^p noise.

3.3.2 The generalized interpolation estimator

Thanks to the surjectivity of ℓ, we may introduce the space

$$V_z := \{v \in V \ : \ \ell(v) = z\},$$

and consider the minimization problem

$$\min_{v^* \in V_z} \min_{\tilde{v} \in V_n} \|v^* - \tilde{v}\|_V.$$

If $(u^*, \tilde{u}) \in V_z \times V_n$ is a minimizing pair, the function u^* is given by

$$u^* = u^*(z) \in \operatorname*{arg\,min}_{\ell(v)=z} \operatorname{dist}(v, V_n)_V,$$

and is called the generalized interpolation estimator, since it exactly matches the data.

Remark 3.7. The best fit and generalized interpolator estimations may be thought of as the two extreme cases, $t \to \infty$ and $t \to 0$, of the penalized estimator

$$u_t \coloneqq \operatorname*{arg\,min}_{v \in V} \|z - \ell(v)\|_Z + t \operatorname{dist}(v, V_n)_V.$$

As explained earlier, the generalized interpolation operator may not be well defined in the general case where the ℓ_i are nonlinear. As opposed to the best fit, or the above penalized estimator u_t when $t > 0$, the generalized interpolation estimator does not involve the choice of a particular norm Z.

On the other hand, we see that $\tilde{u}$ is the solution to the problem

$$\min_{\tilde{v} \in V_n} \operatorname{dist}(\tilde{v}, V_z)_V.$$

Observing that

$$\operatorname{dist}(\tilde{v}, V_z)_V = \min_{\ell(v)=z} \|\tilde{v} - v\|_V = \min_{\ell(v')=\ell(\tilde{v})-z} \|v'\|_V = \|\ell(\tilde{v}) - z\|_W,$$

we thus find that $\tilde{u}$ is precisely the best fit estimator for the Riesz norm $\|\cdot\|_Z \coloneqq \|\cdot\|_W$.

In the Hilbert space setting, the generalized interpolation estimator u^* is therefore the orthogonal projection of this particular best fit estimator $\tilde{u}$ onto the affine space V_z. It may thus also be derived from $\tilde{u}$ by the correction procedure

$$u^* = \tilde{u} + w - P_W \tilde{u},$$

where $w = \arg\min_{\ell(v)=z} \|v\|_V \in W$ is the preimage by ℓ of the measurements z. In the noiseless case when $w = P_W u$, this correction can only improve the approximation since it reduces the component of $u - \tilde{u}$ in the W direction while leaving unchanged the orthogonal component, and so, in view of Theorems 3.3 and 3.5, we are ensured that

$$\|u - u^*\|_V \leqslant C_1 e_n(u),$$

where $C_1 \coloneqq 1 + 2\mu_n^W$.

More generally, in the noisy case, and without the assumption that V is a Hilbert space, there is no guarantee that u^* performs better than $\tilde{u}$, but we still obtain an error estimate on u^* that is similar in nature to that satisfied by $\tilde{u}$.

Theorem 3.8. *The generalized interpolation estimator u^* satisfies the estimate*

$$\|u - u^*\|_V \leqslant C_1 e_n(u) + C_2 \|\eta\|_p, \tag{3.10}$$

where $C_1 \coloneqq 2 + 2\mu_n^W$ and $C_2 \coloneqq (1 + 2\mu_n^W)\beta_W^p$.

Proof. Take $\delta \in \arg\min_{\ell(v)=\eta} \|v\|_V$, so that $\ell(\delta) = \eta$ and $\|\eta\|_W = \|\delta\|_V$. For v and v^* in V_n, decompose

$$\|u - u^*\|_V \leqslant \|u - v\|_V + \|v - v^*\|_V + \|v^* - u^*\|_V. \tag{3.11}$$

For the middle term, using (3.6), we write

$$\begin{aligned}
\|v - v^*\|_V &\leqslant \mu_n^W \|\ell(v - v^*)\|_W \\
&\leqslant \mu_n^W (\|\ell(v - u)\|_W + \|\ell(u - u^*)\|_W + \|\ell(u^* - v^*)\|_W) \\
&\leqslant \mu_n^W (\|v - u\|_V + \|\eta\|_W + \|u^* - v^*\|_V)
\end{aligned}$$

since $\alpha_W = 1$, so the decomposition (3.11) becomes

$$\|u - u^*\|_V \leqslant (1 + \mu_n^W)\|u - v\|_V + \mu_n^W \|\eta\|_W + (1 + \mu_n^W)\|v^* - u^*\|_V.$$

To bound the last term, we optimize over the choice of $v^* \in V_n$ and use the definition of u^* to obtain

$$\inf_{v^* \in V_n} \|v^* - u^*\|_V = \operatorname{dist}(u^*, V_n) \leqslant \operatorname{dist}(u + \delta, V_n) \leqslant \operatorname{dist}(u, V_n) + \|\delta\|_V = e_n(u) + \|\eta\|_W$$

since $\ell(u + \delta) = \ell(u) + \eta = z$. Combining the last two estimates and optimizing over $v \in V_n$ gives

$$\|u - u^*\|_V \leqslant (2 + 2\mu_n^W)e_n(u) + (1 + 2\mu_n^W)\|\eta\|_W,$$

and the result follows from the definition of β_W^p. □

3.4 Shape recovery from cell averages

3.4.1 The shape recovery problem

The problem of reconstructing a function u from its cell averages

$$a_T(u) \coloneqq \frac{1}{|T|} \int_T u, \quad T \in \mathcal{T},$$

where $\mathcal{T}$ is a partition of the domain $\Omega \subset \mathbb{R}^d$ in which u is defined, appears naturally in two areas:

— In $2d$ or $3d$ image processing, it corresponds to the so-called super-resolution problem, that is, reconstructing a high resolution image from its low resolution version defined on the coarse grid $\mathcal{T}$ of pixels or voxels.

— In numerical simulation of hyperbolic conservation laws, it plays a central role when developing finite volume schemes on the computation mesh $\mathcal{T}$.

Standard reconstruction methods are challenged when the function u exhibits jump discontinuities which are not well resolved by the partition $\mathcal{T}$. Such discontinuities correspond to edges in image processing or shocks in conservation laws. Here we may focus on the very simple case of characteristic functions of sets

$$u = \chi_{\mathcal{D}},$$

that already carry the main difficulty. Therefore we are facing a problem of reconstructing a shape $\mathcal{D}$ from local averages of $\chi_{\mathcal{D}}$.

As a simple example we work in the domain $\Omega = [0, 1]^2$ with a uniform grid based on square cells of sidelength $h = \frac{1}{L}$ for some $L \in \mathbb{N}$, therefore of the form

$$\mathcal{T} = \mathcal{T}_h \coloneqq \{T_{i,j} = [(i-1)h, ih] \times [(j-1)h, jh] \ : \ i, j = 1, \ldots, L\}.$$

The cardinality of the grid is therefore

$$n \coloneqq \#(\mathcal{T}) = L^2 = h^{-2}.$$

We consider classes of characteristic functions $\chi_{\mathcal{D}}$ of sets $\mathcal{D} \subset \Omega$ with boundary of a prescribed Hölder smoothness. The definition of these classes requires some precision.

Definition 3.9. For $s \in \mathbb{N}_0$, $s' \in [0, 1]$, $0 < R < 1/2$ and $M > 0$, we define the class $\mathcal{F}_{s,s'}^{R,M}$ as consisting of all characteristic functions $\chi_{\mathcal{D}}$ of domains $\mathcal{D} \subset [R, 1 - R]^2 \subset \Omega$ with the following property: for all $\bar{x} \in \Omega$, there exists an orthonormal system (e_1, e_2) and a function $\psi \in \mathcal{C}^{s,s'}$ with $\|\psi\|_{\mathcal{C}^{s,s'}} \leqslant M$, such that

$$x \in \mathcal{D} \iff \tilde{x}_2 \leqslant \psi(\tilde{x}_1),$$

for any $x = \bar{x} + \tilde{x}_1 e_1 + \tilde{x}_2 e_2$ with $|\tilde{x}_1|, |\tilde{x}_2| \leqslant R$.

Here, we have used the common definition

$$\|\psi\|_{\mathcal{C}^{s,s'}} = \sup_{0\leqslant k\leqslant s} \|\psi^{(k)}\|_{L^\infty([-R,R])} + \sup_{r_1,r_2\in[-R,R]} \frac{|\psi^{(s)}(r_1)-\psi^{(s)}(r_2)|}{|r_1-r_2|^{s'}},$$

for the Hölder norm. In the case $s'=1$, note that $\mathcal{C}^{s,s'}$ denotes functions with Lipschitz derivatives up to order s, so that in particular the case $s=0, s'=1$ corresponds to domains with Lipschitz boundaries.

Remark 3.10. The condition $\mathcal{D}\subset[R,1-R]^2$ imposing that $\mathcal{D}$ remains away from the boundary $\partial\Omega$ might be quite restrictive in some applications; instead, one can assume that the domains $\mathcal{D}$ and Ω are periodic, or symmetrize $\mathcal{D}$ with respect to $\partial\Omega$.

3.4.2 The failure of linear reconstruction methods

The most trivial linear reconstruction method consists in the piecewise constant approximation

$$\tilde{u} = \sum_{T\in\mathcal{T}} a_T(u)\chi_T. \tag{3.12}$$

The approximation rate of this reconstruction over the class $\mathcal{F}^{R,M}_{s,s'}$ is as follows.

Proposition 3.11. *Let $u=\chi_{\mathcal{D}}\in\mathcal{F}^{R,M}_{s,s'}$, its piecewise constant approximation $\tilde{u}$ by average values on each cell, defined in* (3.12)*, satisfies*

$$\|\chi_{\mathcal{D}}-\tilde{u}\|_{L^q} \leqslant Ch^{\frac{1}{q}} = Cn^{-\frac{1}{2q}},$$

where the constant C depends on R and M.

Proof. Let $N=\lceil(\sqrt{2}R)^{-1}\rceil$, and partition the domain $\Omega=[0,1]^2$ into N^2 squares of side $1/N$. Then each subsquare Q is contained in the set $\{\overline{x}+\tilde{x}_1e_1+\tilde{x}_2e_2 \,:\, |\tilde{x}_1|,|\tilde{x}_2|\leqslant R\}$ from Definition 3.9, where $\overline{x}$ is the center of Q. Thus $\partial\mathcal{D}$ is the restriction of the graph of an M-Lipschitz function on Q, so its arc length is bounded by

$$|\partial\mathcal{D}\cap Q| \leqslant \operatorname{diam}(Q)\sqrt{1+M^2} \leqslant 2R\sqrt{1+M^2}.$$

As any curve of arclength h intersects at most four cells from $\mathcal{T}$, $\partial\mathcal{D}\cap Q$ intersects at most $4\lceil 2R\sqrt{1+M^2}/h\rceil$ cells, and summing over all subsquares, $\partial\mathcal{D}$ intersects at most $4N^2\lceil 2R\sqrt{1+M^2}/h\rceil$ cells. Denoting $\mathcal{T}_{\partial\mathcal{D}}$ the set of these cells, and observing that $u|_T\equiv a_T(u)\in\{0,1\}$ for $T\notin\mathcal{T}_{\partial\mathcal{D}}$, we get

$$\|\chi_{\mathcal{D}}-\tilde{u}\|^q_{L^q} = \sum_{T\in\mathcal{T}}\int_T |u-a_T(u)|^q \leqslant \sum_{T\in\mathcal{T}_{\partial\mathcal{D}}} |T| = h^2|\mathcal{T}_{\partial\mathcal{D}}| \leqslant 24\frac{\sqrt{1+M^2}}{R}h$$

for $h\leqslant R$, and this bound also holds for $h>R$ since $\|\chi_{\mathcal{D}}-\tilde{u}\|^q_{L^q}\leqslant 1$. □

The next result shows, for the particular case $q=2$, that no better rate can actually be achieved by any linear method, regardless of the smoothness s of the boundary. We conjecture that a similar result holds for $1\leqslant q\leqslant\infty$. This motivates the use of nonlinear recovery methods, which are the object of the next section.

We recall that the Kolmogorov n-width of a compact set $\mathcal{K}$ from some Banach space V is defined by

$$d_n(\mathcal{K})_V := \inf_{\dim(E)\leqslant n}\operatorname{dist}(\mathcal{K},E)_V,$$

where $\operatorname{dist}(\mathcal{K},E)_V := \max_{u\in\mathcal{K}}\min_{v\in E}\|u-v\|_V$ and the infimum is taken over all finite dimensional spaces E of dimension at most n.

Proposition 3.12. *Let $s\in\mathbb{N}_0$ and $s'\in[0,1]$ be arbitrary. Then for R sufficiently small, and M sufficiently large, there exists $c>0$ such that the Kolmogorov n-widths of the class $\mathcal{F}^{R,M}_{s,s'}$ satisfy*

$$d_n(\mathcal{F}^{R,M}_{s,s'})_{L^2} \geqslant cn^{-\frac{1}{4}}, \quad n\geqslant 1.$$

Proof. The proof of this result relies on similar lower bounds for dictionaries of d-dimensional ridge functions

$$\mathcal{R}_k^d := \{x \mapsto \sigma_k(\omega \cdot x + b) \ : \ \|\omega\|_2 = 1 \ : \ b_1 \leqslant b \leqslant b_2\}$$

where $\sigma_k(t) := \max(0,t)^k$ is the so-called RELU-k function. Here, we work in the space $L^2(B)$ where B is an arbitrary ball of $\mathbb{R}^d$, and the constants (b_1, b_2) are taken as the inf and sup of $\omega \cdot x$ as $x \in B$ and $\|\omega\|_2 = 1$, respectively, that is we take all b such that the line discontinuity of the k-th derivative of $\sigma_k(\omega \cdot x + b)$ crosses the ball B. Theorem 9 from [162], which improves on earlier results from [126], shows that if

$$B_1(\mathcal{R}_k^d) := \overline{\Big\{ \sum_{j=1}^n a_j g_j \ : \ n \in \mathbb{N} \ : \ g_j \in \mathcal{R}_k^d \ : \ \sum_{j=1}^n |a_j| \leqslant 1 \Big\}}$$

denotes the symmetrized convex hull of this dictionary (the closure being taken in $L^2(B)$), then

$$d_n(B_1(\mathcal{R}_k^d))_{L^2(B)} \geqslant c n^{-\frac{2k+1}{2d}}, \quad n \geqslant 1,$$

where c depends on k, d, and the diameter of B.

In our case of interest we work with the value $d = 2$ and $k = 0$, so that the ridge functions are simply the characteristic functions of half-planes. By convexity, we have

$$d_n(\mathcal{R}_0^2))_{L^2(B)} = d_n(B_1(\mathcal{R}_0^2))_{L^2(B)} \geqslant c n^{-\frac{1}{4}}.$$

We take for B the ball of center $(1/2, 1/2)$ and radius $1/4$, which is inside our domain $\Omega = [0,1]^2$. It is then readily seen that for R small enough and M large enough, we can extend any ridge function $g \in \mathcal{R}_0^2$ into a characteristic function $\chi_{\mathcal{D}}$ from $\mathcal{F}_{s,s'}^{R,M}$, as illustrated in Figure 3.1.

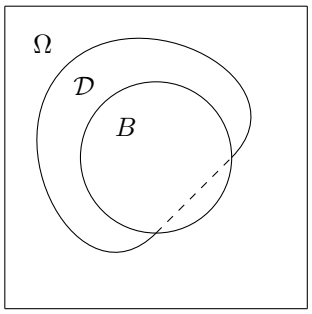


Figure 3.1 – Example of extension of the indicator of a half-plane on B to the indicator of a smooth domain $\mathcal{D}$ on Ω

Observing that if E_Ω is a linear subspace of $L^2(\Omega)$ of dimension at most n, its restriction E_B to B is a linear subspace of $L^2(B)$ of dimension at most n, and one has

$$\operatorname{dist}(\chi_{\mathcal{D}}, E_B)_{L^2(B)} \leqslant \operatorname{dist}(\chi_{\mathcal{D}}, E_\Omega)_{L^2(\Omega)}.$$

By infimizing, it follows that

$$d_n(\mathcal{F}_{s,s'}^{R,M})_{L^2(\Omega)} \geqslant d_n(\mathcal{R}_0^2)_{L^2(B)} \geqslant c n^{-\frac{1}{4}},$$

which concludes the proof. □

Remark 3.13. The fact that we impose conditions on R and M in the above statement is natural since the class $\mathcal{F}_{s,s'}^{R,M}$ becomes empty if R is not small enough or M not large enough, due to the fact that the sets $\mathcal{D}$ are assumed to be contained in the interior of Ω.

Remark 3.14. The above results are easily extended to higher dimension $d \geqslant 2$, with a similar definition for the class $\mathcal{F}_{s,s'}^{R,M}$. The rate of approximation in L^q norm by piecewise constant functions on uniform partitions

is then $n^{-\frac{1}{dq}}$, which in the case $q = 2$ is proved by a similar argument to be the best achievable by any linear reconstruction method. We conjecture that the same holds for more general $1 \leqslant q \leqslant \infty$.

3.5 Shape recovery by nonlinear least-squares

3.5.1 Nonlinear reconstruction on a stencil

We now discuss a nonlinear reconstruction method for $u \in \mathcal{F}^{R,M}_{s,s'}$, whose output $\tilde{u}$ is the indicator of a domain $\widetilde{\mathcal{D}}$ with polygonal boundary : on each cell T, the domain $\widetilde{\mathcal{D}}$ coincides with a certain half plane. In order to define the delimiting line we only use the average values of u on a 3×3 stencil of cells centered at T.

We assume that $h < R$, so that $\mathcal{D}$ does not intersect the boundary cells $T_{i,j}$ with i or j in $\{1, L\}$, and fix indices $1 < i, j < L$. For the cell $T = T_{i,j}$, denote $\overline{x} = ((i - \frac{1}{2})h, (j - \frac{1}{2})h)$ its center, and

$$S = [(i-2)h, (i+1)h] \times [(j-2)h, (j+1)h] = \bigcup_{i-1 \leqslant i' \leqslant i+1,\, j-1 \leqslant j' \leqslant j+1} T_{i'j'}$$

the stencil composed of T and its 8 neighboring cells. We define the nonlinear approximation space

$$V_2 := \left\{ \chi_{\vec{n}\cdot(x-\overline{x}) \geqslant c} \ : \ \vec{n} \in \mathbb{S}^1,\ c \in \mathbb{R} \right\}, \tag{3.13}$$

which is a two-parameter family as each function is determined by $(\arg \vec{n}, c) \in (-\pi, \pi] \times \mathbb{R}$, where $\arg \vec{n}$ is the angle of $\vec{n}$ with respect to the horizontal axis.

Here, our measurements are the average values of u on the cells contained in S

$$\ell(u) = (a_{T'}(u))_{T' \subset S} \in \mathbb{R}^9.$$

In order to find a reconstruction of u in V_2 based on these measurements, we need an inverse stability property of the form (3.6). This is not possible here, since ℓ cancels on all functions $\chi_{\mathcal{D}} \in V_2$ with $\mathcal{D} \cap S = \emptyset$. We therefore restrict the nonlinear family V_2, and consider only indicators of half-planes whose boundary passes through the central cell T:

$$V_{2,T} := \left\{ \chi_{\mathcal{D}} \in V_2 \ : \ \partial\mathcal{D} \cap T \neq \emptyset \right\} = \left\{ \chi_{\vec{n}\cdot(x-\overline{x}) \geqslant c} \ : \ \vec{n} \in \mathbb{S}^1,\ |c| \leqslant \frac{h}{2}|\vec{n}|_1 \right\}. \tag{3.14}$$

In this setting, we prove the existence of the following stability constants for $V = L^1(S)$ and $\|z\|_W = h^2\|z\|_1$, which is the best norm on $\mathbb{R}^m$ in view of Theorem 3.5.

Proposition 3.15. *One has*

$$\|\ell(u)\|_W \leqslant \alpha_W \|u\|_{L^1(S)}, \quad u \in L^1(\Omega), \tag{3.15}$$

and

$$\|u - v\|_{L^1(S)} \leqslant \mu_n^W \|\ell(u - v)\|_W, \quad u, v \in V_{2,T}, \tag{3.16}$$

where $\alpha_W = 1$ and $\mu_n^W = \frac{3}{2}$ are the optimal constants.

The proof of the stability property (3.15) is trivial since

$$\|\ell(u)\|_W = h^2 \sum_{T \subset S} \left| \fint_T u \right| = \sum_{T \subset S} \left| \int_T u \right| \leqslant \|u\|_{L^1(S)},$$

with equality in case u has constant sign in S. The proof of the inverse stability (3.16) is quite technical and left as an appendix at the end of the chapter.

Given the noisy observations

$$z = \ell(u) + \eta \in \mathbb{R}^9,$$

on the stencil S, we define the estimator of u on the cell T by

$$\tilde{u}_T \in \operatorname*{arg\,min}_{v \in V_2} \|z - \ell(v)\|_W. \tag{3.17}$$

Here we minimize over all V_2, that is on all indicators of half planes, but we note that we may restrict to half-planes whose boundary passes through the stencil S.

The following result, which uses Proposition 3.15, shows that the distance from $\tilde{u}_T$ to u in $L^1(T)$ is comparable to the error between u and its best approximation in the $L^1(S)$ norm

$$\overline{u}_S := \operatorname*{arg\,min}_{v \in V_2} \|u - v\|_{L^1(S)}.$$

Lemma 3.16. *For all $u \in \mathcal{F}^{R,M}_{s,s'}$, one has*

$$\|u - \tilde{u}_T\|_{L^1(T)} \leqslant C_1 \|u - \overline{u}_S\|_{L^1(S)} + 2\beta_W^p \mu_n^W \|\eta\|_p,$$

where $C_1 = 1 + 2\alpha_W \mu_n^W = 4$ and $C_2 = 2\beta_W^p \mu_n^W = 3^{3-2/p} h^2$, with α_W, μ_n^W as in Proposition 3.15, and $\beta_W^p = 9^{1-1/p} h^2$ the maximal ratio between $\|\cdot\|_p$ and $\|\cdot\|_W$ norms in $\mathbb{R}^9$.

Proof. We distinguish two cases:

— If $\tilde{u}_T \in V_{2,T}$ and $\overline{u}_S \in V_{2,T}$, that is, both boundaries pass through the central cell T, we apply Theorem 3.8 together with Proposition 3.15

$$\begin{aligned} \|u - \tilde{u}_T\|_{L^1(T)} \leqslant \|u - \tilde{u}_T\|_{L^1(S)} &\leqslant C_1 \min_{v \in V_{2,T}} \|u - v\|_{L^1(S)} + C_2 \|\eta\|_p \\ &= C_1 \|u - \overline{u}_S\|_{L^1(S)} + C_2 \|\eta\|_p. \end{aligned}$$

with $C_1 = 1 + 2\alpha_W \mu_n^W$, $C_2 = 2\beta_W^p \mu_n^W$.

— Otherwise, either $\tilde{u}_T$ or $\overline{u}_S$ has constant value 0 or 1 on T, so $\tilde{u}_T - \overline{u}_S$ has constant sign on T, and thus

$$\begin{aligned} \|\overline{u}_S - \tilde{u}_T\|_{L^1(T)} &= h^2 |a_T(\tilde{u}_T - \overline{u}_S)| \\ &\leqslant \|\ell(\tilde{u}_T - \overline{u}_S)\|_W \\ &\leqslant \|\ell(\overline{u}_S) - z\|_W + \|\ell(\tilde{u}_T) - z\|_W \\ &\leqslant 2\|\ell(\overline{u}_S) - z\|_W \\ &\leqslant 2\|\ell(\overline{u}_S - u)\|_W + 2\|\eta\|_W \\ &\leqslant 2\|u - \overline{u}_S\|_{L^1(S)} + 2\beta_W^p \|\eta\|_p. \end{aligned}$$

By triangle inequality, it follows that

$$\|u - \tilde{u}_T\|_{L^1(T)} \leqslant 3\|u - \overline{u}_S\|_{L^1(S)} + 2\beta_W^p \|\eta\|_p,$$

which has better constants than in the estimate obtained in the first case, since the constant μ_n^W is larger than 1 and $\alpha_W = 1$. □

The order of the best local approximation error $\|u - \overline{u}_S\|_{L^1(S)}$ that appears as a bound for the reconstruction error $\|u - \tilde{u}_T\|_{L^1(T)}$ depends on the smoothness of the boundary, as expressed in the following lemma.

Lemma 3.17. *For all $u \in \mathcal{F}^{R,M}_{s,s'}$, with $R \geqslant \frac{3}{\sqrt{2}} h$, one has*

$$\|u - \overline{u}_S\|_{L^1(S)} \leqslant M(3\sqrt{2}h)^{\min(s+s',2)+1}.$$

Proof. We apply the definition of $\mathcal{F}^{R,M}_{s,s'}$ at point $\overline{x}$: as $R \geqslant \frac{3}{\sqrt{2}} h$, the stencil S is contained in the domain

$$\{\overline{x} + \tilde{x}_1 e_1 + \tilde{x}_2 e_2 \ : \ |\tilde{x}_1|, |\tilde{x}_2| \leqslant R\},$$

so $u|_S$ is the indicator of a domain delimited by a $\mathcal{C}^{s,s'}$ function ψ, with $\|\psi\|_{\mathcal{C}^{s,s'}} \leqslant M$. From the definition of $\mathcal{C}^{s,s'}$, there exists an affine function ξ such that

$$|\psi(\tilde{x}_1) - \xi(\tilde{x}_1)| \leqslant M(3\sqrt{2}h)^{\min(s+s',2)}, \quad |\tilde{x}_1| \leqslant \frac{3}{\sqrt{2}}h.$$

Then the function $v : \overline{x} + \tilde{x}_1 e_1 + \tilde{x}_2 e_2 \mapsto \chi_{\tilde{x}_2 \leqslant \xi(\tilde{x}_1)}$ belongs to V_2, and we have

$$\|u - \overline{u}_S\|_{L^1(S)} \leqslant \|u - v\|_{L^1(S)} \leqslant M(3\sqrt{2}h)^{\min(s+s',2)+1}.$$

□

3.5.2 Global nonlinear reconstruction

We now consider the process of recovering $u \in \mathcal{F}^{R,M}_{s,s'}$ globally from its data

$$z = \ell(u) + \eta,$$

where now $\ell(u) := (a_T(u))_{T\in\mathcal{T}} \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^n$ is the noise vector. Applying to each inner cell $T \in \mathcal{T}$ the previous reconstruction procedure based on the 3×3 stencil S centered at T, we obtain a global recovery $\tilde{u} = \tilde{u}(z)$ such that

$$\tilde{u}|_T = \tilde{u}_T|_T, \quad T = T_{i,j} \in \mathcal{T}, \quad 1 < i, j < L,$$

where $\tilde{u}_T$ is the local estimator from (3.17). On the boundary cells $T = T_{i,j}$ with i or j in $\{1, L\}$, $u|_T$ is zero by Definition 3.9 so we simply set $\tilde{u}|_T = 0$. Note that $\tilde{u}$ is of the form

$$\tilde{u} = \chi_{\widetilde{\mathcal{D}}},$$

where $\widetilde{\mathcal{D}}$ has piecewise linear boundary with respect to the mesh $\mathcal{T}$. The following result gives a global approximation bound, which confirms the improvement over linear methods when $s + s' > 1$.

Theorem 3.18. *For all $u \in \mathcal{F}^{R,M}_{s,s'}$, one has*

$$\|u - \tilde{u}\|_{L^q(\Omega)} \leqslant C_1 n^{-\frac{\min(1,(s+s')/2)}{q}} + C_2 n^{-\frac{1}{pq}} \|\eta\|_p^{\frac{1}{q}}.$$

Proof. First notice that if the result is proved for $p = q = 1$, as $u - v$ has values in $\{-1, 0, 1\}$,

$$\|u - v\|^q_{L^q(\Omega)} = \|u - v\|_{L^1(\Omega)} \leqslant C_1 n^{-s''} + C_2 n^{-1}\|\eta\|_1 \leqslant \left(C_1^{\frac{1}{q}} n^{-\frac{s''}{q}} + C_2^{\frac{1}{q}} n^{-\frac{1}{pq}} \|\eta\|_p^{\frac{1}{q}} \right)^q,$$

where $s'' = \min(1, (s + s')/2)$, so it suffices treat the case $p = q = 1$.

By an argument similar to the proof of Proposition 3.11, $\partial\mathcal{D}$ intersects at most $16N^2\lceil 2R\sqrt{1+M^2}/h\rceil$ stencils of 9 cells. Using the fact that $u = \overline{u}_S$ is a constant on any other stencil, we get

$$\begin{aligned}\|u - \tilde{u}\|_{L^1(\Omega)} &= \sum_{T \text{ inner cell}} \|u - \tilde{u}\|_{L^1(T)} \leqslant \sum_{T \text{ inner cell}} (1 + 2\alpha_W \mu_n^W)\|u - \overline{u}\|_{L^1(S)} + 2\beta_W^p \mu_n^W \|\eta\|_{\ell^1(S)} \\ &\leqslant 16N^2 \left\lceil \frac{2R\sqrt{1+M^2}}{h} \right\rceil M(3\sqrt{2}h)^{2s''+1} + 18\beta_W^p \mu_n^W \|\eta\|_1 \leqslant C_1 h^{2s''} + C_2 h^2 \|\eta\|_1.\end{aligned}$$

We conclude by recalling that $n = h^{-2}$. □

Remark 3.19. Here the convergence rate for the noiseless term $n^{-\min(1/q,(s+s')/2q)}$ is limited due to the use of polygonal domains in the reconstruction. So the best approximation rate $h^{2/q} = n^{-1/q}$ is already attained for $\mathcal{C}^2$ boundaries. When the smoothness parameter $s + s'$ is larger than 2, better rates $n^{-\frac{s+s'}{2q}}$ should be reachable

if we use non-linear approximation spaces that are richer than the space V_2, for example indicator functions of domains with boundary that have a higher order polynomial description rather than straight lines. Of course, the stable identification of these approximants in the sense of (3.6) might require stencils that are of larger size than 3×3.

Remark 3.20. If $\|\eta\|_\infty \leqslant \frac{1}{9}$, then $\tilde{u}$ is exactly equal to u on any cell whose corresponding stencil does not intersect $\partial\mathcal{D}$, so the error is concentrated on $\mathcal{O}(\sqrt{n})$ cells, leading to an improved rate $n^{-\frac{p+1}{2pq}}$ instead of $n^{-\frac{1}{pq}}$ for the noise term.

3.5.3 Numerical illustration

We study the behavior of the above discussed linear and non-linear recovery methods from cell averages for the particular target function $u = \chi_{\mathcal{D}}$, with $\mathcal{D}$ a slightly decentered disk of radius $r = 0.325$.

The linear method consists of the piecewise constant approximation (3.12), referred to as *PiecewiseConstant*. As to the nonlinear method, for the local best fit problem, we use the ℓ^2 norm on $\mathbb{R}^9$ instead of the ℓ^1 norm. By norm equivalence on $\mathbb{R}^9$, the same convergence results can be proved to hold with different constants. This method, which we refer to as *LinearInterface*, does not ensure consistency of the reconstruction in the sense that $a_T(\tilde{u}) = a_T(u)$. One way to approach this consistency property is to modify the ℓ^2 norm by putting a large weight on the central cell. We refer to this variant as *LinearInterfaceCC*, here taking the weight 100.

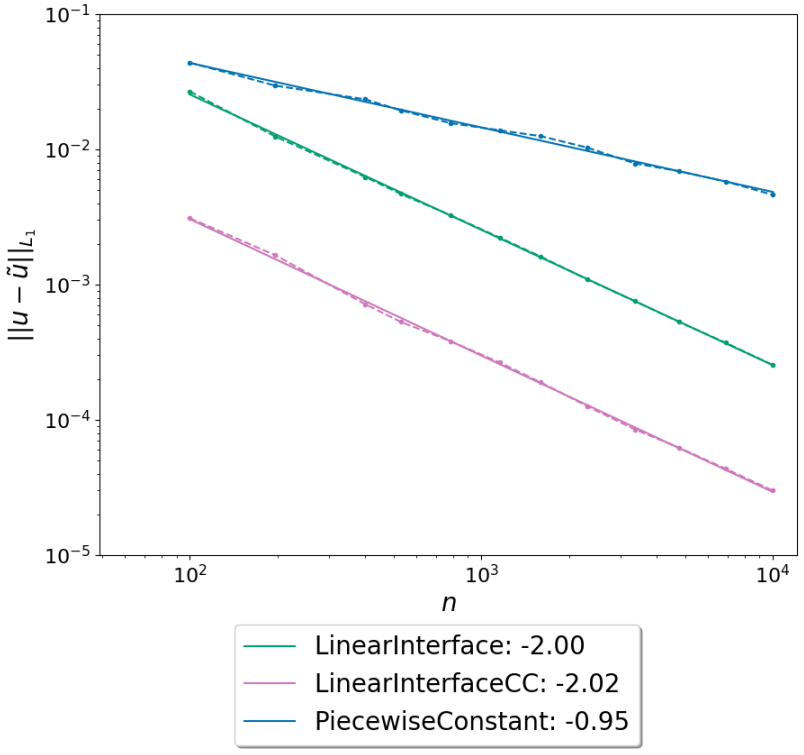


Figure 3.2 – Convergence curves and rates for the linear and nonlinear recovery methods

Figure 3.2 shows the convergence rates of the three methods in the L^1 norm. The expected h^2 decay is observed in both non-linear methods while the linear method lays behind with a decay rate of h. It is relevant to note that although both non-linear methods benefit from the same rate, the associated constants differ by an order of magnitude, showing the practical improvement gained by imposing consistency. This improvement is also visible on Figure 3.3 which shows that in the LinearInterface method, the interfaces that minimize the ℓ_2 error on the 9 surrounding cells lay always inside the circle as the curvature of the boundary pushes them

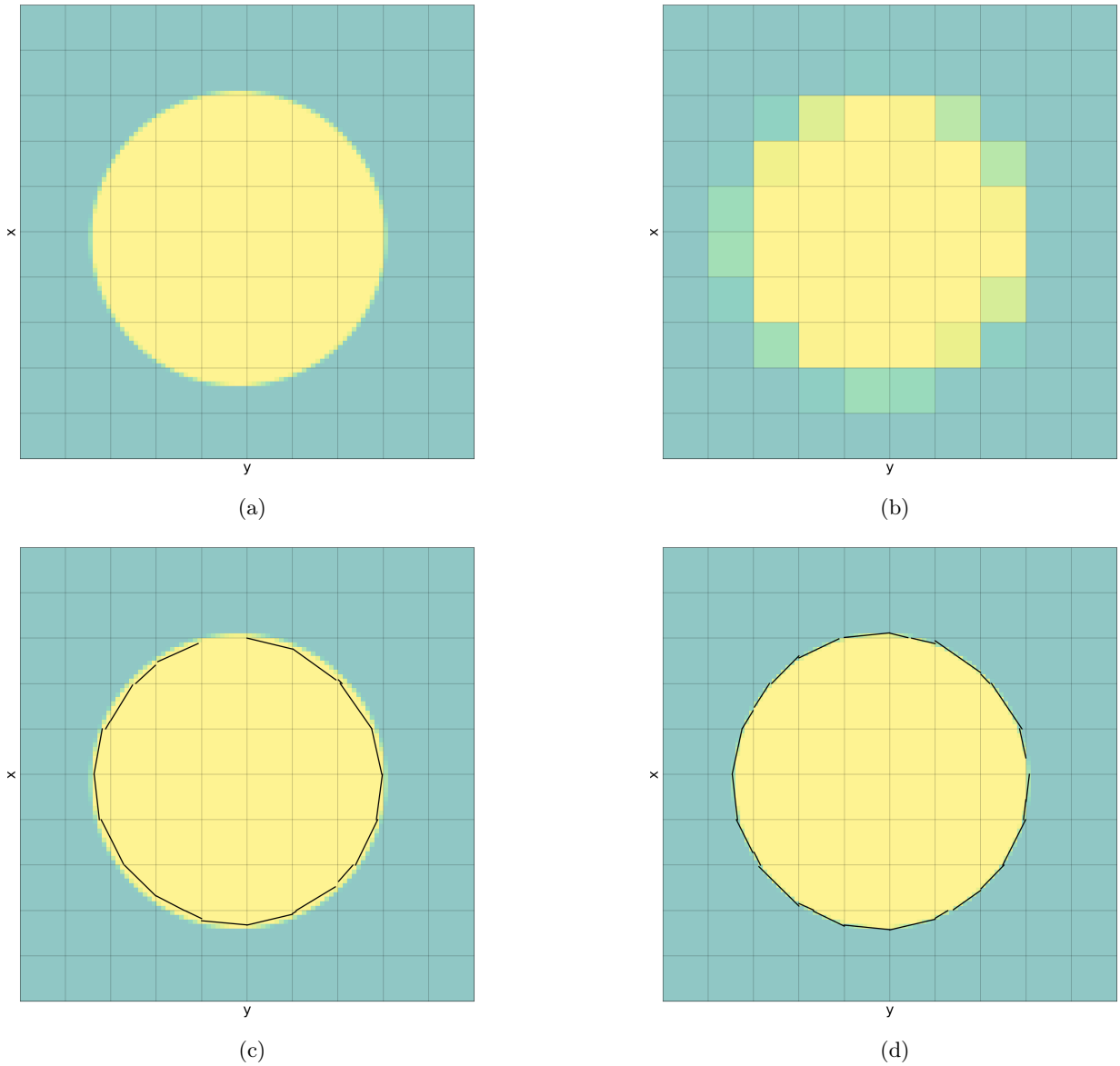


Figure 3.3 – (a) The target function, (b) its recovery by PiecewiseConstant showing the cell-average data, and the recovered boundaries by (c) LinearInterface and (d) LinearInterfaceCC methods.

inwards. On the contrary, LinearInterfaceCC seems to find the right compromise between sticking to the cell average while capturing at the same time the curvature trend hinted by the surrounding cell averages.

3.6 Relation to compressed sensing

3.6.1 Compressed sensing and best n-term approximation

In this section we discuss the application of our setting to the sparse recovery of large vectors from a few linear observations. We thus take

$$V = \mathbb{R}^N,$$

equipped with some given norm $\|\cdot\|_V$ of interest. The linear measurements of $u = (u_1, \ldots, u_N)^\top \in \mathbb{R}^N$ are given by

$$(\ell_1(u), \ldots, \ell_m(u))^\top = \Phi u,$$

where Φ is an $m \times N$ measurement matrix, with typically $m \ll N$.

The topic of compressed sensing deals with sparse recovery of u from such measurements, that is, searching to recover an accurate approximation to u by a vector with only a few non-zero components. We refer to [42, 43] for some first highly celebrated breakthrough results and to [72] for a general treatment.

Remark 3.21. In the case of function recovery, sparse polynomial approximations have been investigated in [155, 156]. It recently turned out [100] that even in the context of approximation in L^2, nonlinear methods based on ℓ^1 minimization can achieve better rates of convergence than least-squares or any linear method, for certain classes of Sobolev functions. We also refer to [62] for sparse recovery results in the same vein.

We define the nonlinear space of n-sparse vectors as

$$V_n := \Big\{ u \in \mathbb{R}^N \ : \ \|u\|_0 := \#\{i \ : \ u_i \neq 0\} \leqslant n \Big\},$$

and the best n-term approximation error in the V norm as

$$e_n(u)_V := \min_{v \in V_n} \|u - v\|_V.$$

One natural question is to understand for which type of measurement matrices Φ does the noise-free measurement $z = \Phi u$ contain enough information, in order to recover any u up to an error $e_n(u)_V$. In other words, one asks if there exists a recovery map $R : \mathbb{R}^m \to \mathbb{R}^N$ such that one has the *instance optimality property* at order n

$$\|u - R(\Phi u)\|_V \leqslant C_0 e_n(u)_V, \quad u \in \mathbb{R}^N, \tag{3.18}$$

with C_0 a fixed constant, which we denote by $IOP(n, C_0)$. This question has been answered in [52] in terms of the null space $\mathcal{N} := \{v \in \mathbb{R}^N \ : \ \Phi v = 0\}$. We say that Φ satisfies the *null space property* at order k with constant C_1, denoted by $NSP(k, C_1)$ if and only if

$$\|v\|_V \leqslant C_1 e_k(v)_V, \quad v \in \mathcal{N}. \tag{3.19}$$

This property quantifies how much vectors from the null space can be concentrated on a few coordinates. One main result of [52] is the equivalence between IOP at order n and NSP at order $2n$ in the following sense.

Theorem 3.22. *One has $IOP(n, C_0) \Rightarrow NSP(2n, C_0)$ and conversely $NSP(2n, C_1) \Rightarrow IOP(n, 2C_1)$.*

One natural question is whether matrices Φ with such properties can be constructed with a number of rows/measurements m barely larger than n. As we recall further the answer to this question is strongly tied to the norm V used on $\mathbb{R}^N$.

3.6.2 Stability and the null space property

The nonlinear estimation results that we have obtained in Section 3.2 and Section 3.3 can be applied to the setting of sparse recovery, offering us a different vehicle than the null space property to establish instance

optimality.

In the present setting, for a given norm $\|\cdot\|_Z$, the stability property (3.5) takes the form

$$\|\Phi u\|_Z \leqslant \alpha_Z \|u\|_V, \quad u \in \mathbb{R}^N \tag{3.20}$$

and the inverse stability property (3.6) takes the form

$$\|v\|_V \leqslant \mu_n^Z \|\Phi v\|_Z, \quad v \in V_{2n}, \tag{3.21}$$

since for sparse vectors we have $V_n^{\mathrm{diff}} = V_n - V_n = V_{2n}$. We refer to these properties as $S(\alpha_Z)$ and $IS(2n, \mu_n^Z)$, respectively.

Application of Theorem 3.3 in the noiseless case immediately gives us that the nonlinear best fit recovery $R(\Phi u) = \tilde{u}$ satisfies the instance optimality bound (3.18) with constant $C_0 = 1 + 2\alpha_Z \mu_n^Z$. In other words

$$S(\alpha_Z) \text{ and } IS(2n, \mu_n^Z) \Rightarrow IOP(n, C_0), \qquad C_0 = 1 + 2\alpha_Z \mu_n^Z. \tag{3.22}$$

The following result shows that (S, IS) is actually equivalent to NSP, and thus to IOS, in the sense that a converse result holds when $\|\cdot\|_Z$ is chosen to be the Riesz norm (3.9).

Theorem 3.23. *For any norm $\|\cdot\|_Z$, one has*

$$S(\alpha_Z) \text{ and } IS(2n, \mu_n^Z) \Rightarrow NSP(2n, C_1), \qquad C_1 = 1 + \alpha_Z \mu_n^Z. \tag{3.23}$$

Conversely, let $\|\cdot\|_W$ be the Riesz norm so that $\|\Phi u\|_W = \min_{\Phi v = \Phi u} \|v\|_V$, then

$$NSP(2n, C_1) \Rightarrow S(\alpha_W) \text{ and } IS(2n, \mu_n^W), \qquad \alpha_W = 1 \text{ and } \mu_n^W = 1 + C_1. \tag{3.24}$$

Proof. Assume that $S(\alpha_Z)$ and $IS(2n, \mu_n^Z)$ hold. Let $v \in \mathcal{N}$ and $\tilde{v}$ its best approximation in V_{2n}, then

$$\begin{aligned} \|v\|_V &\leqslant \|v - \tilde{v}\|_V + \|\tilde{v}\|_V \\ &\leqslant e_{2n}(v)_V + \mu_n^Z \|\Phi \tilde{v}\|_W \\ &= e_{2n}(v)_V + \mu_n^Z \|\Phi(v - \tilde{v})\|_W \leqslant (1 + \alpha_Z \mu_n^Z) e_{2n}(x)_V. \end{aligned}$$

This shows that $NSP(2n, C_1)$ holds with $C_1 = 1 + \alpha_Z \mu_n^Z$.

Conversely, assume that $NSP(2n, C_1)$ holds. From the definition of the Riesz norm, it is immediate that $S(\alpha_W)$ holds with $\alpha_W = 1$. For $v \in V_{2n}$, let $\tilde{v}$ be the minimizer of $\min_{\Phi \tilde{v} = \Phi v} \|\tilde{v}\|_V$. Then, one has

$$\|v\|_V \leqslant \|\tilde{v}\|_V + \|v - \tilde{v}\|_V \leqslant \|\tilde{v}\|_V + C_1 \sigma_{2n}(v - \tilde{v})_V \leqslant (1 + C_1)\|\tilde{v}\|_V,$$

by using v as a sparse approximation to $v - \tilde{v}$. Since $\|\tilde{v}\|_V = \|\Phi v\|_W$, this shows that $S(2n, \mu_n^W)$ holds with $\mu_n^W = 1 + C_1$. □

3.6.3 The case of ℓ^p norms

The range of m allowing the properties to be fulfilled is best understood in the case of the ℓ^p norms, that is $\|\cdot\|_V = \|\cdot\|_p$, as discussed in [52] which points out a striking difference between the cases $p = 2$ and $p = 1$:

1. In the case $p = 2$, it is proved that $NSP(2, C_1)$ cannot hold unless $N \leqslant C_1^2 m$. In other words, instance optimality in ℓ^2 even at order $n = 1$ requires a number of measurements that is proportional to the full space dimension.
2. In the more favorable case $p = 1$, it is proved that for matrices which satisfy the ℓ^2-RIP property of order $3n$
$$(1 - \delta)\|v\|_2^2 \leqslant \|\Phi v\|_2^2 \leqslant (1 + \delta)\|v\|_2^2, \quad v \in V_{3n},$$
with parameter $0 < \delta < \frac{(\sqrt{2}-1)^2}{3}$, the $NSP(2n, C_1)$ holds with C_1 depending on δ. Such matrices are known to exists with $m \sim n \log(N/n)$ rows.

Our setting based on the stability properties S and IS applies more naturally to a different class of matrices built from graphs, which is also known to be well adapted for sparse recovery in the ℓ^1 norm. A bipartite graph with (N, m) left and right vertices, and of left degree d, is an (l, ε)-graph expander if

$$|X| \leqslant l \Rightarrow |N(X)| \geqslant d(1-\varepsilon)|X|, \quad X \subset \{1, \dots, N\},$$

where $N(X) \subset \{1, \dots, m\}$ is the set of vertices connected to X. We necessarily have $|N(X)| \leqslant d|X|$, and $(1-\varepsilon)dl \geqslant m$. From [44], it is known that there exists a $(2n, \frac{1}{2})$-graph expander with $d \sim \log \frac{N}{n}$ and $m \sim nd \sim n \log(N/n)$.

Now denote $\Phi \in \{0,1\}^{m \times N}$ the adjacency matrix of this graph, so that each column of Φ has d nonzero entries. Then

$$\|\Phi x\|_1 \leqslant d\|x\|_1, \quad x \in \mathbb{R}^N,$$

and

$$\|\Phi x\|_1 \geqslant d(1-\varepsilon)\|x\|_1, \quad x \in V_{2n}.$$

Therefore $S(\alpha_1)$ and $IS(2n, \mu_1)$, hold with $\alpha_1 = d$ and $\mu_n^1 = \frac{1}{d(1-\varepsilon)} = \frac{2}{d}$, which by (3.23) and (3.22) gives $NSP(2n, C_1)$ with $C_1 = 3$ and $IOP(n, C_0)$ with $C_0 = 5$.

3.7 Appendix: Proof of Proposition 3.15

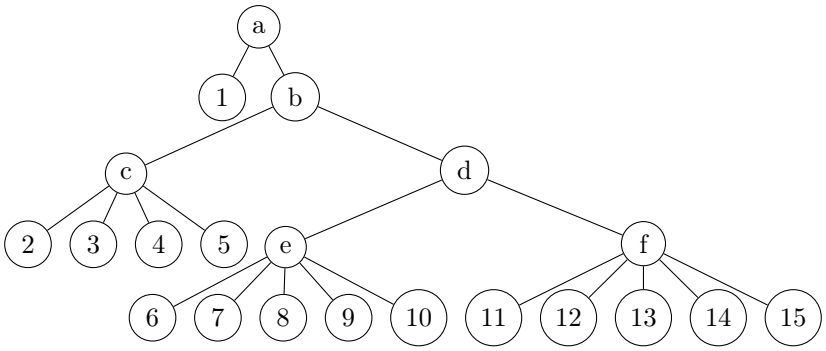


Figure 3.4 – Structure of the proof, each leaf corresponds to a different case, and each node contains a general treatment valid for all its sons

The proof contains 15 cases, represented on a tree in Figure 3.4. These cases correspond to different geometric situations, up to certain symmetries that leave the final relevant quantities $\|\ell(w)\|_W$ and $\|w\|_{L^1(S)}$ unchanged.

Node a: Take $w = u - v \in V_{2,T}^{\text{diff}}$, with $u, v \in V_{2,T}$, and denote $\vec{n}_u$, $\vec{n}_v$ and c_u, c_v the corresponding unit vectors and offsets from the definition 3.14 of $V_{2,T}$. Recalling that $\overline{x} = (\overline{x}_1, \overline{x}_2)$ is the center of S, we also denote

$$\Delta_u = \{x \in \mathbb{R}^2, (x - \overline{x}) \cdot \vec{n}_u = c_u\}$$

the delimiting line between $\{u = 0\}$ and $\{u = 1\}$, and define Δ_v in a similar way.

Case 1: If $\vec{n}_u = \vec{n}_v = \vec{n}$, we have

$$w = \begin{cases} \chi_{c_u \leqslant \vec{n} \cdot (x - \overline{x}) < c_v} & \text{if } c_u \leqslant c_v \\ -\chi_{c_v \leqslant \vec{n} \cdot (x - \overline{x}) < c_u} & \text{otherwise} \end{cases}$$

so w has constant sign, which implies $\|w\|_{L^1(S)} = h^2 \|\ell(w)\|_1 = \|\ell(w)\|_W$.

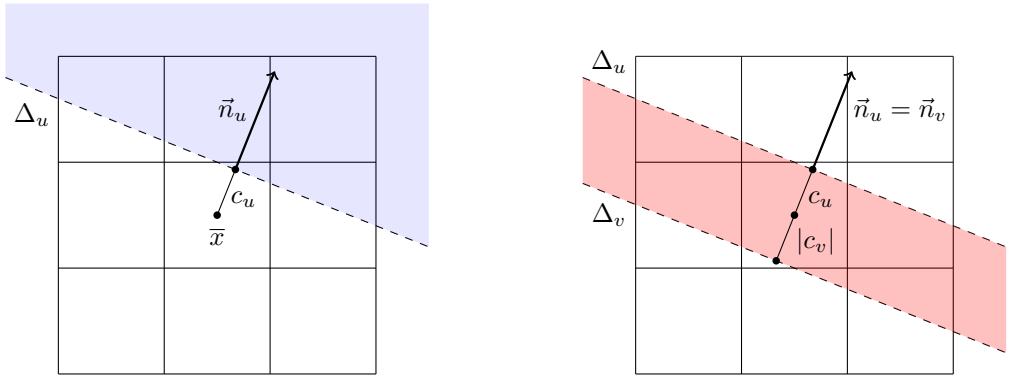


Figure 3.5 – Left: 3×3 stencil S, with $\overline{x}$ its center, and an example of function $u \in V_{2,T}$ with directing vector $\vec{n}_u$ and offset $c_u > 0$. Here the dotted line corresponds to Δ_u, and the shaded region to $u = 1$, while $u = 0$ elsewhere. Right: Representation of Case 1 ($\vec{n}_u = \vec{n}_v$), here $c_v < 0 < c_u$ so $w = -1$ on the shaded region and $w = 0$ elsewhere

Node b: In all other cases, the cones

$$\mathcal{C}_+ = \{x \in \mathbb{R}^2 \ : \ w(x) = 1\} \quad \text{and} \quad \mathcal{C}_- = \{x \in \mathbb{R}^2 \ : \ w(x) = -1\}$$

are non-empty, and we can define the external bisector

$$\Delta = \{x \in \mathbb{R}^2 \ : \ (\vec{n}_u - \vec{n}_v) \cdot (x - \overline{x}) = c_u - c_v\},$$

which is the line of symmetry between $\mathcal{C}_+$ and $\mathcal{C}_-$. We also denote

$$\mathcal{C} = \mathcal{C}_+ \cup \mathcal{C}_- = \{x \in \mathbb{R}^2 \ : \ |w(x)| = 1\}.$$

Observing that

$$\|w\|_{L^1(S)} = |S \cap \mathcal{C}| \tag{3.25}$$

and

$$\|\ell(w)\|_W = \sum_{T \subset S} \Big| |T \cap \mathcal{C}_+| - |T \cap \mathcal{C}_-| \Big|, \tag{3.26}$$

the stability property (3.16) can be rewritten as

$$|S \cap \mathcal{C}| \leqslant \frac{3}{2} \sum_{T \subset S} \Big| |T \cap \mathcal{C}_+| - |T \cap \mathcal{C}_-| \Big| = \frac{3}{2} \left(|S \cap \mathcal{C}| - 2 \sum_{T \subset S} \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|) \right),$$

or equivalently

$$|S \cap \mathcal{C}| \geqslant 6 \sum_{T \subset S} \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|). \tag{3.27}$$

Up to a rotation of S by a multiple of $\frac{\pi}{2}$, we may assume without loss of generality that

$$\arg(\vec{n}_u - \vec{n}_v) \in \left[\frac{\pi}{4}, \frac{3\pi}{4} \right],$$

that is, Δ is at an angle of at most $\frac{\pi}{4}$ with the horizontal axis, and $\mathcal{C}_+$ lies above Δ. Take $(\vec{e}_1, \vec{e}_2)$ the canonical basis of $\mathbb{R}^2$.

Node c: Consider the situation where $(\vec{n}_u \cdot \vec{e}_2)(\vec{n}_v \cdot \vec{e}_2) > 0$. As $\vec{n}_u \neq \vec{n}_v$ and $\vec{n}_u \neq -\vec{n}_v$, the lines Δ_u and Δ_v intersect at one point $X \in \mathbb{R}^2$. Moreover, the above condition implies $X + \vec{e}_2 \notin \mathcal{C}$. Using the fact that

$|\arg(\Delta)| \leqslant \frac{\pi}{4}$, we also get $X + \vec{e}_1 \notin \mathcal{C}$.

Up to a symmetry with respect to the vertical axis, we can assume that $\mathcal{C}_+$ is included in the quadrant $X + \mathbb{R}^2_+$. Now consider a cell $T \subset S$ such that $\min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|) \neq 0$, then there exist points $x \in T \cap \mathcal{C}_-$ and $y \in T \cap \mathcal{C}_+$. As $x_1 \leqslant X_1 \leqslant y_1$ and $x_2 \leqslant X_2 \leqslant y_2$, we get $X \in T$, so there is at most one such cell T, and inequality (3.27) reduces to

$$|S \cap \mathcal{C}| \geqslant 6 \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|).$$

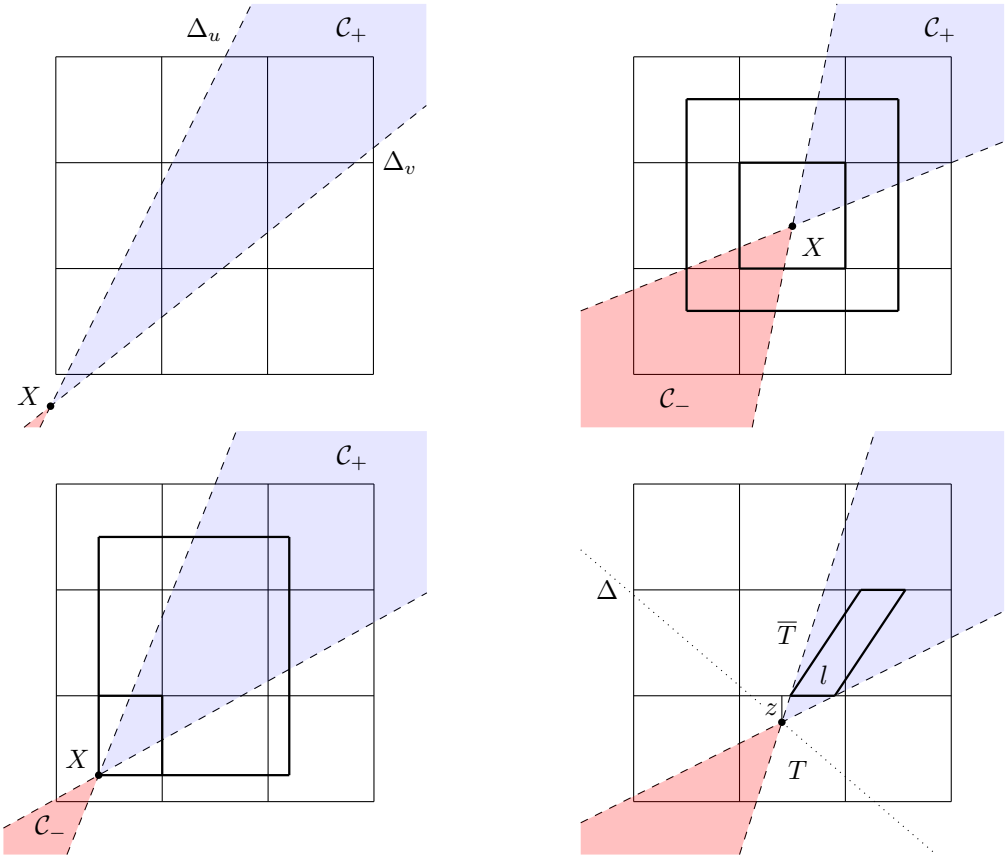


Figure 3.6 – Cases 2, 3, 4, and 5

Case 2: If $X \notin S$, then w has constant sign on S, so $\|w\|_{L^1(S)} = \|\ell(w)\|_W$.

Case 3: If X is in the central cell T, the dilation of T with respect to X by a factor 2 is a subset of S, and the image of $\mathcal{C} \cap T$ is in $\mathcal{C} \cap S$, so

$$|S \cap \mathcal{C}| \geqslant 4|T \cap \mathcal{C}| \geqslant 8 \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|).$$

Case 4: If X is in the lower left cell T, the dilation of $T \cap \mathcal{C}_+$ with respect to X by a factor 3 is in $S \cap \mathcal{C}_+$, so

$$|S \cap \mathcal{C}| \geqslant |S \cap \mathcal{C}_+| \geqslant 9|T \cap \mathcal{C}_+| \geqslant 9 \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|).$$

The same argument holds with $\mathcal{C}_-$ instead of $\mathcal{C}_+$ when X is in the upper right cell. Moreover, as Δ_u and Δ_v go through the central cell, X may not be in the upper left or lower right cells.

Case 5: If X is in the lower central cell T, denote $l = |\partial T \cap \mathcal{C}_+| \in (0, h)$ the distance between Δ_u and Δ_v when

they pass from T to the central cell $\overline{T}$, and $z = \mathrm{dist}(X, \overline{T}) \in (0, h)$ the depth of the point of intersection. Then

$$|T \cap \mathcal{C}_+| = \frac{zl}{2} \quad \text{and} \quad |T \cap \mathcal{C}_-| \leqslant \frac{zl}{2} \left(\frac{h-z}{z} \right)^2,$$

so $\min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|) \leqslant \frac{hl}{4}$. On the other hand, the parallelogram of base $\partial T \cap \mathcal{C}_+$, of height h, and with sides orthogonal to Δ belongs to $(S \setminus T) \cap \mathcal{C}_+$ (it does not escape to the right of S because Δ is close to the horizontal axis, so the sides of the parallelogram are at an angle at most $\frac{\pi}{4}$ with the vertical axis), and has an area hl, which proves that

$$|\mathcal{C} \cap S| \geqslant hl + |\mathcal{C}_+ \cap T| + |\mathcal{C}_- \cap T| \geqslant 6 \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|).$$

A similar construction can be applied to the remaining cases where X is in the upper central, central left or central right cell, which concludes the proof for Node c.

Node d: If now $(\vec{n}_u \cdot \vec{e}_2)(\vec{n}_v \cdot \vec{e}_2) \leqslant 0$, as $\arg(\vec{n}_u - \vec{n}_v) \in \left[\frac{\pi}{4}, \frac{3\pi}{4}\right]$, we get $\vec{n}_u \cdot \vec{e}_2 \geqslant 0 \geqslant \vec{n}_v \cdot \vec{e}_2$. Observe that $\mathcal{C}_+ + \vec{e}_2 \subset \mathcal{C}_+$ since for all $x \in \mathcal{C}_+$,

$$(x + \vec{e}_2 - \overline{x}) \cdot \vec{n}_u \geqslant (x - \overline{x}) \cdot \vec{n}_u \geqslant c_u \quad \text{and} \quad (x + \vec{e}_2 - \overline{x}) \cdot \vec{n}_v \leqslant (x - \overline{x}) \cdot \vec{n}_v < c_v.$$

In the same way, $\mathcal{C}_- - \vec{e}_2 \subset \mathcal{C}_-$. We now divide S into columns separated by the vertical boundaries between cells, and in addition by vertical lines where Δ intersects the two horizontal lines separating cells of S, as illustrated in Figure 3.7.

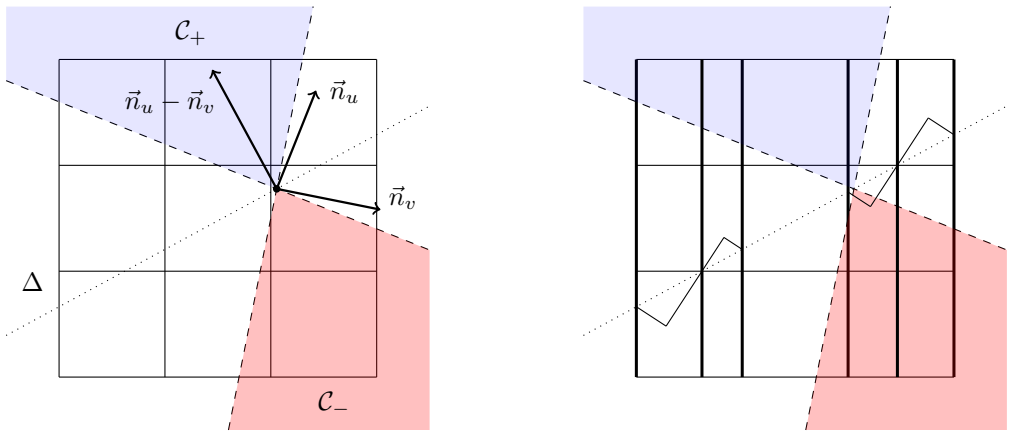


Figure 3.7 – Generic situation for Node d, and partition of S into 5 columns: here, in addition to the 4 vertical lines delimiting the cells of S, we added 2 vertical lines passing through the intersections of Δ with the 2 horizontal cell delimiters

Let U be such a column, and T a cell intersecting U. If $T \cap U \neq T$, Δ intersects either the upper or lower boundary of T, but not both since Δ is at an angle of at most $\frac{\pi}{4}$ with the horizontal axis. If it is the upper boundary, the symmetric of the part of $T \cap U$ above Δ with respect to Δ is in $T \cap U$. If it is the lower boundary, the symmetric of the part of $T \cap U$ below Δ with respect to Δ is in $T \cap U$. Using the fact that $\mathcal{C}_+$ and $\mathcal{C}_-$ are symmetric with respect to Δ, we obtain

$$\min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|) = \min(|T \cap U \cap \mathcal{C}_+|, |T \cap U \cap \mathcal{C}_-|) + \min(|T \cap U^c \cap \mathcal{C}_+|, |T \cap U^c \cap \mathcal{C}_-|).$$

Thanks to this observation, instead of (3.27) we only have to prove the inequality

$$|U \cap \mathcal{C}| \geqslant 6 \sum_{T \subset U} \min(|T \cap U \cap \mathcal{C}_+|, |T \cap U \cap \mathcal{C}_-|) \tag{3.28}$$

on each column U separately. We thus consider only one column U in the sequel, and assume up to a horizontal dilation (which preserves the condition $|\arg(\Delta)| \leqslant \frac{\pi}{4}$) that U has width h and is composed of three full cells.

According to the definition of the columns, there is at most one cell $T \subset U$ such that $T \cap \Delta \neq \emptyset$, and as Δ separates $\mathcal{C}_+$ and $\mathcal{C}_-$, it is only for this cell that we may have $\min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|) \neq 0$. If there is no such cell, (3.28) trivially holds. Otherwise, similar to Node c, we only need to prove

$$|U \cap \mathcal{C}| \geqslant 6 \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|),$$

where $T \subset U$ is the cell containing $\Delta \cap U$. Denoting P_1, P_2, P_3 and P_4 the upper left, upper right, lower left and lower right corner points of T, we observe that the assumptions on Δ and U imply $P_1, P_2 \notin \mathring{\mathcal{C}}_-$ and $P_3, P_4 \notin \mathring{\mathcal{C}}_+$.

Node e: If $U \cap \Delta_u \cap \Delta_v = \emptyset$, that is, if U contains no intersection point between Δ_u and Δ_v, we match 5 cases depending on the position of T in U, and of its corners with respect to $\mathcal{C}$. They are illustrated in Figure 3.8.

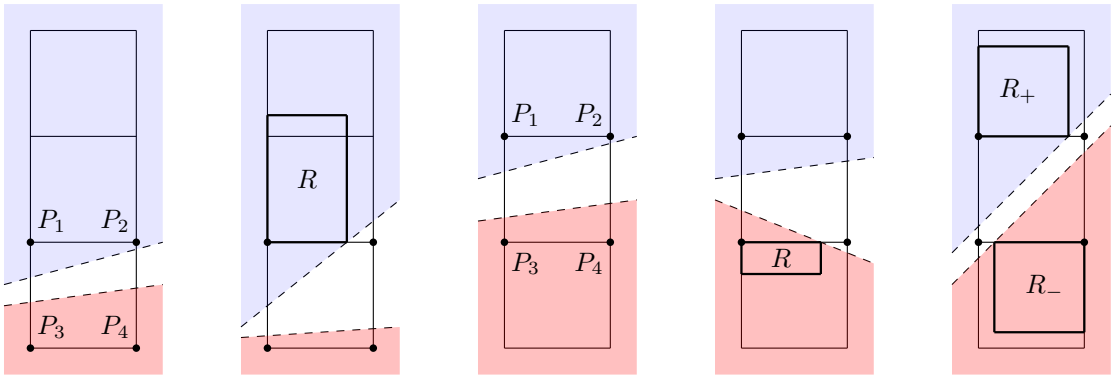


Figure 3.8 – Cases 6, 7, 8, 9 and 10

Case 6: If T is the bottom cell and $P_1, P_2 \in \mathcal{C}_+$, then the two other cells are included in $\mathcal{C}_+$, so

$$|U \cap \mathcal{C}| \geqslant 2h^2 + |T \cap \mathcal{C}| \geqslant 3|T \cap \mathcal{C}| \geqslant 6 \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|).$$

Case 7: If T is the bottom cell and $P_1 \in \mathcal{C}_+$ but $P_2 \notin \mathcal{C}_+$, $T \cap \mathcal{C}_+$ is a triangle of width and height at most h, so there is a rectangle $R \subset (U \setminus T) \cap \mathcal{C}_+$ of same width and twice as high, and thus

$$|U \cap \mathcal{C}| \geqslant |R| + |T \cap \mathcal{C}| = 4|T \cap \mathcal{C}_+| + |T \cap \mathcal{C}| \geqslant 6 \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|).$$

The same argument holds when $P_2 \in \mathcal{C}_+$ but $P_1 \notin \mathcal{C}_+$, and we necessarily have P_1 or P_2 in $\mathcal{C}_+$ since $T \cap \mathcal{C}_+ \neq \emptyset$. If T is the top cell, applying a symmetry with respect to the horizontal axis and exchanging $\mathcal{C}_+$ with $\mathcal{C}_-$ brings us back to Cases 6 and 7.

Case 8: If T is the central cell, $P_1, P_2 \in \mathcal{C}_+$ and $P_3, P_4 \in \mathcal{C}_-$ the two other cells are included in $\mathcal{C}_+$ and $\mathcal{C}_-$, and we conclude as in Case 6.

Case 9: If T is the central cell, $P_1, P_2 \in \mathcal{C}_+$, $P_3 \in \mathcal{C}_-$ but $P_4 \notin \mathcal{C}_-$, the top cell is included in $\mathcal{C}_+$, and there is a rectangle $R \subset \mathcal{C}_-$ of same width and height as $T \cap \mathcal{C}_-$ in the bottom cell, so

$$|U \cap \mathcal{C}| \geqslant h^2 + |T \cap \mathcal{C}| + |R| \geqslant 2|T \cap \mathcal{C}| + 2|T \cap \mathcal{C}_-| \geqslant 6 \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|).$$

The same situation occurs when only three points among $P_1, \ldots, P_4$ are in $\mathcal{C}$.

Case 10: If T is the central cell, only one vertex among P_1, P_2 is in $\mathcal{C}_+$, and only one among P_3, P_4 is

in $\mathcal{C}_-$, both $T \cap \mathcal{C}_+$ and $T \cap \mathcal{C}_-$ are triangles, and there exist rectangles R_+ and R_- of same widths and heights, so

$$|U \cap \mathcal{C}| \geqslant |R_+| + |T \cap \mathcal{C}| + |R_-| \geqslant 3|T \cap \mathcal{C}_+| + 3|T \cap \mathcal{C}_-| \geqslant 6 \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|).$$

As $\mathcal{C}_+$ and $\mathcal{C}_-$ each contain at least one corner of T, we treated all cases for Node e.

Node f: Finally, we consider the situation where there is an intersection point $X \in \Delta_u \cap \Delta_v$ in U, and therefore in T. We again match 5 cases, illustrated in Figure 3.9, depending on the position of T in U, and of its corners with respect to $\mathcal{C}$.

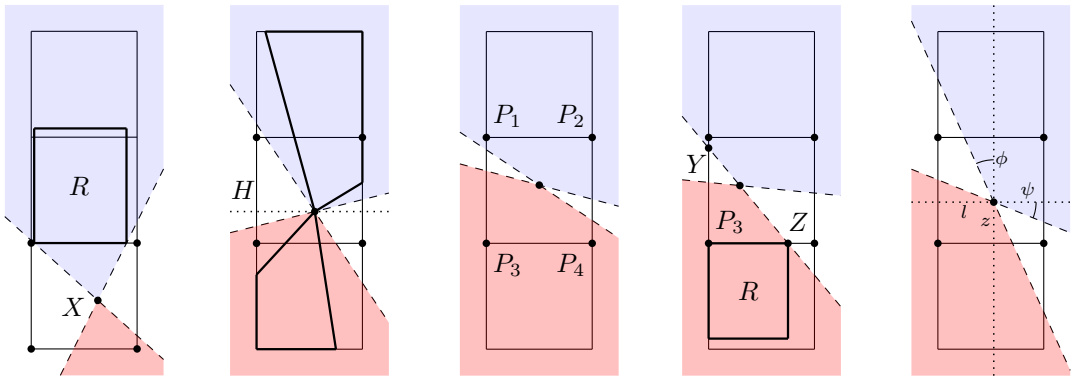


Figure 3.9 – Cases 11, 12, 13, 14 and 15

Case 11: If T is the bottom cell, as Δ_u and Δ_v pass through the central cell of S, U is included in the central column of S, and no corner of T can be in $\mathring{\mathcal{C}}_+$, since otherwise Δ would have to pass through that corner, according to the definition of the columns. As a consequence, Δ_u and Δ_v necessarily pass through the central cell of U, so $T \cap \mathcal{C}_+$ is a triangle, and we proceed as in Case 7. The same happens if T is the top cell, so in the rest of the proof we only consider situations where T is the central cell.

Case 12: If the horizontal line H passing through X does not intersect $\mathcal{C}$ at any other point, $\mathcal{C}_+$ is entirely above H and $\mathcal{C}_-$ entirely below. Denoting $z = X_2 - \overline{x}_2 + \frac{h}{2} \in (0, h)$, the vertical dilation with respect to H by a factor $\frac{2h-z}{h-z}$ sends $T \cap \mathcal{C}_+$ in $U \cap \mathcal{C}_+$, and the vertical dilation with respect to H by a factor $\frac{h+z}{z}$ sends $T \cap \mathcal{C}_-$ in $U \cap \mathcal{C}_-$, so

$$|U \cap \mathcal{C}| \geqslant \frac{2h-z}{h-z}|T \cap \mathcal{C}_+| + \frac{h+z}{z}|T \cap \mathcal{C}_-| \geqslant 6 \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|)$$

because $\frac{2h-z}{h-z} + \frac{h+z}{z} = 2 + \frac{h^2}{z(h-z)} \geqslant 6$ for $z \in (0, h)$.

In the remaining cases, up to a symmetry with respect to the vertical axis, we can assume that $X + \mathbb{R}^2_+ \subset \mathcal{C}_+$ and $X + \mathbb{R}^2_- \subset \mathcal{C}_-$, and in particular $P_2 \in \mathcal{C}_+$ and $P_3 \in \mathcal{C}_-$.

Case 13: If $P_1 \in \mathcal{C}_+$ and $P_4 \in \mathcal{C}_-$, the situation is similar to Case 8.

Case 14: If $P_1 \in \mathcal{C}_+$ and $P_4 \notin \mathcal{C}_-$, the top cell is included in $\mathcal{C}_+$, and one of the lines Δ_u or Δ_v intersects the line segments $[P_1, P_3]$ and $[P_3, P_4]$ at points Y and Z. Then the triangle YP_3Z is included in T and contains $T \cap \mathcal{C}_-$, so there is a rectangle R of same width and height in $(U \setminus T) \cap \mathcal{C}_-$. In the end

$$|U \cap \mathcal{C}| \geqslant h^2 + |T \cap \mathcal{C}| + |R| \geqslant 2|T \cap \mathcal{C}| + 2|T \cap \mathcal{C}_-| \geqslant 6 \min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|).$$

The same approach treats the symmetric case $P_1 \notin \mathcal{C}_+$ and $P_4 \in \mathcal{C}_-$,

Case 15: Finally, if $P_1 \notin \mathcal{C}_+$ and $P_4 \notin \mathcal{C}_-$, denote $l = X_1 - \overline{x}_1 + \frac{h}{2} \in (0, h)$, $z = X_2 - \overline{x}_2 + \frac{h}{2} \in (0, h)$, $\phi \in (0, \frac{\pi}{4})$ the angle between the vertical axis and the line among Δ_u and Δ_v that intersects $[P_1, P_2]$, and

$\psi \in (0, \frac{\pi}{4})$ the angle between the line among Δ_u and Δ_v that intersects $[P_1, P_3]$ and the horizontal axis. As $|\arg(\Delta)| \leqslant \frac{\pi}{4}$, $\phi \geqslant \psi$ so $\tan(\psi) \leqslant \tan(\phi) =: t \leqslant 1$.

We can now compute

$$|T \cap \mathcal{C}_+| = (h-l)(h-z) + \frac{1}{2}(h-l)^2 \tan\psi + \frac{1}{2}(h-z)^2 \tan\phi,$$
$$|T \cap \mathcal{C}_-| = lz + \frac{1}{2}l^2 \tan\psi + \frac{1}{2}z^2 \tan\phi,$$

and

$$|(U \setminus T) \cap \mathcal{C}| \geqslant (h-l)h + (h-z)th + lh + zth = (1+t)h^2.$$

If $l + z \leqslant h$, we get

$$|(U \setminus T) \cap \mathcal{C}| \geqslant (1+t)(l+z)^2 - (1-t)(l-z)^2 = 4lz + 2t(l^2+z^2) \geqslant 4|T \cap \mathcal{C}_-|.$$

Similarly, $l + z \geqslant h$ implies $|(U \setminus T) \cap \mathcal{C}| \geqslant 4|T \cap \mathcal{C}_+|$. In any case, we found

$$|U \cap \mathcal{C}| = |T \cap \mathcal{C}| + |(U \setminus T) \cap \mathcal{C}| \geqslant 6\min(|T \cap \mathcal{C}_+|, |T \cap \mathcal{C}_-|),$$

which concludes the proof.

As a last remark, note that the constants $\alpha_W = 1$ and $\mu_n^W = \frac{3}{2}$ in Proposition 3.15 are sharp, since equality is attained by functions of constant sign on each cell for α_W, and by $w = u - v$ with $\arg(\vec{n}_u) \in \frac{\pi}{4}\mathbb{Z}$, $c_u = 0$ and $v = u - 1$ for μ_n^W.

Part II

Approximation from point values

Chapter 4

Optimal sampling and Christoffel functions on general domains

Abstract. We consider the problem of reconstructing an unknown function $u \in L^2(\Omega, \mu)$ from its evaluations at given sampling points $x^1, \ldots, x^m \in \Omega$, where $\Omega \subset \mathbb{R}^d$ is a general domain and μ a probability measure. The approximation is picked from a linear space V_n of interest where $n = \dim(V_n)$. Recent results [59, 86, 138] have revealed that certain weighted least-squares methods achieve near best (or instance optimal) approximation with a sampling budget m that is proportional to n, up to a logarithmic factor $\ln(2n/\varepsilon)$, where $\varepsilon > 0$ is a probability of failure. The sampling points should be picked at random according to a well-chosen probability measure σ whose density is given by the inverse Christoffel function that depends both on V_n and μ. While this approach is greatly facilitated when Ω and μ have tensor product structure, it becomes problematic for domains Ω with arbitrary geometry since the optimal measure depends on an orthonormal basis of V_n in $L^2(\Omega, \mu)$ which is not explicitly given, even for simple polynomial spaces. Therefore sampling according to this measure is not practically feasible. One computational solution recently proposed in [6] relies on using the restrictions of an orthonormal basis of V_n defined on a simpler bounding domain and sampling according to the original probability measure μ, in turn giving up on the optimal sampling budget $m \sim n$. In this chapter, we discuss practical sampling strategies, which amount to using a perturbed measure $\widetilde{\sigma}$ that can be computed in an offline stage, not involving the measurement of u, as recently proposed in [5, 133]. We show that near best approximation is attained by the resulting weighted least-squares method at near-optimal sampling budget, and we discuss multilevel approaches that preserve optimality of the cumulated sampling budget when the spaces V_n are iteratively enriched. These strategies rely on the knowledge of a-priori upper bounds $B(n)$ on the inverse Christoffel function for the space V_n and the domain Ω. We establish bounds of the form $\mathcal{O}(n^r)$ for spaces V_n of multivariate algebraic polynomials of given total degree, and for general domains Ω. The exact growth rate r is established depending on the regularity of the domain, in particular $r = 1$ for domains with Lipschitz boundaries and $r = 1/d$ for smooth domains.

4.1 Introduction

4.1.1 Reconstruction from point samples

The process of reconstructing an unknown function u defined on a domain $\Omega \subset \mathbb{R}^d$ from its sampled values $z_i \approx u(x^i)$ at a set of points $x^1, \ldots, x^m \in \Omega$ is ubiquitous in data science and engineering. The sampled values may be affected by noise, making critical the stability properties of the reconstruction process. Let us mention three very different settings for such reconstruction problems, that correspond to different areas of applications:

(i) Statistical learning and regression: we observe m independent realizations (x^i, z_i) of a random variable (x, z) distributed according to an unknown measure, where $x \in \Omega$ and $z \in \mathbb{R}$, and we want to recover a function $x \mapsto v(x)$ that makes $|z - v(x)|$ as small as possible in some given sense. If we use the quadratic

loss $\mathbb{E}(|z - v(x)|^2)$, the minimizer is given by the regression function

$$u(x) = \mathbb{E}(z|x).$$

and the observed z_i may be thought of as the observation of $u(x^i)$ affected by noise.

(ii) State estimation from measurements: the function u represents the distribution of a physical quantity (temperature, quantity of a contaminant, acoustic pressure) in a given spatial domain Ω that one is allowed to measure by sensors placed at m locations $x^1, \dots, x^m$. These measurements can be affected by noise reflecting the lack of accuracy of the sensors.

(iii) Design of physical/computer experiments: u is a quantity of interest that depends on the solution f to a parametrized physical problem. For example, $f = f(x)$ could be the solution to a PDE that depends on a vector $x \in \Omega \subset \mathbb{R}^d$ of d physical parameters, and u could be the result of a linear form ℓ applied to f, that is, $u(x) = \ell(f(x))$. We use a numerical solver for this PDE as a black box to evaluate f, and therefore u, at m chosen parameter vectors $x^1, \dots, x^m \in \Omega$, and we now want to approximate u on the whole domain Ω from these computed values z_i. Here, the discretization error of the solver may be considered as a noise affecting the true value $u(x^i)$.

Contrary to statistical learning, in the last two applications (ii) and (iii) the positions of the sample points x^i are not realizations of an unknown probability distribution. They can be selected by the user, which brings out the problem of choosing them in the best possible way. Indeed, measuring u at the sample points may be costly: in (ii) we need a new sensor for each new point, and in (iii) a new physical experiment or run of a numerical solver. Moreover, in certain applications, one may be interested in reconstructing many different instances of functions u. Understanding how to sample in order to achieve the best possible trade-off between the sampling budget and the reconstruction performance is one main motivation of this work. We first make our objective more precise by introducing some benchmarks for the performance of the reconstruction process and sampling budget.

4.1.2 Optimality benchmarks

We are interested in controlling the distance $\|u - \tilde{u}\|_V$ between u and its reconstruction $\tilde{u} = \tilde{u}(z_1, \dots, z_m)$, measured in some given norm $\|\cdot\|_V$, where V is a Banach function space that contains u.

For a given numerical method, the derivation of an error bound is always tied to some prior information on u. One most common way to express such a prior is in terms of membership of u to a restricted class of functions, for example a smoothness class. One alternate way is to express the prior in terms of approximability of u by particular finite dimensional spaces. It is well-known that the two priors are sometimes equivalent: many classical smoothness classes can be characterized in terms of approximability in some given norm by classical approximation spaces such as algebraic or trigonometric polynomials, splines or wavelets [64].

In this chapter, we adopt the second point of view, describing u by its closeness to a given subspace $V_n \subset V$ of dimension n: defining the best approximation error

$$e_n(u)_V := \min_{v \in V_n} \|u - v\|_V,$$

our prior is that $e_n(u)_V \leqslant \varepsilon_n$ for some $\varepsilon_n > 0$. One of our motivations is the rapidly expanding field of *reduced order modeling* in which one searches for approximation spaces V_n which are optimally designed to approximate families of solutions to parametrized PDEs. Such spaces differ significantly from the above-mentioned classical examples. For example in the *reduced basis method*, they are generated by particular instances of solutions to the PDE for well chosen parameter values. We refer to [56] for a survey on such reduced modeling techniques and their approximation capability.

In this context, one first natural objective is to build a reconstruction map

$$(z_1, \dots, z_m) \mapsto \tilde{u} \in V_n,$$

that performs almost as good as the best approximation error. We say that a reconstruction map taking its value in V_n is *instance optimal* with constant $C_0 \geqslant 1$ if and only if

$$\|u - \tilde{u}\|_V \leqslant C_0 \, e_n(u)_V, \tag{4.1}$$

for any $u \in V$.

Obviously, instance optimality implies that if $u \in V_n$, the reconstruction map should return an exact reconstruction $\tilde{u} = u$. For this reason, instance optimality can only be hoped for if the sampling budget m exceeds the dimension n. This leads us to introduce a second notion of optimality: we say that the sample is *budget optimal* with constant $C_1 \geqslant 1$ if

$$m \leqslant C_1\, n. \tag{4.2}$$

Let us stress that in many relevant settings, we do not work with a single space V_n but a sequence of nested spaces

$$V_1 \subset V_2 \subset \cdots \subset V_n \subset \ldots$$

so that $e_n(u)_V$ decreases as n grows. Such a hierarchy could either be fixed in advance (for example when using polynomials of degree n), or adaptively chosen as we collect more samples (for example when using locally refined piecewise polynomials or finite element spaces). Ideally, we may ask that the constants C_0 and C_1 are independent of n. As it will be seen, a more accessible goal is that only one of the two constants is independent of n, while the other grows at most logarithmically in n.

Another way of relaxing instance optimality is to request the weaker property of *rate optimality*, which requires that for any $s > 0$ and $u \in V$,

$$\sup_{n \geqslant 1} n^s \, \|u - \tilde{u}\|_V \leqslant C \sup_{n \geqslant 1} n^s \, e_n(u)_V,$$

where $C \geqslant 1$ is a fixed constant. In other words, the approximant produced by the reconstruction method should converge at the same polynomial rate as the best approximation.

In the context where the spaces V_n are successively refined, even if the reconstruction method is instance and budget optimal for each value of n, the cumulated sampling budget until the n-th refinement step is in principle of the order

$$m(n) \sim 1 + 2 + \cdots + n \sim n^2,$$

if samples are picked independently at each step. A natural question is whether the samples used until stage k can be, at least partially, recycled for the computation of $\tilde{u}_{k+1}$, in such a way that the cumulated sampling budget $m(n)$ remains of the optimal order $\mathcal{O}(n)$. This property will be ensured for example if for each n, the samples are picked at points $\{x^1, \ldots, x^{m(n)}\}$ that are the sections of a unique infinite sequence $\{x^m\}_{m \geqslant 1}$, with $m(n) \sim n$, which means that all previous samples are recycled. We refer to this property as *hierarchical* sampling. It is also referred to as *online machine learning* in the particular above-mentioned application area (i).

4.1.3 Objectives and layout

The design of sampling and reconstruction strategies that combine budget and instance (or rate) optimality, together with the above progressivity prescription, turns out to be a difficult task, even for very classical approximation spaces V_n such as polynomials.

In Section 4.2, we illustrate this difficulty by first discussing the example of reconstruction by *interpolation* for which the sampling budget is optimal but instance optimality with error measured in the L^∞ norm generally fails by a large amount. We then recall recent results [17, 59, 86, 138] revealing that one can get much closer to these optimality objectives by *weighted least-squares* reconstruction methods. In this case, we estimate the approximation error in $V = L^2(\Omega, \mu)$, where μ is an arbitrary but fixed probability measure. The sampling points are picked at random according to a different probability measure σ^* that depends on V_n and μ:

$$d\sigma^*(x) = k_n(x)\, d\mu(x).$$

Here k_n is the *inverse Christoffel function* defined by

$$k_n(x) = \frac{1}{n} \sum_{j=1}^{n} |\varphi_j(x)|^2, \tag{4.3}$$

where $(\varphi_1, \ldots, \varphi_n)$ is any $L^2(\Omega, \mu)$-orthonormal basis of V_n. By Cauchy-Schwarz inequality, it is readily seen

that this function is characterized by the extremality property

$$n\, k_n(x) = \max_{v \in V_n} \frac{|v(x)|^2}{\|v\|_{L^2}^2}, \tag{4.4}$$

where $\|v\|_{L^2} := \|v\|_V = \|v\|_{L^2(\Omega,\mu)}$ (for notational simplicity, throughout the chapter, we ommit the obvious restriction $v \neq 0$ needed when optimizing quotients with numerators and denominators that are null when $v = 0$). Then, instance optimality is achieved in a probabilistic sense with a sampling budget m that is proportional to n, up to a logarithmic factor $\ln(2n/\varepsilon)$, where $\varepsilon > 0$ is a probability of failure which comes as an additional term in the instance optimality estimate

$$\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2) \leqslant C_0\, e_n(u)_{L^2}^2 + \mathcal{O}(\varepsilon).$$

It is important to notice that σ^* differs from μ and that the standard least-squares method using a sample drawn according to μ is generally *not* budget optimal in the sense that instance optimality requires m to be larger than $n \ln n$ times the quantity

$$K_n := \|k_n\|_{L^\infty} = \sup_{x \in \Omega} |k_n(x)| = \frac{1}{n} \max_{v \in V_n} \frac{\|v\|_{L^\infty}^2}{\|v\|_{L^2}^2},$$

which may grow with n, for instance as $\mathcal{O}(n)$ or worse, see [55] as well as Section 4.5.

While these results are in principle attractive since they apply to arbitrary spaces V_n, measures μ and domains Ω, the proposed sampling strategy is highly facilitated when Ω is a tensor-product domain and μ is the tensor-product of a simple univariate measure, so that an $L^2(\Omega,\mu)$-orthonormal basis of V_n can be explicitly provided. This is the case for example when using multivariate algebraic or trigonometric polynomial spaces with μ being the uniform probability measure on $[-1,1]^d$ or $[-\pi,\pi]^d$. For a general domain Ω with arbitrary — possibly irregular— geometry, the orthonormal basis cannot be explicitly computed, even for simple polynomial spaces. Therefore sampling according to the optimal measure σ^* is not feasible.

Non-tensor product domains Ω come out naturally in all the above mentioned settings (i)-(ii)-(iii). For example, in design of physical/computer experiments, this reflects the fact that while the individual parameters $x_1, \ldots, x_d$ could range in intervals $I_1, \ldots, I_d$, not all values x in the rectangle $R = I_1 \times \cdots \times I_d$ are physically admissible. Therefore, the function u is only accessible and searched for in a limited domain $\Omega \subset R$. Here we assume that the domain Ω is known to us, in the sense that membership in Ω of a point $x \in \mathbb{R}^d$ can be assessed at low numerical cost.

One practical solution proposed in [6] consists in sampling according to the measure μ and solving the least-squares problem using the restriction of an orthonormal basis of V_n defined on a simpler tensor product bounding domain, which generally gives rise to a frame. This approach is feasible for example when μ is the uniform probability measure. Due to the use of restricted bases, the resulting Gramian matrix which appears in the normal equations is ill-conditioned or even singular, which is fixed by applying a pseudo-inverse after thresholding the smallest singular values at some prescribed level. Budget optimality is generally lost in this approach since one uses μ as a sampling measure.

In this chapter, we also work under the assumption that we are able to sample according to μ, but we take a different path, which is exposed in Section 4.3. In an *offline* stage, we compute an approximation $\widetilde{k}_n$ to the inverse Christoffel function, which leads to a measure $\widetilde{\sigma}$ that may be thought of as a perturbation of the optimal measure σ^*. We may then use $\widetilde{\sigma}$ to define the sampling points $\{x^1, \ldots, x^m\}$ and weights. In the *online* stage, we perform the weighted least-squares reconstruction strategy based on the measurement of u at these points. Our first result is that if $\widetilde{k}_n$ is equivalent to k_n, we recover the stability and instance optimality results from [59] at near-optimal sampling budget $m \sim n \ln(2n/\varepsilon)$.

One approach for computing $\widetilde{k}_n$, recently proposed in [5, 133], consists in drawing a first sample $\{y^1, \ldots, y^M\}$ according to μ and defining $\widetilde{k}_n$ as the inverse Christoffel function with respect to the discrete measure associated to these points. In order to ensure an equivalence between k_n and $\widetilde{k}_n$ with high probability, the value of M needs to be chosen larger than K_n which is unknown to us. This can be ensured by asking that M is larger than a known upper bound $B(n)$ on K_n. The derivation of such bounds for general domains is one of the objectives of this chapter. We also propose an empirical strategy for choosing M that does not require the knowledge of an upper bound and appears to be effective in our numerical tests. In all cases, the size M of the offline sample

could be of order substantially larger than $\mathcal{O}(n)$. However, this first set of points is only used in the offline stage to perform computations that produce the perturbed measure $\widetilde{\sigma}$, and *not* to evaluate the function u which, as previously explained, is the costly aspect in the targeted applications and could also occur for many instances of u. These more costly evaluations of u only take place in the online stage at the x^i, therefore at near-optimal sampling budget.

In the case where K_n, or its available bound $B(n)$, grows very fast with n, the complexity of the offline stage in this approach becomes itself prohibitive. In order to mitigate this defect, we introduce in Section 4.4 a multilevel approach where the approximation $\widetilde{k}_n$ of k_n is produced by successive space refinements

$$V_{n_1} \subset \cdots \subset V_{n_q}, \quad n_q = n,$$

which leads to substantial computational savings under mild assumptions. This setting also allows us to produce nested sequences of evaluation points $\{x^1, \ldots, x^{m_p}\}$, where m_p grows similar to n_p up to a logarithmic factor, therefore complying with the previously invoked prescription of hierarchical sampling. The analysis of this approach faces the difficulty that the x^i are not anymore identically distributed, and this is solved by using techniques first proposed in [132].

In Section 4.5 we turn to the study of the inverse Christoffel function k_n in the case of algebraic polynomial spaces of given total degree on general multivariate domains $\Omega \subset \mathbb{R}^d$. We establish pointwise and global upper and lower bounds for k_n that depend on the smoothness of the boundary of Ω. We follow an approach adopted in [153] for a particular class of domains with piecewise smooth boundary, namely comparing Ω with simpler reference domains for which the inverse Christoffel function can be estimated. We obtain bounds with growth rate $\mathcal{O}(n)$ for Lipschitz domains and $\mathcal{O}(n^{1/d})$ for smooth domains, and these rates are proved to be sharp. We finally give a systematic approach that also describes the sharp growth rate for domains with cusp singularities.

We close the chapter in Section 4.6 with various numerical experiments that confirm our theoretical investigations. In the particular case of multivariate algebraic polynomials, the sampling points tend to concentrate near to the outward corner or cusp singularities of the domain, while they do not at the re-entrant singularities, as predicted by the previous analysis of the inverse Christoffel function.

4.2 Meeting the optimality benchmarks

4.2.1 Interpolation

One most commonly used strategy to reconstruct functions from point values is interpolation. Here we work in the space $V = \mathcal{C}(\Omega)$ of continuous and bounded functions equipped with the L^∞ norm. For the given space V_n, and n distinct points $x^1, \ldots, x^n \in \Omega$ picked in such way that the map $v \mapsto (v(x^1), \ldots, v(x^n))$ is an isomorphism from V_n to $\mathbb{R}^n$, we define the corresponding interpolation operator $\mathcal{I}_n : \mathcal{C}(\Omega) \to V_n$ by the interpolation condition

$$\mathcal{I}_n u(x^i) = u(x^i), \quad i = 1, \ldots, n.$$

The interpolation operator is also expressed as

$$\mathcal{I}_n u = \sum_{i=1}^{n} u(x^i)\, \psi_i,$$

where $(\psi_1, \ldots, \psi_n)$ is the Lagrange basis of V_n defined by the conditions $\psi_i(x^j) = \delta_{i,j}$. Interpolation is obviously budget optimal since it uses $m = n$ points, that is, $C_1 = 1$ in (4.2). On the other hand, it does not guarantee instance optimality: the constant C_0 in (4.1) is governed by the *Lebesgue constant*

$$\Lambda_n = \|\mathcal{I}_n\|_{L^\infty \to L^\infty} = \max_{x \in \Omega} \sum_{i=1}^{n} |\psi_i(x)|.$$

Indeed, since $\|u - \mathcal{I}_n u\|_{L^\infty} \leqslant \|u - v\|_{L^\infty} + \|\mathcal{I}_n u - \mathcal{I}_n v\|_{L^\infty}$ for any $v \in V_n$, one has

$$\|u - \mathcal{I}_n u\|_{L^\infty} \leqslant (1 + \Lambda_n) e_n(u)_{L^\infty}.$$

The choice of the points x^i is critical to control the growth of Λ_n with n. For example in the elementary case of univariate algebraic polynomials where $\Omega = [-1, 1]$ and $V_n = \mathbb{R}_{n-1}[X]$, it is well known that uniformly spaced x^i result in Λ_n growing exponentially, at least like 2^n, while the slow (and optimal) growth $\Lambda_n \sim \ln(n)$ is ensured when using the Chebychev points $x^i = \cos\left(\frac{2i-1}{2n}\pi\right)$ for $i = 1, \dots n$. Unfortunately, there is no general guideline to ensure such a slow growth for more general hierarchies of spaces $(V_n)_{n\geqslant 1}$ defined on multivariate domains $\Omega \subset \mathbb{R}^d$. As an example, for the space of bivariate algebraic polynomials $V_n = \mathbb{R}_p[X_1, X_2]$ with $n = \frac{(p+1)(p+2)}{2}$, and for a general polygonal domain Ω, a choice of points that would ensure a logarithmic growth of the Lebesgue constant is to our knowledge an open problem.

There exists a general point selection strategy that ensures linear behaviour of the Lebesgue constant for any space V_n spanned by n functions $\{\phi_1, \dots, \phi_n\}$: it consists in choosing $(x^1, \dots, x^n)$ which maximizes over Ω^n the determinant of the collocation matrix

$$A(x^1, \dots, x^n) = (\phi_i(x^j))_{1\leqslant i,j\leqslant n},$$

Since the j-th element of the Lagrange basis is given by

$$\psi_j(x) = \frac{\det(A(x^1, \dots, x^{j-1}, x, x^{j+1}, \dots, x^n))}{\det(A(x^1, \dots, x^n))},$$

the maximizing property gives that $\|\psi_j\|_{L^\infty} \leqslant 1$ and therefore $\Lambda_n \leqslant n$. In the particular case of the univariate polynomials where $\Omega = [-1, 1]$ and $V_n = \mathbb{R}_{n-1}[X]$, this choice corresponds to the Fekete points, which maximize the product $\prod_{i\neq j}(x^i - x^j)$.

While the above strategy guarantees the $\mathcal{O}(n)$ behaviour of Λ_n, its main defect is that it is computationally unfeasible if n or d is large, since it requires solving a non-convex optimization problem in dimension $d\,n$. In addition to this, for a given hierarchy of spaces $(V_n)_{n\geqslant 1}$, the sampling points $S_n = \{x^1, \dots, x^n\}$ generated by this strategy do not satisfy the nestedness property $S_n \subset S_{n+1}$.

A natural alternate strategy that ensures nestedness consists in selecting the points by a stepwise greedy optimization process: given S_{n-1}, define the next point x^n by maximizing over Ω the function

$$x \mapsto \det(A(x^1, \dots, x^{n-1}, x)).$$

This approach was proposed in [123] in the context of reduced basis approximation and termed as *magic points*. It amounts to solving at each step a non-convex optimization problem in the more moderate dimension d, independent of n. However there exists no general bound on Λ_n other than exponential in n. In the univariate polynomial case, this strategy yields the so-called *Leja points* for which it is only known that the Lebesgue constant grows sub-exponentially although numerical investigation indicates that it could behave linearly. In this very simple setting, the bound $\Lambda_n \leqslant n^2$ could be established in [50], however using a variant where the points are obtained by projections of the complex Leja points from the unit circle to the interval $[-1, 1]$.

In summary, while interpolation uses the optimal sampling budget $m = n$, it fails by a large amount in achieving instance optimality, especially when asking in addition for the nestedness of the sampling points, even for simple polynomial spaces.

4.2.2 Weighted least-squares

In order to improve the instance optimality bound, we allow ourselves to collect more data on the function u by increasing the number m of sample points, compared to the critical case $m = n$ studied before, and construct an approximation $\tilde{u}$ by a least-squares fitting procedure. This relaxation of the problem gives more flexibility on the choice of the sample points: for instance, placing two of them too close will only waste one evaluation of u, whereas this situation would have caused ill-conditioning and high values of Λ_n in interpolation. It also leads to more favorable results in terms of instance optimality, as we next recall.

Here, and in the rest of this chapter, we assess the error in the L^2 norm

$$\|v\|_{L^2} = \|v\|_{L^2(\Omega,\mu)},$$

where μ is a fixed probability measure, which can be arbitrarily chosen by the user depending on the targeted

application. For example, if the error has the same significance at all points of Ω, one is naturally led to use the uniform probability measure

$$d\mu := |\Omega|^{-1} dx.$$

In other applications such as uncertainty quantification where the x variable represents random parameters that follow a more general probability law μ, the use of this specific measure is relevant since the reconstruction error may then be interpreted as the mean-square risk

$$\|u - \tilde{u}\|^2_{L^2(\Omega,\mu)} = \mathbb{E}_x(|u(x) - \tilde{u}(x)|^2).$$

Once the evaluations of $u(x^i)$ are performed, the *weighted-least squares methods* defines $\tilde{u}$ as the solution of the minimization problem

$$\min_{v \in V_n} \frac{1}{m} \sum_{i=1}^{m} w(x^i) \, |u(x^i) - v(x^i)|^2, \tag{4.5}$$

where $w(x^1), \dots, w(x^m) > 0$ are position-dependent weights. The solution to this problem is unique under the assumption that no function of $V_n \setminus \{0\}$ vanishes at all the x^i. Notice that in the limit $m = n$, the minimum in (4.5) is zero, and it is attained by the interpolant at the points $x^1, \dots, x^n$, which as previously discussed suffers from a severe lack of instance optimality.

The results from [59] provide with a general strategy to select the points x^i and the weight function w in order to reach instance and budget optimality, in a sense that we shall make precise. In this approach, the points x^i are drawn at random according to a probability measure σ on Ω, that generally differs from μ, but with respect to which μ is absolutely continuous. One then takes for w the corresponding Radon-Nikodym derivative, so that

$$w(x) \, d\sigma(x) = d\mu(x). \tag{4.6}$$

This compatibility condition ensures that we recover a minimization in the continuous norm $\|\cdot\|_{L^2}$ as m tends to infinity:

$$\frac{1}{m} \sum_{i=1}^{m} w(x^i) \, |u(x^i) - v(x^i)|^2 \xrightarrow[m \to \infty]{a.s.} \int_\Omega w \, |u - v|^2 \, d\sigma = \int_\Omega |u - v|^2 \, d\mu = \|u - v\|^2_{L^2}.$$

Here we may work under the sole assumption that u belongs to the space $V = L^2(\Omega, \mu)$, because pointwise evaluations of u and w will be almost surely well-defined. In return, since $\tilde{u}$ is now stochastic, the L^2 estimation error will only be assessed in a probabilistic sense, for example by considering the mean-square error,

$$\mathbb{E}(\|u - \tilde{u}\|^2_{L^2}) = \mathbb{E}_{\otimes^m \sigma}(\|u - \tilde{u}\|^2_{L^2})$$

The weighted least-square approximation may be viewed as the orthogonal projection $\tilde{u} = P_n^m u$ onto V_n for the discrete ℓ^2 norm

$$\|v\|^2_m := \frac{1}{m} \sum_{i=1}^{m} w(x^i) \, |v(x^i)|^2, \tag{4.7}$$

in the same way that the optimal approximation

$$P_n u := \operatorname*{arg\,min}_{v \in V_n} \|u - v\|_{L^2}$$

is the orthogonal projection for the continuous $L^2(\Omega, \mu)$ norm. A helpful object for comparing these two norms on V_n is the Gramian matrix

$$G_m := (\langle \varphi_j, \varphi_k \rangle_m)_{1 \leqslant j,k \leqslant n}, \tag{4.8}$$

where $(\varphi_1, \dots, \varphi_n)$ is any $L^2(\Omega, \mu)$-orthonormal basis of V_n and $\langle \cdot, \cdot \rangle_m$ is the inner product associated with the discrete norm $\|\cdot\|_m$. Indeed, for all $\delta > 0$,

$$\|G_m - I\|_2 \leqslant \delta \iff (1 - \delta)\|v\|^2_{L^2} \leqslant \|v\|^2_m \leqslant (1 + \delta)\|v\|^2_{L^2}, \quad v \in V_n, \tag{4.9}$$

where $\|A\|_2$ denotes the spectral norm of an $n \times n$ matrix A. As noted in [55] in the case of standard least-squares, and in [59] for the weighted case, G_m can be seen as a mean of m independent and identically distributed

rank-one matrices

$$a_i a_i^\dagger := (w(x^i)\,\varphi_j(x^i)\,\overline{\varphi_k(x^i)})_{1\leqslant j,k\leqslant n}$$

satisfying $\mathbb{E}(a_i a_i^\dagger) = I$, so G_m concentrates towards the identity as m grows to infinity. This concentration can be estimated by a matrix Chernoff bound, such as Theorem 1.1 in the survey paper [176]. As observed in [59], for the particular value $\delta = \frac{1}{2}$, this inequality can be rewritten as follows, in our case of interest.

Lemma 4.1. *For any $\varepsilon > 0$, under the sampling budget condition*

$$m \geqslant \gamma\, \|w\, k_n\|_{L^\infty}\, n\, \ln(2n/\varepsilon), \tag{4.10}$$

where $\gamma := (3/2\,\ln(3/2) - 1/2)^{-1} \approx 9.242$, one has $\mathbb{P}(\|G_m - I\|_2 \leqslant 1/2) \geqslant 1 - \varepsilon$.

An estimate comparing the error $\|u - \tilde{u}\|_{L^2}$ with the optimum $e_n(u)_{L^2}$ can be obtained when imposing that $\|G_m - I\|_2 \leqslant 1/2$, as expressed in the following Lemma, which is proved in [59].

Lemma 4.2. *One has*

$$\mathbb{E}\left(\|u - \tilde{u}\|_{L^2}^2 \chi_{\|G_m - I\|_2 \leqslant 1/2}\right) \leqslant \left(1 + 4\frac{n}{m}\|w\, k_n\|_{L^\infty}\right)\, e_n(u)_{L^2}^2. \tag{4.11}$$

On the other hand, the estimator $\tilde{u}$ obtained by solving (4.5) is not reliable in the event where G_m becomes singular, which brings us to modify its definition in various ways:

1. If one is able to compute $\|G_m - I\|_2$, one may condition the estimator to the event $\|G_m - I\|_2 \leqslant \frac{1}{2}$ by defining
$$\tilde{u}^C := \tilde{u}\, \chi_{\|G_m - I\|_2 \leqslant 1/2}, \tag{4.12}$$
that is, we take $\tilde{u}^C = 0$ if $\|G_m - I\|_2 > \frac{1}{2}$.
2. If a uniform bound $\|u\|_{L^\infty(\Omega)} \leqslant \tau$ is known, one may introduce a truncated estimator
$$\tilde{u}^T := T_\tau \circ \tilde{u}, \tag{4.13}$$
where $T_\tau(y) := \min\{\tau, |y|\}\,\mathrm{sgn}(y)$.

The main results from [59], that we slightly reformulate below, show that these estimators are instance optimal in a probabilistic sense. Throughout the rest of the chapter, γ denotes the same constant as in Lemma 4.1.

Theorem 4.3. *Under the sampling budget condition*

$$m \geqslant \gamma\, \|w\, k_n\|_{L^\infty}\, n\, \ln(2n/\varepsilon), \tag{4.14}$$

the weighted least-squares estimator satisfies

$$\mathbb{E}\left(\|u - \tilde{u}\|_{L^2}^2 \chi_{\|G_m - I\|_2 \leqslant 1/2}\right) \leqslant (1 + \eta(m))\; e_n(u)_{L^2}^2. \tag{4.15}$$

The conditionned and truncated estimators satisfy the convergence bounds

$$\mathbb{E}\left(\|u - \tilde{u}^C\|_{L^2}^2\right) \leqslant (1 + \eta(m))\; e_n(u)_{L^2}^2 + \|u\|_{L^2}^2\, \varepsilon, \tag{4.16}$$

and

$$\mathbb{E}\left(\|u - \tilde{u}^T\|_{L^2}^2\right) \leqslant (1 + \eta(m))\; e_n(u)_{L^2}^2 + 4\,\tau^2\, \varepsilon, \tag{4.17}$$

where $\eta(m) = 4\frac{n}{m}\|w\, k_n\|_{L^\infty} \leqslant \frac{4}{\gamma \ln(2n/\varepsilon)} \to 0$, as $n \to \infty$ or $\varepsilon \to 0$.

Proof. The bound (4.15) follows directly from Lemma 4.2 and the assumption on m. In the event $\|G_m - I\|_2 > \frac{1}{2}$, of probability less than ε by Lemma 4.1, one can use the bounds

$$\|u - \tilde{u}^C\|_{L^2}^2 = \|u\|_{L^2}^2 \quad \text{and} \quad \|u - \tilde{u}^T\|_{L^2}^2 \leqslant 4\tau^2.$$

Otherwise, one has

$$\|u - \tilde{u}^C\|_{L^2}^2 \leqslant \|u - \tilde{u}\|_{L^2}^2 \quad \text{and} \quad \|u - \tilde{u}^T\|_{L^2}^2 \leqslant \|u - \tilde{u}\|_{L^2}^2.$$

This leads to (4.16) and (4.17). □

Remark 4.4. The above result shows that the estimators $\tilde{u}^C$ and $\tilde{u}^T$ achieve instance optimality in expectation up to additional error terms of order $\mathcal{O}(\varepsilon)$, accounting for the event $\{\|G_m - I\|_2 > 1/2\}$. Note that ε only influences the constraint on the sampling budget logarithmically. In particular, if $e_n(u)_{L^2}$ decreases like n^{-r} for some $r > 0$, these estimators are rate optimal by taking ε less than n^{-2r}, which thus affects the constraint on sampling budget by a factor $\mathcal{O}(\ln(n))$. Note however that for exponential rates of the form $\exp(-cn^\alpha)$ —that occur for example when approximating analytic functions by polynomials— imposing ε to be of this order results in a sampling budget m of sub-optimal order $n^{1+\alpha}$ up to a logarithmic factor.

Remark 4.5. One way to achieve instance optimality in expectation without an additional error term consists in redrawing the points $\{x^1, \dots, x^m\}$ until one observes that $\|G_m - I\|_2 \leqslant \frac{1}{2}$, as proposed in [85]. We denote by u^* the weighted least-square estimator corresponding to this conditionned draw. In other words u^* is the weighted least-square estimator $\tilde{u}$ conditionned to the event $\{\|G_m - I\|_2 \leqslant \frac{1}{2}\}$. Since, by Baye's rule,

$$\mathbb{P}\left(\|G_m - I\|_2 \leqslant \frac{1}{2}\right) \mathbb{E}\left(\|u - \tilde{u}\|_{L^2}^2 \;\middle|\; \|G_m - I\|_2 \leqslant \frac{1}{2}\right) = \mathbb{E}\left(\|u - \tilde{u}\|_{L^2}^2 \chi_{\|G_m - I\|_2 \leqslant \frac{1}{2}}\right),$$

we find that under the sampling budget (4.14), one has

$$\mathbb{E}(\|u - u^*\|_{L^2}^2) = \mathbb{E}\left(\|u - \tilde{u}\|_{L^2}^2 \;\middle|\; \|G_m - I\|_2 \leqslant \frac{1}{2}\right) \leqslant \frac{1}{1-\varepsilon} \mathbb{E}\left(\|u - \tilde{u}\|_{L^2}^2 \chi_{\|G_m - I\|_2 \leqslant \frac{1}{2}}\right),$$

and thus

$$\mathbb{E}(\|u - u^*\|_{L^2}^2) \leqslant \frac{1}{1-\varepsilon}(1 + \eta(m)) e_n(u)_{L^2}^2. \tag{4.18}$$

The sampling budget condition also ensures a probabilistic control on the number of required redraws, since the probability that the event $\{\|G_m - I\|_2 \leqslant \frac{1}{2}\}$ did not occur after k redraws is less than ε^k.

Now the natural objective is to find a weight function w that makes $\|w\, k_n\|_{L^\infty}$ small in order to minimize the sampling budget. Since

$$\|w\, k_n\|_{L^\infty} \geqslant \int_\Omega w\, k_n \, d\sigma = \int_\Omega k_n \, d\mu = 1,$$

with equality attained for the weight function

$$w^* := \frac{1}{k_n} = \left(\sum_{j=1}^n |\varphi_j|^2\right)^{-1},$$

this theorem shows that the choice of sampling measure

$$d\sigma^* = \frac{1}{w^*} \, d\mu = k_n \, d\mu$$

is optimal, in the sense that the above instance optimality results are achieved with a near-optimal sampling budget $m \sim n$ up to logarithmic factors.

As already explained in the introduction, when working on a general domain Ω, we face the difficulty that the orthonormal basis $(\varphi_1, \dots, \varphi_n)$ cannot be exactly computed, and therefore the optimal w^* and σ^* are out of reach. The next section discusses computable alternatives $\widetilde{w}$ and $\widetilde{\sigma}$ that still yield similar instance optimality results at near-optimal sampling budget.

4.3 Near-optimal sampling strategies on general domains

4.3.1 Two steps sampling strategies

The sampling and reconstruction strategies that we discuss proceed in two steps:

1. In an offline stage we search for an approximation to the Christoffel function k_n. For this purpose, we sample $y^1, \dots, y^M \in \Omega$ according to μ, use these sampling points to compute an orthonormal basis $(\widetilde{\varphi}_1, \dots, \widetilde{\varphi}_n)$ with respect to the induced discrete inner product. The approximation to the Christoffel function is then $\widetilde{k}_n = \frac{1}{n}\sum_{j=1}^n |\widetilde{\varphi}_j|^2$. As we explain further, one objective is to guarantee that $\widetilde{k}_n$ and k_n are pointwise equivalent. We define the sampling measure $\widetilde{\sigma}$ as proportional to $\widetilde{k}_n\,\mu$ and draw the points $x^1, \dots, x^m$ according to this measure.
2. In an online stage, we evaluate u at the sampling points x^i and construct an estimate $\tilde{u}$ by the weighted least-squares method.

In the offline stage M could be much larger than n, however it should be understood that the function u is only evaluated in the online stage at the m point x^i which will be seen to have optimal cardinality $m \sim n$ up to logarithmic factors.

The two main requirements in these approaches are the data of a (non-orthogonal) basis $(\phi_1, \dots, \phi_n)$ of V_n and the ability to sample according to measure μ. When $\Omega \subset \mathbb{R}^d$ is a general multivariate domain, one typical setting for this second assumption to be valid is the following:

— There is a set R containing Ω such that μ is the restriction of a measure μ_R which can easily be sampled.
— Membership of a point x to the set Ω can be efficiently tested, that is, χ_Ω is easily computed.

This includes for instance the uniform probability measure on domains described by general systems of algebraic inequalities (such as polyhedrons, ellipsoids..), by including such domains Ω in a rectangle $R = I_1 \times \cdots \times I_d$ on which sampling according to the uniform measure can be done componentwise. Then the y^i are produced by sampling according to μ_R and rejecting the samples that do not belong to Ω. The offline stage is described more precisely as follows.

Algorithm 1 Offline stage of a two-step sampling strategy

1: Draw a certain amount M of points $y^1, \dots, y^M$ independently according to μ
2: Define the inner product
$$\langle u, v\rangle_M := \frac{1}{M}\sum_{i=1}^M u(y^i)\,\overline{v(y^i)} \tag{4.19}$$
3: Construct from $(\phi_j)_{1\leqslant j\leqslant n}$ an orthonormal basis $(\widetilde{\varphi}_j)_{1\leqslant j\leqslant n}$ of V_n with respect to $\|\cdot\|_M$
4: Define the approximate inverse Christoffel function
$$\widetilde{k}_n(x) = \frac{1}{n}\sum_{j=1}^n |\widetilde{\varphi}_j(x)|^2$$
5: Define the normalization factor $Z = \int_\Omega \widetilde{k}_n\,d\mu$, and the sampling measure
$$d\widetilde{\sigma} := \frac{1}{Z}\,\widetilde{k}_n\,d\mu$$

Note that the factor Z is unknown to us but its value is not needed in typical sampling strategies, such as rejection sampling or MCMC. In contrast to k_n, the function $\widetilde{k}_n$ is stochastic since it depends on the drawing of the y^i. In the online stage, we sample $x^1, \dots, x^m$ independently according to $d\widetilde{\sigma}$. We then measure u at the points x^i, and define the estimator $\tilde{u} \in V_n$ as the solution to the weighted least-squares problem

$$\min_{v\in V_n} \frac{1}{m}\sum_{i=1}^m \widetilde{w}(x^i)\,|u(x^i) - v(x^i)|^2, \tag{4.20}$$

with $\widetilde{w} = Z/\widetilde{k}_n$. This least-squares problem can be solved explicitly by computing $\tilde{u} = P_n^m u$ as the orthogonal projection of u on V_n with respect to the inner product from (4.7)

$$\langle u, v\rangle_m := \frac{1}{m}\sum_{i=1}^m \widetilde{w}(x^i)\,u(x^i)\,\overline{v(x^i)}. \tag{4.21}$$

Remark 4.6. There are now two levels of stochasticity: the draw of the y^i and the subsequent draw of the x^i. We sometimes use the symbols $\mathbb{E}_y$ and $\mathbb{P}_y$ referring to the first draw, and $\mathbb{E}_x$ and $\mathbb{P}_x$ referring to the second draw given the first one, while $\mathbb{E}$ and $\mathbb{P}$ refer to both draws.

We keep the notations G_m and $\tilde{u}^T$ from (4.8) and (4.13). In the following section, we establish instance optimal convergence results under near optimal sample complexity m similar to (4.15) and (4.17) in Theorem 4.3. On the other hand, we do not consider the conditioned estimator $\tilde{u}^C$ any further since we do not have access to the matrix G_m, which would require the knowledge of the functions φ_j. The derivation of a computable estimator that satisfies a similar estimate as $\tilde{u}^C$ is an open question. We also discuss the required sample complexity M of the offline stage.

4.3.2 Convergence bounds and sample complexity

Our principle objective is to ensure the uniform framing

$$c_1\, k_n(x) \leqslant \widetilde{k}_n(x) \leqslant c_2\, k_n(x), \quad x \in \Omega, \tag{4.22}$$

for some known constants $0 < c_1 \leqslant c_2$. Our motivation is that instance optimal convergence bounds with near-optimal sampling budget hold under this framing, as expressed by the following result.

Theorem 4.7. *Assume that* (4.22) *holds for some* $0 < c_1 \leqslant c_2$. *Then, under the sampling budget condition*

$$m \geqslant \frac{c_2}{c_1}\, \gamma\, n\, \ln(2n/\varepsilon), \tag{4.23}$$

one has $\mathbb{P}_x\left(\|G_m - I\|_2 \geqslant \frac{1}{2}\right) \leqslant \varepsilon$. *In addition, one has the convergence bounds*

$$\mathbb{E}_x\left(\|u - \tilde{u}\|_{L^2}^2 \chi_{\|G_m - I\|_2 \leqslant \frac{1}{2}}\right) \leqslant (1 + \eta(m))\, e_n(u)_{L^2}^2, \tag{4.24}$$

and

$$\mathbb{E}_x(\|u - \tilde{u}^T\|_{L^2}^2) \leqslant (1 + \eta(m))\, e_n(u)_{L^2}^2 + 4\,\varepsilon\,\tau^2, \tag{4.25}$$

where $\eta(m) = 4\,\frac{c_2}{c_1}\,\frac{n}{m} \leqslant \frac{4}{\gamma \ln(2n/\varepsilon)}$.

Proof. It is an immediate application of the results from § 4.2.2. Indeed

$$\|\widetilde{w}\, k_n\|_{L^\infty} = \left\|\frac{k_n Z}{\widetilde{k}_n}\right\|_{L^\infty} = \left\|\frac{k_n}{\widetilde{k}_n}\right\|_{L^\infty} \int_\Omega \widetilde{k}_n\, d\mu \leqslant \left\|\frac{k_n}{\widetilde{k}_n}\right\|_{L^\infty} \left\|\frac{\widetilde{k}_n}{k_n}\right\|_{L^\infty} \int_\Omega k_n\, d\mu \leqslant \frac{c_2}{c_1}.$$

Therefore, the sampling condition (4.23) implies $m \geqslant \gamma\, \|\widetilde{w}\, k_n\|_{L^\infty}\, n\, \ln(2n/\varepsilon)$, and the results follow by direct application of Lemma 4.1 and Theorem 4.3. □

We now concentrate our attention on the offline procedure which should be tuned in order to ensure that (4.22) holds with high probability. For this purpose, we introduce the Gramian matrix

$$G_M := (\langle \varphi_j, \varphi_k \rangle_M)_{1 \leqslant j,k \leqslant n},$$

where $\langle \cdot, \cdot \rangle_M$ is the inner product defined by (4.19) that uses the intermediate samples y^i, which are i.i.d. according to μ. This matrix should not be confused with G_m defined in (4.8), that uses the inner product $\langle \cdot, \cdot \rangle_m$ based on the final samples x^i.

Lemma 4.8. *For any pair of constants* $0 < c_1 \leqslant c_2$, *the matrix framing property*

$$c_2^{-1}\, I \leqslant G_M \leqslant c_1^{-1}\, I, \tag{4.26}$$

implies the uniform framing (4.22).

Proof. We use the fact that, similar to k_n, the function $\widetilde{k}_n$ is characterized by the extremality property

$$\widetilde{k}_n(x) = \frac{1}{n} \max_{v \in V_n} \frac{|v(x)|^2}{\|v\|_M^2}.$$

For any $x \in \Omega$ and $v \in V_n$, one has on the one hand

$$|v(x)|^2 \leqslant n\, \widetilde{k}_n(x)\, \|v\|_M^2 \leqslant \frac{n}{c_1}\, \widetilde{k}_n(x)\, \|v\|_{L^2}^2,$$

where the last inequality results from the upper one in (4.26). This shows that $c_1\, k_n(x) \leqslant \widetilde{k}_n(x)$. On the other hand, using the lower inequality in (4.26), we find that

$$|v(x)|^2 \leqslant n\, k_n(x)\, \|v\|_{L^2}^2 \leqslant c_2\, n\, k_n(x)\, \|v\|_M^2,$$

which shows that $\widetilde{k}_n(x) \leqslant c_2\, k_n(x)$. □

Remark 4.9. The matrix framing (4.26) implies the uniform framing (4.22) but the converse does not seem to hold. Finding an algebraic condition equivalent to (4.22) is an open question.

Lemma 4.1 indicates that if the amount of offline samples satisfies the condition

$$M \geqslant \gamma\, K_n\, n \ln(2n/\varepsilon), \quad K_n := \|k_n\|_{L^\infty(\Omega)}, \tag{4.27}$$

then, we are ensured that

$$\mathbb{P}_y\left(\|G_M - I\|_2 \geqslant 1/2\right) \leqslant \varepsilon,$$

and therefore the framing (4.26) holds with probability greater than $1 - \varepsilon$, for the particular values $c_1 = 2/3$ and $c_2 = 2$. Bearing in mind that k_n is unknown to us, we assume at least that we know an upper estimate for its L^∞ norm

$$K_n \leqslant B(n).$$

Explicit values for $B(n)$ for general domains Ω are established in Section 4.5 for spaces V_n of algebraic polynomials. Therefore, given such a bound, taking M such that

$$M \geqslant \gamma\, B(n) \ln(2n/\varepsilon), \tag{4.28}$$

guarantees a similar framing with probability greater than $1 - \varepsilon$. We obtain the following result as a direct consequence of Theorem 4.7.

Corollary 4.10. *Assume that the amount of sample points M used in the offline stage described by Algorithm 1 satisfies* (4.28) *for some given $\varepsilon > 0$. Then, under the sampling budget condition*

$$m \geqslant 3\, \gamma\, n \ln(2n/\varepsilon), \tag{4.29}$$

for the online stage, the event $E := \{\|G_m - I\|_2 \leqslant \frac{1}{2}$ and $\|G_M - I\|_2 \leqslant \frac{1}{2}\}$ satisfies $\mathbb{P}(E^c) \leqslant 2\varepsilon$. In addition, one has the convergence bounds

$$\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2 \chi_E) \leqslant (1 + \eta(m))\, e_n(u)_{L^2}^2, \tag{4.30}$$

and

$$\mathbb{E}(\|u - \tilde{u}^T\|_{L^2}^2) \leqslant (1 + \eta(m))\, e_n(u)_{L^2}^2 + 8\, \varepsilon\, \tau^2, \tag{4.31}$$

where $\eta(m) = 12\, \frac{n}{m} \leqslant \frac{4}{\gamma \ln(2n/\varepsilon)}$.

Proof. The estimate on $\mathbb{P}(E^c)$ follows from a union bound. Since $\|G_M - I\|_2 \geqslant \frac{1}{2}$ ensures the framing (4.22) with $c_1 = 2/3$ and $c_2 = 2$, the bound (4.30) follows from (4.24) in Theorem 4.7. Finally, the bound (4.31) follows from (4.30) and the probability estimate on E^c by the same argument as in the proof of Theorem 4.3. □

4.3.3 An empirical determination of the value of M

In many situations, the best available bound $B(n)$ on K_n could be overestimated by a large amount. Moreover, the theoretical requirement $M \geqslant \gamma\, K_n\, n \ln(2n/\varepsilon)$ is only a sufficient condition that guarantees that $\|G_M - I\|_2 \leqslant \frac{1}{2}$ with probability larger than $1-\varepsilon$. It could happen that for smaller values of M, the matrix G_M satisfies the framing (4.26) with constants c_1 and c_2 that have moderate ratio c_2/c_1.

Since the computational cost of the offline stage is proportional to M, it would desirable to use such a smaller value of M. If we could compute the matrix G_M, it would suffice to increase M until the condition number

$$\kappa(G_M) = \frac{\lambda_{\max}(G_M)}{\lambda_{\min}(G_M)},$$

has value smaller than a prescribed threshold $c^* > 1$, so that (4.26) holds with $c_2/c_1 = \kappa(G_M) \leqslant c^*$.

However, since the exact orthonormal basis elements φ_j are generally unknown to us, we cannot compute the matrix G_M. As an alternate strategy, we propose the following method that provides an empirical determination of the value M that should be used in Algorithm 1: start from the minimal value $M = n$, and draw points $z_1, \dots, z^M$ and $y^1, \dots y^M$ independently according to μ. Then, defining

$$\langle u, v\rangle_y = \frac{1}{M}\sum_{i=1} u(y^i)\,\overline{v(y^i)} \quad \text{and} \quad \langle u, v\rangle_z = \frac{1}{M}\sum_{i=1} u(z^i)\,\overline{v(z^i)},$$

compute an orthonormal basis (φ_j^y) with respect to $\langle\cdot,\cdot\rangle_y$, and define the test matrix

$$T := (\langle \varphi_j^y, \varphi_k^y\rangle_z)_{1\leqslant j,k\leqslant n}.$$

If $\kappa(T) \geqslant c^*$, increase the value of M by some fixed amount, and repeat this step until $\kappa(T) \leqslant c^*$. For this empirically found value $M = M_{\mathrm{emp}}(n)$, use the points $\{y^1, \dots, y^M\}$ in the offline stage described by Algorithm 1, and the ratio $c_2/c_1 = c^*$ in the sampling budget condition (4.23) used in the online stage.

The rationale for this approach is that if G_M is well conditioned with high probability, then T should also be, as shown in the following result.

Proposition 4.11. *If M is chosen in such a way that $\mathbb{P}(\kappa(G_M) \geqslant c) \leqslant \varepsilon$ for some $c > 1$, then*

$$\mathbb{P}(\kappa(T) \geqslant c^2) \leqslant 2\,\varepsilon.$$

Proof. Since both matrices $G_y = (\langle\varphi_j, \varphi_k\rangle_y)_{1\leqslant j,k\leqslant n}$ and $G_z = (\langle\varphi_j, \varphi_k\rangle_z)_{1\leqslant j,k\leqslant n}$ are realizations of G_M, we obtain by a union bound that, with probability at least $1-2\varepsilon$, both G_y and G_z have condition numbers less than c. Under this event,

$$\begin{aligned}
\lambda_{\max}(T) &= \sup_{\alpha\in\mathbb{R}^n} \frac{\|\sum_{j=1}^n \alpha_j \varphi_j^y\|_z^2}{|\alpha|^2} \leqslant \sup_{\alpha\in\mathbb{R}^n} \frac{\|\sum_{j=1}^n \alpha_j \varphi_j^y\|_{L^2}^2}{|\alpha|^2} \sup_{v\in V_n} \frac{\|v\|_z^2}{\|v\|_{L^2}^2} \\
&= \sup_{v\in V_n} \frac{\|v\|_{L^2}^2}{\|v\|_y^2} \sup_{v\in V_n} \frac{\|v\|_z^2}{\|v\|_{L^2}^2} = \left(\inf_{v\in V_n} \frac{\|v\|_y^2}{\|v\|_{L^2}^2}\right)^{-1} \sup_{v\in V_n} \frac{\|v\|_z^2}{\|v\|_{L^2}^2} = \frac{\lambda_{\max}(G_z)}{\lambda_{\min}(G_y)},
\end{aligned}$$

and

$$\begin{aligned}
\lambda_{\min}(T) &= \inf_{\alpha\in\mathbb{R}^n} \frac{\|\sum_{j=1}^n \alpha_j \varphi_j^y\|_z^2}{|\alpha|^2} \geqslant \inf_{\alpha\in\mathbb{R}^n} \frac{\|\sum_{j=1}^n \alpha_j \varphi_j^y\|_{L^2}^2}{|\alpha|^2} \inf_{v\in V_n} \frac{\|v\|_z^2}{\|v\|_{L^2}^2} \\
&= \inf_{v\in V_n} \frac{\|v\|_{L^2}^2}{\|v\|_y^2} \inf_{v\in V_n} \frac{\|v\|_z^2}{\|v\|_{L^2}^2} = \left(\sup_{v\in V_n} \frac{\|v\|_y^2}{\|v\|_{L^2}^2}\right)^{-1} \inf_{v\in V_n} \frac{\|v\|_z^2}{\|v\|_{L^2}^2} = \frac{\lambda_{\min}(G_z)}{\lambda_{\max}(G_y)},
\end{aligned}$$

which implies that $\kappa(T) \leqslant \kappa(G_y)\,\kappa(G_z) \leqslant c^2$. □

The above proposition shows that a good conditioning of G_M with high probability implies the same property for T. There is of course no theoretical guarantee that the value of M provided by the above empirical approach

is sufficient to achieve good conditioning of G_M, unless the resulting M satisfies (4.28). However, in the numerical experiments of Section 4.6, we will check that the values of M for which $\kappa(T) \leqslant c$ holds also ensure a similar bound for $\kappa(G_M)$.

4.4 Multilevel strategies

The sampling strategy described by Algorithm 1 provides instance optimal reconstructions of u with an optimal sampling budget up to a multiplicative factor $\ln(2n/\varepsilon)$. Thus, the execution time of the online stage, dominated by the m evaluations of u at points x^i, cannot be significantly improved. On the other hand, the complexity of the offline stage is dominated by the computation of the Gramian matrix in order to derive the basis $(\widetilde{\varphi}_1, \ldots, \widetilde{\varphi}_n)$, and is therefore of order $\mathcal{O}(Mn^2)$. In particular, it depends linearly on the number of points M, which could be very large if K_n grows fast with n, or if its available bound $B(n)$ is over-estimated.

In this section we discuss a multilevel approach aiming at improving this offline computational cost: we produce an approximation to k_n in several iterations, by successive refinements of this function as the dimension of V_n increases. We consider a family of nested spaces $(V_{n_p})_{p \geqslant 1}$ of increasing dimension n_p and take an orthonormal basis $(\varphi_j)_{j \geqslant 1}$ adapted to this hierarchy, in the sense that

$$V_{n_p} = \operatorname{span}\{\varphi_1, \ldots, \varphi_{n_p}\}, \quad n_p \geqslant 1.$$

As previously, the exact functions k_{n_p} are out of reach, since we do not have access to the continuous inner product by which we would compute the basis $(\varphi_j)_{1 \leqslant j \leqslant n_p}$. The offline stage described in Section 4.3 computes approximations $\widetilde{\varphi}_j$ by orthogonalizing with respect to a discrete inner product with points y^i drawn according to $d\mu$. We know that a more efficient sample for performing this orthogonalization could be obtained by drawing according to $d\sigma = \frac{k_{n_p}}{n_p}\, d\mu$, which is however unknown to us.

The idea for breaking this dependency loop is to replace k_{n_p} with $\widetilde{k}_{n_{p-1}}$, which was computed at the previous step. Our analysis of this strategy is based on the following assumption of proximity between $k_{n_{p-1}}$ and k_{n_p}: there exists a known constant $\kappa > 1$ such that

$$k_{n_1}(x) \leqslant 3\,\kappa \text{ and } n_p\, k_{n_p}(x) \leqslant n_{p+1}\, k_{n_{p+1}}(x) \leqslant \kappa\, n_p\, k_{n_p}(x), \quad p \geqslant 1, \quad x \in \Omega. \tag{4.32}$$

The validity of this assumption can be studied through lower and upper estimates for k_n, such as those discussed in the next section. For example, Theorem 4.27 allows one to establish (4.32) for bivariate polynomial spaces of total degree p, therefore with $n_p = (p+1)(p+2)/2$, on bidimensional domains with piecewise smooth boundary. Note that (4.32) allows up to exponential growth of K_n, if we simply take $n_p = p$.

Assuming that the targeted space V_n is a member of this hierarchy, that is,

$$n = n_q, \quad \text{for some} \quad q > 1,$$

we modify the offline stage as indicated in Algorithm 2 below.

The online stage remains unchanged: the samples $x^1, \ldots, x^m$ for evaluation of u are drawn i.i.d. according to $\widetilde{\sigma}$, and we solve the weighed least-squares problem (4.20). The sample size M of the offline stage is now replaced by $\overline{M} = M_1 + \cdots + M_q$. Denote by $G_p := (\langle \varphi_j, \varphi_k \rangle_p)_{1 \leqslant j,k \leqslant n_p}$ the Gramian matrices for the inner products (4.33). The following result shows that the conditions imposed on the M_p are less stringent than those that were imposed on M.

Theorem 4.12. *Let $\varepsilon_p > 0$ such that $\varepsilon := \sum_{p=1}^{q} \varepsilon_p < 1$, and assume that the amount of offline samples used in Algorithm 2 satisfies*

$$M_p \geqslant 3\,\kappa\,\gamma\, n_p\, \ln \frac{2n_p}{\varepsilon_p}, \quad p = 1, \ldots, q,$$

with κ the constant in the assumption (4.32)*. Then if $m \geqslant 3\,\gamma\, n\, \ln \frac{2n}{\varepsilon}$, the same convergence bounds* (4.30) *and* (4.31) *as in Corollary 4.10 hold, with $E := \{\|G_m - I\|_2 \leqslant \frac{1}{2} \text{ and } \|G_q - I\|_2 \leqslant \frac{1}{2}\}$ that satisfies $\mathbb{P}(E^c) \leqslant 2\varepsilon$.*

Algorithm 2 Offline stage of a multi-step sampling strategy

1: Start with $\widetilde{w}_0 = 1$ and $\widetilde{\sigma}_0 = \mu$

2: **for** $p = 1$ **to** q **do**

3: Draw a certain amount M_p of points $y^{p,1}, \dots, y^{p,M_p}$ independently according to $\widetilde{\sigma}_{p-1}$

4: Define the inner product

$$\langle u, v \rangle_p := \frac{1}{M_p} \sum_{i=1}^{M_p} \widetilde{w}_{p-1}(y^{p,i})\, u(y^{p,i})\, \overline{v(y^{p,i})} \tag{4.33}$$

5: Construct from $(\phi_j)_{1 \leqslant j \leqslant n_p}$ an orthonormal basis $(\widetilde{\varphi}_j^p)_{1 \leqslant j \leqslant n_p}$ of V_{n_p} with respect to $\| \cdot \|_{M_p}$

6: Define the approximate inverse Christoffel function, weight and density

$$\widetilde{k}_{n_p} = \frac{1}{n_p} \sum_{j=1}^{n_p} |\widetilde{\varphi}_j^p|^2, \quad Z_p = \int_\Omega \widetilde{k}_{n_p} d\mu, \quad \widetilde{w}_p = \frac{Z_p}{\widetilde{k}_{n_p}} \quad \text{and} \quad d\widetilde{\sigma}_p := \frac{1}{Z_p}\, \widetilde{k}_{n_p}\, d\mu$$

7: **end for**

8: Define the final Christoffel density for V_n as $\widetilde{k}_n = \widetilde{k}_{n_q}$, as well as $\widetilde{w} = \widetilde{w}_q$ and $\widetilde{\sigma} = \widetilde{\sigma}_q$

Proof. We show by induction on p that the event

$$B_p := \left\{ \|G_1 - I\|_2 \leqslant \frac{1}{2}, \dots, \|G_p - I\|_2 \leqslant \frac{1}{2} \right\}$$

occurs with probability at least $1 - \varepsilon_1 - \cdots - \varepsilon_p$. As

$$M_1 \geqslant 3\,\kappa\,\gamma\,n_1\,\ln\frac{2n_1}{\varepsilon_1} \geqslant \gamma\,\|\widetilde{w}_0\,k_{n_1}\|_{L^\infty}\,n_1\,\ln\frac{2n_1}{\varepsilon_1},$$

by Lemma 4.1,

$$\mathbb{P}(B_1) \geqslant 1 - \varepsilon_1.$$

For $1 \leqslant p < q$, under the event B_p, Lemma 4.8 gives

$$\frac{2}{3}\,k_{n_p}(x) \leqslant \widetilde{k}_{n_p}(x) \leqslant 2\,k_{n_p}(x), \quad x \in \Omega.$$

Therefore, using assumption (4.32), we find that

$$\left\|\widetilde{w}_p\,k_{n_{p+1}}Z_p\right\|_{L^\infty} = n_{p+1}Z_p \left\| \frac{k_{n_{p+1}}}{\widetilde{k}_{n_p}} \right\|_{L^\infty} \leqslant n_{p+1} \left\| \frac{\widetilde{k}_{n_p}}{k_{n_p}} \right\|_{L^\infty} \left\| \frac{k_{n_p}}{\widetilde{k}_{n_p}} \right\|_{L^\infty} \left\| \frac{k_{n_{p+1}}}{k_{n_p}} \right\|_{L^\infty} \leqslant 3\,\kappa\,n_p.$$

As $M_{p+1} \geqslant 3\,\kappa\,\gamma\,n_p\,\ln\frac{2\,n_p}{\varepsilon_{p+1}}$, Lemma 4.1 applies, and combining this with the induction hypothesis:

$$\begin{aligned}\mathbb{P}(B_{p+1}) &= \mathbb{P}(B_p)\,\mathbb{P}\left(\|G_{p+1} - I\|_2 \leqslant \frac{1}{2} \,\middle|\, B_p \right) \\ &\geqslant (1 - \varepsilon_1 - \cdots - \varepsilon_p)(1 - \varepsilon_{p+1}) \geqslant 1 - \varepsilon_1 - \cdots - \varepsilon_p - \varepsilon_{p+1}.\end{aligned}$$

Use Lemma 4.8 one last time to write, in the event B_q,

$$\frac{2}{3}\,k_{n_q}(x) \leqslant \widetilde{k}_{n_q}(x) \leqslant 2\,k_{n_q}(x), \quad x \in \Omega,$$

which is the framing (4.26) for the particular values $c_1 = 2/3$ and $c_2 = 2$. Since B_q has probability larger than $1 - \varepsilon$, we conclude by the exact same arguments used in the proof Corollary 4.10. □

We now comment on the gain of complexity by using Algorithm 2 in two different situations:

1. Exponential growth of K_n: the property (4.32) might be satisfied even when K_n grows exponentially with n, by taking the choice $n_p = p$. Then, the complexity of Algorithm 1 is of order

$$\mathcal{O}(M\,n^2) \gtrsim \mathcal{O}(K_n\,n^3\ln(2n/\varepsilon)),$$

which grows exponentially in n. In contrast, the total amount of sampling in Algorithm 2 is

$$\overline{M} = M_1 + \cdots + M_n \leqslant n\,M_n = \mathcal{O}(n^2\ln(2n/\varepsilon)),$$

so the first stage remains of polynomial complexity $\mathcal{O}(\overline{M}n^2) = \mathcal{O}(n^4\ln(2n/\varepsilon))$.

2. Algebraic growth of K_n: if $K_n \sim n^r$ only grows algebraically in n, one may choose $n_p = 2^p$, in which case the total number of sample points $\overline{M}$ rewrites as $M_{n_0} + \cdots + M_{n_q} \sim M_{n_q}$, giving an optimal complexity $\mathcal{O}(n^3\ln(2n/\varepsilon))$ for the first stage. This is smaller than the complexity

$$\mathcal{O}(K_n\,n^3\ln(2n/\varepsilon)) = \mathcal{O}(n^{3+r}\ln(2n/\varepsilon))$$

encountered in Algorithm 1.

While Algorithm 2 presents a computational gain for providing a near-optimal measure $\widetilde{\sigma}$, the resulting sample $x^1,\dots,x^m$ is specifically targeted at approximating u in the space $V_n = V_{n_q}$. As explained in Section 4.1.2, it is sometimes desirable to obtain optimal weighted least-squares approximations $\tilde{u}_{n_p}$ for each space V_{n_p} while maintaining the cumulated number of evaluations of u until step p of the optimal order n_p up to logarithmic factors. Therefore, we would like to recycle the evaluation points $x^1,\dots,x^{m_{p-1}}$ drawn until step $p-1$ when assembling the new sample $x^1,\dots,x^{m_p}$, for some well chosen sequence $(m_p)_{p\geqslant 1}$ growing as $(n_p)_{p\geqslant 1}$ up to logarithmic factors.

Here, we assume that the family $(V_{n_p})_{p\geqslant 1}$ has been fixed, independently of the target function u, in contrast to being generated in an adaptive manner. Adaptive space generation brings out new difficulties: the space V_{n_p} depends on the approximation computed in the previous space $V_{n_{p-1}}$ and therefore the new evaluation points $x^{m_{p-1}+1},\dots,x^{m_p}$ will not be independant from the previous ones. Maintaining optimal sample complexity in the adaptive context is, to our knowledge, an open problem.

Intuitively, since the sample should have a density proportional to k_{n_p}, most of the new points we draw at step p should be distributed according a density proportional to $n_p\,k_{n_p} - n_{p-1}\,k_{n_{p-1}} = \sum_{j=n_{p-1}+1}^{n_p}|\varphi_j|^2$. This leads us to the following algorithm.

Algorithm 3 Offline stage of a multi-step adaptive sampling strategy

1: Start with $\widetilde{w}_0 = 1$, $\widetilde{\sigma}_0 = \mu$ and $m_0 = 0$

2: **for** $p = 1, 2, \dots$ **do**

3: Generate $y^{n_p,i}$ and compute $\widetilde{w}_{n_p}$, $\widetilde{\sigma}_{n_p}$ and $\widetilde{k}_{n_p}$ as in Algorithm 2

4: Create the orthonormal basis $(\varphi_j^{n_p})$, compatibly with the inclusion $V_{n_{p-1}} \subset V_{n_p}$, in the sense that

$$\operatorname{span}\{\varphi_1^{n_p},\dots,\varphi_{n_{p-1}}^{n_p}\} = V_{n_{p-1}}.$$

5: Having already defined $x^1,\dots,x^{m_{p-1}}$, draw the new evaluation points $x^{m_{p-1}+1},\dots,x^{m_p}$ according to

$$d\rho_p := \frac{1}{Z'_p}\frac{1}{m_p - m_{p-1}}\left(\frac{m_p}{n_p}\sum_{j=1}^{n_p}|\varphi_j^{n_p}|^2 - \frac{m_{p-1}}{n_{p-1}}\sum_{j=1}^{n_{p-1}}|\varphi_j^{n_p}|^2\right)d\mu, \tag{4.34}$$

6: **end for**

Remark 4.13. Note that the non-negativity of ρ_p is only guaranteed when $(m_p/n_p)_{p\geqslant 1}$ is non-decreasing, a condition which is easily met since m_p has to grow as $n_p\ln n_p$. If we had taken m_p exactly linear with respect to the dimension n_p, the terms with $j \leqslant n_{p-1}$ in the expression (4.34) would cancel, hence $d\rho_p$ would only be an approximation of the probability density $\frac{1}{n_p - n_{p-1}}\sum_{j=n_{p-1}+1}^{n_p}|\varphi_j|^2\,d\mu$.

Remark 4.14. Another approach to hierarchical sampling was proposed in [132] and consists in drawing constant proportions of samples according to the measure $|\varphi_j|^2 d\mu$. It was adapted in [5] to our setting of interest where the φ_j cannot be exactly computed.

Remark 4.15. In the above algorithm, the various subsets $\{x^{m_{k-1}+1}, \dots, x^{m_k}\}$ of $\{x^1, \dots, x^{m_p}\}$, for k between 1 and p, are drawn according to different probability measures. The sample $\{x^1, \dots, x^{m_p}\}$ is thus not i.i.d. anymore, which affects the proof of the convergence theorem given below. Instead it may be thought as a deterministic mixture of collections of i.i.d. samples, as in Theorem 2 of [132].

At any iteration q, we use the evaluations of u at all points $x^1, \dots, x^m$ as follows to compute a least-squares approximation $\tilde{u}_n \in V_n$, where $n := n_q$ and $m := m_q$. We denote by w the weight function defined by

$$w(x) \sum_{p=1}^{q} (m_p - m_{p-1})\, d\rho_p = m\, d\mu, \tag{4.35}$$

and solve the weighted least square problem (4.5). The following result shows that instance optimality is maintained at every step q, with a cumulated sampling budget m_q that is near-optimal.

Theorem 4.16. *Take numbers $\delta_p, \varepsilon_p \in (0,1)$ such that $\varepsilon := \sum_{p=1}^q \varepsilon_p < 1$ and $\delta := \sum_{p=1}^q \delta_p < 1/2$, and define $c_\delta = ((1+\delta)\ln(1+\delta) - \delta)^{-1}$. Assume that, for all $p \geqslant 1$,*

$$M_{n_p} \geqslant 2\,\kappa\, c_{\delta_p}\, n_p \ln \frac{2n_p}{\varepsilon_p} \quad \text{and} \quad m_p \geqslant \frac{\gamma}{1-2\delta}\, n_p \ln \frac{2n_p}{\varepsilon},$$

with κ the constant in the assumption (4.32), and that m_p/n_p is a non-decreasing function of p. Then, with $n := n_q$ and $m := m_q$, the convergence bounds (4.30) and (4.31) simultaneously hold for all $q \geqslant 1$, with

$$\eta(m) = \frac{4}{(1-2\delta)} \frac{n}{m} \leqslant \frac{4}{\gamma \ln(2n/\varepsilon)}$$

and $E := \{\|G_m - I\|_2 \leqslant \frac{1}{2} \text{ and } \|G_p - I\|_2 \leqslant \delta_p \text{ for } p \geqslant 1\}$, which satisfies $\mathbb{P}(E^c) \leqslant 2\varepsilon$.

The proof of this theorem requires a refinement of Lemma 4.1, due to the fact that the x^i are not anymore identically distributed. This uses the following tail bound, directly obtained from the matrix Chernoff bound in [176].

Proposition 4.17. *Consider a finite sequence $(a_i a_i^\dagger)_{1 \leqslant i \leqslant m}$ of independent, random, rank-one self-adjoint matrices with dimension n. Assume that each matrix satisfies $0 \preccurlyeq a_i a_i^\dagger \preccurlyeq N\, I$ almost surely, and that $\frac{1}{m}\sum_{i=1}^m \mathbb{E}(a_i a_i^\dagger) = I$. Then for all $\delta \in (0,1)$,*

$$\mathbb{P}\left(\left\| \frac{1}{m} \sum_{i=1}^m a_i a_i^\dagger - I \right\|_2 > \delta \right) \leqslant 2\, n \exp\left(-\frac{m}{c_\delta\, N} \right),$$

where $c_\delta = ((1+\delta)\ln(1+\delta) - \delta)^{-1}$ as in Theorem 4.16.

Proof of Theorem 4.16. By the same argument as in Theorem 4.12, we find that the event

$$B = \{\|G_p - I\|_2 \leqslant \delta_p \ : \ p \geqslant 1\}$$

has probability larger than $1 - \varepsilon$, where the G_p are as in the proof of Theorem 4.12.

We then fix a value of q and for $n = n_q$ and $m = m_q$, we study the Gramian matrix G_m which is the sum of the independent, but not identically distributed, matrices

$$a_i a_i^\dagger := w(x^i)\, (\varphi_j(x^i)\overline{\varphi_k(x^i)})_{1 \leqslant j,k \leqslant n}, \quad i = 1, \dots, m.$$

Then, with the notation $H(x) = (\varphi_j(x)\overline{\varphi_k(x)})_{1\leqslant j,k\leqslant n}$,

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(a_i a_i^\dagger) = \sum_{p=1}^{q}(m_p - m_{p-1})\int_\Omega \frac{1}{m}\,w(x)\,H(x)\,d\rho_p(x) = \int_\Omega H(x)\,d\mu(x) = I,$$

and

$$\|a_i a_i^\dagger\|_2 = w(x^i)\sum_{j=1}^{n}|\varphi_j(x^i)|^2\|w\,k_n\|_{L^\infty} n =: N.$$

One also has, under the event B, $\int_\Omega |\varphi_j^{n_p}|^2\,d\mu \leqslant \frac{1}{1-\delta_p}$ for $j = 1,\dots,n_p$ so $Z_p' \geqslant 1-\delta_p$, and consequently

$$\begin{aligned}
\frac{m}{w} &= \sum_{p=1}^{q}(m_p - m_{p-1})\,\frac{d\rho_p}{d\mu}\\
&= \sum_{p=1}^{q}\frac{1}{Z_p'}\left(\frac{m_p}{n_p}\sum_{j=1}^{n_p}|\varphi_j^{n_p}|^2 - \frac{m_{p-1}}{n_{p-1}}\sum_{j=1}^{n_{p-1}}|\varphi_j^{n_p}|^2\right)\\
&\geqslant \sum_{p=1}^{q}\left(m_p\frac{1-\delta_p}{1+\delta_p}\,k_{n_p} - m_{p-1}\,k_{n_{p-1}}\right)\\
&\geqslant m\,k_n - \sum_{p=1}^{q} m_p\,\frac{2\delta_p}{1+\delta_p}\,k_{n_p}\\
&\geqslant (1-2\delta)\,m\,k_n,
\end{aligned}$$

so $N = \|w\,k_n\|_{L^\infty} n \leqslant n/(1-2\delta)$. Applying Proposition 4.17, we find that

$$\mathbb{P}_x\left(\|G_m - I\|_2 > \frac{1}{2}\;\middle|\; B\right) \leqslant 2n\,\exp\left(-\frac{1}{\gamma\,N}\right) \leqslant 2n\,\exp\left(-\frac{1-2\delta}{\gamma n}\right) \leqslant \varepsilon.$$

Therefore, since $E := B \cap \left\{\|G_m - I\|_2 \leqslant \frac{1}{2}\right\}$, we find that $\mathbb{P}(E) \geqslant 1-2\varepsilon$.

In order to prove the convergence bounds (4.30) and (4.31), we cannot proceed as in Corollary 4.10 by simply invoking Theorem 4.7, because the x^i are not identically distributed. This leads us to modify the statement of Lemma 4.2 and its proof given in [59], following a strategy proposed in [132]. First, using similar arguments as in [59], we find that

$$\mathbb{E}(\|u-\tilde{u}\|_{L^2}^2\,\chi_E) \leqslant e_n(u)_{L^2}^2 + 4\,\mathbb{E}\left(\sum_{k=1}^{n}|\langle\varphi_k, g\rangle_m|^2\,\chi_E\right), \tag{4.36}$$

where $g = u - P_n u$ is the projection error. For each $k = 1,\dots,n$, we define $g_k := w\,\varphi_k\,\overline{g}$ and write

$$\begin{aligned}
\mathbb{E}\left(|\langle\varphi_k, g\rangle_m|^2\,\chi_E\right) &\leqslant \mathbb{E}\left(|\langle\varphi_k, g\rangle_m|^2\,\chi_B\right)\\
&= \frac{1}{m^2}\sum_{1\leqslant i,j\leqslant m}\mathbb{E}\left(g_k(x^i)\,\overline{g_k(x^j)}\,\chi_B\right)\\
&= \frac{1}{m^2}\,\mathbb{E}_y\left(\chi_B\left(\sum_{1\leqslant i\leqslant m}\mathbb{E}_x\left(|g_k(x^i)|^2\right) + \sum_{i\neq j}\mathbb{E}_x\left(g_k(x^i)\,\overline{g_k(x^j)}\right)\right)\right)\\
&\leqslant \frac{1}{m^2}\,\mathbb{E}_y\left(\chi_B\left(\sum_{1\leqslant i\leqslant m}\mathbb{E}_x\left(|g_k(x^i)|^2\right) + \left|\sum_{1\leqslant i\leqslant m}\mathbb{E}_x\left(g_k(x^i)\right)\right|^2\right)\right)\\
&= \mathbb{E}_y\left(\chi_B\left(\frac{1}{m}\,\mathbb{E}_t\left(|g_k(t)|^2\right) + \left|\,\mathbb{E}_t\left(g_k(t)\right)\right|^2\right)\right),
\end{aligned}$$

where t is a random variable distributed according to $\sum_{p=1}^{q} \frac{m_p - m_{p-1}}{m} \, d\rho_p = \frac{1}{w} \, d\mu$. We then note that

$$\mathbb{E}_t(g_k(t)) = \int_\Omega \varphi_k \, \overline{g} \, d\mu = 0$$

since $g \in V_n^\perp$, and that $\sum_{k=1}^{n} |g_k(t)|^2 = w(t)^2 \, g(t)^2 \, k_n(t)$. Therefore

$$\mathbb{E}\left(\sum_{k=1}^{n} |\langle \varphi_k, g \rangle_m|^2 \, \chi_E \right) \leqslant \mathbb{E}_y \left(\chi_B \, \frac{1}{m} \int_\Omega w \, k_n \, g^2 d\mu \right) \leqslant \mathbb{E}_y \left(\chi_B \, N \, \|g\|_{L^2}^2 \right) \leqslant \frac{n}{(1-2\delta)} \, e_n(u)_{L^2}^2.$$

Combining this with (4.36), we finally obtain

$$\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2 \, \chi_E) \leqslant \left(1 + \frac{4}{(1-2\delta)} \, \frac{n}{m} \right) e_n(u)_{L^2}^2$$

□

Remark 4.18. If a stopping time q is known in advance, the simplest choice is to take $\varepsilon_p = \varepsilon/q$ and $\delta_p = \delta/q$. If the stopping time q is not known in advance, we can take for instance $\varepsilon_p = \frac{6}{\pi^2} \, \frac{\varepsilon}{p^2}$ and $\delta_p = \frac{6}{\pi^2} \, \frac{\delta}{p^2}$. As $c_\delta \sim \frac{2}{\delta^2}$ when $\delta \to 0$, this choice only increases the number M_p of sample points y^i by a factor p^4, which is satisfying in view the previous remarks.

4.5 Estimates on the inverse Christoffel function

We have seen that the success of Algorithm 1 is based on the offline sampling condition (4.28), which means that a uniform upper bound $B(n)$ on the inverse Christoffel function k_n is needed in the first place. Likewise, the multilevel Algorithms 2 and 3 from Section 4.4 are based on the assumption (4.32), whose verification requires pointwise upper and lower estimates on $k_n(x)$. In this section we establish such bounds on general domains, when the V_n are spaces of algebraic multivariate polynomials of varying total degree. Throughout this section, we assume that

$$\mu = \mu_\Omega = |\Omega|^{-1} \, \chi_\Omega \, dx$$

is the uniform measure over Ω, which is thus assumed to have finite Lebesgue measure $|\Omega|$.

4.5.1 Comparison strategies

Our vehicle for estimating the Christoffel function is a general strategy, first introduced in [115], which consists in comparing Ω with reference domains R for which the Christoffel function can be estimated. For simplicity, we use the notation

$$L^2(R) = L^2(R, \mu_R),$$

for any domain R, where $\mu_R = |R|^{-1} \, \chi_R \, dx$ is the uniform measure over R. In order to make clear the dependence on the domain, we define

$$k_{n,R}(x) = \frac{1}{n} \max_{v \in V_n} \frac{|v(x)|^2}{\|v\|_{L^2(R)}^2},$$

and

$$K_{n,R} = \|k_n\|_{L^\infty(R)} = \frac{1}{n} \max_{v \in V_n} \frac{\|v\|_{L^\infty(R)}^2}{\|v\|_{L^2(R)}^2},$$

We first state a pointwise comparison result.

Lemma 4.19. *For $x \in \Omega$, let R be such that $x \in R \subset \Omega$ and $\beta \, |\Omega| \leqslant |R|$ for some $\beta \in (0, 1]$. Then*

$$k_{n,\Omega}(x) \leqslant \beta^{-1} \, k_{n,R}(x).$$

Conversely, let S be such that $\Omega \subset S$ and $\beta\,|S| \leqslant |\Omega|$ for some $\beta \in (0,1]$. Then

$$k_{n,\Omega}(x) \geqslant \beta\, k_{n,S}(x), \quad x \in \Omega.$$

Proof. For any $v \in V_n$, we have

$$\frac{1}{n}|v(x)|^2 \leqslant k_{n,R}(x)\,\|v\|^2_{L^2(R)} \leqslant k_{n,R}(x)\,\frac{|\Omega|}{|R|}\,\|v\|^2_{L^2(\Omega)},$$

and

$$\frac{1}{n}|v(x)|^2 \leqslant k_{n,\Omega}(x)\,\|v\|^2_{L^2(\Omega)} \leqslant k_{n,\Omega}(x)\,\frac{|S|}{|\Omega|}\,\|v\|^2_{L^2(S)}.$$

Optimizing over v gives the upper and lower estimates of $k_{n,\Omega}(x)$. □

Obviously, a framing on $K_{n,\Omega}$ can be readily derived as follows, by application of the above lemma to any point in Ω.

Proposition 4.20. *Assume that there exist a family $\mathcal{R}$ of reference domains with the following properties:*
(i) For all $x \in \Omega$ there exist $R_x \in \mathcal{R}$ such that $x \in R_x \subset \Omega$.
(ii) There exists a constant $\beta \in (0,1]$ such that $|R| \geqslant \beta\,|\Omega|$ for all $R \in \mathcal{R}$.
Then, one has

$$K_{n,\Omega} \leqslant \beta^{-1} \sup_{x\in\Omega} k_{n,R_x}(x) \leqslant \beta^{-1} \sup_{R\in\mathcal{R}} K_{n,R}.$$

Likewise, for any $S \in \mathcal{R}$ such that $\Omega \subset S$ and $|\Omega| \geqslant \beta\,|S|$, one has

$$K_{n,\Omega} \geqslant \beta \sup_{x\in\Omega} k_{n,S}(x).$$

In what follows, we apply this strategy to spaces V_n of multivariate algebraic polynomials. Throughout this section, we consider

$$V_n = \mathbb{R}_p[X_1,\dots,X_d] := \operatorname{span}\{X_1^{\nu_1}\dots X_d^{\nu_d} \;:\; |\nu| = \nu_1+\cdots+\nu_d \leqslant p\}, \tag{4.37}$$

the space of polynomials with total degree less or equal to p, for which we have

$$n = \binom{d+p}{p}.$$

We assume Ω is a bounded open set of $\mathbb{R}^d$.

It is important to note that V_n is invariant by affine transformation. As a consequence, if A is any affine transformation, one has

$$R' = A(R) \implies k_{n,R'}(A(x)) = k_{n,R}(x), \quad x \in R,$$

and in particular $K_{n,R'} = K_{n,R}$.

4.5.2 Lipschitz domains

In the case of the cube $Q = [-1,1]^d$, we may express $k_{n,Q}$ by using tensorized Legendre polynomials, that is

$$k_{n,Q}(x) = \sum_{|\nu|\leqslant p} |\varphi_\nu(x)|^2, \quad \varphi_\nu(x) = \varphi_{\nu_1}(x_1)\dots\varphi_{\nu_d}(x_d),$$

where the univariate polynomials $t \mapsto \varphi_j(t)$ are normalized in $L^2([-1,1],\frac{dt}{2})$. Using this expression, it can be proved by induction on the dimension d that

$$K_{n,Q} \leqslant n, \quad n \geqslant 1,$$

see Lemma 1 in [48]. Therefore, by affine invariance,

$$K_{n,R} \leqslant n, \quad n \geqslant 1, \tag{4.38}$$

for any d-dimensional parallelogram R. Using this result, we may bound the growth of Christoffel functions from above for a general class of domains.

Definition 4.21. An open set $\Omega \subset \mathbb{R}^d$ satisfies the inner cone condition if there exist $\bar{r} > 0$ and $\theta \in (0, \pi)$, such that for all $x \in \overline{\Omega}$, there exists a unit vector $\vec{u}$ such that the cone

$$\mathcal{C}_{\bar{r},\theta}(x, \vec{u}) := \{x + r\,\vec{v},\, 0 \leqslant r \leqslant \bar{r},\, |\vec{v}| = 1,\, \vec{u} \cdot \vec{v} \geqslant \cos(\theta)\}$$

is contained in $\overline{\Omega}$. In particular, any Lipschitz domain $\Omega \subset \mathbb{R}^d$ satisfies the inner cone condition (see e.g. §4.11 in [1]).

Theorem 4.22. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain that satisfies the inner cone condition. Then, one has*

$$K_n \leqslant C_\Omega\, n, \quad n \geqslant 1, \tag{4.39}$$

where C_Ω depends on d, $|\Omega|$, and on $\bar{r}$ and θ in the previous definition.

Proof. The uniform cone condition ensures that there exists $\kappa = \kappa(\bar{r}, \theta, d) > 0$ such that for any $x \in \Omega$, there exists a parallelogram R such that $x \in R \subset \Omega$ and $|R| = \kappa$. Therefore, applying Proposition 4.20 with $\mathcal{R}$ the family of all parallelograms of area κ, one obtains (4.39) with $C_\Omega = |\Omega|/\kappa$. □

Remark 4.23. The bound $K_{n,Q} \leqslant n$ is actually established in [48] for the more general class of polynomial spaces of the form

$$V_n = \mathbb{R}_\Lambda[X_1, \ldots, X_d] := \operatorname{span}\{X_1^{\nu_1} \ldots X_d^{\nu_d} \;:\; \nu \in \Lambda\}, \quad \#(\Lambda) = n,$$

where $\Lambda \in \mathbb{N}^d$ is downward closed, i.e. such that

$$\nu \in \Lambda \quad \text{and} \quad \widetilde{\nu} \leqslant \nu \implies \widetilde{\nu} \in \Lambda.$$

These spaces are however not invariant by affine transformation, and so one cannot apply the above method to treat general domains with inner cone condition. On the other hand, these spaces are invariant by affine transformation of the form $x \mapsto x_0 + Ax$ where A is a diagonal matrix, therefore transforming the cube Q into an arbitrary rectangle R aligned with the coordinate axes. As observed in [6], this leads to a bound of the form (4.39) for any domain Ω that satisfies the following geometrical property: for all $x \in \Omega$ there exists a rectangle R aligned with the coordinate axes such that $x \in R \subset \Omega$ and $|R| \geqslant \beta\,|\Omega|$. Note that this property does not readily follow from a smoothness property of the boundary, in particular there exist smooth domains for which this property does not hold.

4.5.3 Smooth domains

We next investigate smooth domains. For this purpose, we replace parallelograms by ellipsoids as reference domains. In the case of the unit ball $B := \{|x| \leqslant 1\}$, it is known [191] that the Christoffel function reaches its maximum on the unit sphere $S := \{|x| = 1\}$, where we have

$$n\, k_{n,B}(x) = \binom{p+d+1}{p} + \binom{p+d-2}{p-1}.$$

In order to estimate how this quantity scales with $n = \binom{p+d}{p}$ we use the fact that for any integer m, one has

$$e\left(\frac{m}{e}\right)^m \leqslant m! \leqslant m^m.$$

For the lower bound, we bound from below the first term

$$\binom{p+d+1}{p} = \binom{p+d}{p}\,\frac{p+d+1}{d+1} = n\,\frac{p+d+1}{d+1}$$
$$\geqslant \frac{n}{d\,e^{1/d}}(p+d+1) \geqslant \frac{n}{e\,(d!)^{1/d}}\left(\frac{(p+d)!}{p!}\right)^{1/d} = e^{-1}\,n^{\frac{d+1}{d}},$$

which leads to

$$K_{n,B} \geqslant k_{n,B}(x) \geqslant \frac{1}{e}\,n^{1/d}, \quad x \in S. \tag{4.40}$$

For the upper bound, we write

$$\binom{p+d+1}{p} + \binom{p+d-2}{p-1} = \binom{p+d}{p}\left(\frac{p+d+1}{d+1} + \frac{pd}{(p+d)(p+d-1)}\right)$$
$$\leqslant n\left(\frac{p+d+1}{d+1}+1\right) = n\left(\frac{p}{d+1}+2\right),$$

Since

$$n^{1/d} = (d!)^{-1/d}\left(\frac{(p+d)!}{p!}\right)^{1/d} \geqslant \frac{p+1}{d} \geqslant \frac{p}{d+1},$$

we find that

$$k_{n,B}(x) \leqslant n^{1/d} + 2 \leqslant 3\,n^{1/d}.$$

By affine invariance, we thus obtain

$$\frac{1}{e}\,n^{1/d} \leqslant K_{n,E} \leqslant 3\,n^{1/d}, \tag{4.41}$$

for all ellipsoids E. This leads to the following result.

Theorem 4.24. *Assume $\Omega \subset \mathbb{R}^d$ is a bounded domain with $\mathcal{C}^2$ boundary. Then, one has*

$$K_n \leqslant C_\Omega\,n^{1/d}, \quad n \geqslant 1, \tag{4.42}$$

where C_Ω depends on Ω.

Proof. Since the boundary of Ω has finite curvature, we are ensured that there exists a $\beta > 0$ such that for any $x \in \Omega$, there exist an ellipsoid E such that $x \in E \subset \Omega$ and $|E| \geqslant \beta\,|\Omega|$. Therefore, applying Proposition 4.20 with $\mathcal{R}$ the family of ellipsoids with area larger than $\beta\,|\Omega|$, we obtain (4.42) with $C_\Omega = 3\,\beta^{-1}$. □

Remark 4.25. In the above argument, one could simply use balls instead of ellipsoids, however at the price of diminishing the value of β and thus raising the constant C_Ω.

We next give a general lower bound for K_n showing that the above rate for smooth domains is sharp.

Theorem 4.26. *Let $\Omega \subset \mathbb{R}^d$ be an arbitrary bounded domain, and let B be its Chebychev ball, that is, the smallest closed ball that contains Ω. Then, one has*

$$K_{n,\Omega} \geqslant \frac{1}{e}\,\frac{|\Omega|}{|B|}\,n^{1/d}, \quad n \geqslant 1.$$

Proof. As $\overline{\Omega}$ is compact and B is the smallest possible ball containing $\overline{\Omega}$, there exists a point $x \in \overline{\Omega} \cap \partial B$, and by Lemma 4.19 one has

$$K_{n,\Omega} \geqslant k_{n,\Omega}(x) \geqslant \frac{|\Omega|}{|B|}\,k_{n,B}(x) \geqslant \frac{1}{e}\,\frac{|\Omega|}{|B|}\,n^{1/d},$$

where the last inequality follows from (4.40) and affine invariance. □

4.5.4 Pointwise bounds for piecewise smooth domains

As already observed, it may be needed to get sharper bounds on $k_n(x)$ that depend on the point x, in particular when checking the validity of (4.32). In the case of algebraic polynomials in dimension $d = 2$, for which $n = \frac{(p+1)(p+2)}{2}$, such bounds have been obtained for a particular class of piecewise smooth domains with outward corners, in the following result from [153].

Theorem 4.27. *Let $\Omega \subset \mathbb{R}^2$ be a bounded open such that $\partial\Omega = \cup_{i=1}^K \Gamma_i$, where the Γ_i are one-to-one C^2 curves that intersect only at their extremities, at which points the interior angles belong to $(0, \pi)$. Then, there exists a constant C_Ω that only depends on Ω such that, for all $x \in \Omega$,*

$$C_\Omega^{-1}\, k_n(x) \leqslant \min_{(i,j)\in S} \rho_i(x)\, \rho_j(x) \leqslant C_\Omega\, k_n(x), \quad n \geqslant 1,$$

where S consists of the (i, j) such that Γ_i and Γ_j intersects, and $\rho_i(x) := \min\big(p, d(x, \Gamma_i)^{-1/2}\big)$.

For the square domain $\Omega = Q = [-1, 1]^2$, this implies that $k_{n,Q}(x) \sim p^2 \sim n$ when x is close enough to a corner, and we retrieve the bound $K_{n,Q} \leqslant C_\Omega\, n$ from (4.39). Theorem 4.27 also proves that for bidimensional domains with $\mathcal{C}^2$ boundary, it holds $k_n(x) \sim \min\big(p, d(x, \partial\Omega)^{-1/2}\big)$, which is consistent with the global bound (4.42) in the case $d = 2$.

4.5.5 Rate of growth of $K_{n,\Omega}$ and order of cuspitality

We end this section by a more technical but systematic approach which allows us to estimate the rate of growth of the inverse Christoffel function in a sharp way for domains Ω that could either be smooth, of α-Hölder boundary, or even with cusps of a given order. It is based on using the following more elaborate reference domain that describes a certain order of smoothness at the origin.

Definition 4.28. For $\alpha_1, \ldots, \alpha_{d-1} \in (0, 2]$, denote $R_{\alpha_1,\ldots,\alpha_{d-1}}$ the reference domain

$$R_{\alpha_1,\ldots,\alpha_{d-1}} := \left\{ x \in [\,-1, 1]^d,\ \max_{1\leqslant i\leqslant d-1} |x_i|^{\alpha_i} \leqslant x_d \right\}.$$

We shall establish upper and lower bounds for $K_{n,\Omega}$ based on comparisons between Ω and affine transformations of this reference domain, by adapting certain techniques and results from [66]. The upper bound is as follows.

Theorem 4.29. *Let Ω be a bounded domain. Assume there exist $\alpha_1, \ldots, \alpha_{d-1} \in (0, 2]$ and $\beta > 0$ such that, for all $x \in \Omega$, one can find an affine map A such that $A(0) = x$, $A(R_{\alpha_1,\ldots,\alpha_{d-1}}) \subset \overline{\Omega}$ and $|A(R_{\alpha_1,\ldots,\alpha_{d-1}})| \geqslant \beta\, |\Omega|$. Then*

$$K_{n,\Omega} \leqslant C_\Omega\, n^{\frac{1}{d}\left(1+\sum_{i=1}^{d-1} \frac{2-\alpha_i}{\alpha_i}\right)}, \tag{4.43}$$

where C_Ω is a constant depending only on Ω.

This result is obtained with the extension strategy proposed in [66], which consists in combining Proposition 4.31 below with a comparison of domains. Such a method was applied in the same paper to the case of smooth domains, polytopes, some 2-dimensional domains, and ℓ^α balls in $\mathbb{R}^d$, which all correspond to the situation $\alpha_1 = \cdots = \alpha_{d-1} \in [1, 2]$ in our theorem. We give below a series of intermediate results that lead to the proof of Theorem 4.29.

Lemma 4.30. *For $\alpha \in (0, 2]$ and $n \geqslant 1$, the function $f : x \mapsto \frac{1}{9p^2} + \beta x^2 - |x|^\alpha$ remains non-negative on $\mathbb{R}$ as soon as $\beta \geqslant \frac{\alpha}{2}\left(\frac{9}{2}\,(2-\alpha)\,p^2\right)^{\frac{2-\alpha}{\alpha}}$.*

Proof. As f is even, one only has to consider this function on $\mathbb{R}_+$. For $x > 0$, $f'(x) = 2\beta x - \alpha x^{\alpha-1}$ cancels only at $x_0 = \left(\frac{\alpha}{2\beta}\right)^{\frac{1}{2-\alpha}}$, so

$$\min_{x\in\mathbb{R}} f(x) = f(x_0) = \frac{1}{9p^2} - \frac{2-\alpha}{2}\left(\frac{\alpha}{2\beta}\right)^{\frac{\alpha}{2-\alpha}},$$

which is non-negative if and only if $\beta \geqslant \frac{\alpha}{2}\left(\frac{9}{2}\,(2-\alpha)\,p^2\right)^{\frac{2-\alpha}{\alpha}}$. □

The following result is Theorem 5.2 from [66].

Proposition 4.31. *Suppose $\Omega \subset \mathbb{R}^d$ is a compact set and T is an affine transformation of $\mathbb{R}^d$ such that $T(B(0,1)) \subset \Omega$. Then*

$$k_{n,\Omega}\left(T\left(0,\dots,0,1+\frac{1}{3p^2}\right)\right) \leqslant c\,|\det T|^{-1}p.$$

where c depends only on d.

Lemma 4.32. *For $\alpha_1,\dots,\alpha_{d-1} \in (0,2]$, one has*

$$k_{n,R_{\alpha_1,\dots,\alpha_{d-1}}}(0) \leqslant Cp^{1+\sum_{i=1}^{d-1}\frac{2-\alpha_i}{\alpha_i}},$$

where C depends only on d and $\alpha_1,\dots,\alpha_{d-1}$.

Proof. Define $\beta_i = \frac{\alpha_i}{2}\left(\frac{9}{2}(2-\alpha_i)\,p^2\right)^{\frac{2-\alpha_i}{\alpha_i}}$ for $1 \leqslant i \leqslant d-1$, and let T be the affine transformation

$$T : x = (x_1,\dots,x_d) \mapsto \left(\frac{x_1}{\sqrt{3\beta_1}},\dots,\frac{x_{d-1}}{\sqrt{3\beta_{d-1}}},\frac{1}{3}\left(1+\frac{1}{3p^2}-x_d\right)\right).$$

Then, for all $x \in B(0,1)$, $T(x)_d \in [0,1]$ and $1 \leqslant i \leqslant d-1$, using Lemma 4.30,

$$T(x)_d = \frac{1}{3}\left(1+\frac{1}{3p^2}-x_d\right) \geqslant \frac{1}{3}\left(\frac{1}{3p^2}+x_i^2\right) = \frac{1}{9p^2}+\beta_i T(x)_i^2 \geqslant T(x)_i^{\alpha_i},$$

so $\max_{1\leqslant i\leqslant d-1}|T(x)_i|^{\alpha_i} \leqslant T(x)^d$, which implies that $T(B(0,1)) \subset R_{\alpha_1,\dots,\alpha_{d-1}}$.

As $T\left(0,\dots,0,1+\frac{1}{3p^2}\right) = 0$, a direct application of Proposition (4.31) gives

$$k_{n,R_{\alpha_1,\dots,\alpha_{d-1}}}(0) \leqslant c\,p\,|\det T|^{-1} = 3\,c\,p\prod_{i=1}^{d-1}\sqrt{3\beta_i} \leqslant Cp^{1+\sum_{i=1}^{d-1}\frac{2-\alpha_i}{\alpha_i}}$$

□

Proof of Theorem 4.29. One simply applies Proposition 4.20 to the family $\mathcal{R}$ of all domains of the form $A(R_{\alpha_1,\dots,\alpha_{d-1}})$ where A is an affine map such that $|\det A|\,|R_{\alpha_1,\dots,\alpha_{d-1}}| > \beta\,|\Omega|$. As $p \leqslant \bar{C}\,n^{1/d}$ for some $L > 0$, we obtain (4.43) with $C_\Omega = C\,\beta^{-1}\,\bar{C}^{\,2+\sum_{i=1}^{d-1}2/\alpha_i}$, the constant C coming from the lemma above. □

We now prove a lower bound based on the same reference domain.

Theorem 4.33. *Let Ω be a bounded domain. Assume there exist $\bar{x} \in \overline{\Omega}$, $0 < r_1 \leqslant r_2$, $\alpha_1,\dots,\alpha_{d-1} \in (0,2]$ and an affine transformation A with $A(0) = \bar{x}$ such that*

$$\Omega \subset A(R_{\alpha_1,\dots,\alpha_{d-1}}) \cup \left(\overline{B}(\bar{x},r_2) \setminus B(\bar{x},r_1)\right).$$

Then

$$K_{n,\Omega} \geqslant c_\Omega\,n^{\frac{1}{d}\left(1+\sum_{i=1}^{d-1}\frac{2-\alpha_i}{\alpha_i}\right)},$$

where c_Ω is a constant depending only on Ω.

The proof follows the same path as in Theorem 8.1 and Remark 8.4 of [66], but with a radial polynomial centered at x instead of a planar polynomial, that is a univariate polynomial composed with an affine function. This small improvement shows that for a point x and a domain Ω satisfying the conditions of Theorems 4.29 and 4.33 with the same α_i, the asymptotic behavior of $k_{n,\Omega}(x)$ only depends on Ω in a neighborhood of x.

We first recall Lemma 6.1 from the same article:

Lemma 4.34. *For any $p, m \geqslant 1$ and $y \in [-1,1]$, there exists a univariate polynomial $P_{p,m,y}$ of degree at most p such that $P_{p,m,y}(y) = 1$ and*

$$|P_{p,m,y}(x)| \leqslant c(m) \left(\frac{1 + p\sqrt{1-y^2}}{1 + p\sqrt{1-y^2} + p^2\,|x-y|} \right)^m, \quad x \in [-1,1].$$

Taking $y = -1$ and applying a change of variable $x \mapsto \frac{x+1}{2}$, we get as an immediate consequence:

Lemma 4.35. *For any $p, m \geqslant 1$, there exists a univariate polynomial $P_{p,m}$ of degree at most p such that $P_{p,m}(0) = 1$ and*

$$|P_{p,m}(x)| \leqslant c(m) \min\left(1, \frac{1}{p^{2m}|x|^m}\right), \quad x \in [0,1].$$

We also need a bound on the volume of $R_{\alpha_1,\ldots,\alpha_{d-1}}$.

Lemma 4.36. *For all $r > 0$, $|R_{\alpha_1,\ldots,\alpha_{d-1}} \cap B(0,r)| \leqslant c\, r^{1+\sum_{i=1}^{d-1} 1/\alpha_i}$.*

Proof. Given $r \in [0,1]$, $R_{\alpha_1,\ldots,\alpha_{d-1}} \cap \{x_d = r\} = [-r^{1/\alpha_1}, r^{1/\alpha_1}] \times \cdots \times [-r^{1/\alpha_{d-1}}, r^{1/\alpha_{d-1}}] \times \{r\}$ has a $(d-1)$-volume equal to $\prod_{i=1}^{d-1} 2\, r^{-\alpha_i}$, so

$$|R_{\alpha_1,\ldots,\alpha_{d-1}} \cap \{0 \leqslant x_d \leqslant r\}| = \int_0^r \prod_{i=1}^{d-1} 2\, x_d^{1/\alpha_i}\, dx_d = c\, r^{1+\sum_{i=1}^{d-1} 1/\alpha_i}.$$

As for all $r > 0$, $R_{\alpha_1,\ldots,\alpha_{d-1}} \cap B(0,r) \subset R_{\alpha_1,\ldots,\alpha_{d-1}} \cap \{0 \leqslant x_d \leqslant \min(1,r)\}$, we obtain the desired result. □

Proof of Theorem 4.33. Take $p_0 = \lfloor \frac{p}{2} \rfloor$, $m \geqslant \frac{1}{2}\sum_{i=1}^{d-1} \frac{1}{\alpha_i}$ and $r_3 \geqslant r_2$ such that $T(R_{\alpha_1,\ldots,\alpha_{d-1}}) \subset \overline{B}(\bar{x}, r_3)$, and define the multivariate polynomial

$$P(x) = P_{p_0,m}\left(\frac{|x-\bar{x}|^2}{r_3^2}\right), \quad x \in \mathbb{R}^d.$$

Then P has degree at most $2p_0 \leqslant p$ in each variable, $P(\bar{x}) = 1$, and Lemma 4.35 bounds P from above since $\Omega \subset \overline{B}(\bar{x}, r_3)$. It remains to compute an upper bound of $\|P\|_{L^2(\Omega)}$. For $0 < r < r_1$, one has:

$$\begin{aligned}
|\Omega \cap B(\bar{x},r)| &= |T(R_{\alpha_1,\ldots,\alpha_{d-1}}) \cap B(\bar{x},r)| \\
&\leqslant |\det T|\, |R_{\alpha_1,\ldots,\alpha_{d-1}} \cap T^{-1}(B(\bar{x},r))| \\
&\leqslant |\det T|\, |R_{\alpha_1,\ldots,\alpha_{d-1}} \cap B(0, r\,\lambda_{\max}(T^{-1}))| \\
&\leqslant c' r^{1+\sum_{i=1}^{d-1} 1/\alpha_i},
\end{aligned}$$

where in the last line we used Lemma 4.36, and with $c' = \frac{c\,|\det T|}{\lambda_{\min}(T)^{1+\sum_{i=1}^{d-1} 1/\alpha_i}}$. Therefore, one can compute

$$\begin{aligned}
\|P\|_{L^2(\Omega)}^2 &\leqslant \left\| c(m) \min\left(1, \frac{1}{p^{2m}|x|^m}\right) \right\|_{L^2(\Omega)}^2 \\
&\leqslant c(m)^2 \left(\int_{\Omega \cap B(\bar{x},p^{-2})} dx + \int_{B(\bar{x},r_3)} \frac{dx}{p^{4m} r_1^{2m}} + \int_{\Omega \cap B(\bar{x},r_1) \setminus B(\bar{x},p^{-2})} \left(\frac{1}{|x|^{2m}} - \frac{1}{r_1^{2m}} \right) \frac{dx}{p^{4m}} \right) \\
&= c(m)^2 \left(\left|\Omega \cap B\left(\bar{x}, p^{-2}\right)\right| + \frac{|B(\bar{x},r_3)|}{p^{4m} r_1^{2m}} + \int_{p^{-2}}^{r_1} \frac{2m}{p^{4m} r^{2m+1}} \,|\Omega \cap B(\bar{x},r)|\, dr \right) \\
&\leqslant c(m)^2 \left(c' p^{-2-\sum_{i=1}^{d-1} 2/\alpha_i} + \frac{|B(\bar{x},r_3)|}{p^{4m} r_1^{2m}} + c'\, \frac{2m}{p^{4m}} \int_{p^{-2}}^{r_1} r^{\sum_{i=1}^{d-1} 1/\alpha_i - 2m} dr \right) \\
&\leqslant c'' \max\left(p^{-2-\sum_{i=1}^{d-1} 2/\alpha_i}, p^{-4m}, p^{-4m-2(\sum_{i=1}^{d-1} 1/\alpha_i - 2m + 1)} \right) \\
&= c'' p^{-2-\sum_{i=1}^{d-1} 2/\alpha_i},
\end{aligned}$$

and conclude that $K_{n,\Omega} \geqslant k_{n,\Omega}(\bar{x}) \geqslant \frac{1}{n}\frac{|P(\bar{x})|^2}{\|P\|^2_{L^2(\Omega)}} \geqslant c_\Omega\, p^{1+\sum_{i=1}^{d-1}\frac{2-\alpha_i}{\alpha_i}}$, with $c_\Omega = p^d/c''n$. □

Remark 4.37. These theorems include the case of smooth domains : indeed, taking $\alpha_1 = \cdots = \alpha_{d-1} = 2$ and $e_d = (0, \ldots, 0, 1)$, one has

$$B\left(\frac{1}{2}e_d, \frac{1}{2}\right) \subset R_{2,\ldots,2} \subset \left\{x \in [\,-1, 1]^d,\ \frac{1}{d-1}\sum_{i=1}^{d-1}|x_i|^2 \leqslant x_d\right\} \subset B\left((d-1)e_d, (d-1)\right),$$

so one can recover the results 4.24 and 4.26, without explicit constants. Similarly, Lipschitz boundaries correspond to the particular values $\alpha_1 = \cdots = \alpha_{d-1} = 1$.

Example 4.38. It becomes useful to take distinct values for the α_i in the case of domains with edges but no corners. For instance, consider $\Omega = \frac{\sqrt{3}}{2}e_d + B\left(\frac{1}{2}e_1, 1\right) \cap B(-\frac{1}{2}e_1, 1)$. Then $0 \in A(R_{1,2\ldots,2}) \subset \overline{\Omega} \subset B(R_{1,2\ldots,2})$, where A and B are the linear maps defined by

$$A(x_1, \ldots, x_d) = \left(\frac{1}{4}\,x_1, \frac{1}{2\sqrt{d-2}}\,x_2, \ldots, \frac{1}{2\sqrt{d-2}}\,x_{d-1}, \frac{\sqrt{3}}{2}\,x_d\right)$$

and

$$B(x_1, \ldots, x_d) = \left(3\,x_1, \frac{1}{\sqrt{3}}\,x_2, \ldots, \frac{1}{\sqrt{3}}\,x_{d-1}, \sqrt{3}\,x_d\right).$$

Thus $K_{n,\Omega} \geqslant k_{n,\Omega}(0) \sim p^2$. Moreover, for all $x \in \Omega$ there exists an affine transformation T such that $\det T \geqslant 2^{-d}$, $T(\Omega) \subset \Omega$ and $T(0) = x$, so $K_{n,\Omega} \sim p^2$.

Remark 4.39. It is easily seen that for domains having a cusp that points outside, the value of K_n may grow as fast as any polynomial, depending on the order of cuspitality. For instance, given $\alpha \in (0, 2]$, according to Theorems 4.29 and 4.33, one has

$$k_{n,R_{\alpha,\ldots,\alpha}}(0) \sim p^{1+\frac{2-\alpha}{\alpha}(d-1)},$$

so that $K_{n,R_{\alpha,\ldots,\alpha}} \geqslant c\, p^{1+\frac{2-\alpha}{\alpha}(d-1)}$.

4.6 Numerical illustration

In this section we give numerical illustrations of the offline and online sampling strategies in the particular case of algebraic polynomials and for different domains. As in the previous section, we consider spaces of polynomials of fixed total degree $V_n = \mathbb{R}_p[X_1, X_2]$ as defined by (4.37).

The three considered domains are

1. $\Omega := \{x \in [-1, 1]^2 \ :\ |x|^2 \leqslant \frac{2}{\pi}\}$, the ball of area 2.
2. $\Omega := \{x \in [-1, 1]^2 \ :\ 0 \leqslant |x_1| - x_2 \leqslant 1\}$, a polygon with a re-entrant corner at $(0, 0)$.
3. $\Omega := \{x \in [-1, 1]^2 \ :\ 0 \leqslant \sqrt{|x_1|} - x_2 \leqslant 1\}$, a domain with a re-entrant cusp at $(0, 0)$.

The measure μ for the error metric $L^2(\Omega, \mu)$ is the uniform probability measure $\frac{1}{|\Omega|}dx = \frac{1}{2}dx$ on the considered domains. In all three cases, the domain Ω is included in the unit cube $Q = [-1, 1]^2$, and described by algebraic inequalities. Thus, sampling according to μ is readily performed by uniform sampling on Q, which is done separately on the two coordinates, followed by rejection when $x \notin \Omega$.

The above three domains are instances of smooth, Lipschitz and cuspital domains, respectively. They are meant to illustrate how the smoothness of the boundary affects the amount of sample needed in the offline state, as rigorously analyzed in the previous section. On these particular domains, we are actually able to exactly integrate polynomials, and therefore in principle to compute the exact orthogonal polynomials φ_j up to round-off error due to the orthogonalization procedure. In our numerical tests, the considered total degrees are $p = 0, 1, \ldots, 20$, therefore $n = n_p = 1, 3, 6, \ldots, 231$. The intermediate values of n between n_p and n_{p+1} are treated by complementing the space V_n with the monomials $X_1^{\alpha_1}X_2^{\alpha_2}$ for $\alpha_1 + \alpha_2 = p + 1$ in the order

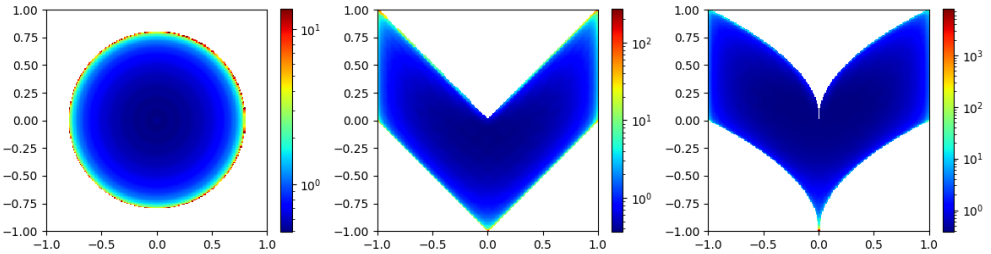


Figure 4.1 – The three domains (disc, polygon, cusp) and the function k_n for $n = 231$.

$\alpha_2 = 0, \ldots, p+1$. For such values, we could compute the φ_j using Cholesky factorization with quadruple precision, and check that $|\langle \varphi_j, \varphi_k \rangle - \delta_{j,k}| \leqslant 10^{-16}$, that is, orthonormality holds up to double precision.

We may thus compute for each value of n the exact inverse Christoffel function k_n and optimal measure $\sigma^* = k_n \, \mu$. Figure 4.1 displays the three domains and the value of k_n for the maximal value $n = 231$ which, as explained by the results in Section 4.5, grows near to the boundary, faster at the outward corners (and even faster at outward cusps), and slower in smooth regions or at re-entrant singularities.

This exact computation allows us to compare the optimal sampling strategy based on σ^* and the more realistic strategy based on $\widetilde{\sigma}$ which is computed from the approximate inverse Christoffel function $\widetilde{k}_n$ derived in the offline stage. We next show that both strategies perform similarly well in terms of instance optimality at near-optimal sampling budget. We stress however that for more general domains where exact integration of polynomials is not feasible, only the second strategy based on $\widetilde{k}_n$ is viable.

4.6.1 Sample complexity of the offline stage

We first illustate the sample complexity M in the offline stage of Algorithm 1. As discussed in § 4.3.2, a sufficient condition to ensure the framing (4.22) between k_n and $\widetilde{k}_n$ is the matrix framing property (4.26) which expresses the fact that the condition number of G_M satisfies the bound

$$\kappa(G_M) \leqslant \frac{c_2}{c_1}.$$

For the constants $c_1 = 2/3$ and $c_2 = 2$, this occurs with high probability when M is larger than K_n, or a known upper bound $B(n)$, multiplied by logarithmic factors, as expressed by (4.27).

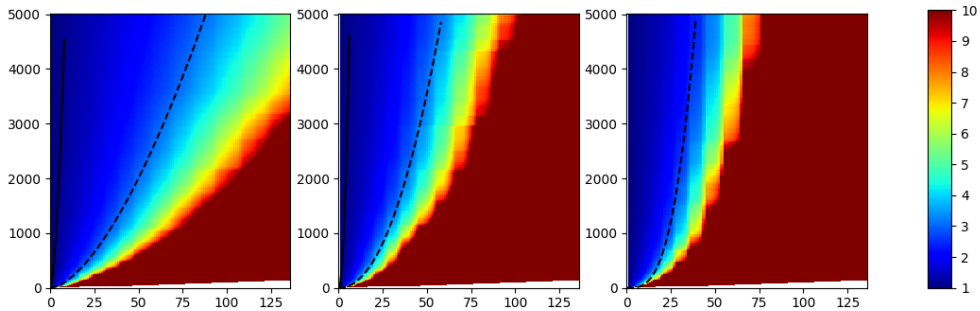


Figure 4.2 – Conditioning of the matrix G_M for the disc (left), polygon (center) and cusp (right) domains, averaged over 100 realizations, with theoretical value of $M_{\mathrm{suf}}(n)$ (full curve) and adjusted value $M_{\mathrm{adj}}(n)$ (dashed curve). The x-coordinate stands for n and the y-coordinate for M; moreover, the plotted values are saturated at 10 since we are only interested in small condition numbers.

Figure 4.2 displays the condition number $\kappa(G_M)$, averaged over 100 realizations of the offline sample $\{y^1, \dots, y^M\}$, as a function of n and $M \geqslant n$, for the three considered domains. We observe a transition region that illustrates the minimal offline sampling budget $M_{\min}(n)$ that should be practically invested in order for G_M to be well conditioned. For example, if $M_{\min}(n)$ is defined as the minimal value of M such that $\mathbb{E}(\kappa(G_M)) \leqslant 3$, this value can be estimated and visualized on Figure 4.2 as the transition to the dark blue color.

We also draw in full line the value of the sufficient value

$$M_{\text{suf}}(n) := \gamma\, B(n) \ln(2n/\varepsilon),$$

for $\varepsilon = 10^{-2}$ where $B(n)$ is the upper bound for K_n derived from the theoretical analysis of Section 4.5. This upper bound is $3n^{3/2}$ for the disc in view of (4.41) and $2n^2$ for the polygonal domain by application of Proposition 4.20 with $\beta = \frac{1}{2}$, since Ω is the union of two parallelograms of equal size. While the sampling budget $m = M_{\text{suf}}(n)$ guarantees that $\kappa(G_M) \leqslant 3$ with high probability —here 0.99— the plots reveal that this budget is by far an over-estimation of $M_{\min}(n)$.

We draw in dashed line the adjusted values $M_{\text{adj}}(n) = C_{\text{adj}}\, M_{\text{suf}}(n)$ where the multiplicative constant is picked as small as possible with the constraint of still fitting the requirement $\mathbb{E}(\kappa(G_M)) \leqslant 3$, thus better fitting the minimal budget $M_{\min}(n)$. We find that constant C_{adj} is approximately $\frac{1}{45}$ for the disc and $\frac{1}{120}$ for the polygon. It is even smaller for the cusp domain, for which Theorem 4.29 with $\alpha_1 = \frac{1}{2}$ yields an upper bound of the form $B(n) = Cn^3$ with a constant C that can be numerically estimated but turns out to be very pessimistic.

In summary, the offline sampling budget $M_{\text{suf}}(n)$ suggested by the theoretical analysis is always pessimistic by a large multiplicative constant. Let us remind that the value $M_{\min}(n)$ is typically not accessible to us since G_M and its condition number cannot be exactly evaluated for more general domains Ω.

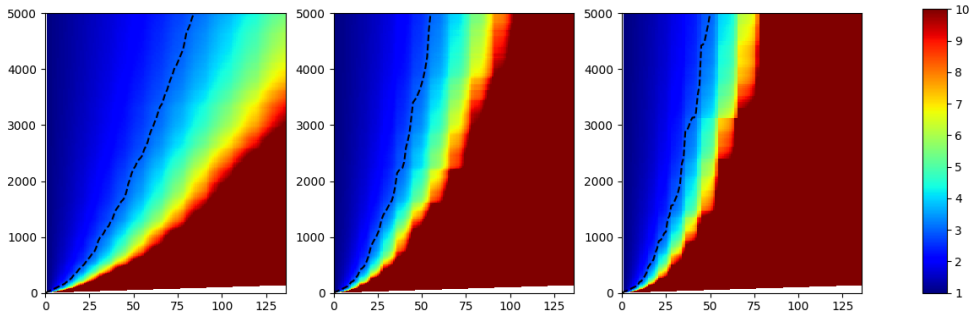


Figure 4.3 – Conditioning of the matrix T for the disc (left), polygon (center) and cusp (right) domains, and value of $M_{\text{emp}}(n)$ (dashed curve), averaged over 100 realizations.

This state of affair justifies the use of the empirical method outlined in § 4.3.3 for selecting a good value of M. Recall that this approach consists in raising M until the conditioning of the computable matrix T becomes less than some prescribed value, for example $\kappa(T) \leqslant 3$. Figure 4.3 displays the conditioning $\kappa(T)$ again averaged over 100 realizations of the offline sample, as well as the curve showing the empirical value $M_{\text{emp}}(n)$ which corresponds to the smallest value of M such that $\kappa(T) \leqslant 3$. It reveals the relevance of the empirical approach: due to the very good fit between $\kappa(T)$ and $\kappa(G_M)$, the value $M_{\text{emp}}(n)$ appears as a much sharper estimate for $M_{\min}(n)$ than $M_{\text{suf}}(n)$.

4.6.2 Sample complexity of the online stage

We next study the sample complexity m of the online stage of Algorithm 1 through the conditioning of the matrix $G_m = (\langle \varphi_j, \varphi_k \rangle_m)_{1 \leqslant j,k \leqslant n}$, where $\langle \cdot, \cdot \rangle_m$ is the inner product associated to the discrete norm

$$\|v\|_m^2 := \frac{1}{m} \sum_{i=1}^{m} w(x^i)\, |v(x^i)|^2.$$

For the sampling measure σ and weight w, we both consider:

(i) The optimal sampling measure $d\sigma^* := k_n \, d\mu$ and weight $w^* = 1/k_n$, which, for these particular domains, can be exactly computed from the φ_j, but are not accessible for more general domains.

(ii) The empirical sampling measure $d\widetilde{\sigma} := \widetilde{k}_n d\mu$ and weight $\widetilde{w} = 1/\widetilde{k}_n$ where $\widetilde{k}_n$ has been obtained from the offline stage, using the previously described empirical choice of M.

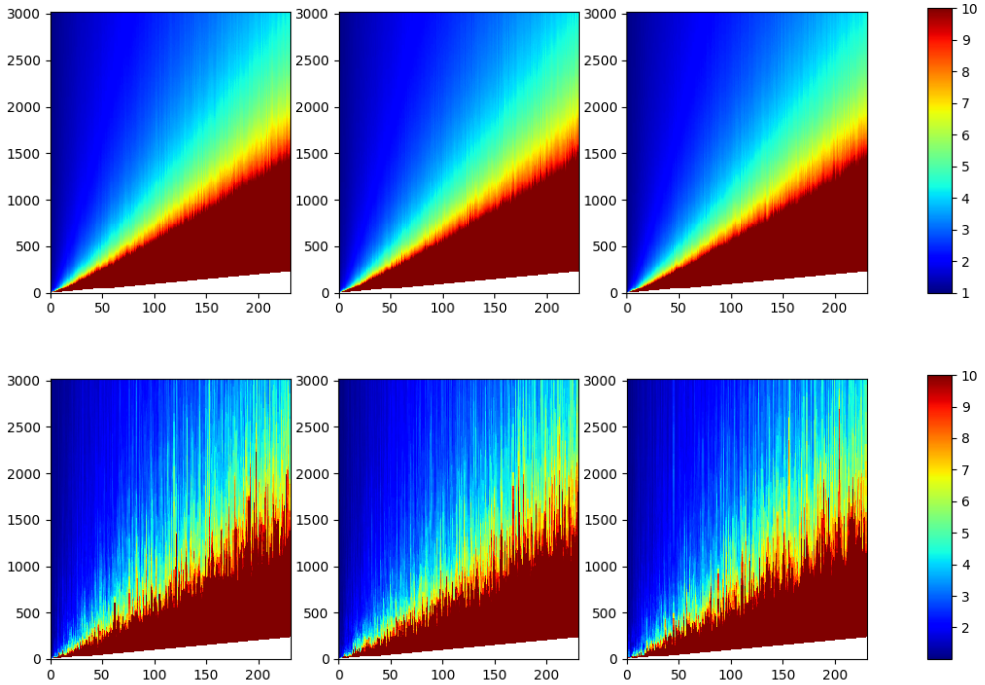


Figure 4.4 – Conditioning of $G_m = \langle \varphi_j, \varphi_k \rangle_m$ depending on m and n for the disc (left), polygon (center) and cusp (right) domains, using an average over 100 realizations with k_n (up), or a single realization with the estimated $\widetilde{k}_n$ (down).

Figure 4.4 displays the condition number $\kappa(G_m)$, as a function of m and n, for both choices and the three domains. In order to illustrate the fluctuations of $\kappa(G_m)$, we display an averaging over 100 realizations when using k_n, and one single realization when using $\widetilde{k}_n$. While the behaviour for a single realization is more chaotic, we find that in both case, as expected, the online sampling budget $m(n)$ which ensures that G_m is well conditioned, for example $\kappa(G_m) \leqslant 3$, grows linearly with n (up to logarithmic factors), now independently of the domain shape.

4.6.3 Instance and budget optimality

In order to illustrate the achievement of our initial goal of instance and budget optimality, we consider the approximation in a polynomial space $V_n = \mathbb{R}_p[X_1, X_2]$ of a function u that consists of a polynomial part $P_n u \in V_n$ and a residual part $P_n^\perp u \in V_n^\perp$ that are both explicitly given in terms of their expansions

$$P_n u = \sum_{j=1}^{n} c_j \varphi_j,$$

and

$$P_n^\perp u = \sum_{j \geqslant n+1} c_j \varphi_j.$$

For numerical testing, we take only finitely many non-zero c_j in this second expansion and adjust them so that $\sum_{j \geqslant n+1} |c_j|^2 = 10^{-4}$. Thus, the best approximation error has value

$$e_n(u)_{L^2} = \|u - P_n u\|_{L^2} = \|P_n^{\perp} u\|_{L^2} = 10^{-2}.$$

We study the mean-square error $\mathbb{E}(\|u - P_n^m u\|_{L^2}^2)$ as a function of m, and compare the different sampling strategies through their ability to reach this ideal benchmark.

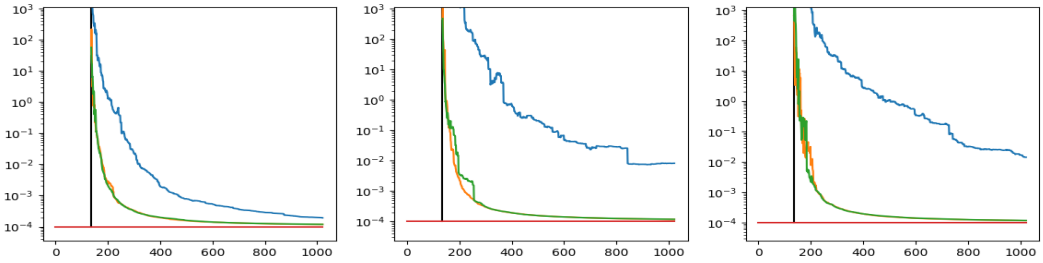


Figure 4.5 – Mean-square reconstruction error for the disc (left), polygon (center) and cusp (right) domains, with total polynomial degree $p = 15$, and sampling measures μ (blue), σ^* (orange), $\widetilde{\sigma}$ (green). Horizontal red line: best approximation error $e_n(u)_{L^2}^2 = 10^{-4}$. Vertical black line: polynomial dimension $n = 136$.

Figure 4.5 displays the error curves (obtained by averaging $\|u - P_n^m u\|_{L^2}^2$ over 100 realizations) for the three domains and polynomial degree $p = 15$ that corresponds to the dimension $n = 136$. For all domains, we observe that the best approximation error is attained up to multiplicative factor 2 with a sampling budget m that is thrice larger than n, when using either the optimal sampling measure σ^* based on k_n or the measure $\widetilde{\sigma}$ based on $\widetilde{k}_n$ obtained in the offline stage Algorithm 1. This does not occur when sampling according to the uniform measure μ: the error remains orders of magnitude above the best approximation error and this effect is even more pronounced as the domain becomes singular. This reflects the fact that with the uniform sampling, the budget m needs to be larger than $n\,K_n$, which has faster growth with n for singular domains.

Chapter 5

Optimal pointwise sampling for L^2 approximation

Abstract. Given a function $u \in L^2 = L^2(\Omega, \mu)$, where μ is a measure on a set Ω, and a linear subspace $V_n \subset L^2$ of dimension n, we show that near-best approximation of u in V_n can be computed from a near-optimal budget of Cn pointwise evaluations of u, with $C > 1$ a universal constant. The sampling points are drawn according to some random distribution, the approximation is computed by a weighted least-squares method, and the error is assessed in expected L^2 norm. This result improves on the results in [59, 85] which require a sampling budget that is sub-optimal by a logarithmic factor, thanks to a sparsification strategy introduced in [128, 140]. As a consequence, we obtain for any compact class $\mathcal{K} \subset L^2$ that the sampling number $\rho_{Cn}^{\mathrm{rand}}(\mathcal{K})_{L^2}$ in the randomized setting is dominated by the Kolmogorov n-width $d_n(\mathcal{K})_{L^2}$. While our result shows the existence of a randomized sampling with such near-optimal properties, we discuss remaining issues concerning its generation by a computationally efficient algorithm.

5.1 Introduction

We study the approximation of a function $u \in L^2(\Omega, \mu)$, where μ is a measure on a set Ω, by an element $\tilde{u}$ of V_n, a subspace of $L^2(\Omega, \mu)$ of finite dimension n, based on pointwise data of of u. Therefore, to construct $\tilde{u}$, we are allowed to evaluate u on a sample of m points $(x^1, \dots, x^m) \in \Omega^m$. In addition, we consider randomized sampling and reconstruction, in the sense that the sample will be drawn according to a distribution σ_m over Ω^m, so the error $u - \tilde{u}$ should be evaluated in some probabilistic sense. For the sake of notational simplicity, having fixed Ω and μ, we write throughout the chapter

$$L^2 := L^2(\Omega, \mu) \quad \text{and} \quad \|v\|_{L^2} := \left(\int_\Omega |v|^2 d\mu \right)^{1/2},$$

as well as

$$e_n(u)_{L^2} := \min_{v \in V_n} \|u - v\|_{L^2}.$$

One typical applicative setting is the reconstrution of multivariate functions, which corresponds to Ω being a domain in $\mathbb{R}^d$. Our main result is the following:

Theorem 5.1. *For some universal constants $C, K \geqslant 1$, and for any n-dimensional space $V_n \subset L^2$, there exists a random sample $x^1, \dots, x^m$ with $m \leqslant Cn$ and a reconstruction map $R : \Omega^m \times \mathbb{C}^m \mapsto V_n$, such that for any $u \in L^2$, it holds*

$$\mathbb{E}\left(\|u - \tilde{u}\|_{L^2}^2 \right) \leqslant K e_n(u)_{L^2}^2 \tag{5.1}$$

where $\tilde{u} := R(x^1, \dots, x^m, u(x^1), \dots, u(x^m))$.

The reconstruction map R is obtained through a weighted least-squares method introduced in [59], which has already been discussed in several papers, see [5, 85, 86, 132, 133, 138, c]. The weights involved are given by

the expression

$$w : x \in \Omega \mapsto n \min_{v \in V_n} \frac{\|v\|_{L^2}^2}{|v(x)|^2} = \frac{n}{\sum_{j=1}^n |\varphi_j(x)|^2}, \tag{5.2}$$

where the last formula holds for any L^2-orthonormal basis $(\varphi_1, \dots, \varphi_n)$ of V_n. Up to the factor n, w is the *Christoffel function* associated to the space V_n and the space $L^2(\Omega, \mu)$. The weighted least-squares solution is then simply defined as

$$\tilde{u} := \arg\min_{v \in V_n} \frac{1}{m} \sum_{i=1}^m w(x^i) |u(x^i) - v(x^i)|^2.$$

Introducing the discrete ℓ^2 norm

$$\|v\|_m^2 := \frac{1}{m} \sum_{i=1}^m w(x^i) |v(x^i)|^2$$

and its associated scalar product $\langle \cdot, \cdot \rangle_m$, we get a computable formula for $\tilde{u}$:

$$\tilde{u} = \arg\min_{v \in V_n} \|u - v\|_m^2 = P_n^m u,$$

where P_n^m denotes the orthogonal projection on V_n with respect to $\langle \cdot, \cdot \rangle_m$. Note that, strictly speaking, $\| \cdot \|_m$ is not a norm over L^2, however the existence and uniqueness of P_n^m will be ensured by the second condition in Lemma 5.2 below, see Remark 5.6. Therefore our main achievements lie in the particular choice of the random sample $(x^1, \dots, x^m)$ ensuring near-optimal approximation error and sampling budget in Theorem 5.1.

The proof of Theorem 5.1 relies on two conditions: first, the expectation of $\| \cdot \|_m^2$ has to be bounded by $\| \cdot \|_{L^2}^2$ up to a constant. Second, an inverse bound should hold almost surely, instead of just in expectation, for functions in V_n. More precisely, one has:

Lemma 5.2. *Assume that m and the law σ_m of $(x^1, \dots, x^m)$ are such that*

$$\mathbb{E}(\|v\|_m^2) \leqslant \alpha \|v\|_{L^2}^2, \qquad v \in L^2, \tag{5.3}$$

and

$$\|v\|_{L^2}^2 \leqslant \beta \|v\|_m^2 \quad a.s., \qquad v \in V_n. \tag{5.4}$$

Then

$$\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2) \leqslant (1 + \alpha\beta)\, e_n(u)_{L^2}^2. \tag{5.5}$$

Proof. Denote $P_n u$ the orthogonal projection of u on V_n with respect to the $L^2(\Omega, \mu)$ norm. Applying Pythagoras theorem both for $\| \cdot \|_{L^2}$ and $\| \cdot \|_m$, one obtains

$$\begin{aligned}
\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2) &= \|u - P_n u\|_{L^2}^2 + \mathbb{E}(\|P_n u - \tilde{u}\|_{L^2}^2) \\
&\leqslant \|u - P_n u\|_{L^2}^2 + \beta\, \mathbb{E}(\|P_n u - \tilde{u}\|_m^2) \\
&= \|u - P_n u\|_{L^2}^2 + \beta\, \mathbb{E}(\|P_n u - u\|_m^2 - \|u - \tilde{u}\|_m^2) \\
&\leqslant \|u - P_n u\|_{L^2}^2 + \beta\, \mathbb{E}(\|P_n u - u\|_m^2) \\
&\leqslant (1 + \alpha\beta)\, \|u - P_n u\|_{L^2}^2,
\end{aligned}$$

which proves (5.5) since $\|u - P_n u\|_{L^2} = e_n(u)_{L^2}$. □

In Section 5.2, we recall how both conditions (5.3) and (5.4) can be obtained with m quasi-linear in n, that is, of order $n \log n$. We reduce this budget to m of order n in Section 5.3, by randomly subsampling the set of evaluation points, based on results from [128, 140]. The proof of Theorem 5.1 follows. We compare it to the recent results [114, 121, 137] in Section 5.4, in particular regarding the domination of sampling numbers by n-widths. We conclude in Section 5.5 by a discussion on the *offline* computational cost for practically generating the sample $x^1, \dots, x^m$.

5.2 Weighted least-squares

A first approach consists in drawing the x^i independently according to the same distribution σ, that is, taking $\sigma_m = \sigma^{\otimes m}$. The natural choice for σ is $d\sigma = \frac{1}{w} d\mu$, which is a probability measure since

$$\int_\Omega \frac{1}{w} d\mu = \frac{1}{n} \sum_{j=1}^n \int_\Omega |\varphi_j(x)|^2 d\mu(x) = \frac{1}{n} \sum_{j=1}^n \|\varphi_j\|_{L^2}^2 = 1.$$

With this sampling measure,

$$\mathbb{E}(\|v\|_m^2) = \frac{1}{m} \sum_{i=1}^m \int_\Omega w(x)|v(x)|^2 d\sigma = \int_\Omega |v|^2 d\mu = \|v\|_{L^2}^2,$$

so condition (5.3) is ensured with $\alpha = 1$. To study the second condition, we introduce the Hermitian positive semi-definite Gram matrix

$$G_m := (\langle \varphi_j, \varphi_k \rangle_m)_{1 \leqslant j,k \leqslant n}$$

and notice that (5.4) is equivalent to

$$|\nu|^2 = \Big\| \sum_{j=1}^n \nu_j \varphi_j \Big\|_{L^2}^2 \leqslant \beta \Big\| \sum_{j=1}^n \nu_j \varphi_j \Big\|_m^2 = \beta\, \nu^\dagger G_m \nu, \qquad \nu \in \mathbb{C}^n,$$

which in turn rewrites as $\lambda_{\min}(G_m) \geqslant \beta^{-1}$.

By the central limit theorem, as m tends to infinity, the scalar products $\langle \varphi_j, \varphi_k \rangle_m$ converge almost surely to $\langle \varphi_j, \varphi_k \rangle = \delta_{j,k}$, so G_m converges to the identity matrix, and we expect that $\lambda_{\min}(G_m) \geqslant \beta^{-1}$ holds for $\beta > 1$ with high probability as m gets large. A quantitative formulation can be obtained by studying the concentration of G_m around I in the matrix spectral norm

$$\|A\|_2 := \max\{|Ax| \ : \ |x| = 1\}.$$

This is based on the matrix Chernoff bound, see [7, 176] for the original inequality and [c], Lemma 2.1, for its application to our problem:

Lemma 5.3. *For $m \geqslant 10n \ln(\frac{2n}{\varepsilon})$, if $(x^1, \dots, x^m) \sim \sigma^{\otimes m}$, then*

$$\mathbb{P}\left(\|G_m - I\|_2 \leqslant \frac{1}{2} \right) \geqslant 1 - \varepsilon.$$

In particular, $\mathbb{P}\left(\lambda_{\min}(G_m) \geqslant \frac{1}{2}\right) \geqslant 1 - \varepsilon$.

Thus assumption (5.4) is satisfied with $\beta = 2$, but only with probability $1 - \varepsilon$. As we would like it to hold almost surely, we condition the random sample points to the event

$$E := \left\{ \|G_m - I\|_2 \leqslant \frac{1}{2} \right\}.$$

In practice, the conditional sample can be obtained through a rejection method, which consists in discarding the whole sample $(x^1, \dots, x^m)$ and redrawing it according to the same probability measure $\sigma_m = \sigma^{\otimes m}$, as many times as needed, until event E is attained. We then define $\tilde{u}$ as the weighted least-square estimator based on this conditioned sample, that is

$$\tilde{u} := \mathbb{E}(P_n^m u | E). \tag{5.6}$$

This approach was introduced and analyzed in [85], see in particular Theorem 3.6 therein. A simpler version of their result, sufficient for our purposes, is the following:

Lemma 5.4. *For $m \geqslant 10n\ln(4n)$, if $(x^1, \dots, x^m)$ is drawn according to the conditional law $\sigma^{\otimes m}|E$, then*

$$\|G_m - I\|_2 \leqslant \frac{1}{2},$$

and $\tilde{u}$ defined in (5.6) *satisfies*

$$\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2) \leqslant 5\, e_n(u)_{L^2}^2.$$

Proof. The first part immediately results from the definition of E, and implies condition (5.4) with $\beta = 2$. Moreover, $\mathbb{P}_{\sigma^{\otimes m}}(E) \geqslant \frac{1}{2}$ by Lemma 5.3 with $\varepsilon = \frac{1}{2}$, so for any $v \in L^2(\Omega, \mu)$,

$$\mathbb{E}_{(\sigma^{\otimes m}|E)}(\|v\|_m^2) = \mathbb{E}_{\sigma^{\otimes m}}(\|v\|_m^2|E) = \frac{\mathbb{E}_{\sigma^{\otimes m}}(\|v\|_m^2 \chi_E)}{\mathbb{P}_{\sigma^{\otimes m}}(E)} \leqslant \frac{\mathbb{E}_{\sigma^{\otimes m}}(\|v\|_m^2)}{\mathbb{P}_{\sigma^{\otimes m}}(E)} \leqslant 2\|v\|_{L^2}^2,$$

so condition (5.3) holds with $\alpha = 2$. The conclusion follows from Lemma 5.2. □

Remark 5.5. The number of redraws k needed to obtain the conditional sample follows a geometric law of expectation $\mathbb{E}(k) = \mathbb{P}(E)^{-1} = (1-\varepsilon)^{-1}$, that is $\mathbb{E}(k) \leqslant 2$ for the particular choice of m in the above lemma. It should be noted that u is not evaluated at the intermediately generated samples not complying with event E. This part of the sampling algorithm thus counts as an offline cost only.

Remark 5.6. Given a sample $x^1, \dots, x^m$ satisfying E, the fact that the Gramian G_m is non-singular implies that we can uniquely define

$$P_n^m u = \sum_{k=1}^n c_k \varphi_k,$$

since $c = (c_1, \dots, c_n)^\dagger$ solves the system of normal equations

$$G_m c = b,$$

where the right-side vector has coordinates

$$b_j = \langle \varphi_j, u \rangle_m = \frac{1}{m} \sum_{i=1}^m w(x^i) \varphi_j(x^i) \overline{u(x^i)}.$$

If u is in L^2, the $u(x^i)$ are only defined up to a representer, however since two representers u' and u'' coincide μ-almost surely, we find that $\tilde{u} = \mathbb{E}(P_n^m u|E)$ is well defined almost surely.

5.3 Random subsampling

With Lemma 5.4, we already have an error bound similar to that of Theorem 5.1. However, the sampling budget is larger than n by a logarithmic factor, which we seek to remove in this section. To do so, we partition the conditional sample into subsets of size comparable to n, and randomly pick one of these subsets to define a reduced sample. An appropriate choice of the partitioning is needed to circumvent the main obstacle, namely the preservation of condition (5.4). It relies on the following lemma, taken from Corollary B of [140], itself a consequence of Corollary 1.5 in [128]. The relevance of these two results to sampling problems were exploited in [137] and noticed in [104], respectively.

Lemma 5.7. *Let $a_1, \dots, a_m \in \mathbb{C}^n$ be vectors of norm $|a_i|^2 \leqslant \delta m$ for $i = 1, \dots, m$, satisfying*

$$\alpha I \preccurlyeq \frac{1}{m} \sum_{i=1}^m a_i a_i^\dagger \preccurlyeq \beta I$$

for some constants $\delta < \alpha \leqslant \beta$. Then there exists a partition of $\{1, \dots, m\}$ into two sets S_1 and S_2 such that

$$\frac{1 - 5\sqrt{\delta/\alpha}}{2}\alpha\, I \preccurlyeq \frac{2}{m}\sum_{i\in S_s} a_i a_i^\dagger \preccurlyeq \frac{1 + 5\sqrt{\delta/\alpha}}{2}\beta\, I, \qquad s = 1, 2.$$

In Lemma 2 of [140] this result is applied inductively in order to find a smaller set $J \subset \{1, \dots, m\}$ of cardinality $|J| \leqslant cn$ such that

$$C^{-1} I \preccurlyeq \frac{1}{n}\sum_{i\in J} a_i a_i^\dagger \preccurlyeq CI,$$

for some universal constants $c, C > 1$. We adapt this approach in order to obtain a complete partition of $\{1, \dots, m\}$ by sets having such properties.

Lemma 5.8. *Let $a_1, \dots, a_m \in \mathbb{C}^n$ be vectors of norm $|a_i|^2 = n$ for $i = 1, \dots, m$, satisfying*

$$\frac{1}{2} I \preccurlyeq \frac{1}{m}\sum_{i=1}^{m} a_i a_i^\dagger \preccurlyeq \frac{3}{2} I.$$

Then there exists an integer L and a partition of $\{1, \dots, m\}$ into 2^L sets $J_1, \dots, J_{2^L}$ such that

$$c_0 I \preccurlyeq \frac{1}{n}\sum_{i\in J_s} a_i a_i^\dagger \preccurlyeq C_0 I, \qquad 1 \leqslant s \leqslant 2^L, \tag{5.7}$$

with universal constants c_0 and C_0. In addition, each set J_s satisfies

$$|J_s| \leqslant C_0 n. \tag{5.8}$$

Proof. The cardinality estimate (5.8) follows from the upper inequality in (5.7) by taking the trace

$$C_0 n = \mathrm{Tr}(C_0 I) \geqslant \frac{1}{n}\sum_{i\in J_s} \mathrm{Tr}(a_i a_i^\dagger) = \frac{1}{n}\sum_{i\in J_s} |a_i|^2 = |J_s|.$$

For the proof of (5.7), if $n/m \geqslant 1/200$, then the result holds with $L = 0$, $J_1 = \{1, \dots, m\}$, $c_0 = 1/2$ and $C_0 = 300$. Now assuming $\delta := n/m < 1/200$, define by induction $\alpha_0 = \frac{1}{2}$, $\beta_0 = \frac{3}{2}$, and

$$\alpha_{\ell+1} := \alpha_\ell \frac{1 - 5\sqrt{\delta/\alpha_\ell}}{2}, \qquad \beta_{\ell+1} := \beta_\ell \frac{1 + 5\sqrt{\delta/\alpha_\ell}}{2}, \qquad \ell \geqslant 0.$$

As $\alpha_{\ell+1} \leqslant \frac{\alpha_\ell}{2}$, the minimal integer L such that $\alpha_L \leqslant 100\,\delta$ is well defined, and satisfies

$$\alpha_L = \alpha_{L-1}\frac{1 - 5\sqrt{\delta/\alpha_{L-1}}}{2} > 100\,\delta\,\frac{1 - 5\sqrt{1/100}}{2} = 25\,\delta.$$

Moreover $\alpha_\ell \geqslant 2^{L-\ell-1}\alpha_{L-1} \geqslant 2^{L-\ell-1}\,100\,\delta$ for $\ell = 0, \dots, L-1$, so

$$\beta_L = 3\,\alpha_L \prod_{\ell=0}^{L-1} \frac{1 + 5\sqrt{\delta/\alpha_\ell}}{1 - 5\sqrt{\delta/\alpha_\ell}} \leqslant C\delta,$$

with $C := 300 \prod_{\ell\geqslant 2} \frac{1+\sqrt{2}^{-\ell}}{1-\sqrt{2}^{-\ell}}$.

Finally, we inductively define partitions $\{S_1^\ell, \dots, S_{2^\ell}^\ell\}$ for $0 \leqslant \ell \leqslant L$. We start with $S_1^0 = \{1, \dots, m\}$, and for any $\ell < L$ and $1 \leqslant s \leqslant 2^\ell$, noticing that

$$\alpha_\ell I \preccurlyeq \frac{1}{m}\sum_{i\in S_s^\ell} a_i a_i^\dagger \preccurlyeq \beta_\ell I,$$

we apply Lemma 5.7 to split S_s^ℓ into subsets $S_{2s-1}^{\ell+1}$ and $S_{2s}^{\ell+1}$ satisfying the same property. At the last step, we define

$$J_s = S_s^L.$$

The framing (5.7) thus holds with $c_0 = \alpha_L/\delta \geqslant 25$ and $C_0 = \beta_L/\delta \leqslant 11000$. □

Proof of Theorem 5.1. Consider $(x^1, \dots, x^m) \sim (\sigma^{\otimes m}|E)$ the conditioned sample introduced in the previous section, and define

$$a_i = \left(\sqrt{w(x^i)}\varphi_j(x^i)\right)_{1\leqslant j\leqslant n}$$

the corresponding normalised random vectors. As E holds almost surely,

$$\frac{1}{2}I \preccurlyeq G_m = \frac{1}{m}\sum_{i=1}^m a_i a_i^\dagger \preccurlyeq \frac{3}{2}I,$$

and

$$|a_i|^2 = w(x^i)\sum_{j=1}^n |\varphi_j(x^i)|^2 = n$$

thanks to the choice of weights (5.2), so the assumptions of Lemma 5.8 are satisfied. Applying this lemma, we obtain sets $J_1, \dots, J_{2^L}$ partitioning $\{1, \dots, m\}$. Let r be a random variable taking value $s \in \{1, \dots, 2^L\}$ with probability $p_s = |J_s|/m$, and randomly subsample $\{x^1, \dots, x^m\}$ by keeping only the points

$$\{x^i \,:\, i \in J_r\}.$$

Then the budget condition $|J_r| \leqslant C_0 n$ is satisfied according to (5.8). Here, we define the discrete norm as

$$\|v\|_{J_r}^2 := \frac{1}{|J_r|}\sum_{i\in J_r} w(x^i)|v(x^i)|^2,$$

and the associated Gram matrix

$$G_{J_r} := (\langle\varphi_j, \varphi_k\rangle_{J_r})_{1\leqslant j,k\leqslant n} = \frac{1}{|J_r|}\sum_{i\in J_r} a_i a_i^\dagger.$$

The weighted least-squares estimate is now defined as

$$\tilde{u} := \operatorname*{arg\,min}_{v\in V_n} \frac{1}{|J_r|}\sum_{i\in J_r} w(x^i)|u(x^i) - v(x^i)|^2,$$

and it thus depends on the random draws of both $(x^1, \dots, x^m) \in \Omega^m$ and $1 \leqslant r \leqslant 2^L$. Condition (5.4) follows from the lower inequality in (5.7) with $\beta = \frac{C_0}{c_0}$ since

$$G_{J_r} \geqslant \frac{c_0 n}{|J_r|} I \geqslant \frac{c_0}{C_0} I.$$

Finally, we have for any $v \in L^2(\Omega, \mu)$

$$\mathbb{E}_{(\sigma^{\otimes m}|E),\mathcal{L}(r)}(\|v\|_{J_r}^2) = \mathbb{E}_{(\sigma^{\otimes m}|E)}\left(\sum_{s=1}^{2^L}\frac{p_s}{|J_s|}\sum_{i\in J_s} w(x^i)|v(x^i)|^2\right) = \mathbb{E}_{(\sigma^{\otimes m}|E)}(\|v\|_m^2) \leqslant 2\, e_n(u)_{L^2}^2,$$

so condition (5.3) holds with $\alpha = 2$. Applying Lemma 5.2, we conclude that (5.1) holds with $C = C_0$ and $K = 1 + 2\frac{C_0}{c_0}$. □

5.4 Comparison with related results

In order to compare Theorem 5.1 with several recent results [104, 113, 137, 168], we consider its implication when the target function u belongs to a certain class of functions $\mathcal{K}$ that describes some prior information on u, such as smoothness.

Recall that if V is a Banach space of functions defined on Ω and $\mathcal{K} \subset V$ is a compact set, its *Kolmogorov n-widths* are defined by

$$d_n(\mathcal{K})_V := \inf_{\dim V_n = n} \sup_{u \in \mathcal{K}} \inf_{v \in V_n} \|u - v\|_V,$$

where the first infimum is taken over all linear spaces $V_n \subset V$ of dimension n. This quantity thus describes the best approximation error that can be achieved uniformly over the class $\mathcal{K}$ by an n-dimensional linear space.

On the other hand, building a best approximation of u requires in principle full knowledge on u, and we want to consider the situation where we only have access to a limited number of point evaluations. This leads us to the notion of *sampling numbers*, also called *optimal recovery numbers*, both in the deterministic and randomized settings.

For deterministic samplings, we define the (linear) sampling numbers

$$\rho_m(\mathcal{K})_V := \inf_{x^1,\dots,x^m} \inf_{R \in \mathcal{L}(\mathbb{C}^m, V)} \max_{u \in \mathcal{K}} \|u - R(u(x^1), \dots, u(x^m))\|_V,$$

where the infimum is taken over all samples $(x^1, \dots, x^m) \in \Omega^m$ and linear reconstruction maps $R : \mathbb{C}^m \to V$. For random samplings, we may define similar quantities by

$$\rho_m^{\mathrm{rand}}(\mathcal{K})_V^2 := \inf_{\sigma_m} \inf_{R : \Omega^m \times \mathbb{C}^m \to V} \max_{u \in \mathcal{K}} \mathbb{E}\left(\|u - R_{(x^1,\dots,x^m)}(u(x^1), \dots, u(x^m))\|_V^2\right),$$

where the infimum is taken over all random sampling laws $\sigma_m \in \mathrm{Prob}(\Omega^m)$ and linear reconstruction maps $R_{(x^1,\dots,x^m)} \in \mathcal{L}(\mathbb{C}^m, V)$. Note that a deterministic sample can be viewed as a particular choice of random sample following a Dirac distribution in Ω^m, and therefore

$$\rho_m^{\mathrm{rand}}(\mathcal{K})_V \leqslant \rho_m(\mathcal{K})_V.$$

Sampling numbers may also be defined without imposing the linearity of R, leading to smaller quantities. In what follows, we shall establish upper bounds on the linear sampling numbers, which in turn are upper bounds for the nonlinear ones. We refer to [142] for an introduction and study of sampling numbers in the context of general linear measurements, and to [143, 144] that focus on point evaluations, also termed as *standard information*.

By optimizing the choice of the space V_n used in Theorem 5.1, we deduce that, for $V = L^2 = L^2(\Omega, \mu)$, the sampling numbers in the randomized setting are dominated by the Kolmogorov n-widths.

Corollary 5.9. *For any compact set $\mathcal{K} \subset L^2$, one has*

$$\rho_{Cn}^{\mathrm{rand}}(\mathcal{K})_{L^2} \leqslant K d_n(\mathcal{K})_{L^2}, \tag{5.9}$$

where C and K are the same constants as in Theorem 5.1.

Remark 5.10. The bound (5.9) cannot be attained with independent and identically distributed sampling points $x^1, \dots, x^m$. Indeed, consider the simple example, already evoked in [176], where $\Omega = [0,1]$, μ is the Lebesgue measure,

$$V_n = \left\{ \sum_{i=1}^n c_i \chi_{\left[\frac{i-1}{n}, \frac{i}{n}\right]} \ : \ (c_1, \dots, c_n) \in \mathbb{C}^n \right\}$$

is a space of piecewise constant functions, and $\mathcal{K} = \{u \in V_n \ : \ \|u\|_{L^\infty} \leqslant 1\}$. Then $\mathcal{K} \subset V_n$ so $d_n(\mathcal{K})_{L^2} = 0$, and an exact reconstruction $Ru = u$ is possible if and only if the sample contains at least one point in each interval $\left[\frac{i-1}{n}, \frac{i}{n}\right[$. Thus $\rho_n^{\mathrm{det}}(\mathcal{K})_{L^2} = 0$, but in the case of i.i.d measurements, m has to grow like $n \log n$ to ensure this constraint, due to the coupon collector's problem.

Remark 5.11. In [106], a result similar to Theorem 5.1 is obtained under the extra assumption of a uniform bound on $e_n(u)_{L^2}/e_{2n}(u)_{L^2}$, yielding the validity of (5.9) assuming a uniform bound on $d_n(\mathcal{K})_{L^2}/d_{2n}(\mathcal{K})_{L^2}$.

The recovery method used in [106] is not of least-square type, but rather an elaboration of the pseudo-spectral approach that would simply approximate the inner products $\langle u, \varphi_j \rangle = \int_\Omega u \, \overline{\varphi_j} \, d\mu$ by a quadrature, using a hierarchical approach introduced in [185].

Ideally, one would like a "worst case" or "uniform" version of Theorem 5.1, of the form

$$\rho_{Cn}(\mathcal{K})_{L^2} \leqslant K d_n(\mathcal{K})_{L^2}, \tag{5.10}$$

but it is easily seen that such an estimate cannot be expected for general compact sets of L^2, due to the fact that pointwise evaluations are not continuous in L^2 norm.

It is however possible to recover such uniform estimates by mitigating the non-achievable estimate (5.10) in various ways. One first approach, developed in [121, 168], gives an inequality similar to (5.10), with $d_n(\mathcal{K})_{L^2}$ replaced by $d_n(\mathcal{K})_{L^\infty}$. It is based on the following lemma, see Theorem 2.1 in [168], which we recall for comparison with our Lemma 5.2:

Lemma 5.12. *Assume that μ is a measure of finite mass $\mu(\Omega) < \infty$, that the constant functions belong to V_n, and that there exists a sample $\{x^1, \ldots, x^m\}$ and weights w_i such that the discrete norm*

$$\|v\|_m^2 = \frac{1}{m} \sum_{i=1}^m w_i |v(x^i)|^2$$

satisfies a framing

$$\beta^{-1} \|v\|_{L^2}^2 \leqslant \|v\|_m^2 \leqslant \alpha \|v\|_{L^2}^2, \quad v \in V_n. \tag{5.11}$$

Then

$$\|u - P_n^m u\|_{L^2} \leqslant \sqrt{\mu(\Omega)} \left(1 + \sqrt{\alpha\beta}\right) e_n(u)_{L^\infty},$$

where $e_n(u)_{L^\infty} = \min_{v \in V_n} \|u - v\|_{L^\infty}$.

Proof. For any $v \in L^2$, we have $\|v\|_{L^2}^2 \leqslant \mu(\Omega) \|v\|_{L^\infty}^2$, and as $1 \in V_n$,

$$\|v\|_m^2 \leqslant \|1\|_m^2 \, \|v\|_{L^\infty}^2 \leqslant \alpha \, \|1\|_{L^2}^2 \, \|v\|_{L^\infty}^2 = \alpha \, \mu(\Omega) \, \|v\|_{L^\infty}^2.$$

Hence

$$\begin{aligned} \|u - P_n^m u\|_{L^2} &\leqslant \|u - v\|_{L^2} + \|v - P_n^m u\|_{L^2} \\ &\leqslant \|u - v\|_{L^2} + \sqrt{\beta} \, \|v - P_n^m u\|_m \\ &\leqslant \|u - v\|_{L^2} + \sqrt{\beta} \, \|v - u\|_m \\ &\leqslant \left(\sqrt{\mu(\Omega)} + \sqrt{\alpha\beta\mu(\Omega)}\right) \|u - v\|_{L^\infty}, \end{aligned}$$

and we conclude by optimizing over $v \in V_n$. □

Remark 5.13. The assumption that the constant functions belong to the space V_n can be avoided by adding a constant term in the density, which results in bounded weights. The constant $\alpha\mu(\Omega)$ is then replaced by the bound on the weights, see [23].

Here, in contrast to the derivation of (5.5) in Lemma 5.2, one only uses the framing property (5.11), and does not need the condition $\mathbb{E}(\|v\|_m^2) \leqslant \alpha \|v\|_{L^2}^2$. For this reason, one may achieve the above objective with a simpler sparsification approach proposed in [24] and adapted in [121], which performs a greedy selection of the points x^i within the sample $\{x^1, \ldots, x^m\}$, and defines adapted weights w_i. If the initial sample satisfies

$$\frac{1}{2} I \preccurlyeq G_m \preccurlyeq \frac{3}{2} I,$$

then, for any $r > 1$ the selection algorithm produces a sample with at most rn points such that (5.11) holds with $\alpha = \frac{3}{2} \left(1 + 1/\sqrt{r}\right)^2$ and $\beta^{-1} = \frac{1}{2} \left(1 - 1/\sqrt{r}\right)^2$.

Optimizing the choice of V_n (but imposing that constant functions are contained in this space), this leads to the following comparison result between deterministic optimal recovery numbers in L^2 and n-widths in L^∞: for any compact set $\mathcal{K} \in \mathcal{C}(\Omega)$, one has

$$\rho_{cn}(\mathcal{K})_{L^2} \leqslant C\sqrt{\mu(\Omega)}\, d_{n-1}(\mathcal{K})_{L^\infty}, \tag{5.12}$$

where C depends on $c > 1$. For $c = 2$, one can take $C = 11$. We refer to [121, 168] where this type of result is established.

Another approach consists in making pointwise evaluations continuous by restriction to the case where $\mathcal{K} = B_H$ is the unit ball of a separable reproducing kernel Hilbert space $H \subset L^2$, and assuming that the sequence $(d_n(B_H)_{L^2})_{n\geqslant 1}$ is ℓ^2-summable. The following result from [137], also based on the sparsification techniques from [128], improves on a bound found in [113]

$$\rho_{Cn}(B_H)^2_{L^2} \leqslant K\frac{\log n}{n}\sum_{k\geqslant n} d_k(B_H)^2_{L^2}, \tag{5.13}$$

More general compact classes $\mathcal{K}$ of L^2, such that point evaluations are well defined on functions of $\mathcal{K}$, are considered in [114], where the following general result is established: if

$$d_n(\mathcal{K})^2_{L^2} \leqslant Cn^{-\alpha}\ln(n+1)^\beta, \quad n \in \mathbb{N},$$

for some $\alpha > 1$ and $\beta \in \mathbb{R}$, then

$$\rho_n(\mathcal{K})^2_{L^2} \leqslant C'n^{-\alpha}\ln(n+1)^{\beta+1}, \quad n \in \mathbb{N}. \tag{5.14}$$

In the above results, the additional logarithmic factor appears as a residual of the result obtained before sparsification, contrarily to the bounds (5.9) and (5.12), which do not explicitely depend on the size of the initial sample Y. This results in a gap of a factor $\log n$ between (5.13) or (5.14) and known lower bounds for $\rho_n(\mathcal{K})^2_{L^2}$, see [137].

5.5 Computational aspects

The various results (5.9), (5.12), (5.13), (5.14) ensure the existence of good sampling and reconstruction algorithms in various settings. We end by a discussion on the computational cost of these strategies.

For the weighted least-squares methods from Section 5.2, the most expensive step consists in assembling the matrix G_m as a sum of m matrices of size n, so the algorithmic complexity is of order $\mathcal{O}(mn^2) = \mathcal{O}(n^3\log n)$. Besides, to obtain $\mathbb{E}(G_m|E)$, this step may need to be repeated a few times, as explained in Remark 5.5, but this only affects the offline complexity by a small random factor.

Note that we assumed that an orthogonal basis $(\varphi_1, \ldots, \varphi_n)$ of V_n is explicitly known, which might not be the case for irregular domains Ω. However, under reasonable assumptions on Ω or V_n, one can compute an approximately orthogonal basis $(\widetilde{\varphi}_1, \ldots, \widetilde{\varphi}_n)$, either by performing a first discretization of Ω with a large number of points, or by using a hierarchical method on a sequence of nested spaces $V_1 \subset \cdots \subset V_n$, see [5, 17, 85, 132, 133] and [c]. These additional steps have complexities $\mathcal{O}(K_n n^3)$ and $\mathcal{O}(n^4)$ respectively, where K_n is the maximal value of the inverse Christoffel function $\frac{1}{n}\sum_{j=1}^n |\varphi_j|^2$, which might grow with n for certain choices of spaces V_n. Results similar to Lemma 5.4 have been obtained in the above references, with $(\varphi_j)_{1\leqslant j\leqslant n}$ replaced by $(\widetilde{\varphi}_j)_{1\leqslant j\leqslant n}$.

One could stop at this point and compute the approximation $\tilde{u} = \mathbb{E}(P_n^m u|E)$, which satisfies error bounds both in expectation when comparing to $e_n(u)_{L^2}$, see Lemma 5.4, or uniformly when comparing to $e_n(u)_{L^\infty}$, see Theorem 1 (iii) in [59]. Once the measurements of u are performed, the computation of $\tilde{u}$ requires to solve a $n \times n$ linear system as in Remark 5.6, so the online stage takes a time $\mathcal{O}(\tau n\log n + n^3)$, where τ is the cost of each measurement of u.

However, in applications where the evaluation cost τ becomes very high (for example when each evaluation $x \mapsto u(x)$ requires solving a PDE by some numerical code, or running a physical experiment), further reduction of the size of the sample may prove interesting, and justifies the interest for sparsification methods. The greedy

selection method from [24], which is used in [168] and leads to (5.12), has a complexity in $\mathcal{O}(mn^3) = \mathcal{O}(n^4 \log n)$, but it can only be applied to the worst-case setting, with the uniform error bound $e_n(u)_{L^\infty}$.

On the other hand, the iterative splitting method that we have used in this chapter following the ideas from [128, 137] is not easily implemented, and one obvious method consists in testing all partitions of $\{1, \dots, m\}$ into sets S_1 and S_2 when applying Lemma 5.7. Note that this lemma is in practice used L times, with $L = \mathcal{O}(\log \log n)$ since $2^L = \mathcal{O}(\frac{m}{n}) = \mathcal{O}(\log n)$. The algorithm consisting in subdividing the sample L times, each time checking that the Gram matrices corresponding to S_1 and S_2 are well conditioned, and keeping one such subset at random, thus has an exponential complexity $\mathcal{O}(2^m n^3) = \mathcal{O}(n^{cn})$. Having a different strategy that would produce the random sample in polynomial time is currently an open problem to us. Note that the hierarchical Monte-Carlo approaches from [106, 185] have similar optimal error bounds with an optimal sampling budget, and without exponential complexity in the generation of samples, however under the additional assumption that is described in Remark 5.11.

We summarise these computational observations in the following table, which illustrates the conflicts between reducing the sampling budget, ensuring optimal approximation results, and maintaining a reasonable cost for sample generation.

sampling algorithm	sample cardinality m	offline complexity	$\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2) \leqslant C e_n(u)_{L^2}^2$	$\|u - \tilde{u}\|_{L^2}^2 \leqslant C e_n(u)_{L^\infty}^2$
conditioned $\sigma^{\otimes m} \,\|\, E$	$10\, n \log(4n)$	$\mathcal{O}(n^3 \log n)$	✓	✓
+ deterministic sparsification [24]	$(1+\varepsilon)n$	$\mathcal{O}(n^4 \log n)$	✗	✓
+ random sparsification [128]	Cn	$\mathcal{O}(n^{cn}) \to \mathcal{O}(n^r)$?	✓	✓

As a final remark, let us emphasize that although the results presented in this chapter are mainly theorical and not practically satisfactory, due both to the computational complexity of the sparsification, and to the high values of the numerical constants C and K in Theorem 5.1, they provide some intuitive justification to the boosted least-squares methods presented in [85], which consist in removing points from the initial sample as long as the corresponding Gram matrix G_m remains well conditioned. For instance, Lemma 5.7 allows to keep splitting the sample even after L steps, if one still has a framing $\frac{1}{2} I \preccurlyeq G_J \preccurlyeq \frac{3}{2} I$ and a sufficiently large ratio $|J|/n$. Nevertheless, it would be of much interest to find a randomized version of [24] giving a bound of the form (5.9), since this would give algorithmic tractability, smaller values for C and K, and the possibility to balance these constants in Theorem 5.1.

Chapter 6

A sharp upper bound for sampling numbers in L^2

Abstract. For a class $\mathcal{K}$ of complex-valued functions on a set Ω, we denote by $\rho_n(\mathcal{K})_{L^2}$ its sampling numbers, i.e., the minimal worst-case error on $\mathcal{K}$, measured in L^2, that can be achieved with a recovery algorithm based on n function evaluations. We prove that there is a universal constant $c \in \mathbb{N}$ such that, if $\mathcal{K}$ is the unit ball of a separable reproducing kernel Hilbert space, then

$$\rho_{cn}(\mathcal{K})_{L^2}^2 \;\le\; \frac{1}{n} \sum_{k \geqslant n} d_k(\mathcal{K})_{L^2}^2,$$

where $d_k(\mathcal{K})_{L^2}$ are the Kolmogorov widths (or approximation numbers) of $\mathcal{K}$ in L^2. We also obtain similar upper bounds for more general classes $\mathcal{K}$, including all compact subsets of the space of continuous functions on a bounded domain $\Omega \subset \mathbb{R}^d$, and show that these bounds are sharp by providing examples where the converse inequality holds up to a constant. The results rely on the solution to the Kadison-Singer problem, which we extend to the subsampling of a sum of infinite rank-one matrices.

6.1 Introduction and main results

The general question of how well point-wise evaluations perform for approximating a function, which is often called *sampling recovery* or approximation using *standard information*, is a classical question in theoretical and applied mathematics. A historical treatment and various basics may be found in the monographs [60, 64, 67, 167, 189] for general approximation theory and in [142–144] for information-based complexity. It is of particular interest to compare the *power of function evaluations* with the power of optimal linear measurements (which could be Fourier coefficients or derivatives), since the latter are well understood in many cases and easier to handle from a theoretical point of view, while the first are of larger practical relevance. The quest for a systematic comparison has attracted much attention recently. We will describe the history and related results below after presenting the setting and the main results, see also Section 6.1.1.

The *power* of a given class of measurements is often expressed in terms of the minimal error achievable with a given amount of such information. Here, we consider L^2-approximation in a worst-case setting, so that these minimal errors correspond to sampling numbers and Kolmogorov (or approximation) numbers, as we summarize below.

Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $L^2 := L^2(\Omega, \mathcal{A}, \mu)$ be the space of square-integrable complex-valued functions on Ω. Let $\mathcal{K}$ be a set of functions contained in L^2. The *Kolmogorov widths* of $\mathcal{K}$ in L^2 are defined by

$$d_n(\mathcal{K})_{L^2} \;:=\; \inf_{\substack{\ell_1,\dots,\ell_n \colon \mathcal{K} \to \mathbb{C} \\ \phi_1,\dots,\phi_n \in L^2}} \sup_{u \in \mathcal{K}} \Big\| u - \sum_{j=1}^{n} \ell_j(u)\, \phi_j \Big\|_{L^2}.$$

This is the worst-case error of an optimal approximation within a linear space of dimension n. It coincides

with the nth approximation number (or linear width) of $\mathcal{K}$, which is the worst-case error of an optimal linear algorithm that uses at most n linear functionals as information, see Remark 6.5. On the other hand, the *sampling numbers* are given by

$$\rho_m(\mathcal{K})_{L^2} := \inf_{\substack{x^1,\dots,x^m \in \Omega \\ \psi_1,\dots,\psi_m \in L^2}} \sup_{u \in \mathcal{K}} \left\| u - \sum_{i=1}^m u(x^i)\, \psi_i \right\|_{L^2},$$

i.e., $\rho_m(\mathcal{K})_{L^2}$ is the minimal worst-case error of linear algorithms based on m function evaluations. Therefore, the task is to compare the numbers $d_n(\mathcal{K})_{L^2}$ and $\rho_m(\mathcal{K})_{L^2}$.

It is clear that we have $\rho_n(\mathcal{K})_{L^2} \geq d_n(\mathcal{K})_{L^2}$. Here, we aim for an upper bound on $\rho_m(\mathcal{K})_{L^2}$ in terms of the numbers $d_n(\mathcal{K})_{L^2}$. We first describe the situation where $\mathcal{K}$ is the unit ball of a separable reproducing kernel Hilbert space (RKHS). A priori, it is not clear whether such a bound is even possible. And indeed, there can be no such bound in the case that $(d_n(\mathcal{K})_{L^2}) \notin \ell^2$. More precisely, it is shown in [97] that for any non-negative and non-increasing sequence $(\sigma_n) \notin \ell^2$ and any sequence (τ_m) tending to infinity, e.g. $\tau_m = \ln \ln m$, there exists a RKHS with unit ball $\mathcal{K}$ such that $d_n(\mathcal{K})_{L^2} = \sigma_n$ for all n but $\limsup_{m\to\infty} \tau_m \cdot \rho_m(\mathcal{K})_{L^2} > 0$.

The situation is completely different when $(d_n(\mathcal{K})_{L^2}) \in \ell^2$, which is equivalent to assuming that the kernel K of the Hilbert space has finite trace

$$\int_\Omega K(x,x)\, d\mu(x) < \infty, \tag{6.1}$$

see, e.g., [135]. Under this assumption, first upper bounds on $\rho_m(\mathcal{K})_{L^2}$ in terms of the numbers $d_n(\mathcal{K})_{L^2}$ were obtained more than 20 years ago in [184]. These upper bounds were later improved in [113, 117, 137]. On the other hand, a lower bound from [95, Theorem 2] tells us how far these improvements might go: for every non-negative and non-increasing $(\sigma_n) \in \ell^2$, there exists a separable RKHS with unit ball $\mathcal{K}$ such that $d_n(\mathcal{K})_{L^2} = \sigma_n$ for all $n \in \mathbb{N}$ and

$$\rho_{\lfloor n/8 \rfloor}(\mathcal{K})_{L^2} \geqslant \sqrt{\frac{1}{n} \sum_{k \geqslant n} d_k(\mathcal{K})_{L^2}^2} \tag{6.2}$$

for infinitely many values of $n \in \mathbb{N}$. Anctually, it turns out that this is already the worst possible scenario. The main result of this chapter is an upper bound, which matches the above lower bound (6.2) up to a universal constant, and which is true for any separable reproducing kernel Hilbert space.

Theorem 6.1. *There is a universal constant $c \in \mathbb{N}$ such that the following holds. Let μ be a measure on a set Ω and let $\mathcal{K} \subset L^2(\mu)$ be the unit ball of a separable RKHS on Ω such that the finite trace assumption* (6.1) *holds. Then, for all $n \in \mathbb{N}$, we have*

$$\rho_{cn}(\mathcal{K})_{L^2} \leq \sqrt{\frac{1}{n} \sum_{k \geqslant n} d_k(\mathcal{K})_{L^2}^2}.$$

This settles the question on the power of standard information compared to general linear information for the problem of L^2-approximation on Hilbert spaces, and solves the open problems from [95, 113], Open Problem 140 in [144], as well as Outstanding Open Problem 1.4 in [67] for L^2-approximation. The latter is discussed in Example 6.28, where we consider tensor product spaces. We note that the case of L^p-approximation ($p \neq 2$) is widely open. A slightly stronger version of Theorem 6.1 and explicit constants are given in Theorem 6.23.

Let us add that, in principle, Theorem 6.1 does only imply the *existence* of (linear) sampling algorithms achieving the error bound. However, all upper bounds on $\rho_m(\mathcal{K})_{L^2}$ will be obtained by a suitable (unregularized) *least squares method*, see Remark 6.7 and Section 6.5.

Theorem 6.1 is a direct continuation of the series of works initiated in [113], in which the sampling numbers were bounded by

$$\rho_{\lfloor cn \ln n \rfloor}(\mathcal{K})_{L^2} \leq \sqrt{\frac{1}{n} \sum_{k \geqslant n} d_k(\mathcal{K})_{L^2}^2},$$

see also [104, 177], and an improvement from [137], where the logarithmic oversampling was removed in exchange for an additional factor $\sqrt{\ln n}$ on the right hand side.

The ingredients for the proof are still the existence of good point sets with $\mathcal{O}(n \ln n)$ points from [113], and a subsampling of $\mathcal{O}(n)$ points based on the solution to the Kadison-Singer problem [128]. The Kadison-Singer

subsampling has already been applied for the related problem of sampling discretization in [121] (see [105] for a survey) and was subsequently introduced to the study of sampling numbers in [137, 168]. In these papers, the subsampling was, roughly speaking, only performed for a finite-dimensional sub-problem which resulted in the excessive factor $\sqrt{\ln n}$ in [137]. The new ingredient here is an infinite-dimensional version of the subsampling theorem that might be of independent interest, see Proposition 6.17.

If we apply Theorem 6.1 and the lower bound from [95] to sequences with polynomial decay, we obtain the following characterization.

Corollary 6.2. *Let $\mathcal{K}$ be the unit ball of a separable RKHS with*

$$d_n(\mathcal{K})_{L^2} \lesssim n^{-\alpha} \ln^{-\beta} n \tag{6.3}$$

for some $\alpha \geq 1/2$, $\beta \in \mathbb{R}$ and $c > 0$. Then

$$\rho_m(\mathcal{K})_{L^2} \lesssim \begin{cases} m^{-\alpha} \ln^{-\beta} m & \text{if } \alpha > 1/2, \\ m^{-\alpha} \ln^{-\beta+1/2} m & \text{if } \alpha = 1/2 \text{ and } \beta > 1/2. \end{cases} \tag{6.4}$$

Moreover, there exist classes $\mathcal{K}$ such that these bounds are sharp.

Here, $a_n \lesssim b_n$ means that there is a constant $c > 0$ such that $a_n \leq c\, b_n$ for all but finitely many $n \in \mathbb{N}$; later we will also use the symbols $\gtrsim$ and $\asymp$, which are defined accordingly. It is clear from Theorem 6.1 that the hidden constant in (6.4) is given by the product of the hidden constant in (6.3) and a constant that only depends on α and β.

We now turn to general function classes $\mathcal{K}$ that are assumed to satisfy the following assumption.

Assumption A. Let $\mathcal{K}$ be a class of complex-valued functions on a set Ω and let μ be a measure on Ω. We say that $\mathcal{K}$ and μ satisfy Assumption A, if there is a metric on $\mathcal{K}$ such that $\mathcal{K}$ is continuously embedded into L^2, separable, and function evaluation $u \mapsto u(x)$ is, for each $x \in \Omega$, continuous on $\mathcal{K}$.

Note that Assumption A is satisfied, for example, if

— $\mathcal{K}$ is a separable subset of the space of bounded functions equipped with the maximum distance and the measure μ is finite, **or**

— $\mathcal{K}$ is the unit ball of a separable normed space that is continuously embedded in L^2 and on which function evaluation at each point is a continuous functional, **or**

— $\mathcal{K}$ is a countable set of square-integrable functions, equipped with the discrete metric.

In this setting, we prove the following bound.

Theorem 6.3. *Let $0 < p < 2$. There is a constant $c_p \in \mathbb{N}$, depending only on p, such that for any $\mathcal{K}$ and μ that satisfy Assumption A and all $n \in \mathbb{N}$,*

$$\rho_{c_p n}(\mathcal{K})_{L^2} \leq \left(\frac{1}{n} \sum_{k \geqslant n} d_k(\mathcal{K})_{L^2}^p \right)^{1/p}.$$

Theorem 6.3 is an improvement over [114], where again we removed the excessive logarithmic factor. We will also show that the result is not true for $p = 2$, see Example 6.31. However, we provide a variant of Theorem 6.3, under the weaker condition $\big((\ln n)^s d_n(\mathcal{K})_{L^2}\big) \in \ell^2$ for some $s > 1/2$, in Section 6.6.2. In Proposition 11 of the very recent paper [109], one can find a more general result encompassing Theorems 6.3 and 6.27. This leads to the following corollary.

Corollary 6.4. *Let $\mathcal{K}$ and μ satisfy Assumption A and*

$$d_n(\mathcal{K})_{L^2} \lesssim n^{-\alpha} \ln^{-\beta} n \tag{6.5}$$

for some $\alpha > 0$ and $\beta \in \mathbb{R}$. Then

$$\rho_m(\mathcal{K})_{L^2} \lesssim \begin{cases} m^{-\alpha} \ln^{-\beta} m & \text{if } \alpha > 1/2, \\ m^{-\alpha} \ln^{-\beta+1} m & \text{if } \alpha = 1/2 \text{ and } \beta > 1, \\ 1 & \text{otherwise.} \end{cases} \tag{6.6}$$

Moreover, there exist classes $\mathcal{K}$ such that these bounds are sharp.

Again, the hidden constant in (6.6) is given by the product of the hidden constant in (6.5) and a constant that only depends on α and β. The difference compared to unit balls of RKHSs is the case $\alpha = 1/2$, where we need $\beta > 1$ instead of $\beta > 1/2$, and lose a factor $\ln n$ instead of $\sqrt{\ln n}$, see Example 6.30. In addition, if $(d_n(\mathcal{K})_{L^2}) \notin \ell^2$, then $\rho_m(\mathcal{K})_{L^2}$ might be bounded below by a constant, opposite to the RKHS setting where $\rho_m(\mathcal{K})_{L^2}$ tends to zero as soon as $d_n(\mathcal{K})_{L^2}$ does, see [97]. However, for $\alpha > 1/2$, the results for general classes are just as strong as before.

6.1.1 Remarks and related literature

We want to add several remarks on the history of the result and related topics.

Remark 6.5 (Equivalent widths)**.** There are several quantities to measure the "width" of a set $\mathcal{K}$. Although we work here with the Kolmogorov numbers $d_n(\mathcal{K})_{L^2}$ as benchmark, let us add that these quantities coincide in L^2 with the *approximation numbers* of $\mathcal{K}$, i.e.

$$d_n(\mathcal{K})_{L^2} \;=\; a_n(\mathcal{K})_{L^2} \;:=\; \inf_{\substack{\ell_1,\dots,\ell_n\colon \mathcal{K}\to\mathbb{C} \text{ linear} \\ \phi_1,\dots,\phi_n\in L^2}} \sup_{u\in\mathcal{K}} \Big\| u - \sum_{j=1}^{n} \ell_j(u)\,\phi_j \Big\|_{L^2},$$

as the infimum in the definition of $d_n(\mathcal{K})_{L^2}$ for given $\phi_1,\dots,\phi_n$ is attained by the L^2-orthogonal projection onto their span, which is linear in any case. The approximation numbers of a class represent the worst-case error of an optimal linear algorithm that uses at most n linear functionals as information. If $\mathcal{K}$ is the unit ball of some Hilbert space H, then the approximation numbers agree with the *singular values* of the identity $\mathrm{Id}\colon H \to L^2$. In this case, the $d_n(\mathcal{K})_{L^2}$ also coincide with the *Gelfand n-widths* $g_n(\mathcal{K})_{L^2}$, which represent the minimal worst-case error of (possibly non-linear) algorithms based on n arbitrary linear functionals, see, e.g., Chapter 4 in [142].

Remark 6.6 (Extreme classes $\mathcal{K}$)**.** It is interesting to note that the lower bound (6.2) from [95] is attained for univariate Sobolev spaces of periodic functions. By Theorem 6.1, this means that these basic classes already represent the most difficult RKHSs for sampling recovery when the numbers $d_n(\mathcal{K})_{L^2}$ are fixed.

Remark 6.7 (Least squares methods)**.** The upper bounds in Theorem 6.1 and 6.3 are proved for a weighted least squares algorithm using samples from a set of $c\,n$ points that is subsampled from a set of $c\,n\ln n$ i.i.d. random points, see Section 6.5. Depending on the function class $\mathcal{K}$, the algorithm using the full set of random points may be constructive but the subsampling is based on an existence result from [128] and is therefore not constructive. It would be very interesting to make the subsampling constructive, see Remark 6.21.

Remark 6.8 (Spline algorithm)**.** Let $\mathcal{K}$ be the unit ball of a RKHS H. If we fix the sampling points $x^1, ..., x^m$, it is known that the smallest possible worst case error is achieved by the spline algorithm

$$S_m(u) \;:=\; \operatorname*{argmin}_{v\in H\colon\, v(x^i)=u(x^i)} \|v\|_H,$$

that is,

$$\inf_{\psi_1,\dots,\psi_m\in L^2} \sup_{u\in\mathcal{K}} \Big\| u - \sum_{i=1}^{m} u(x^i)\,\psi_i \Big\|_{L^2} \;=\; \sup_{u\in\mathcal{K}} \Big\| u - S_m(u) \Big\|_{L^2},$$

see e.g. [175, Theorem 5.1]. The function $S_m(u)$ is also known as the minimal norm interpolant and, by the famous *representer theorem*, can be expressed as a linear combination of the kernel functions $K(x^i,\cdot)$, see e.g. [182, Proposition 12.32]. Therefore, our upper bounds are true not only for the least squares algorithm, but also for the kernel-based approximation $S_m(u)$. Both types of algorithms are common in the context of learning, see e.g. the seminal paper [60].

Remark 6.9 (The power of i.i.d. sampling)**.** It is remarkable that, up to a logarithmic factor, the upper bound from Theorem 6.1 is achieved with high probability for i.i.d. random sampling points, see [113, 177]. In regard of the personal history of the authors DK and MU, Theorem 6.1 is a byproduct of a series of work on the power of i.i.d. sampling for approximation and integration problems that started in [93, 94] and was also continued in [98, 108, 111, 112].

Remark 6.10 (Expected error)**.** A different approach to L^2-approximation is by using randomized algorithms and taking the worst case expected error instead of a worst case deterministic error. The results in this randomized setting are quite different; the error of optimal algorithms does not depend on the tail of the sequence $(d_n(\mathcal{K})_{L^2})$. We refer to [59, 106, 122, 144, 185, d].

Remark 6.11 (Upper bounds for infinite trace)**.** We note that our bounds make sense also if $d_n(\mathcal{K})_{L^2}$ is infinite for small n, but they are useless if the *tail* of $(d_n(\mathcal{K})_{L^2})$ is not square-summable, which is the case, e.g., if $\mathcal{K}$ is the unit ball of a RKHS with infinite trace, see (6.1).
An alternative approach is to bound the numbers $\rho_m(\mathcal{K})_{L^2}$ by the Kolmogorov widths $d_n(\mathcal{K})_{L^\infty}$ in L^∞: it is shown in [168] that there is a universal constant $c \in \mathbb{N}$ such that $\rho_{cn}(\mathcal{K})_{L^2} \le c\, d_n(\mathcal{K})_{L^\infty}$ for probability spaces $(\Omega, \mathcal{A}, \mu)$. Although this bound is sometimes weaker than Theorem 6.3 (see Example 1 in [114]), it has the great advantage that it may be applied in situations where the Kolmogorov widths in L^2 are not square-summable, see, e.g., [171, 173]. It would be very interesting to see whether it is possible to unify the two approaches.

Remark 6.12 (Tractability)**.** Assume now that a whole sequence of classes $\mathcal{K}_d$ is given, where d could be the dimension of the underlying domain. For some classes we know that the curse of dimensionality is present, if only standard information (function values) is allowed, while the problem is tractable for general linear information, see e.g., [96, 141, 181]. However, since the constants from Theorems 6.1 and 6.3 are independent of the dimension, it is possible to transfer certain tractability properties from linear information to standard information [104, 110, 144].

Remark 6.13 (Separability of $\mathcal{K}$)**.** Contrarily to the ℓ^2-summability of the Kolmogorov widths, it should be possible to remove the separability assumption on the class $\mathcal{K}$, at least in Theorem 6.1, by adding a term $\mathrm{tr}_0(K)/n$ inside the square root in the right-hand side, as done in [135].

Remark 6.14 (Discretization of continuous frames)**.** A related problem is the question whether a *continuous frame* for a Hilbert space may be sampled to obtain a frame, see [51] for details. This problem, which was originally posed in the physics book [11], has only recently been solved in [73], see also the survey [38]. Although seemingly independent, this line of research uses remarkably similar methods. We leave it to future research to better understand and expand the connections.

6.1.2 Outline

The rest of the chapter can be outlined as follows. Sections 6.2–6.5 form the proof of Theorem 6.1. In Section 6.2, we collect some basics on the RKHS setting. In Section 6.3, we obtain our initial sample of $\mathcal{O}(n \ln n)$ points based on a concentration inequality for infinite matrices. The subsampling is performed in Section 6.4, which applies the solution to the Kadison-Singer problem in a slightly original way, leading to the core of the proof in Section 6.5. In Section 6.6, we prove our results for general function classes by constructing a suitable RKHS, on which a local version of Theorem 6.1 (Theorem 6.23) can be applied. Finally, in Section 6.7, we present examples, applying our result to tensor product problems and showing that our upper bounds are sharp.

6.2 Hilbert space setting

We first consider the case where $\mathcal{K}$ is the unit ball of a separable Hilbert space H with reproducing kernel $K \in \mathbb{C}^{\Omega\times\Omega}$. We refer to [135] and references therein for theoretical background on RKHSs.

Thanks to the finite trace assumption (6.1), we know that the identity map $\mathrm{Id}\colon H \to L^2$ is Hilbert-Schmidt, thus its left and right singular vectors $(\varphi_n)_{n\in\mathbb{I}}$ and $(\sigma_n\varphi_n)_{n\in\mathbb{I}}$ are orthonormal families in L^2 and H, respectively. Here, we only list the singular vectors with respect to the nonzero singular values $\sigma_n > 0$, and the index set is of the form $\mathbb{I} = \{n \in \mathbb{N}_0 \,:\, n < N\}$ with $N \in \mathbb{N} \cup \{\infty\}$. The singular vectors satisfy

$$\langle u, \varphi_n\rangle_{L^2} = \langle u, \sigma_n^2\varphi_n\rangle_H \quad \text{for all } u \in H \text{ and } n \in \mathbb{I}.$$

We use the convention that $\mathbb{N}_0 := \{0, 1, 2, \ldots\}$ and the singular values are arranged in a non-increasing order. In particular, $\sum_{k\in\mathbb{I}} \sigma_k^2 < \infty$ and the Kolmogorov width $d_n(\mathcal{K})_{L^2} = \sigma_n$ is attained by the L^2-orthogonal projection

P_n onto $V_n = \operatorname{span}\{\varphi_k \,:\, k < n\}$. Moreover, the separability of H ensures that the equality

$$K(x,y) = \sum_{n\in\mathbb{I}} \sigma_n^2 \, \varphi_n(x)\overline{\varphi_n(y)}$$

holds for all $x, y \in \Omega_0$ with some set $\Omega_0 \subset \Omega$ satisfying $\mu(\Omega \setminus \Omega_0) = 0$. We therefore have the identity

$$u(x) \,=\, \sum_{n\in\mathbb{I}} \langle u, \varphi_n\rangle_{L^2} \, \varphi_n(x) \quad \text{for all } u \in H \text{ and } x \in \Omega_0. \tag{6.7}$$

Our sampling points will be contained in the set Ω_0.

As a consequence of the following lemma, we only have to show the validity of Theorem 6.1 for all $1 \leqslant n < N$.

Lemma 6.15. *Let $N = \min\{n \in \mathbb{N} \,:\, d_n(\mathcal{K})_{L^2} = 0\} < \infty$. Then we have $\rho_m(\mathcal{K})_{L^2} = 0$ for all $m \geq N$.*

Proof. For $x \in \Omega_0$, we write $\varphi(x) = (\varphi_0(x), \dots, \varphi_{N-1}(x))$. Then there are points $x^0, \dots, x^{N-1} \in \Omega_0$ such that every $\varphi(x)$ is contained in the span of the vectors $\varphi(x^i)$. We write $\varphi(x) = \sum \psi_i(x)\varphi(x^i)$ with coefficients $\psi_i(x) \in \mathbb{C}$. By (6.7), we have

$$u(x) = \sum_{n<N} \langle u, \varphi_n\rangle_{L^2} \sum_{i<N} \psi_i(x)\varphi_n(x^i) = \sum_{i<N} u(x^i)\psi_i(x),$$

for all $x \in \Omega_0$ and $u \in H$. Thus, the identity $u = \sum u(x^i)\psi_i$ holds almost everywhere. Moreover, the functions $\varphi_0, \dots, \varphi_{N-1}$ restricted to Ω_0 form a basis of $\operatorname{span}\{\psi_i \,:\, i < N\}$, and thus $\psi_i \in L^2$. □

We fix an integer $1 \leq n < N$ for the rest of the proof of Theorem 6.1.

6.3 Concentration inequality

As proposed in [113] and applied in [104, 114, 135, 137, 177], we define the probability density

$$\kappa_n(x) = \frac{1}{2}\left(\frac{1}{n}\sum_{k<n} |\varphi_k(x)|^2 + \frac{\sum_{k\geqslant n} \sigma_k^2 \, |\varphi_k(x)|^2}{\sum_{k\geqslant n} \sigma_k^2}\right),$$

and draw i.i.d. random points $x^1, \dots, x^m \in \Omega$ according to this density. We define the N-dimensional vectors $a_1, \dots, a_m$ by

$$(a_i)_k = \begin{cases} \kappa_n(x^i)^{-1/2} \, \varphi_k(x^i) & \text{if } 0 \leq k < n, \\ \kappa_n(x^i)^{-1/2} \, \gamma_n^{-1} \, \sigma_k \, \varphi_k(x^i) & \text{if } n \leqslant k < N, \end{cases}$$

where

$$\gamma_n \,:=\, \max\left\{\sigma_n, \sqrt{\frac{1}{n}\sum_{k\geqslant n} \sigma_k^2}\right\} > 0.$$

Note that $\kappa_n(x^i) > 0$ almost surely. It follows from these definitions that $a_i \in \ell^2(\mathbb{I})$ with

$$\|a_i\|_2^2 = \kappa_n(x^i)^{-1}\left(\sum_{k<n} |\varphi_k(x^i)|^2 + \gamma_n^{-2}\sum_{k\geqslant n} \sigma_k^2 \, |\varphi_k(x^i)|^2\right) \leqslant 2n,$$

and

$$\mathbb{E}(a_i a_i^\dagger) = \operatorname{diag}(1, \dots, 1, \sigma_n^2/\gamma_n^2, \sigma_{n+1}^2/\gamma_n^2, \dots) =: E,$$

with $\|E\|_{2\to 2} = 1$ since $\sigma_k^2/\gamma_n^2 \leqslant 1$ for $k \geqslant n$. Here, $\operatorname{diag}(v)$ denotes a diagonal matrix with diagonal v, and $\|\cdot\|_{2\to 2}$ denotes the spectral norm of a matrix.

We apply the following concentration inequality for infinite matrices, which was proved by Mendelson and Pajor in [130, Theorem 2.1]. We use a version of this result from [135, Theorem 1.1] and [137, Theorem 5.3].

Lemma 6.16. *Let $m \geqslant 3$ and $a_1, \dots, a_m$ be i.i.d. random sequences from $\ell^2(\mathbb{I})$ satisfying $\|a_i\|_2^2 \leqslant 2n$ almost surely and $\|E\|_{2\to 2} \leqslant 1$, with $E = \mathbb{E}(a_i a_i^\dagger)$. Then, for $0 \leq t \leq 1$,*

$$\mathbb{P}\left(\left\| \frac{1}{m} \sum_{i=1}^m a_i a_i^\dagger - E \right\|_{2\to 2} > t \right) \leqslant 2^{3/4}\, m \exp\left(-\frac{mt^2}{42\, n} \right).$$

For $t = 1/2$, this probability is less than $1/2$ as soon as $\frac{m}{\ln(4m)} \geqslant 168\, n$. In the sequel we take

$$m = \lfloor C_0\, n \ln(n+1) \rfloor,$$

with C_0 large enough, so that the previous inequality holds true. (One can take $C_0 = 10^4$, for instance.) Thanks to Lemma 6.16, we know that there exists a realization $x^1, \dots, x^m \in \Omega_0$ of the random sampling such that the corresponding family $a_1, \dots, a_m$ satisfies

$$\left\| \frac{1}{m} \sum_{i=1}^m a_i a_i^\dagger - E \right\|_{2\to 2} \leqslant \frac{1}{2}. \tag{6.8}$$

We fix such a sequence for the rest of the proof of Theorem 6.1.

6.4 Subsampling of infinite vectors

We now want to apply the solution to the Kadison-Singer problem, or specifically to Weaver's conjecture, to the sum of rank-one matrices

$$\frac{1}{m} \sum_{i=1}^m a_i a_i^\dagger,$$

in order to find a subsampling of order n preserving the spectral properties of the sum. The original result comes from the celebrated paper [128] by Marcus, Spielman and Srivastava, and has already been applied numerous times in approximation theory, see for instance [106, 113, 114, 135, 137, 140, 168, d]. However, the original subsampling strategy only works for finite matrices. The main result of this section is the following infinite-dimensional variant, which might be of independent interest.

Proposition 6.17. *There are absolute constants $c_1 \leqslant 43200$, $c_2 \geqslant 50$, $c_3 \leqslant 21600$, with the following properties. Let $m, n \in \mathbb{N}$ and $a_1, \dots, a_m$ be vectors from $\ell^2(\mathbb{N}_0)$ satisfying $\|a_i\|_2^2 \leqslant 2n$ and*

$$\left\| \frac{1}{m} \sum_{i=1}^m a_i a_i^\dagger - \begin{pmatrix} I_n & 0 \\ 0 & \Lambda \end{pmatrix} \right\|_{2\to 2} \leq \frac{1}{2}, \tag{6.9}$$

for some Hermitian matrix Λ with $\|\Lambda\|_{2\to 2} \leqslant 1$, where $I_n \in \mathbb{C}^{n\times n}$ denotes the identity. Then, there is a subset $J \subset \{1, \dots, m\}$ with $|J| \leq c_1 n$, such that

$$\left(\frac{1}{n} \sum_{i\in J} a_i a_i^\dagger \right)_{<n} \succcurlyeq c_2\, I_n \qquad \text{and} \qquad \frac{1}{n} \sum_{i\in J} a_i a_i^\dagger \preccurlyeq c_3\, I,$$

where $A_{<n} := (A_{k,l})_{k,l<n}$ and $A \preccurlyeq B$ denotes the Loewner order of Hermitian matrices A and B.

The conclusion can be understood as an upper bound on the largest eigenvalue of $A = \sum_{i\in J} a_i a_i^\dagger$ and a lower bound on the smallest eigenvalue of $A_{<n}$. Note that the constants in Proposition 6.17, and hence also the final sampling size, are independent of m, the original sampling size. The rest of this section is devoted to the proof of this proposition.

6.4.1 Reduction to finite dimension

Let U_0 be a matrix whose columns form an orthonormal basis of

$$\operatorname{span}\{(a_i)_{\geqslant n} \,:\, i = 1, \dots, m\} \subset \ell^2,$$

where $(a_i)_{\geq n} = ((a_i)_k)_{k \geq n}$. Clearly, U_0 has at most m columns. Then we have that $U_0^\dagger U_0$ is the identity matrix and in particular the spectral norm of U_0 and $U_0^\dagger$ equals one. We set

$$U = \begin{pmatrix} I_n & 0 \\ 0 & U_0 \end{pmatrix},$$

which is a matrix that satisfies $U^\dagger U = I_p$, where $p \leq m + n$, and therefore also U and $U^\dagger$ have unit norm. We choose vectors $b_i \in \mathbb{C}^p$ that satisfy $Ub_i = a_i$ for all $i \leq m$. Such vectors exist since a_i is contained in the span of the columns of U. Then we also have $b_i = U^\dagger U b_i = U^\dagger a_i$.

Let $E = \begin{pmatrix} I_n & 0 \\ 0 & \Lambda \end{pmatrix}$ be the matrix from Proposition 6.17. We define

$$\hat{E} = U^\dagger \, E \, U = \begin{pmatrix} I_n & 0 \\ 0 & E' \end{pmatrix} \quad \text{where} \quad \|E'\|_{2\to 2} \leq \|E\|_{2\to 2} \leqslant 1.$$

With the norm bounds on U and $U^\dagger$, equation (6.9) gives

$$\Big\| \frac{1}{m} \sum_{i=1}^m b_i b_i^\dagger - \hat{E} \Big\|_{2\to 2} = \Big\| U^\dagger \left(\frac{1}{m} \sum_{i=1}^m a_i a_i^\dagger - E \right) U \Big\|_{2\to 2} \leqslant \frac{1}{2}.$$

6.4.2 Approximating the identity

In addition to finite dimension, the result from [128] requires the matrix $\frac{1}{m}\sum_{i=1}^m b_i b_i^\dagger$ to be close to the identity in spectral norm, and this is not ensured here. To circumvent this obstacle, we artificially add rank-one matrices $b_i b_i^\dagger \in \mathbb{C}^{p\times p}$ for $i = m+1, \dots, q$ in the following way.

As $I_p - \hat{E}$ is positive semi-definite, we can decompose it as a sum of rank-one matrices

$$I_p - \hat{E} = \begin{pmatrix} 0 & 0 \\ 0 & I_{p-n} - E' \end{pmatrix} = \sum_{j=1}^{p-n} t_j t_j^\dagger,$$

where $t_j \in \mathbb{C}^p$. We now choose

$$b_i = \sqrt{\frac{m}{m_{j(i)}}}\, t_{j(i)}, \quad m_j = \left\lceil \frac{m}{2n} \, \|t_j\|_2^2 \right\rceil,$$

with $j(i) \in \{1, \dots, p-n\}$ such that $\{b_i \,:\, i = m+1, \dots, q\}$ contains exactly m_j copies of each $\sqrt{m/m_j}\, t_j$. In this way, for $i > m$, the first n entries of b_i are zero since this is true of the t_j,

$$\|b_i\|_2^2 = \frac{m}{m_{j(i)}} \, \|t_{j(i)}\|_2^2 \leqslant 2n,$$

and

$$\begin{aligned} \Big\| \frac{1}{m} \sum_{i=1}^q b_i b_i^\dagger - I_p \Big\|_{2\to 2} &= \Big\| \frac{1}{m} \sum_{i=1}^m b_i b_i^\dagger + \sum_{j=1}^{p-n} t_j t_j^\dagger - I_p \Big\|_{2\to 2} \\ &= \Big\| \frac{1}{m} \sum_{i=1}^m b_i b_i^\dagger - \hat{E} \Big\|_{2\to 2} \leqslant \frac{1}{2}. \end{aligned}$$

Remark 6.18. As $\|t_j\|_2^2 \leqslant \|I_p - \hat{E}\|_{2\to 2} \leqslant 1$, we count

$$q = m + \sum_{j=1}^{p-n} m_j \leqslant m + \sum_{j=1}^{p-n} \Big(1 + \frac{m}{2n}\Big) \leqslant m + (p-n)\frac{m}{n} = \frac{mp}{n}.$$

Conversely, taking traces in $\mathbb{C}^{p\times p}$, we find

$$\frac{p}{2} = \mathrm{Tr}\left(\frac{1}{2}\, I_p\right) \leqslant \mathrm{Tr}\left(\frac{1}{m}\sum_{i=1}^{q} b_i b_i^\dagger\right) = \frac{1}{m}\sum_{i=1}^{q} \|b_i\|_2^2 \leqslant \frac{2nq}{m}.$$

So, we obtain $m/n \geqslant q/p \geq m/4n$. Recall that, given n the dimension of the approximation space V_n, we took $m = \mathcal{O}(n \ln n)$ initial sample points, and vectors b_i of size $p = \mathcal{O}(n \ln n)$. Hence, the number of such vectors is $q = \mathcal{O}(n \ln^2 n)$. Surprisingly, we do not use estimates on p and q in the rest of the argument.

Remark 6.19. In fact, we did not need an exponential speed of convergence in the concentration inequality. The reduction of the sample size to $\mathcal{O}(n)$ points works for any initial set of sampling points satisfying (6.8). If the cardinality of the initial sample is $m = n\,\ell(n)$, where $\ell(n)$ is any positive function of n, we get $p = \mathcal{O}(n\,\ell(n))$ and $q = \mathcal{O}(n\,\ell(n)^2)$.

6.4.3 Reduction of the sample size

We can now use the Kadison-Singer solution from [128] in an iterated way, as proposed in Lemma 3 of [140], and later used in [113, 114, 121, 135, 137, 168, d]. The following lemma is obtained from Corollary B and Lemma 1 in [140].

Lemma 6.20. *Let $b_1, \dots, b_q \in \mathbb{C}^p$ with $\|b_i\|_2^2 \leq \delta$ and*

$$\alpha I_p \preccurlyeq \sum_{i=1}^{q} b_i b_i^\dagger \preccurlyeq \beta I_p$$

for some $\beta \geq \alpha > 100\,\delta > 0$. Then there is a partition of $\{1, \dots, q\}$ into sets $J_1, \dots, J_t$ such that, for all $s \leq t$, we have

$$25\,\delta I_p \preccurlyeq \sum_{i\in J_s} b_i b_i^\dagger \preccurlyeq 3600\,\frac{\beta}{\alpha}\,\delta I_p.$$

Proof. Since the matrix $M = \sum_{i=1}^q b_i b_i^\dagger$ is positive, we may define $\tilde{b}_i = M^{-1/2} b_i$. Then we have $\sum_{i=1}^q \tilde{b}_i \tilde{b}_i^\dagger = I_p$ and $\|\tilde{b}_i\|_2^2 \leq \delta/\alpha =: \delta' < 1/100$. By Corollary B and Lemma 1 in [140], noting that the constant C from Lemma 1 is at most 36, we get a partition of $\{1, \dots, q\}$ into sets $J_1, \dots, J_t$ such that, for all $s \leq t$, we have

$$25\,\delta'\, I_p \preccurlyeq \sum_{i\in J_s} \tilde{b}_i \tilde{b}_i^\dagger \preccurlyeq 3600\,\delta'\, I_p.$$

Now, using

$$\sum_{i\in J_s} b_i b_i^\dagger = M^{1/2} \sum_{i\in J_s} \tilde{b}_i \tilde{b}_i^\dagger\, M^{1/2},$$

we get the statement. □

Note that one could obtain better constants by adapting the proof of Theorem 2.3 from [137]. In our case, we have $\delta = 2n$, $\alpha = m/2$ and $\beta = 3m/2$. The relation $\alpha > 100\,\delta$ is satisfied. We thus obtain

$$50\,n I_p \preccurlyeq \sum_{i\in J_s} b_i b_i^\dagger \preccurlyeq 21600\,n I_p$$

for every J_s from the partition. Moreover, the inequality

$$\frac{m}{2} I_p \preccurlyeq \sum_{i=1}^{q} b_i b_i^\dagger = \sum_{s=1}^{t} \sum_{i \in J_s} b_i b_i^\dagger \preccurlyeq 21600\, tn I_p$$

implies that one of the sets $J' = J_s$ from the partition must satisfy

$$|J' \cap \{1, \dots, m\}| \leq \frac{m}{t} \leq 43200\, n.$$

After applying Lemma 6.20 and removing the indices from $J' \cap \{m+1, \dots, q\}$ corresponding to artificially added vectors, we are left with a set $J := J' \cap \{1, \dots, m\}$ of cardinality

$$|J| \leq 43200\, n.$$

It remains to show that the artificial vectors do not interfere with our desired properties. For this, recall that $(b_i)_k = (a_i)_k$ for $k < n$ and $i \leqslant m$, whereas the first n entries of $b_i \in \mathbb{C}^p$ are zero for $i > m$. Hence,

$$\left(\sum_{i \in J} a_i a_i^\dagger \right)_{<n} = \left(\sum_{i \in J'} b_i b_i^\dagger \right)_{<n} \succcurlyeq 50\, n\, I_n,$$

where we use a simple linear algebra fact on self-adjoint matrices A:

$$\lambda_{\min}(A_{<n}) = \inf_{\substack{z \in \mathbb{C}^p,\, \|z\|_2 = 1 \\ z_k = 0 \text{ for } k \geqslant n}} z^\dagger A z \geqslant \inf_{z \in \mathbb{C}^p,\, \|z\|_2 = 1} z^\dagger A z = \lambda_{\min}(A).$$

Similarly, and using positive definiteness, we have

$$\sum_{i \in J} b_i b_i^\dagger \preccurlyeq \sum_{i \in J'} b_i b_i^\dagger \preccurlyeq 21600\, n\, I_p.$$

With the orthogonal transformation U from Section 6.4.1, we get

$$\Big\| \sum_{i \in J} a_i a_i^\dagger \Big\|_{2 \to 2} = \Big\| U \left(\sum_{i \in J} b_i b_i^\dagger \right) U^\dagger \Big\|_{2 \to 2} \leq \Big\| \sum_{i \in J} b_i b_i^\dagger \Big\|_{2 \to 2} \leqslant 21600\, n.$$

This proves Proposition 6.17.

□

Remark 6.21. It would be an interesting improvement to use the result of Batson, Spielman and Srivastava, see [24], instead of [128] for the subsampling. This earlier paper is applied to approximation theory in e.g. [121, 139, 170] and more recently in [23]. It presents a slightly less powerful method, requiring additional weights, but comes with an almost linear algorithmic complexity, see [119], and much smaller constants, which could make the bound presented here sharp also in terms of numerical values. Another approach would consist in using randomized sampling strategies similar to the early work [164], but with correlated inputs, which aim at avoiding the subsequent logarithmic oversampling. This optimality gap has for instance been reduced to $O(\log(n)/\log(\log(n)))$ in [91].

Remark 6.22. We recently learned that it might be possible to use results from [73], which work directly in an infinite-dimensional setting, to avoid the reduction to a finite dimension in § 6.4.1. However, as the core of our method is [128], we decided to keep our more direct deduction.

6.5 Proof of the main theorem

We now have all the tools for proving Theorem 6.1. To obtain our sampling points, we combine (6.8) for our initial vectors $a_i \in \ell^2(\mathbb{I})$ with Proposition 6.17. Clearly, Proposition 6.17 stays true if we replace $\mathbb{N}_0$ by the

possibly finite index set $\mathbb{I}$. We obtain points $x^1, \dots, x^m \in \Omega_0$ with $m \le 43200\, n$ such that the vectors

$$(a_i)_k = \begin{cases} \kappa_n(x^i)^{-1/2}\, \varphi_k(x^i) & \text{if } 0 \le k < n, \\ \kappa_n(x^i)^{-1/2}\, \gamma_n^{-1}\, \sigma_k\, \varphi_k(x^i) & \text{if } n \leqslant k < N, \end{cases}$$

satisfy

$$\left(\sum_{i=1}^m a_i a_i^\dagger\right)_{<n} \succcurlyeq 50\, n\, I \quad \text{and} \quad \left(\sum_{i=1}^m a_i a_i^\dagger\right)_{\ge n} \preccurlyeq 21600\, n\, I,$$

where we use the notation $A_{\ge n} = (A_{k,l})_{k,l \ge n}$ for a matrix A.

As in earlier papers, we use the *weighted least squares estimator*

$$P_n^m u := \operatorname*{argmin}_{v \in V_n} \sum_{i=1}^m \frac{|u(x^i) - v(x^i)|^2}{\kappa_n(x^i)}$$

with V_n and κ_n as defined in Sections 6.2 and 6.3, respectively, see [113]. This algorithm may be written as

$$P_n^m u = \sum_{k=1}^n (D^+ F u)_k\, \varphi_k$$

where $F\colon \mathcal{K} \to \mathbb{C}^m$ with $F(u) := \big(\kappa_n(x^i)^{-1/2} u(x^i)\big)_{i \leqslant m}$ is the *information mapping* and $D^+ \in \mathbb{C}^{n \times m}$ is the Moore-Penrose inverse of the *design matrix*

$$D := \Big(\kappa_n(x^i)^{-1/2}\, \varphi_k(x^i)\Big)_{i \le m,\, k \le n} \in \mathbb{C}^{m \times n}.$$

Since we have the identity $\overline{D^\dagger D} = (\sum_{i=1}^m a_i a_i^\dagger)_{<n}$, the matrix D has full rank and the spectral norm of D^+ is bounded by $(50n)^{-1/2}$. In particular, the argmin in the definition of P_n^m is uniquely defined and P_n^m satisfies $P_n^m u = u$ for all $u \in V_n$.
Denoting with Q_n the L^2-orthogonal projection onto $\operatorname{span}\{\varphi_k \,:\, k \ge n\}$, we obtain for any $u \in H$ that

$$\begin{aligned} \|u - P_n^m u\|_{L^2}^2 &= \|u - P_n u\|_{L^2}^2 + \|P_n u - P_n^m u\|_{L^2}^2 \\ &= \|Q_n u\|_{L^2}^2 + \|P_n^m (u - P_n u)\|_{L^2}^2 \\ &= \|Q_n u\|_{L^2}^2 + \big\|D^+ F(u - P_n u)\big\|_{\ell^2(\mathbb{C}^n)}^2 \\ &\le \sigma_n^2\, \|Q_n u\|_H^2 + \big\|D^+\big\|_{2\to 2}^2 \cdot \|F(u - P_n u)\|_{\ell^2(\mathbb{C}^m)}^2 . \end{aligned}$$

By (6.7) we have $F(u - P_n u) = \Phi \xi_u$, where

$$\Phi = \Big(\kappa_n(x^i)^{-1/2} \sigma_k\, \varphi_k(x^i)\Big)_{i \le m,\, k \geqslant n} \quad \text{and} \quad \xi_u = (\langle u, \sigma_k\, \varphi_k \rangle_H)_{k \geqslant n}.$$

The matrix Φ satisfies

$$\overline{\Phi^* \Phi} = \gamma_n^2 \left(\sum_{i=1}^m a_i a_i^\dagger\right)_{\ge n}$$

and therefore its spectral norm is bounded by $\big(21600\, n\, \gamma_n^2\big)^{1/2}$. Thus,

$$\|F(u - P_n u)\|_{\ell^2(\mathbb{C}^m)}^2 \le 21600\, n\, \gamma_n^2 \|\xi_u\|_2^2 = 21600\, n\, \gamma_n^2\, \|Q_n u\|_H^2.$$

In summary, as $1 + 21600/50 = 433$, we obtain for all $1 \le n < N$ the bound

$$\|u - P_n^m u\|_{L^2}^2 \leqslant 433 \max\left\{\sigma_n^2,\, \frac{1}{n} \sum_{k \geqslant n} \sigma_k^2\right\} \|Q_n u\|_H^2 . \tag{6.10}$$

for all $u \in H$ and some $m \leq 43200\, n$. Taking the supremum over $u \in \mathcal{K}$ and using that

$$\max\left\{\sigma_n^2,\, \frac{1}{n}\sum_{k\geqslant n}\sigma_k^2\right\} \,\leq\, \frac{2}{n}\sum_{k\geqslant \lceil n/2\rceil}\sigma_k^2,$$

we obtain

$$\rho_{43200\, n}(\mathcal{K})_{L^2}^2 \,\leq\, \frac{866}{n}\sum_{k\geqslant \lceil n/2\rceil}\sigma_k^2.$$

This finishes the proof of Theorem 6.1 with $c = 43200 \cdot 866$.

□

In fact, equation (6.10) provides a local upper bound which is sometimes superior to Theorem 6.1. We therefore state it separately.

Theorem 6.23. *Let μ be a measure on a set Ω and let $\mathcal{K} \subset L^2(\mu)$ be the unit ball of a separable RKHS H such that the finite trace assumption* (6.1) *holds. For $n \in \mathbb{N}$, let P_n be the orthogonal projection onto the span V_n of the singular vectors corresponding to the n largest singular values of the embedding of H into L^2. Then there exist $x^1, \ldots, x^m \in \Omega$ and $\psi_1, \ldots, \psi_m \in V_n$, where $m \leq 43200\, n$, such that, for all $u \in H$,*

$$\Big\|u - \sum_{i=1}^{m} u(x^i)\,\psi_i\Big\|_{L^2}^2 \leqslant 433 \max\left\{d_n(\mathcal{K})_{L^2}^2,\, \frac{1}{n}\sum_{k\geqslant n} d_k(\mathcal{K})_{L^2}^2\right\}\|u - P_n u\|_H^2.$$

Remark 6.24. For the purpose of Theorem 6.1 it was enough to bound $\|u - P_n u\|_H \leq \|u\|_H$. However, Theorem 6.23 will be of advantage later for the study of general classes since it is able to see additional decay of the Fourier coefficients $\langle u, \varphi_k\rangle_{L^2}$ compared to the decay implied by $u \in H$. Note that faster decay of the Fourier coefficients often corresponds to higher smoothness of the function. In a certain sense, this means that the algorithm is universal. The error has the optimal rate of decay for any smoothness higher than the smoothness of H.

Remark 6.25. The condition on the point sets can also be given by finite matrices that are related to the kernel K of the Hilbert space. For this, let us define $k_n(x,y) := \frac{1}{n}\sum_{k<n}\varphi_k(x)\varphi_k(y)$, and $r_n(x,y) := \frac{1}{n}\sum_{k\geq n}\sigma_k^2\,\varphi_k(x)\varphi_k(y)$. The non-zero singular values of $DD^\dagger$ are the same as those of $D^\dagger D$, and the non-zero singular values of $\Phi\Phi^\dagger$ are the same as those of $\Phi^\dagger\Phi$, where D and Φ are from above. Hence, the algorithm P_n^m based on points $x^1, \ldots, x^m$ satisfies the error bound above (up to a constant) if

$$c \,\leq\, \frac{1}{n}\,\lambda_n\left(DD^\dagger\right) \,=\, \lambda_n\left(\left(\frac{k_n(x^i,x^j)}{\sqrt{\kappa_n(x^i)\kappa_n(x^j)}}\right)_{1\leqslant i,j\leqslant m}\right)$$

and

$$\left(\frac{r_n(x^i,x^j)}{\sqrt{\kappa_n(x^i)\kappa_n(x^j)}}\right)_{1\leqslant i,j\leqslant m} \,=\, \frac{1}{n}\,\Phi\Phi^\dagger \,\preccurlyeq\, C\,\gamma_n^2\, I$$

for some constants $c, C > 0$, where λ_n denotes the nth eigenvalue. It would be interesting to find a property that only involves the kernel K directly (instead of the truncated kernels k_n and r_n above), or to verify that a similar property characterizes *good* point sets, in a way similar to Proposition 1 of [96] for integration.

6.5.1 Proof of Corollary 6.2

For the given bounds on the sampling numbers for sequences of polynomial decay, we only need to note that

$$\frac{1}{n}\sum_{k\geq n} k^{-\alpha}\ln^{-\beta} k \,\lesssim\, \begin{cases} n^{-\alpha}\ln^{-\beta} n & \text{if } \alpha > 1,\ \beta \in \mathbb{R}, \\ n^{-\alpha}\ln^{-\beta+1} n & \text{if } \alpha = 1,\ \beta > 1. \end{cases}$$

Hence, Corollary 6.2 immediately follows from Theorem 6.1, and the existence of $\mathcal{K}$ where the bounds are attained comes from (6.2), see [95]. □

6.6 General function classes

We now prove all results related to general function classes.

6.6.1 Proof of Theorem 6.3

We will make use of the following observation from [114, Lemma 3]. We copy its proof for completeness.

Lemma 6.26. *Let $\mathcal{K} \subset L^2$ and let L^2 be infinite-dimensional. There is an orthonormal system $(\varphi_k)_{k\in\mathbb{N}_0}$ in L^2 such that for all $n \geq 1$, the orthogonal projection P_n onto $V_n = \operatorname{span}\{\varphi_k \,:\, k < n\}$ satisfies*

$$\sup_{u\in\mathcal{K}} \|u - P_n u\|_{L^2} \;\leq\; 2\, d_{\lfloor n/4 \rfloor}(\mathcal{K})_{L^2}. \tag{6.11}$$

Proof. Clearly it is enough to find an increasing sequence of subspaces of L^2,

$$U_1 \subseteq U_2 \subseteq U_3 \subseteq \ldots, \qquad \dim(U_n) \leq n,$$

such that the projection P_n onto U_n satisfies (6.11). By the definition of $d_k(\mathcal{K})_{L^2}$, $k \in \mathbb{N}_0$, there is a subspace $W_k \subset L^2$ of dimension k and a mapping $T_k \colon \mathcal{K} \to W_k$ such that

$$\sup_{u\in\mathcal{K}} \|u - T_k u\|_{L^2} \;\leq\; 2\, d_k(\mathcal{K})_{L^2}.$$

This is also true if $d_k(\mathcal{K})_{L^2} = 0$. We let U_n be the space that is spanned by the union of the spaces W_{2^ℓ} over all $\ell \in \mathbb{N}_0$ such that $2^\ell \leq n/2$. Note that U_n contains a subspace W_k with $k \geq \lfloor n/4 \rfloor$. Therefore, $P_n u$ is at least as close to u as $T_k u$ for some $k \geq \lfloor n/4 \rfloor$, which implies (6.11). □

We now turn to the proof of Theorem 6.3. The basic idea is to construct a suitable reproducing kernel Hilbert space H that contains a dense subset of $\mathcal{K}$ and apply Theorem 6.23 to this Hilbert space. It will be important to use the local bound from Theorem 6.23 instead of the global bound from Theorem 6.1.

Proof of Theorem 6.3. Without loss of generality, we assume that L^2 is infinite-dimensional. Moreover, we assume that $d_k(\mathcal{K})_{L^2}$ is finite for $k \geq k_0$ and that $(d_k(\mathcal{K})_{L^2})_{k\geq k_0} \in \ell^p$. Otherwise, the statement is trivial.

By Lemma 6.26, there is an orthonormal system $(\varphi_k)_{k\in\mathbb{N}_0}$ such that (6.11) is satisfied for all $n \in \mathbb{N}$. We will consider φ_k as a function, where we fix an arbitrary representer from the equivalence class in L^2. We call

$$\hat{u}(k) := \langle u, \varphi_k \rangle_{L^2}$$

the kth Fourier coefficient of u. Moreover, we fix a countable dense subset $\mathcal{K}_0$ of $\mathcal{K}$ and set $\sigma_k = \max\{1, k\}^{-\alpha}$ for all $k \in \mathbb{N}_0$ and some $\alpha \in (1/2, 1/p)$. Then we have $(\sigma_k) \in \ell^2$.

We now want to define a RKHS on a set $\Omega_0 \subset \Omega$, with $\mu(\Omega \setminus \Omega_0) = 0$, which admits the orthonormal basis $(\sigma_k \varphi_k)$ and contains the set $\mathcal{K}_0$. Such a Hilbert space will have the reproducing kernel

$$K(x, y) \;=\; \sum_{k\in\mathbb{N}_0} \sigma_k^2 \varphi_k(x) \overline{\varphi_k(y)}.$$

To find a suitable set Ω_0, we first note that

$$\int_\Omega K(x, x)\, d\mu(x) \;=\; \sum_{k\in\mathbb{N}_0} \sigma_k^2 < \infty \tag{6.12}$$

and thus $K(x, x)$ is finite for all $x \in \Omega \setminus E$ with a null set $E \subset \Omega$. Moreover, for all $u \in \mathcal{K}_0$, we have

$$\sum_{k\geq 1} k\, |\hat{u}(k)|^2 \;=\; \sum_{n\geq 0} \sum_{k>n} |\hat{u}(k)|^2 \;=\; \sum_{n\geq 0} \|u - P_n u\|_{L^2}^2 < \infty,$$

where we use (6.11) and the assumptions on $\mathcal{K}$. The Rademacher-Menchov Theorem, see e.g. [159], now implies that the Fourier series of u at x converges to $u(x)$ for all $x \in \Omega \setminus E_u$ with a null set $E_u \subset \Omega$. We put $\Omega_0 := \Omega \setminus E_0$,

where $E_0 := E \cup \bigcup_{u\in\mathcal{K}_0} E_u$ is a null set. Then for all $x \in \Omega_0$ and $u \in \mathcal{K}_0$, we have

$$K(x,x) < \infty \quad \text{and} \quad u(x) = \sum_{k\in\mathbb{N}_0} \hat{u}(k)\,\varphi_k(x).$$

We now define the space H as the set of all square-integrable functions $u\colon \Omega_0 \to \mathbb{C}$ which are point-wise represented by their Fourier series $\sum_k \hat{u}(k)\,\varphi_k$ and which satisfy

$$\|u\|_H^2 := \sum_{k\in\mathbb{N}_0} \frac{|\hat{u}(k)|^2}{\sigma_k^2} < \infty.$$

Then H is a separable reproducing kernel Hilbert space on Ω_0 since

$$|u(x)|^2 \leqslant K(x,x)\|u\|_H^2 \quad \text{for all } x \in \Omega_0 \text{ and } u \in H,$$

and $(\sigma_k\varphi_k)_{k\in\mathbb{N}_0}$ is an orthonormal basis of H. The reproducing kernel is K, which has finite trace from (6.12).

We now show that $\mathcal{K}_0$ (with functions restricted to Ω_0) is a subset of H. Recall that any $u \in \mathcal{K}_0$ is point-wise represented by its Fourier series. Moreover, note that the Kolmogorov widths of $\mathcal{K}_0$ and $\mathcal{K}$ are the same. We use

$$d_{2n}(\mathcal{K})_{L^2} = \big(d_{2n}(\mathcal{K})_{L^2}^p\big)^{1/p} \leq \left(\frac{1}{n}\sum_{k\geq n} d_k(\mathcal{K})_{L^2}^p\right)^{1/p}$$

and obtain for any $n \in 8\mathbb{N}$ and $u \in \mathcal{K}_0$ that

$$\begin{aligned}
\|u - P_n u\|_H^2 &= \sum_{k\geq n} k^{2\alpha}\,|\hat{u}(k)|^2 \leq \sum_{\ell\in\mathbb{N}_0} \big(n2^{\ell+1}\big)^{2\alpha} \sum_{k=n2^\ell}^{n2^{\ell+1}-1} |\hat{u}(k)|^2 \\
&\leq 4\sum_{\ell\in\mathbb{N}_0} \big(n2^{\ell+1}\big)^{2\alpha}\, d_{n2^{\ell-2}}(\mathcal{K})_{L^2}^2 \\
&\leq 4\sum_{\ell\in\mathbb{N}_0} \big(n2^{\ell+1}\big)^{2\alpha} \left(\frac{1}{n2^{\ell-3}} \sum_{k\geq n2^{\ell-3}} d_k(\mathcal{K})_{L^2}^p\right)^{2/p} \\
&\leqslant 2^{2+2\alpha+6/p} n^{2\alpha-2/p} \sum_{\ell\in\mathbb{N}_0} 2^{(2\alpha-2/p)\ell} \left(\sum_{k\geq n/8} d_k(\mathcal{K})_{L^2}^p\right)^{2/p}.
\end{aligned}$$

The last expression is finite for $n \geq 8k_0$, since $2\alpha - 2/p < 0$. This implies that $u \in H$ and

$$\|u - P_n u\|_H \leqslant C\, n^\alpha \left(\frac{1}{n}\sum_{k\geq n/8} d_k(\mathcal{K})_{L^2}^p\right)^{1/p}, \tag{6.13}$$

where $C > 0$ only depends on $p \in (0,2)$ and $\alpha \in (\frac{1}{2}, \frac{1}{p})$.

We now apply Theorem 6.23 to the newly constructed Hilbert space H to find $m \leq 43200\,n$ and a linear algorithm P_n^m of the form

$$P_n^m u = \sum_{i=1}^m u(x^i)\psi_i, \quad x^i \in \Omega_0,\ \psi_i \in L^2,$$

such that

$$\|u - P_n^m u\|_{L^2(\Omega_0,\mu)}^2 \leq 433 \max\left\{\sigma_n^2,\, \frac{1}{n}\sum_{k\geqslant n}\sigma_k^2\right\} \|u - P_n u\|_H^2 \tag{6.14}$$

for all $u \in H$ and thus, for all $u \in \mathcal{K}_0$. Clearly, in the last inequality, Ω_0 can be replaced with Ω. If we now

insert the estimate (6.13) and the estimate

$$\max\left\{\sigma_n^2,\, \frac{1}{n}\sum_{k\geqslant n}\sigma_k^2\right\} \lesssim n^{-2\alpha}, \tag{6.15}$$

into (6.14), we obtain that

$$\|u - P_n^m u\|_{L^2}^2 \leq \left(\frac{\tilde{c}_p}{n}\sum_{k\geq n/8} d_k(\mathcal{K})_{L^2}^p\right)^{2/p}$$

for all $u \in \mathcal{K}_0$ and some $\tilde{c}_p > 0$ that only depends on p. Since $\mathcal{K}_0$ is dense in $\mathcal{K}$ and both $\mathrm{id}\colon \mathcal{K} \to L^2$ and $P_n^m\colon \mathcal{K} \to L^2$ are continuous, the last bound is true for all $u \in \mathcal{K}$. This finishes the proof of Theorem 6.3 with $c_p = 43200\,\max(\tilde{c}_p, 8)$.

□

6.6.2 The boundary case

We provide a variant of Theorem 6.3 under a weaker condition than $(d_k(\mathcal{K})_{L^2}) \in \ell^p$ for $p < 2$. In fact, we show that the condition $\big((\ln k)^s d_k(\mathcal{K})_{L^2}\big) \in \ell^2$ for some $s > 1/2$ is enough for a comparison of the sampling and the Kolmogorov widths, while the same assumption for $s = 1/2$ is not enough, see Example 6.31.

Theorem 6.27. *Let $s > 1/2$. There is a universal constant $c \in \mathbb{N}$ and a constant $c_s > 0$, depending only on s, such that for every $\mathcal{K}$ and μ that satisfy Assumption A and all $n \geq 2$,*

$$\rho_{cn}(\mathcal{K})_{L^2}^2 \leq c_s\, n^{-1} \ln^{-2s+1} n \sum_{k\geq n} d_k(\mathcal{K})_{L^2}^2 \cdot \ln^{2s} k.$$

Proof. The proof follows the same lines as the proof of Theorem 6.3. The only difference is that we now choose $\sigma_k = k^{-1/2}\ln^{-s} k$ for $k \geq 2$. Then, inequality (6.13) becomes

$$\begin{aligned}\|u - P_n u\|_H^2 &= \sum_{k\geq n} |\hat{u}(k)|^2 k \ln^{2s}(k) \leq \sum_{k\geq n} |\hat{u}(k)|^2 \sum_{n\leq r\leq 2k} \ln^{2s}(r)\\ &\leq \sum_{r\geq n} \ln^{2s}(r) \sum_{k\geq r/2} |\hat{u}(k)|^2 \leq 4\sum_{r\geq n} \ln^{2s}(r)\, d_{\lfloor r/8\rfloor}(\mathcal{K})_{L^2}^2\\ &\leq 32 \sum_{k\geq \lfloor n/8\rfloor} \ln^{2s}(8k+7)\, d_k(\mathcal{K})_{L^2}^2.\end{aligned}$$

Likewise, inequality (6.15) becomes

$$\max\left\{\sigma_n^2,\, \frac{1}{n}\sum_{k\geqslant n}\sigma_k^2\right\} \lesssim n^{-1}\ln^{-2s+1} n$$

and the stated inequality is obtained. □

6.6.3 Proof of Corollary 6.4

Using the same bound as in the proof of Corollary 6.2, the case $\alpha > 1/2$ immediately follows from Theorem 6.3 if we choose $1/\alpha < p < 2$, and the case $\alpha = 1/2$, $\beta > 1$ from Theorem 6.27 if we choose $1/2 < s < \beta - 1/2$.

All bounds are attained with the same classes $\mathcal{K}$ as in Corollary 6.2 for the first case, and with the constructions from the next section for the two other cases.

□

6.7 Examples

We first apply Theorem 6.1 to tensor product spaces.

Example 6.28. Let H be a RKHS on Ω that is compactly embedded into L^2 and let $\mathcal{K}$ be its unit ball. We consider L^2-approximation on the unit ball $\mathcal{K}_d$ of the d-fold tensor product H_d of H, which is a RKHS on the domain Ω^d. We assume that $\rho_m(\mathcal{K})_{L^2} \lesssim m^{-\alpha}$ for some $\alpha > 0$. The famous Smolyak algorithm, see [163], gives the estimate

$$\rho_m(\mathcal{K}_d) \lesssim m^{-\alpha} \ln^{(\alpha+1)(d-1)} m. \tag{6.16}$$

An example of such tensor product spaces are the spaces of dominating mixed smoothness $\alpha > 1/2$, see [67]. For these spaces, it is known that the error bound (6.16) for the Smolyak algorithm can be improved [161]; the exponent of the logarithm can be reduced to $(\alpha + 1/2)(d-1)$. With Corollary 6.2 and known results on the approximation numbers of tensor product operators, see [18, 134], we now obtain

$$\rho_m(\mathcal{K}_d) \lesssim m^{-\alpha} \ln^{\alpha(d-1)} m \quad \text{if} \quad \alpha > 1/2. \tag{6.17}$$

This bound is asymptotically optimal for the spaces of mixed smoothness, see [172, Theorem 1] or [167, Theorem 6.4.3]. More generally, it is known that $d_n(\mathcal{K})_{L^2} \asymp n^{-\alpha}$ implies $d_n(\mathcal{K}_d) \asymp n^{-\alpha} \ln^{\alpha(d-1)} n$ (see e.g. [107]) and therefore the asymptotic bound (6.17) is optimal whenever the approximation numbers in the univariate case are of order $n^{-\alpha}$. Let us note, however, that also preasymptotic estimates on the sampling numbers (say, for $m < d^d$) are of interest, especially if the dimension d is high, see [107, 116, 186].

Remark 6.29. Note that, for Sobolev spaces with mixed smoothness $r > 1/p$ and $1 < p < 2$, the nonlinear sampling numbers in L^2 decay faster than the linear sampling numbers (and the Kolmogorov widths in L^2) if the dimension d is large, see [100].

We now present two examples that show that our upper bounds cannot be improved without further assumptions on the class $\mathcal{K}$.

First, we show that the worst possible behavior of the sampling numbers in the case $d_n(\mathcal{K})_{L^2} \lesssim n^{-1/2} \ln^{-\beta} n$ with $\beta > 1$ is indeed $m^{-1/2} \ln^{-\beta+1} m$.

Example 6.30. For $\ell \in \mathbb{N}_0$ and $k \in \{1, \dots, 2^\ell\}$, define the interval $I_{\ell,k} = [(k-1)2^{-\ell}, k2^{-\ell})$ and denote $\chi_{\ell,k}$ the indicator function of $I_{\ell,k}$. Let $\beta > 1$. We set

$$\mathcal{C}_\beta := \left\{ \mathbf{c} = (c_{\ell,k})_{\ell \in \mathbb{N}_0,\, 1 \le k \le 2^\ell} \,:\, \sum_{k=1}^{2^\ell} |c_{\ell,k}|^2 \le (\ell+1)^{-2\beta} \text{ for all } \ell \in \mathbb{N}_0 \right\}$$

and consider the class

$$\mathcal{K}_\beta := \left\{ u_{\mathbf{c}} = \sum_{\ell \in \mathbb{N}_0} \sum_{k=1}^{2^\ell} c_{\ell,k} \chi_{\ell,k} \,:\, \mathbf{c} \in \mathcal{C}_\beta \right\}.$$

Note that the series $u_{\mathbf{c}}$ converge uniformly, since the inner sum is bounded by $(\ell+1)^{-\beta}$. If $\mathcal{K}_\beta$ is equipped with the maximum distance on $[0,1)$, it is a separable metric space, function evaluation is continuous, and the embedding in $L^2([0,1))$ is continuous.

For every $L \in \mathbb{N}_0$, the span V_L of the functions $\chi_{\ell,k}$ with $\ell \le L$ has dimension 2^L. If P_L is the L^2-orthogonal projection onto V_L, we have for all $\mathbf{c} \in \mathcal{C}_\beta$ that

$$\begin{aligned} \|u_{\mathbf{c}} - P_L u_{\mathbf{c}}\|_2 &\le \Big\| \sum_{(\ell,k)\,:\,\ell > L} c_{\ell,k} \chi_{\ell,k} \Big\|_2 \le \sum_{\ell > L} \Big\| \sum_{k=1}^{2^\ell} c_{\ell,k} \chi_{\ell,k} \Big\|_2 \\ &= \sum_{\ell > L} \left(\sum_{k=1}^{2^\ell} c_{\ell,k}^2 \|\chi_{\ell,k}\|_2^2 \right)^{1/2} \le \sum_{\ell > L} 2^{-\ell/2} (\ell+1)^{-\beta} \lesssim 2^{-L/2} L^{-\beta}, \end{aligned}$$

and thus

$$d_{2^L}(\mathcal{K}_\beta) \lesssim 2^{-L/2} L^{-\beta},$$

or equivalently

$$d_n(\mathcal{K}_\beta) \lesssim n^{-1/2} \ln^{-\beta} n.$$

We now show a lower bound for the sampling numbers. Let $x^1, \dots, x^m \in [0,1)$. For all $\ell \in \mathbb{N}_0$, we let J_ℓ be the set of indices $1 \le k \le 2^\ell$ such that $I_{\ell,k}$ contains at least one of these points. Clearly, the cardinality of J_ℓ is at most m. We choose $L \in \mathbb{N}_0$ of order $\ln m$ and define

$$u_L := \sum_{\ell > L} |J_\ell|^{-1/2} (\ell+1)^{-\beta} \sum_{k \in J_\ell} \chi_{\ell,k}.$$

This function is contained in $\mathcal{K}_\beta$ and for all $i \le m$, we have

$$h := u_L(x^i) = \sum_{\ell > L} |J_\ell|^{-1/2} (\ell+1)^{-\beta} \gtrsim m^{-1/2} \ln^{-\beta+1} m,$$

where h is independent of i. On the other hand, as shown by our previous calculation,

$$\left| \int_0^1 u_L(x)\,dx \right| \le \|u_L\|_2 \lesssim 2^{-L/2} L^{-\beta} \lesssim m^{-1/2} \ln^{-\beta} m.$$

Thus, if we set $u = h - u_L$, the function is contained in $\mathcal{K}_\beta$, vanishes at all points $x^1, \dots, x^m$, and satisfies

$$\|u\|_2 \ge \int_0^1 u(x)\,dx \ge h - \left| \int_0^1 u_L(x)\,dx \right| \gtrsim m^{-1/2} \ln^{-\beta+1} m.$$

This shows $\rho_m(\mathcal{K}_\beta) \gtrsim m^{-1/2} \ln^{-\beta+1} m$.

□

The next example shows that, in the case $d_n(\mathcal{K})_{L^2} \lesssim n^{-1/2} \ln^{-\beta} n$ with $\beta \le 1$, no general statement on the sampling numbers is possible.

Example 6.31. Similar to Example 6.30, we define

$$\mathcal{C} := \left\{ \mathbf{c} = (c_{\ell,k})_{\ell \in \mathbb{N}_0,\, 1 \le k \le 2^\ell} \,:\, \sum_{k=1}^{2^\ell} |c_{\ell,k}|^2 \le (\ell+1)^{-2} \ln(\ell+e)^{-2} \text{ for all } \ell \in \mathbb{N}_0 \right\}$$

and consider the class

$$\mathcal{K} := \left\{ u_{\mathbf{c}} = \sum_{\ell \in \mathbb{N}_0} \sum_{k=1}^{2^\ell} c_{\ell,k} \chi_{\ell,k} \,:\, \mathbf{c} \in \mathcal{C},\, \mathbf{c} \text{ finite} \right\}.$$

The finiteness of the sequences ensures that $\mathcal{K}$, equipped with the maximum distance, is still a separable metric space, where function evaluation is continuous, and the embedding in $L^2([0,1))$ is continuous. As above, we obtain

$$d_n(\mathcal{K})_{L^2} \lesssim n^{-1/2} (\ln n)^{-1} (\ln \ln n)^{-1}.$$

In particular, we have $(d_n(\mathcal{K})_{L^2} \ln^{1/2} n) \in \ell^2$. On the other hand, given $x^1, \dots, x^m$ and $\varepsilon > 0$, we choose $L \in \mathbb{N}_0$ with

$$\sum_{\ell > L} 2^{-\ell/2} (\ell+1)^{-1} (\ln(\ell+e))^{-1} \le \varepsilon,$$

define the sets J_ℓ as above, and choose $L' \in \mathbb{N}_0$ such that

$$h := \sum_{\ell = L+1}^{L'} |J_\ell|^{-1/2} (\ell+1)^{-1} (\ln(\ell+e))^{-1} \ge 1.$$

The function

$$u_L := \frac{1}{h} \sum_{\ell = L+1}^{L'} |J_\ell|^{-1/2} (\ell+1)^{-1} (\ln(\ell+e))^{-1} \sum_{k \in J_\ell} \chi_{\ell,k},$$

is contained in $\mathcal{K}$, its integral is at most ε, and it satisfies $u_L(x^i) = 1$ for all $i \le m$. Then $u = 1 - u_L$ is contained

in $\mathcal{K}$, vanishes at all points $x^1, \dots, x^m$, and satisfies

$$\|u\|_2 \geq \int_0^1 u(x)\,dx \geq 1 - \left| \int_0^1 u_L(x)\,dx \right| \geq 1 - \varepsilon.$$

This shows $\rho_m(\mathcal{K})_{L^2} \geq 1$ for all $m \in \mathbb{N}_0$.

□

We note that the lower bounds in Example 6.30 and 6.31 already hold for the easier problem of numerical integration on $\mathcal{K}_\beta$. Thus, the upper bounds from Corollary 6.4 are also sharp for the minimal error of quadrature rules on probability spaces.

Chapter 7

Randomized least-squares with minimal oversampling

Abstract. When approximating functions based on point values, least-squares methods provide more stability than interpolation methods, at the expense of increasing the sampling budget. We show that near-optimal approximation results can nevertheless be achieved, in an expected L^2 sense, as soon as the sample size m is larger than the dimension n of the approximation space by a constant ratio. On the other hand, for $m = n$, we obtain an interpolation strategy with a stability factor of order n in L^2 and $n^{3/2}\|k_n\|_{L^\infty}^{1/2}$ in L^∞, with k_n the normalized inverse Christoffel function. The sampling algorithm is a greedy procedure based on [24] and [119], with polynomial computational complexity.

7.1 Introduction and main results

Let $(\Omega, \mathcal{A}, \mu)$ be a probability space. We consider the problem of estimating an unknown function $u : \Omega \to \mathbb{C}$ from observations $(u(x^i))_{1\leqslant i\leqslant m}$ of u at chosen points $x^1, \dots, x^m \in \Omega$. We assess the error between u and its estimator $\tilde{u}$ either in the $L^2(\Omega, \mu)$ norm

$$\|v\|_{L^2} := \left(\int_\Omega |v(x)|^2 d\mu(x)\right)^{1/2}, \tag{7.1}$$

or in the uniform norm $\|v\|_{L^\infty} = \|v\|_{L^\infty(\Omega,\mu)}$.

Given a subspace V_n of $L^2(\Omega, \mu)$ such that $\dim(V_n) = n$, or a sequence $(V_n)_{n\geqslant 1}$ of such spaces, we would like to compute the best approximation of u in V_n. This is given by the $L^2(\Omega, \mu)$ orthogonal projection onto V_n, which we denote by P_n, i.e. $P_n u$ is the unique solution to the optimization problem

$$P_n u = \operatorname*{arg\,min}_{v\in V_n} \|u - v\|_{L^2}. \tag{7.2}$$

In general, we may not have access to any information about u apart from its point evaluations. In this case we cannot explicitly compute $P_n u$. A natural approach in this setting is to consider a solution to the weighted least-squares problem

$$P_n^m u \in \operatorname*{arg\,min}_{v\in V_n} \frac{1}{m}\sum_{i=1}^m w_i |u(x^i) - v(x^i)|^2, \tag{7.3}$$

where $w_1, \dots, w_m > 0$ are weights chosen in order to account for the difference between $d\mu$ and the density of sample points. We are interested in the case where $m \geqslant n$, which is the regime where this problem may admit a unique solution.

Compared to the orthogonal projection P_n for the L^2 norm (7.1), the operator P_n^m is the orthogonal projector

onto V_n with respect to

$$\|v\|_m := \left(\frac{1}{m} \sum_{i=1}^{m} w_i |v(x^i)|^2 \right)^{1/2}. \tag{7.4}$$

The above norm is the discrete ℓ^2 norm for the empirical measure $\frac{1}{m}\sum_{i=1}^m w_i \delta_{x^i}$. Analogously, we denote by $\langle \cdot, \cdot \rangle_m$ the associated discrete inner product.

It is well known that least squares approximations may be inaccurate even when the measured samples are noiseless. For example, if V_n is the space $\mathbb{R}_{n-1}[X]$ of algebraic polynomials of degree less than n over the interval $[-1, 1]$ and if we choose $m = n$, this corresponds to Lagrange interpolation. This setting is known to be highly unstable, failing to converge towards u when given values at uniformly spaced samples, even when u is infinitely smooth. This is the so-called *Runge phenomenon* [158]. When considering non-uniform points, for instance Fekete [71] or Leja [120] sequences, better results are obtained, however the Lebesgue constant still increases polynomially with n on general domains, see [14, 47, 49].

Regularization by taking m larger than n is therefore required to achieve optimality up to a constant. We present below our main theorem, which provides a new bound on this constant, depending on the ratio between m and n, and on a constant $\gamma \in [0, 1]$ which can be picked arbitrarily by the user.

Theorem 7.1. *For any $m \geqslant n$ and $\gamma \in [0, 1]$, the weighted least-squares approximation $\tilde{u} \in V_n$ provided by Algorithm 5 using m evaluations of u at points selected by Algorithm 4 simultaneously satisfies*

$$\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2) \leqslant \left(1 + \frac{1}{1-\gamma} \frac{1}{(1-\sqrt{r})^2} \right) \min_{v \in V_n} \|u - v\|_{L^2}^2 \tag{7.5}$$

and

$$\|u - \tilde{u}\|_{L^2}^2 \leqslant \left(1 + \frac{1}{\gamma} \frac{1}{(1-\sqrt{r})^2} \right) \min_{v \in V_n} \|u - v\|_{L^\infty}^2 \quad a.s., \tag{7.6}$$

where $r = (n-1)/m < 1$ is the oversampling ratio.

In particular, one can take $m = n$ in (7.5) and (7.6), leading to a statement on interpolation in L^2.

Corollary 7.2. *For $m = n$ and any $\gamma \in [0, 1]$, the interpolation $\tilde{u} \in V_n$ of u at random points $x^1, \ldots, x^n$ selected by Algorithm 4 achieves the accuracy bounds*

$$\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2) \leqslant \frac{4n^2}{1-\gamma} \min_{v \in V_n} \|u - v\|_{L^2}^2 \tag{7.7}$$

and

$$\|u - \tilde{u}\|_{L^2}^2 \leqslant \frac{4n^2}{\gamma} \min_{v \in V_n} \|u - v\|_{L^\infty}^2 \quad a.s. \tag{7.8}$$

Instance optimality statements in expected L^2 norm such as (7.5) and (7.7), of the form

$$\mathbb{E}(\|u - \tilde{u}\|_{L^2}^2) \leqslant C \, \|u - P_n u\|_{L^2}^2,$$

have already been obtained in [55] for i.i.d. sampling according to μ, but with a sample size m growing polynomially with n in classical approximation settings. With the weighted least-squares introduced in [59], only a logarithmic oversampling is needed. In [d], such an inequality is reached for some universal constants C and r. However these constants are quite large, and the proof involves the Kadison-Singer solution [128], resulting in exponential computational complexity with respect to n.

On the other hand, uniform bounds such as (7.6) and (7.8) follow the approach developed in [62, 121, 152, 168], attaining sharper constants, especially in the case where m is close to n. Note that by allowing the right-hand side to be infinite, we do not require u and functions from V_n to belong to L^∞.

A third context where similar bounds can also be obtained is the deterministic L^2 setting, in which more regularity is assumed on u through a nested sequence of approximation spaces $(V_n)_{n \geqslant 1}$. If $(\|u - P_n u\|_{L^2}^2)_{n \geqslant 1}$ is summable, the approximation error can be bounded almost surely by tails of this sequence in a Hilbert space

setting [23, 104, 113, 135, 137], and by tails possibly involving an additional logarithmic factor in general Banach spaces [114, e].

An important application in the case $m = n$ is interpolation in L^∞, which is obtained by combining (7.8) with an inverse inequality between L^2 and L^∞ in V_n. Here we assume that $V_n \subset L^4(\Omega, \mu)$, which is the case as soon as the natural assumption $V_n \subset L^\infty(\Omega, \mu)$ is met.

Theorem 7.3. *For $m = n$ and $\gamma = 1$, the interpolation $\tilde{u} \in V_n \subset L^4$ of u at points $x^1, \dots, x^n$ achieves the accuracy*

$$\|u - \tilde{u}\|_{L^\infty} \leqslant (1 + 2n\|nk_n\|_{L^\infty}^{1/2}) \min_{v \in V_n} \|u - v\|_{L^\infty} \quad a.s., \tag{7.9}$$

where nk_n is the inverse Christoffel function associated to V_n on $L^2(\Omega, \sigma)$ for any probability measure σ.

Note that in the case $m = n$, the values of the weights have no importance given that the minimum in (7.3) is zero. In fact, we exhibit a constructive set of points such that the Lebesgue stability constant

$$\Lambda_n = \max_{v \in V_n} \frac{\|P_n^n v\|_{L^\infty}}{\|v\|_{L^\infty}}$$

is at most $2n\|nk_n\|_{L^\infty}^{1/2}$. Although Fekete points achieve $\Lambda_n = n$, resulting in a factor only $1 + n$ in (7.9) (see [144], Theorem 29.7, for a discussion of this result in the context of Information Based Complexity), their computational complexity is exponential in n. Greedy strategies, based on Leja points [14, 47], have polynomial complexity but only achieve $\Lambda_n \sim n^{13/4}$ or $\Lambda_n \sim n^{1+\log_2(3)}$, even in one-dimensional settings.

Remark 7.4. One can also use an inverse inequality between L^2 and L^∞ in the least-squares regime $m > n$. This is investigated in [109] for general classes of functions in Banach spaces, with error bounds in any L^p norm. We also refer [79], where implications in the field of Information Based Complexity are drawn, in a specific Hilbert setting. These two very recent papers rely on the earlier work [152], and on the infinite-dimensional adaptation [e] of the result from [128], see also [73]. This adaptation is itself based on the pioneering works [113, 114].

In Section 7.2 we introduce our randomized algorithm and prove that it achieves stability and continuity bounds for sampling discretization of functions in V_n. This allows in Section 7.3 our main results. The strategy we adopt is in the same flavor as [d] and [168].

7.2 Randomized sampling algorithm

Let $\varphi = (\varphi_j)_{1 \leqslant j \leqslant n} \in L^2(\Omega, \mathbb{C}^n)$ be an orthonormal basis of V_n. Notice that

$$\int_\Omega \varphi(x)\varphi(x)^\dagger d\mu(x) = I, \tag{7.10}$$

where $\varphi(x)^\dagger$ stands for the row vector $\big(\overline{\varphi_1(x)}, \dots, \overline{\varphi_n(x)}\big)$. Taking the inner product against φ_k, the least-squares solution $\tilde{u} = \sum_{j=1}^n c_j \varphi_j \in V_n$ from (7.3) can be characterized by

$$0 = (\langle \tilde{u} - u, \varphi_k \rangle_m)_{1 \leqslant k \leqslant n} = D^\dagger D c - D^\dagger f, \tag{7.11}$$

where $f = (\sqrt{w_i} u(x^i))_{1 \leqslant i \leqslant m}$ and

$$D = \Big(\sqrt{w_i}\varphi_j(x^i)\Big)_{\substack{1 \leqslant i \leqslant m \\ 1 \leqslant j \leqslant n}} \in \mathbb{C}^{m \times n}$$

is the *design matrix* associated to the basis (φ_j) and sample (x^i).

Our goal is to obtain a stability property on the linear system (7.11), or equivalently a lower bound on the eigenvalues of the *Gram matrix*

$$G_m = \frac{1}{m}\overline{D^\dagger D} = \frac{1}{m}\sum_{i=1}^m w_i \varphi(x^i)\varphi(x^i)^\dagger,$$

while asking a certain control on the norm of each term. This lower bound can be rewritten in the following equivalent ways.

— Matrix formulation:

$$G_m \succcurlyeq \alpha I,$$

where $A \succcurlyeq B$ means that the hermitian matrix $A - B$ is positive semi-definite.

— Frame inequality:

$$\frac{1}{m}\sum_{i=1}^{m} w_i \left|\sum_{j=1}^{n} c_j \varphi_j(x^i)\right|^2 \geqslant \alpha |c|^2, \quad c \in \mathbb{C}^n.$$

— Marcinkiewicz-Zygmund inequality:

$$\frac{1}{m}\sum_{i=1}^{m} w_i |u(x^i)|^2 \geqslant \alpha \|u\|_{L^2}^2, \quad u \in V_n.$$

All three versions have been extensively used in the litterature for emphasizing the relations with subsampling of frames and discretization of continuous norms, see [73, 121, 139, 140]. Here we adopt the matrix formulation for concision.

We start with the following greedy Algorithm 4, inspired by [119] and [118], which are themselves randomized versions of [24]. The algorithm ouputs points x^i and weigths w_i from which one can compute the discrete inner product $\|\cdot\|_m$ and the Gram matrix $G_m = \frac{1}{m}(A_{m+1} + (m\delta - n)I)$. Given a matrix R, one main tool in the analysis is the effective resistance $\varphi(x)^\dagger R \varphi(x)$, coined after [164], which quantifies the interaction between $\varphi(x)$ and the eigenvectors of R. During the entire process, the parameters $\gamma \in [0,1]$ and $\delta \in (0,1)$ are fixed.

Algorithm 4 Randomized sampling

1: Initialize $A_1 = nI$, $0 \leqslant \gamma \leqslant 1$, $0 < \delta < 1$
2: **for** $i = 1$ **to** m **do**
3: Let $B_i = A_i - \delta I$
4: Define $R_i = \left(\operatorname{Tr} B_i^{-1} - \operatorname{Tr} A_i^{-1}\right)^{-1} B_i^{-2} - B_i^{-1}$
5: Define $\kappa_i(x) = \varphi(x)^\dagger R_i \varphi(x) \chi_{\varphi(x)^\dagger R_i \varphi(x) \geqslant \gamma \frac{1-\delta}{\delta}}$
6: Draw x^i according to $\frac{\kappa_i(x) d\mu(x)}{\int \kappa_i d\mu}$
7: Let $w_i = \frac{1}{\kappa_i(x^i)}$
8: Update $A_{i+1} = B_i + w_i \varphi(x^i)\varphi(x^i)^\dagger$
9: **end for**

Before analyzing the algorithm, a few comments are in order.

— Contrarily to most variations on the algorithm presented in [24], no upper potential is used to bound the eigenvalues of A_i from above. In fact, we will only need an upper bound on $\mathbb{E}(\|v\|_m^2)$ in the randomized setting, and a bound on the weights w_i in the deterministic setting.

— The initialization $A_0 = n$ is arbitrary, this particular choice is made in order to have $\operatorname{Tr}(A_0^{-1}) = 1$ and $\delta < 1$. By a scaling invariance, one could multiply all A_i, B_i, w_i and δ by the same factor, while dividing all R_i, κ_i and γ accordingly, without changing the results.

— In comparison to [24], we exploited a translation invariance to replace the lower barrier ℓ and lower potential $\Phi_\ell(A) = \operatorname{Tr}(A - \ell I)^{-1}$ by 0 and $\operatorname{Tr}(A^{-1})$ respectively. This choice is only made to simplify notations.

Remark 7.5. An important application is the case of a finite set $\Omega = \{x^1, \dots, x^N\}$, with μ the uniform measure on Ω. Then the orthonormality (7.10) of the basis φ rewrites

$$\frac{1}{N}\sum_{i=1}^{N} \varphi_i \varphi_i^\dagger = I,$$

where $\varphi_i = \varphi(x^i)$, and we recover a randomized subsampling algorithm for sums of rank-one matrices. The idea of using the algorithm of [24] with non-discrete measures seems to originate in [61], which investigates Marcinkiewicz-type discretization theorems.

We now address the performance of the algorithm. We start with the following loop invariant.

Lemma 7.6. *The algorithm is well-defined, and for any $1 \leqslant i \leqslant m+1$, the matrices A_i and B_i are positive definite with*

$$\operatorname{Tr}(A_i^{-1}) = 1.$$

Proof. We use an induction on index i. Observe that $\operatorname{Tr}(A_1^{-1}) = 1$. For $1 \leqslant i \leqslant m$, assume that A_i is positive definite with $\operatorname{Tr}(A_i^{-1}) = 1$. Then

$$\lambda_{\min}(A_i) = \lambda_{\max}(A_i^{-1})^{-1} \geqslant (\operatorname{Tr} A_i^{-1})^{-1} \geqslant 1 > \delta,$$

so B_i is positive definite. As $B_i \prec A_i$ implies $A_i^{-1} \prec B_i^{-1}$, the denominator in the definition of R_i is positive. Next, using (7.10),

$$\int_\Omega \varphi(x)^\dagger R_i \varphi(x) d\mu(x) = \int_\Omega \operatorname{Tr}(R_i \varphi(x)\varphi(x)^\dagger) d\mu(x) = \operatorname{Tr}(R_i).$$

We rewrite the computations of [24], Lemma 3.5 and Claim 3.6. Using the eigendecomposition $A_i = \sum_{j=1}^n \lambda_j u_j u_j^\dagger$ and denoting

$$Y = A_i^{-1} = \sum_{j=1}^n \frac{1}{\lambda_j} u_j u_j^\dagger \quad \text{and} \quad Z = B_i^{-1} = \sum_{j=1}^n \frac{1}{\lambda_j - \delta} u_j u_j^\dagger,$$

it holds

$$\operatorname{Tr} Z - \operatorname{Tr} Y = \sum_{j=1}^n \frac{1}{\lambda_j - \delta} - \frac{1}{\lambda_j} = \sum_{j=1}^n \frac{\delta}{\lambda_j(\lambda_j - \delta)} = \delta \operatorname{Tr}(YZ) \tag{7.12}$$

and

$$\begin{aligned} \operatorname{Tr} Z^2 - \operatorname{Tr}(YZ) &= \sum_{j=1}^n \frac{1}{(\lambda_j - \delta)^2} - \frac{1}{\lambda_j(\lambda_j - \delta)} \\ &= \sum_{j=1}^n \frac{\delta}{\lambda_j(\lambda_j - \delta)^2} = \delta \operatorname{Tr}(YZ^2). \end{aligned} \tag{7.13}$$

As a consequence,

$$\begin{aligned} \operatorname{Tr} R_i &= \frac{\operatorname{Tr} Z^2}{\operatorname{Tr} Z - \operatorname{Tr} Y} - \operatorname{Tr} Z \\ &= \frac{\operatorname{Tr}(YZ) + \delta \operatorname{Tr}(YZ^2)}{\delta \operatorname{Tr}(YZ)} - \operatorname{Tr} Y - \delta \operatorname{Tr}(YZ) \\ &> \frac{1}{\delta} + \frac{\operatorname{Tr}(YZ^2)}{\operatorname{Tr}(YZ)} - 1 - \frac{\operatorname{Tr}(YZ)}{\operatorname{Tr} Y} \geqslant \frac{1-\delta}{\delta}, \end{aligned}$$

where we used the definition of R_i in the first line, equations (7.12) and (7.13) to go to the second line, the fact that $\delta < 1$ and $\operatorname{Tr} Y = 1$ to reach the third, and finished with a Cauchy-Schwarz inequality. As $\gamma \leqslant 1$, this proves that

$$\mu\left(\left\{x \in \Omega \ : \ \varphi(x)^\dagger R_i \varphi(x) \geqslant \gamma \frac{1-\delta}{\delta}\right\}\right) > 0,$$

hence $\int_\Omega \kappa_i \, d\mu > 0$ and it is possible to draw x^i according to its prescribed law. Finally, by the Shermann-Morrison formula,

$$\operatorname{Tr}(A_{i+1}^{-1}) = \operatorname{Tr} B_i^{-1} - \frac{\varphi(x^i)^\dagger B_i^{-2} \varphi(x^i)}{1/w_i + \varphi(x^i)^\dagger B_i^{-1} \varphi(x^i)} = \operatorname{Tr} A_i^{-1} = 1,$$

which concludes the induction. □

The main result of this section is the following.

Proposition 7.7. *At the end of Algorithm 4,*

$$\|v\|_m^2 \geqslant \frac{m\delta - n + 1}{m}\|v\|_{L^2}^2$$

for any function $v \in V_n$, whereas

$$\|u\|_m^2 \leqslant \frac{\delta}{\gamma(1-\delta)}\|u\|_{L^\infty}^2 \quad a.s.$$

for any function $u \in L^\infty(\Omega, \mu)$ and

$$\mathbb{E}\left(\|u\|_m^2\right) \leqslant \frac{\delta}{(1-\gamma)(1-\delta)}\|u\|_{L^2}^2$$

for any function $u \in L^2(\Omega, \mu)$.

Proof. At the end of the algorithm, we have that $\mathrm{Tr}(A_{m+1}^{-1}) = 1$ and hence $A_{m+1} \succcurlyeq I$. This implies that

$$G_m = \frac{1}{m}\sum_{i=1}^m w_i \varphi(x^i)\varphi(x^i)^\dagger = \frac{1}{m}A_{m+1} + \frac{m\delta - n}{m} I \succcurlyeq \frac{m\delta - n + 1}{m} I.$$

Thus, for any $v = \sum_{j=1}^n \nu_j \varphi_j \in V_n$,

$$\|v\|_m^2 = \frac{1}{m}\sum_{i=1}^m w_i |v(x^i)|^2 = \nu^\dagger G_m \nu \geqslant \lambda_{\min}(G_m)|\nu|^2 \geqslant \frac{m\delta - n + 1}{m}\|v\|_{L^2}^2.$$

For the second statement, note that almost surely,

$$\frac{1}{w_i} = \kappa_i(x^i) \geqslant \gamma\frac{1-\delta}{\delta},$$

and hence

$$\|u\|_m^2 = \frac{1}{m}\sum_{i=1}^m w_i |u(x^i)|^2 \leqslant \frac{1}{m}\sum_{i=1}^m \frac{\delta}{\gamma(1-\delta)}|u(x^i)|^2 \leqslant \frac{\delta}{\gamma(1-\delta)}\|u\|_{L^\infty}^2.$$

Finally we consider the third statement. We denote

$$\Omega_i = \{x \in \Omega \ : \ \kappa_i(x) > 0\} = \left\{x \in \Omega \ : \ \varphi(x)^\dagger R_i \varphi(x) \geqslant \gamma\frac{1-\delta}{\delta}\right\}$$

and observe that

$$\begin{aligned}\int_\Omega \kappa_i d\mu &= \int_\Omega \varphi(x)^\dagger R_i \varphi(x) d\mu(x) - \int_{\Omega_i^c} \varphi(x)^\dagger R_i \varphi(x) d\mu(x) \\ &\geqslant \mathrm{Tr}(R_i) - \int_\Omega \gamma\frac{1-\delta}{\delta} d\mu \geqslant (1-\gamma)\frac{1-\delta}{\delta}.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}\left(\|u\|_m^2\right) &= \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left(w_i|u(x^i)|^2\right)\\ &= \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left(\int_{\Omega_i}\frac{1}{\kappa_i(x)}|u(x)|^2\frac{\kappa_i(x)}{\int_\Omega \kappa_i d\mu}d\mu(x)\right)\\ &\leqslant \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left(\frac{\|u\|_{L^2}^2}{\int_\Omega \kappa_i d\mu}\right)\\ &\leqslant \frac{\delta}{(1-\gamma)(1-\delta)}\|u\|_{L^2}^2,\end{aligned}$$

which concludes the proof. □

7.3 Weighted least-squares

Recall that the least-squares solution can be computed by solving the linear system (7.11). We use the sample $(x^1,\dots,x^m)$ and weights $(w_1,\dots,w_m)$ provided by Algorithm 4.

Algorithm 5 Weighted least-squares approximation

1: Evaluate $u(x^1),\dots,u(x^m)$
2: Compute $\frac{1}{m}D^\dagger f = \frac{1}{m}\sum_{i=1}^m w_i u(x^i)\overline{\varphi(x^i)}$
3: Solve $\overline{G_m}\,c = \frac{1}{m}D^\dagger f$
4: Return $\tilde{u} = \sum_{j=1}^n c_j\varphi_j$

Proof of Theorem 7.1. We follow an approach similar to Lemma 1 in [d], which consists in using the results from Proposition 7.7 to exchange continuous norms $\|\cdot\|_{L^2}$ and discrete norms $\|\cdot\|_m$. Denote $g = u - P_n u$ the residual, for which according to (7.2) we have

$$\|g\|_{L^2}^2 = \|u - P_n u\|_{L^2}^2 = \min_{v\in V_n}\|u-v\|_{L^2}^2.$$

By Pythagoras theorem one has

$$\|u-\tilde{u}\|_{L^2}^2 = \|g\|_{L^2}^2 + \|\tilde{u}-P_n u\|_{L^2}^2,$$

so that we only need to bound the term

$$\|\tilde{u}-P_n u\|_{L^2}^2 = \|P_n^m(u-P_n u)\|_{L^2}^2 = \|P_n^m g\|_{L^2}^2$$

in expectation. This is achieved thanks to Proposition 7.7 using the estimate

$$\begin{aligned}\frac{\mathbb{E}(\|P_n^m g\|_{L^2}^2)}{\|g\|_{L^2}^2} &= \frac{\mathbb{E}(\|P_n^m g\|_{L^2}^2)}{\mathbb{E}\left(\|P_n^m g\|_m^2\right)}\frac{\mathbb{E}\left(\|P_n^m g\|_m^2\right)}{\mathbb{E}\left(\|g\|_m^2\right)}\frac{\mathbb{E}\left(\|g\|_m^2\right)}{\|g\|_{L^2}^2}\\ &\leqslant \frac{m}{m\delta-n+1}\times 1\times\frac{\delta}{(1-\gamma)(1-\delta)}.\end{aligned}$$

Recalling that $r=(n-1)/m$, this last bound is equal to $1/(1-\gamma)$ times

$$\frac{m\delta}{(m\delta-n+1)(1-\delta)} = \frac{1}{(1-r/\delta)(1-\delta)} = \frac{1}{\left(1-\sqrt{r}\right)^2}$$

if we take $\delta = \sqrt{r}$. This proves inequality (7.5). In order to deal with (7.6), we denote by

$$P_n^\infty u = \operatorname*{arg\,min}_{v \in V_n} \|u - v\|_{L^\infty}$$

the Chebyshev projection of u onto V_n with respect to the Banach norm $\|\cdot\|_{L^\infty}$, and take $g = u - P_n^\infty u$ the associated residual. By an argument similar to [59] and [168], we have

$$\|u - P_n^m u\|_{L^2} - \|u - P_n^\infty u\|_{L^2} \leqslant \|P_n^m u - P_n^\infty u\|_{L^2} = \|P_n^m g\|_{L^2}$$

and we conclude by bounding the operator norm $\|P_n^m\|_{L^\infty \to L^2}$ by

$$\frac{\|P_n^m g\|_{L^2(\sigma)}}{\|g\|_{L^\infty}} = \frac{\|P_n^m g\|_{L^2(\sigma)}}{\|P_n^m g\|_m} \frac{\|P_n^m g\|_m}{\|g\|_m} \frac{\|g\|_m}{\|g\|_{L^\infty}} \leqslant \frac{1}{\sqrt{\gamma}} \frac{1}{1 - \sqrt{r}}, \tag{7.14}$$

where the last inequality stems from the same reasons as above. □

Proof of Corollary 7.2. Taking $m = n$, we simply observe that $r = 1 - \frac{1}{n}$ and thus $\sqrt{r} \leqslant 1 - \frac{1}{2n}$. This implies that $(1 - \sqrt{r})^{-2} \leqslant 4n^2$. □

Remark 7.8. Some intuition is given in [24], where an explanation is given on why their algorithm cannot achieve a better bound than the so-called twice-Ramanujan bound

$$\left(\frac{1 + \sqrt{r}}{1 - \sqrt{r}}\right)^2.$$

This in some sense implies that the analysis of Algorithms 4 and 5 yields at least a factor $(1 - \sqrt{r})^{-2}$ in Theorem 7.1 when r is close to 1, and therefore a factor of order $4n^2$ in Corollary 7.2. When r is close to 1, the factor $(1 - \sqrt{r})^{-2}$ is also slightly better than the $(1 - r)^3$ obtained in [23].

Proof of Theorem 7.3. Let σ be a probability measure on Ω, and define the associated inverse Christoffel function

$$nk_n(x) := \max_{v \in V_n} \frac{|v(x)|^2}{\|v\|^2_{L^2(\Omega,\sigma)}}.$$

Then, for any $v \in V_n$, it holds

$$\|v\|_{L^\infty} \leqslant \|nk_n\|^{1/2}_{L^\infty} \|v\|_{L^2(\Omega,\sigma)},$$

with

$$\|nk_n\|^{1/2}_{L^\infty} := \max_{v \in V_n} \frac{\|v\|_{L^\infty}}{\|v\|_{L^2(\Omega,\sigma)}}.$$

The rest of the proof follows the same path as made previously: by letting $g = u - P_n^\infty u$, we have

$$\|u - P_n^m u\|_{L^\infty} - \|u - P_n^\infty u\|_{L^\infty} \leqslant \|P_n^m u - P_n^\infty u\|_{L^\infty} = \|P_n^m g\|_{L^\infty}$$

and according to (7.14), with μ replaced by σ, we also that

$$\frac{\|P_n^m g\|_{L^\infty}}{\|g\|_{L^\infty}} = \frac{\|P_n^m g\|_{L^\infty}}{\|P_n^m g\|_{L^2(\Omega,\sigma)}} \frac{\|P_n^m g\|_{L^2(\Omega,\sigma)}}{\|g\|_{L^\infty}} \leqslant 2n \, \|nk_n\|^{1/2}_{L^\infty},$$

which concludes the proof. □

Note that in Theorem 7.3 the choice of σ is left to the user. An important question is whether one can find a probability measure σ making $\|k_n\|_{L^\infty}$ bounded for general spaces Ω and V_n, which would result in a Lebesgue interpolation constant of order $\mathcal{O}(n^{3/2})$. To our knowledge, this is an open problem.

Remark 7.9. A nice feature of the randomized and deterministic results is that they are obtained with the same algorithm, up to a choice of parameter γ. If one takes an intermediate value $\gamma = \frac{1}{2}$, we attain both

estimates at the same time, up to a factor 2 in the stability constants, compared to the optimal choices $\gamma = 0$ in the randomized setting and $\gamma = 1$ in the deterministic setting.

Remark 7.10. In the randomized setting, we do not need $(\Omega, \mathcal{A}, \mu)$ to be a probability space, a measure space is sufficient. The uniform results also hold when μ is a measure of finite mass, if one multiplies all the L^∞ norms by a factor $\mu(\Omega)$. Lastly, Theorem 7.3 holds for any sigma-finite measure μ, by rescaling μ with a density of mass 1.

Remark 7.11. In the deterministic setting, it is not necessary to draw each point x^i with a density proportional to $\kappa_i(x)$. If g is uniformly bounded (and not just essentially bounded), one can drop the "almost sure" limitation, and the only requirement is that

$$\varphi(x^i)^\dagger R_i \varphi(x^i) \geqslant \gamma \frac{1-\delta}{\delta}, \quad 1 \leqslant i \leqslant m,$$

which may be easier from a numerical point a view. However, it seems that searching for x^i by rejection sampling works better in practice than sorting Ω and looking for the first point that achieves this condition.

Remark 7.12. Moreover, in the deterministic setting, one can replace each weight w_i by its upper bound $\frac{\delta}{\gamma(1-\delta)}$, resulting in an unweighted discrete norm

$$\|g\|_m^2 = \frac{1}{m} \sum_{i=1}^m |g(x^i)|^2,$$

which still satisfies

$$\|v\|_m \geqslant \sqrt{\gamma}(1-\sqrt{r})\|v\|_{L^2} \quad \text{and} \quad \|g\|_m \leqslant \|g\|_{L^\infty} \quad a.s.$$

for any $v \in V_n$ and $g \in L^\infty$, and therefore achieves the bound (7.6). We refer to [23] for earlier results on subsampling of frames with unweighted discrete norms.

Publications

[a] A. Cohen, W. Dahmen, M. Dolbeault, and A. Somacal. « Reduced order modeling for elliptic problems with high contrast diffusion coefficients ». In: *arXiv preprint arXiv:2304.10971* (2023).

[b] A. Cohen, M. Dolbeault, O. Mula, and A. Somacal. « Nonlinear approximation spaces for inverse problems ». In: *Anal. Appl. (Singap.)* 21.1 (2023), pp. 217–253.

[c] A. Cohen and M. Dolbeault. « Optimal sampling and Christoffel functions on general domains ». In: *Constructive Approximation* (2021), pp. 1–43.

[d] A. Cohen and M. Dolbeault. « Optimal pointwise sampling for L^2 approximation ». In: *Journal of Complexity* 68 (2022), p. 101602.

[e] M. Dolbeault, D. Krieg, and M. Ullrich. « A sharp upper bound for sampling numbers in L_2 ». In: *Appl. Comput. Harmon. Anal.* 63 (2023), pp. 113–134.

Bibliography

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*. Second. Vol. 140. Pure and Applied Mathematics (Amsterdam). Elsevier/Academic Press, Amsterdam, 2003, pp. xiv+305.

[2] B. Adcock, A. C. Hansen, and C. Poon. « Beyond consistent reconstructions: optimality and sharp bounds for generalized sampling, and application to the uniform resampling problem ». In: *SIAM Journal on Mathematical Analysis* 45.5 (2013), pp. 3132–3167.

[3] B. Adcock, S. Brugiapaglia, N. Dexter, and S. Moraga. « On efficient algorithms for computing near-best polynomial approximations to high-dimensional, Hilbert-valued functions from limited samples ». In: *arXiv preprint arXiv:2203.13908* (2022).

[4] B. Adcock, S. Brugiapaglia, N. Dexter, and S. Moraga. « Near-optimal learning of Banach-valued, high-dimensional functions via deep neural networks ». In: *arXiv preprint arXiv:2211.12633* (2022).

[5] B. Adcock and J. M. Cardenas. « Near-optimal sampling strategies for multivariate function approximation on general domains ». In: *SIAM J. Math. Data Sci.* 2.3 (2020), pp. 607–630.

[6] B. Adcock and D. Huybrechs. « Approximating smooth, multivariate functions on irregular domains ». In: *Forum Math. Sigma* 8 (2020), Paper No. e26, 45.

[7] R. Ahlswede and A. Winter. « Strong converse for identification via quantum channels ». In: *IEEE Trans. Inform. Theory* 48.3 (2002), pp. 569–579.

[8] M. Ainsworth. « Robust a posteriori error estimation for nonconforming finite element approximation ». In: *SIAM J. Numer. Anal.* 42.6 (2005), pp. 2320–2341.

[9] B. Aksoylu, I. G. Graham, H. Klie, and R. Scheichl. « Towards a rigorously justified algebraic preconditioner for high-contrast diffusion problems ». In: *Comput. Vis. Sci.* 11.4-6 (2008), pp. 319–331.

[10] B. Aksoylu and Z. Yeter. « Robust multigrid preconditioners for cell-centered finite volume discretization of the high-contrast diffusion equation ». In: *Comput. Vis. Sci.* 13.5 (2010), pp. 229–245.

[11] S. T. Ali, J.-P. Antoine, and J.-P. Gazeau. *Coherent states, wavelets, and their generalizations*. Second. Theoretical and Mathematical Physics. Springer, New York, 2014, pp. xviii+577.

[12] Z. Allen-Zhu, Y. Li, A. Singh, and Y. Wang. « Near-optimal discrete optimization for experimental design: a regret minimization approach ». In: *Math. Program.* 186.1-2, Ser. A (2021), pp. 439–478.

[13] J. Anderson. « Extensions, restrictions, and representations of states on C^*-algebras ». In: *Trans. Amer. Math. Soc.* 249.2 (1979), pp. 303–329.

[14] V. Andrievskii and F. Nazarov. « A simple upper bound for Lebesgue constants associated with Leja points on the real line ». In: *J. Approx. Theory* 275 (2022), Paper No. 105699, 13.

[15] F. Arandiga, A. Cohen, R. Donat, and N. Dyn. « Interpolation and approximation of piecewise smooth functions ». In: *SIAM Journal on Numerical Analysis* 43.1 (2005), pp. 41–57.

[16] J.-P. Argaud, B. Bouriquet, F. de Caso, H. Gong, Y. Maday, and O. Mula. « Sensor placement in nuclear reactors based on the generalized empirical interpolation method ». In: *Journal of Computational Physics* 363 (2018), pp. 354–370.

[17] B. Arras, M. Bachmayr, and A. Cohen. « Sequential sampling for optimal weighted least squares approximations in hierarchical spaces ». In: *SIAM J. Math. Data Sci.* 1.1 (2019), pp. 189–207.

[18] K. I. Babenko. « Approximation of periodic functions of many variables by trigonometric polynomials ». In: *Soviet Math. Dokl.* 1 (1960), pp. 513–516.

[19] I. Babuška, F. Nobile, and R. Tempone. « A stochastic collocation method for elliptic partial differential equations with random input data ». In: *SIAM J. Numer. Anal.* 45.3 (2007), pp. 1005–1034.

[20] M. Bachmayr and A. Cohen. « Kolmogorov widths and low-rank approximations of parametric elliptic PDEs ». In: *Mathematics of Computation* 86.304 (2017), pp. 701–724.

[21] M. Bachmayr, A. Cohen, and G. Migliorati. « Sparse polynomial approximation of parametric elliptic PDEs. Part I: Affine coefficients ». In: *ESAIM Math. Model. Numer. Anal.* 51.1 (2017), pp. 321–339.

[22] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone. « Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison ». In: *Spectral and high order methods for partial differential equations*. Vol. 76. Lect. Notes Comput. Sci. Eng. Springer, Heidelberg, 2011, pp. 43–62.

[23] F. Bartel, M. Schäfer, and T. Ullrich. « Constructive subsampling of finite frames with applications in optimal function recovery ». In: *Appl. Comput. Harmon. Anal.* 65 (2023), pp. 209–248.

[24] J. Batson, D. A. Spielman, and N. Srivastava. « Twice-ramanujan sparsifiers ». In: *SIAM Journal on Computing* 41.6 (2012).

[25] B. Battisti, T. Blickhan, G. Enchery, V. Ehrlacher, D. Lombardi, and O. Mula. « Wasserstein model reduction approach for parametrized flow problems in porous media ». In: (2022).

[26] J. Beck, F. Nobile, L. Tamellini, and R. Tempone. « Implementation of optimal Galerkin and collocation approximations of PDEs with random coefficients ». In: *CANUM 2010,* 40^{e} *Congrès National d'Analyse Numérique*. Vol. 33. ESAIM Proc. EDP Sci., Les Ulis, 2011, pp. 10–21.

[27] J. Beck, F. Nobile, L. Tamellini, and R. Tempone. « Convergence of quasi-optimal stochastic Galerkin methods for a class of PDES with random coefficients ». In: *Comput. Math. Appl.* 67.4 (2014), pp. 732–751.

[28] P. Benner, A. Cohen, M. Ohlberger, and K. Willcox. *Model Reduction and Approximation: Theory and Algorithms*. Vol. 15. SIAM, 2017.

[29] P. Berger, K. Gröchenig, and G. Matz. « Sampling and reconstruction in distinct subspaces using oblique projections ». In: *J. Fourier Anal. Appl.* 25.3 (2019), pp. 1080–1112.

[30] C. Bernardi and R. Verfürth. « Adaptive finite element methods for elliptic equations with non-smooth coefficients ». In: *Numer. Math.* 85.4 (2000), pp. 579–608.

[31] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. « Convergence Rates for Greedy Algorithms in Reduced Basis Methods ». In: *SIAM Journal on Mathematical Analysis* 43.3 (2011), pp. 1457–1472.

[32] P. Binev, A. Cohen, O. Mula, and J. Nichols. « Greedy Algorithms for Optimal Measurements Selection in State Estimation Using Reduced Models ». In: *SIAM/ASA Journal on Uncertainty Quantification* 6.3 (2018), pp. 1101–1126.

[33] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. « Data assimilation in reduced modeling ». In: *SIAM/ASA J. Uncertain. Quantif.* 5.1 (2017), pp. 1–29.

[34] M.-R. Blel, V. Ehrlacher, and T. Lelièvre. « Influence of sampling on the convergence rates of greedy algorithms for parameter-dependent random variables ». In: *arXiv preprint arXiv:2105.14091* (2021).

[35] B. Bojanov. « Optimal recovery of functions and integrals ». In: *First European Congress of Mathematics*. Springer. 1994, pp. 371–390.

[36] A. Bonito, A. Cohen, R. DeVore, D. Guignard, P. Jantsch, and G. Petrova. « Nonlinear methods for model reduction ». In: *ESAIM Math. Model. Numer. Anal.* 55.2 (2021), pp. 507–531.

[37] P. Borwein and T. Erdélyi. *Polynomials and polynomial inequalities*. Vol. 161. Graduate Texts in Mathematics. Springer-Verlag, New York, 1995, pp. x+480.

[38] M. Bownik. « Continuous frames and the Kadison-Singer problem ». In: *Coherent states and their applications*. Vol. 205. Springer Proc. Phys. Springer, Cham, 2018, pp. 63–88.

[39] J. Bruna, P. Sprechmann, and Y. LeCun. « Super-resolution with deep convolutional sufficient statistics ». In: *4th International Conference on Learning Representations, ICLR 2016*. 2016.

[40] A. Buffa, Y. Maday, A. T. Patera, C. Prud'homme, and G. Turinici. « A priori convergence of the greedy algorithm for the parametrized reduced basis method ». In: *ESAIM: Mathematical modelling and numerical analysis* 46.3 (2012), pp. 595–603.

[41] V. M. Calo, Y. Efendiev, and J. Galvis. « Asymptotic expansions for high-contrast elliptic equations ». In: *Mathematical Models and Methods in Applied Sciences* 24.03 (2014), pp. 465–494.

[42] E. J. Candès, J. K. Romberg, and T. Tao. « Stable signal recovery from incomplete and inaccurate measurements ». In: *Comm. Pure Appl. Math.* 59.8 (2006), pp. 1207–1223.

[43] E. J. Candes and T. Tao. « Decoding by linear programming ». In: *IEEE Trans. Inform. Theory* 51.12 (2005), pp. 4203–4215.

[44] M. Capalbo, O. Reingold, S. Vadhan, and A. Wigderson. « Randomness conductors and constant-degree lossless expanders ». In: *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. 2002, pp. 659–668.

[45] B. Carl. « Entropy numbers, s-numbers, and eigenvalue problems ». In: *J. Functional Analysis* 41.3 (1981), pp. 290–306.

[46] A. Chatterjee. « An introduction to the proper orthogonal decomposition ». In: *Current science* (2000), pp. 808–817.

[47] A. Chkifa and A. Cohen. « On the stability of polynomial interpolation using hierarchical sampling ». In: *Sampling theory, a renaissance*. Appl. Numer. Harmon. Anal. Birkhäuser/Springer, Cham, 2015, pp. 437–458.

[48] A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, and R. Tempone. « Discrete least squares polynomial approximation with random evaluations—application to parametric and stochastic elliptic PDEs ». In: *ESAIM Math. Model. Numer. Anal.* 49.3 (2015), pp. 815–837.

[49] A. Chkifa, A. Cohen, and C. Schwab. « High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs ». In: *Found. Comput. Math.* 14.4 (2014), pp. 601–633.

[50] M. A. Chkifa. « On the Lebesgue constant of Leja sequences for the complex unit disk and of their real projection ». In: *J. Approx. Theory* 166 (2013), pp. 176–200.

[51] O. Christensen. *An introduction to frames and Riesz bases*. Second. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, [Cham], 2016, pp. xxv+704.

[52] A. Cohen, W. Dahmen, and R. DeVore. « Compressed sensing and best k-term approximation ». In: *Journal of the American mathematical society* 22.1 (2009), pp. 211–231.

[53] A. Cohen, W. Dahmen, R. DeVore, J. Fadili, O. Mula, and J. Nichols. « Optimal reduced model algorithms for data-based state estimation ». In: *SIAM Journal on Numerical Analysis* 58.6 (2020), pp. 3355–3381.

[54] A. Cohen, W. Dahmen, O. Mula, and J. Nichols. « Nonlinear reduced models for state and parameter estimation ». In: *SIAM/ASA Journal on Uncertainty Quantification* 10.1 (2022), pp. 227–267.

[55] A. Cohen, M. A. Davenport, and D. Leviatan. « On the stability and accuracy of least squares approximations ». In: *Found. Comput. Math.* 13.5 (2013), pp. 819–834.

[56] A. Cohen and R. DeVore. « Approximation of high-dimensional parametric PDEs ». In: *Acta Numerica* 24 (2015), pp. 1–159.

[57] A. Cohen, R. DeVore, G. Petrova, and P. Wojtaszczyk. « Optimal stable nonlinear approximation ». In: *Found. Comput. Math.* 22.3 (2022), pp. 607–648.

[58] A. Cohen, R. Devore, and C. Schwab. « Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's ». In: *Anal. Appl. (Singap.)* 9.1 (2011), pp. 11–47.

[59] A. Cohen and G. Migliorati. « Optimal weighted least-squares methods ». In: *SMAI J. Comput. Math.* 3 (2017), pp. 181–203.

[60] F. Cucker and S. Smale. « On the mathematical foundations of learning ». In: *Bull. Amer. Math. Soc. (N.S.)* 39.1 (2002), pp. 1–49.

[61] F. Dai, A. Prymak, A. Shadrin, V. Temlyakov, and S. Tikhonov. « Entropy numbers and Marcinkiewicz-type discretization ». In: *J. Funct. Anal.* 281.6 (2021), Paper No. 109090, 25.

[62] F. Dai and V. Temlyakov. « Random points are good for universal discretization ». In: *arXiv preprint arXiv:2301.12536* (2023).

[63] R. A. DeVore. « Nonlinear approximation ». In: *Acta numerica* 7 (1998), pp. 51–150.

[64] R. A. DeVore and G. G. Lorentz. *Constructive approximation*. Vol. 303. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1993, pp. x+449.

[65] R. DeVore, G. Petrova, and P. Wojtaszczyk. « Greedy algorithms for reduced bases in Banach spaces ». In: *Constr. Approx.* 37.3 (2013), pp. 455–466.

[66] Z. Ditzian and A. Prymak. « On Nikol'skii inequalities for domains in $\mathbb{R}^d$ ». In: *Constr. Approx.* 44.1 (2016), pp. 23–51.

[67] D. Dũng, V. Temlyakov, and T. Ullrich. *Hyperbolic cross approximation*. Advanced Courses in Mathematics. CRM Barcelona. Edited and with a foreword by Sergey Tikhonov. Birkhäuser/Springer, Cham, 2018, pp. xi+218.

[68] J. L. Eftang, A. T. Patera, and E. M. Rønquist. « An "*hp*" certified reduced basis method for parametrized elliptic partial differential equations ». In: *SIAM J. Sci. Comput.* 32.6 (2010), pp. 3170–3200.

[69] V. Ehrlacher, D. Lombardi, O. Mula, and F.-X. Vialard. « Nonlinear model reduction on metric spaces. Application to one-dimensional conservative PDEs in Wasserstein spaces ». In: *ESAIM M2AN* 54.6 (2020), pp. 2159–2197.

[70] M.-J. Fadili, J.-L. Starck, and F. Murtagh. « Inpainting and zooming using sparse representations ». In: *The Computer Journal* 52.1 (2009), pp. 64–79.

[71] M. Fekete. « Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten ». In: *Math. Z.* 17.1 (1923), pp. 228–249.

[72] S. Foucart and H. Rauhut. « An invitation to compressive sensing ». In: *A mathematical introduction to compressive sensing*. Springer, 2013, pp. 1–39.

[73] D. Freeman and D. Speegle. « The discretization problem for continuous frames ». In: *Adv. Math.* 345 (2019), pp. 784–813.

[74] F. Galarce, D. Lombardi, and O. Mula. « State estimation with model reduction and shape variability. Application to biomedical problems ». In: *SIAM Journal on Scientific Computing* 44.3 (2022), B805–B833.

[75] J. Galvis and Y. Efendiev. « Domain decomposition preconditioners for multiscale flows in high contrast media: reduced dimension coarse spaces ». In: *Multiscale Model. Simul.* 8.5 (2010), pp. 1621–1644.

[76] M. J. Gander, L. Halpern, and K. Santugini-Repiquet. « On optimal coarse spaces for domain decomposition and their approximation ». In: *Domain decomposition methods in science and engineering XXIV*. Vol. 125. Lect. Notes Comput. Sci. Eng. Springer, Cham, 2018, pp. 271–280.

[77] M. J. Gander, A. Loneland, and T. Rahman. « Analysis of a new harmonically enriched multiscale coarse space for domain decomposition methods ». In: *arXiv preprint arXiv:1512.05285* (2015).

[78] M. J. Gander and T. Vanzan. « Heterogeneous optimized Schwarz methods for second order elliptic PDEs ». In: *SIAM J. Sci. Comput.* 41.4 (2019), A2329–A2354.

[79] J. Geng and H. Wang. « On the power of standard information for tractability for L_∞ approximation of periodic functions in the worst case setting ». In: *arXiv preprint arXiv:2304.14748* (2023).

[80] C. Greif and K. Urban. « Decay of the Kolmogorov N-width for wave problems ». In: *Applied Mathematics Letters* 96 (2019), pp. 216–222.

[81] K. Gröchenig. « Sampling, Marcinkiewicz-Zygmund inequalities, approximation, and quadrature rules ». In: *J. Approx. Theory* 257 (2020), pp. 105455, 20.

[82] P. Grohs and F. Voigtlaender. « Proof of the theory-to-practice gap in deep learning via sampling complexity bounds for neural network approximation spaces ». In: *arXiv preprint arXiv:2104.02746* (2021).

[83] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[84] B. Haasdonk. « Reduced basis methods for parametrized PDEs—a tutorial introduction for stationary and instationary problems ». In: *Model reduction and approximation*. Vol. 15. Comput. Sci. Eng. SIAM, Philadelphia, PA, 2017, pp. 65–136.

[85] C. Haberstich, A. Nouy, and G. Perrin. « Boosted optimal weighted least-squares ». In: *Math. Comp.* 91.335 (2022), pp. 1281–1315.

[86] M. Hadigol and A. Doostan. « Least squares polynomial chaos expansion: A review of sampling strategies ». In: *Computer Methods in Applied Mechanics and Engineering* 332 (2018), pp. 382–407.

[87] A.-L. Haji-Ali, F. Nobile, R. Tempone, and S. Wolfers. « Multilevel weighted least squares polynomial approximation ». In: *ESAIM Math. Model. Numer. Anal.* 54.2 (2020), pp. 649–677.

[88] J. K. Hammond, R. Chakir, F. Bourquin, and Y. Maday. « PBDW: A non-intrusive Reduced Basis Data Assimilation method and its application to an urban dispersion modeling framework ». In: *Applied Mathematical Modelling* 76 (2019), pp. 1–25.

[89] D. Hardin and E. Saff. « Discretizing manifolds via minimum energy points ». In: *Notices of the AMS* 51.10 (2004), pp. 1186–1194.

[90] A. Harten. « ENO schemes with subcell resolution ». In: *Journal of Computational Physics* 83.1 (1989), pp. 148–184.

[91] N. J. A. Harvey and N. Olver. « Pipage rounding, pessimistic estimators and matrix concentration ». In: *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. ACM, New York, 2014, pp. 926–945.

[92] J. S. Hesthaven, G. Rozza, and B. Stamm. « Certified Reduced Basis Methods for Parametrized Partial Differential Equations ». In: *SpringerBriefs in Mathematics* (2015).

[93] A. Hinrichs, D. Krieg, E. Novak, J. Prochno, and M. Ullrich. « On the power of random information ». In: *Multivariate Algorithms and information-based complexity* 27 (2020), pp. 43–64.

[94] A. Hinrichs, D. Krieg, E. Novak, J. Prochno, and M. Ullrich. « Random sections of ellipsoids and the power of random information ». In: *Trans. Amer. Math. Soc.* 374.12 (2021), pp. 8691–8713.

[95] A. Hinrichs, D. Krieg, E. Novak, and J. Vybíral. « Lower bounds for integration and recovery in L_2 ». In: *J. Complexity* 72 (2022), Paper No. 101662, 15.

[96] A. Hinrichs, D. Krieg, E. Novak, and J. Vybíral. « Lower bounds for the error of quadrature formulas for Hilbert spaces ». In: *J. Complexity* 65 (2021), Paper No. 101544, 20.

[97] A. Hinrichs, E. Novak, and J. Vybíral. « Linear information versus function evaluations for L_2-approximation ». In: *J. Approx. Theory* 153.1 (2008), pp. 97–107.

[98] A. Hinrichs, J. Prochno, and M. Sonnleitner. « Random sections of ℓ_p-ellipsoids, optimal recovery and Gelfand numbers of diagonal operators ». In: *arXiv preprint arXiv:2109.14504* (2021).

[99] T. Hrycak and K. Gröchenig. « Pseudospectral Fourier reconstruction with the modified inverse polynomial reconstruction method ». In: *Journal of Computational Physics* 229.3 (2010), pp. 933–946.

[100] T. Jahn, T. Ullrich, and F. Voigtlaender. « Sampling numbers of smoothness classes via ℓ_1-minimization ». In: *arXiv preprint arXiv:2212.00445* (2022).

[101] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik. *Homogenization of differential operators and integral functionals*. Springer Science & Business Media, 2012.

[102] M. Kadets and M. Snobar. « Certain functionals on the Minkowski compactum ». In: *Mat. Zametki* 10.453-457 (1971), p. 9.

[103] R. V. Kadison and I. M. Singer. « Extensions of pure states ». In: *Amer. J. Math.* 81 (1959), pp. 383–400.

[104] L. Kämmerer, T. Ullrich, and T. Volkmer. « Worst-case recovery guarantees for least squares approximation using random samples ». In: *Constr. Approx.* 54.2 (2021), pp. 295–352.

[105] B. Kashin, E. Kosov, I. Limonova, and V. Temlyakov. « Sampling discretization and related problems ». In: *J. Complexity* 71 (2022), Paper No. 101653, 55.

[106] D. Krieg. « Optimal Monte Carlo methods for L^2-approximation ». In: *Constr. Approx.* 49.2 (2019), pp. 385–403.

[107] D. Krieg. « Tensor power sequences and the approximation of tensor product operators ». In: *J. Complexity* 44 (2018), pp. 30–51.

[108] D. Krieg, E. Novak, and M. Sonnleitner. « Recovery of Sobolev functions restricted to iid sampling ». In: *Math. Comp.* 91.338 (2022), pp. 2715–2738.

[109] D. Krieg, K. Pozharska, M. Ullrich, and T. Ullrich. « Sampling recovery in the uniform norm ». In: *arXiv preprint arXiv:2305.07539* (2023).

[110] D. Krieg, M. Siedlecki Paweł and Ullrich, and H. Woźniakowski. « Exponential tractability of L_2-approximation with function values ». In: *Adv. Comput. Math.* 49.2 (2023), Paper No. 18, 13.

[111] D. Krieg and M. Sonnleitner. « Function recovery on manifolds using scattered data ». In: *arXiv preprint arXiv:2109.04106* (2021).

[112] D. Krieg and M. Sonnleitner. « Random points are optimal for the approximation of Sobolev functions ». In: *arXiv preprint arXiv:2009.11275* (2020).

[113] D. Krieg and M. Ullrich. « Function values are enough for L_2-approximation ». In: *Found. Comput. Math.* 21.4 (2021), pp. 1141–1151.

[114] D. Krieg and M. Ullrich. « Function values are enough for L_2-approximation: Part II ». In: *J. Complexity* 66 (2021), Paper No. 101569, 14.

[115] A. Kroó. « Christoffel functions on convex and starlike domains in $\mathbb{R}^d$ ». In: *J. Math. Anal. Appl.* 421.1 (2015), pp. 718–729.

[116] T. Kühn, W. Sickel, and T. Ullrich. « Approximation of mixed order Sobolev functions on the d-torus: asymptotics, preasymptotics, and d-dependence ». In: *Constr. Approx.* 42.3 (2015), pp. 353–398.

[117] F. Y. Kuo, G. W. Wasilkowski, and H. Woźniakowski. « On the power of standard information for multivariate approximation in the worst case setting ». In: *J. Approx. Theory* 158.1 (2009), pp. 97–125.

[118] Y. T. Lee and H. Sun. « An SDP-based algorithm for linear-sized spectral sparsification ». In: *STOC'17—Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2017, pp. 678–687.

[119] Y. T. Lee and H. Sun. « Constructing linear-sized spectral sparsification in almost-linear time ». In: *SIAM J. Comput.* 47.6 (2018), pp. 2315–2336.

[120] F. Leja. « Sur certaines suites liées aux ensembles plans et leur application à la représentation conforme ». In: *Annales Polonici Mathematici*. 1 4. 1957, pp. 8–13.

[121] I. Limonova and V. Temlyakov. « On sampling discretization in L_2 ». In: *J. Math. Anal. Appl.* 515.2 (2022), Paper No. 126457, 14.

[122] W. Lu and H. Wang. « On the power of standard information for tractability for L_2-approximation in the average case setting ». In: *J. Complexity* 70 (2022), Paper No. 101618, 22.

[123] Y. Maday, N. C. Nguyen, A. T. Patera, and G. S. H. Pau. « A general multipurpose interpolation procedure: the magic points ». In: *Commun. Pure Appl. Anal.* 8.1 (2009), pp. 383–404.

[124] Y. Maday, A. T. Patera, J. D. Penn, and M. Yano. « A parameterized-background data-weak approach to variational data assimilation: formulation, analysis, and application to acoustics ». In: *International Journal for Numerical Methods in Engineering* 102.5 (2015), pp. 933–965.

[125] Y. Maday and B. Stamm. « Locally adaptive greedy approximations for anisotropic parameter reduced basis spaces ». In: *SIAM J. Sci. Comput.* 35.6 (2013), A2417–A2441.

[126] Y. Makovoz. « Random approximants and neural networks ». In: *Journal of Approximation Theory* 85.1 (1996), pp. 98–109.

[127] A. W. Marcus, D. A. Spielman, and N. Srivastava. « Interlacing families I: Bipartite Ramanujan graphs of all degrees ». In: *Ann. of Math. (2)* 182.1 (2015), pp. 307–325.

[128] A. W. Marcus, D. A. Spielman, and N. Srivastava. « Interlacing families II: Mixed characteristic polynomials and the Kadison-Singer problem ». In: *Ann. of Math. (2)* 182.1 (2015), pp. 327–350.

[129] A. Marquina and S. J. Osher. « Image super-resolution by TV-regularization and Bregman iteration ». In: *Journal of Scientific Computing* 37.3 (2008), pp. 367–382.

[130] S. Mendelson and A. Pajor. « On singular values of matrices with independent rows ». In: *Bernoulli* 12.5 (2006), pp. 761–773.

[131] C. A. Micchelli and T. J. Rivlin. *A survey of optimal recovery*. Springer, 1977.

[132] G. Migliorati. « Adaptive approximation by optimal weighted least-squares methods ». In: *SIAM J. Numer. Anal.* 57.5 (2019), pp. 2217–2245.

[133] G. Migliorati. « Multivariate approximation of functions on irregular domains by weighted least-squares methods ». In: *IMA J. Numer. Anal.* 41.2 (2021), pp. 1293–1317.

[134] B. S. Mitjagin. « Approximation of functions in L^p and C spaces on the torus ». In: *Mat. Sb. (N.S.)* 58 (100) (1962), pp. 397–414.

[135] M. Moeller and T. Ullrich. « L_2-norm sampling discretization and recovery of functions from RKHS with finite trace ». In: *Sampl. Theory Signal Process. Data Anal.* 19.2 (2021), Paper No. 13, 31.

[136] O. Mula. « Inverse problems: A deterministic approach using physics-based reduced models ». In: *arXiv preprint arXiv:2203.07769* (2022).

[137] N. Nagel, M. Schäfer, and T. Ullrich. « A new upper bound for sampling numbers ». In: *Foundations of Computational Mathematics* 22.2 (2022), pp. 445–468.

[138] A. Narayan, J. D. Jakeman, and T. Zhou. « A Christoffel function weighted least squares algorithm for collocation approximations ». In: *Math. Comp.* 86.306 (2017), pp. 1913–1947.

[139] S. Nitzan, A. Olevskii, and A. Ulanovskii. « A few remarks on sampling of signals with small spectrum ». In: *Tr. Mat. Inst. Steklova* 280.Ortogonalńye Ryady, Teoriya Priblizhenii i Smezhnye Voprosy (2013), pp. 247–254.

[140] S. Nitzan, A. Olevskii, and A. Ulanovskii. « Exponential frames on unbounded sets ». In: *Proc. Amer. Math. Soc.* 144.1 (2016), pp. 109–118.

[141] E. Novak and H. Woźniakowski. « Tractability of multivariate problems for standard and linear information in the worst case setting: Part I ». In: *J. Approx. Theory* 207 (2016), pp. 177–192.

[142] E. Novak and H. Woźniakowski. *Tractability of multivariate problems. Vol. 1: Linear information*. Vol. 6. EMS Tracts in Mathematics. European Mathematical Society (EMS), Zürich, 2008, pp. xii+384.

[143] E. Novak and H. Woźniakowski. *Tractability of multivariate problems. Volume II: Standard information for functionals*. Vol. 12. EMS Tracts in Mathematics. European Mathematical Society (EMS), Zürich, 2010, pp. xviii+657.

[144] E. Novak and H. Woźniakowski. *Tractability of multivariate problems. Volume III: Standard information for operators*. Vol. 18. EMS Tracts in Mathematics. European Mathematical Society (EMS), Zürich, 2012, pp. xviii+586.

[145] M. Ohlberger and S. Rave. « Reduced Basis Methods: Success, Limitations and Future Challenges ». In: *Proceedings of the Conference Algoritmy*. 2016, pp. 1–12.

[146] G. Peyré, S. Bougleux, and L. D. Cohen. « Non-local regularization of inverse problems ». In: *Inverse Problems and Imaging* 5.2 (2011), pp. 511–530.

[147] A. Pietsch. *Eigenvalues and s-numbers*. Vol. 43. Mathematik und ihre Anwendungen in Physik und Technik [Mathematics and its Applications in Physics and Technology]. Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig, 1987, p. 360.

[148] J. E. Pilliod Jr and E. G. Puckett. « Second-order accurate volume-of-fluid algorithms for tracking material interfaces ». In: *Journal of Computational Physics* 199.2 (2004), pp. 465–502.

[149] J. E. Pilliod. *An analysis of piecewise linear interface reconstruction algorithms for volume-of-fluid methods*. U. of Calif., Davis, 1992.

[150] A. Pinkus. *n-widths in approximation theory*. Vol. 7. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]. Springer-Verlag, Berlin, 1985, pp. x+291.

[151] G. Pisier. *The volume of convex bodies and Banach space geometry*. Vol. 94. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1989, pp. xvi+250.

[152] K. Pozharska and T. Ullrich. « A note on sampling recovery of multivariate functions in the uniform norm ». In: *SIAM J. Numer. Anal.* 60.3 (2022), pp. 1363–1384.

[153] A. Prymak and O. Usoltseva. « Christoffel functions on planar domains with piecewise smooth boundary ». In: *Acta Math. Hungar.* 158.1 (2019), pp. 216–234.

[154] E. G. Puckett. « A volume-of-fluid interface tracking algorithm with applications to computing shock wave refraction ». In: *proceedings of the fourth international symposium on Computational Fluid Dynamics*. 1991, pp. 933–938.

[155] H. Rauhut and R. Ward. « Interpolation via weighted ℓ_1 minimization ». In: *Appl. Comput. Harmon. Anal.* 40.2 (2016), pp. 321–351.

[156] H. Rauhut and R. Ward. « Sparse Legendre expansions via ℓ_1-minimization ». In: *J. Approx. Theory* 164.5 (2012), pp. 517–533.

[157] G. Rozza, D. B. P. Huynh, and A. T. Patera. « Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations: application to transport and continuum mechanics ». In: *Arch. Comput. Methods Eng.* 15.3 (2008), pp. 229–275.

[158] C. Runge. « Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten ». In: *Zeitschrift für Mathematik und Physik* 46.224-243 (1901), p. 20.

[159] R. Salem. « A new proof of a theorem of Menchoff ». In: *Duke Math. J.* 8 (1941), pp. 269–272.

[160] S. Sen. « Reduced-basis approximation and a posteriori error estimation for many-parameter heat conduction problems ». In: *Numerical Heat Transfer, Part B: Fundamentals* 54.5 (2008), pp. 369–389.

[161] W. Sickel and T. Ullrich. « The Smolyak algorithm, sampling on sparse grids and function spaces of dominating mixed smoothness ». In: *East J. Approx.* 13.4 (2007), pp. 387–425.

[162] J. W. Siegel and J. Xu. « Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks ». In: *Found. Comput. Math.* (2022), pp. 1–57.

[163] S. A. Smolyak. « Quadrature and interpolation formulas for tensor products of certain classes of functions ». In: *Doklady Akademii Nauk*. Vol. 148. Russian Academy of Sciences. 1963, pp. 1042–1045.

[164] D. A. Spielman and N. Srivastava. « Graph sparsification by effective resistances ». In: *SIAM J. Comput.* 40.6 (2011), pp. 1913–1926.

[165] E. M. Stein. *Singular integrals and differentiability properties of functions*. Princeton Mathematical Series, No. 30. Princeton University Press, Princeton, N.J., 1970, pp. xiv+290.

[166] R. Stevenson. « An optimal adaptive finite element method ». In: *SIAM journal on numerical analysis* 42.5 (2005), pp. 2188–2217.

[167] V. Temlyakov. *Multivariate approximation*. Vol. 32. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2018, pp. xvi+534.

[168] V. Temlyakov. « On optimal recovery in L_2 ». In: *J. Complexity* 65 (2021), Paper No. 101545, 11.

[169] V. N. Temlyakov. « Nonlinear Kolmogorov widths ». In: *Mat. Zametki* 63.6 (1998), pp. 891–902.

[170] V. N. Temlyakov. « The Marcinkiewicz-type discretization theorems ». In: *Constr. Approx.* 48.2 (2018), pp. 337–369.

[171] V. Temlyakov and T. Ullrich. « Bounds on Kolmogorov widths and sampling recovery for classes with small mixed smoothness ». In: *J. Complexity* 67 (2021), Paper No. 101575, 19.

[172] V. N. Temlyakov. « On a way of obtaining lower estimates for the errors of quadrature formulas ». In: *Mathematics of the USSR-Sbornik* 71.1 (1992), p. 247.

[173] V. N. Temlyakov and T. Ullrich. « Approximation of functions with small mixed smoothness in the uniform norm ». In: *J. Approx. Theory* 277 (2022), Paper No. 105718, 23.

[174] H. Tran, C. G. Webster, and G. Zhang. « Analysis of quasi-optimal polynomial approximations for parameterized PDEs with deterministic and stochastic coefficients ». In: *Numerische Mathematik* 137.2 (2017), pp. 451–493.

[175] J. F. Traub and H. Woźniakowsi. *A general theory of optimal algorithms*. ACM Monograph Series. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1980, pp. xiv+341.

[176] J. A. Tropp. « User-friendly tail bounds for sums of random matrices ». In: *Found. Comput. Math.* 12.4 (2012), pp. 389–434.

[177] M. Ullrich. « On the worst-case error of least squares algorithms for L_2-approximation with high probability ». In: *J. Complexity* 60 (2020), pp. 101484, 6.

[178] K. Veroy, C. Prud'Homme, D. Rovas, and A. Patera. « A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations ». In: *16th AIAA Computational Fluid Dynamics Conference*. 2003, p. 3847.

[179] K. Veroy, D. V. Rovas, and A. T. Patera. « A posteriori error estimation for reduced-basis approximation of parametrized elliptic coercive partial differential equations:"convex inverse" bound conditioners ». In: *ESAIM: Control, Optimisation and Calculus of Variations* 8 (2002), pp. 1007–1028.

[180] S. Volkwein. « Proper orthogonal decomposition: Theory and reduced-order modelling ». In: *Lecture Notes, University of Konstanz* 4.4 (2013), pp. 1–29.

[181] J. Vybíral. « A variant of Schur's product theorem and its applications ». In: *Adv. Math.* 368 (2020), pp. 107140, 9.

[182] M. J. Wainwright. *High-dimensional statistics*. Vol. 48. Cambridge Series in Statistical and Probabilistic Mathematics. A non-asymptotic viewpoint. Cambridge University Press, Cambridge, 2019, pp. xvii+552.

[183] Z. Wang, J. Chen, and S. C. Hoi. « Deep learning for image super-resolution: A survey ». In: *IEEE transactions on pattern analysis and machine intelligence* 43.10 (2020), pp. 3365–3387.

[184] G. W. Wasilkowski and H. Woźniakowski. « On the power of standard information for weighted approximation ». In: *Found. Comput. Math.* 1.4 (2001), pp. 417–434.

[185] G. W. Wasilkowski and H. Woźniakowski. « The power of standard information for multivariate approximation in the randomized setting ». In: *Math. Comp.* 76.258 (2007), pp. 965–988.

[186] G. W. Wasilkowski and H. Woźniakowski. « Explicit cost bounds of algorithms for multivariate tensor product problems ». In: *J. Complexity* 11.1 (1995), pp. 1–56.

[187] N. Weaver. « The Kadison-Singer problem in discrepancy theory ». In: *Discrete Math.* 278.1-3 (2004), pp. 227–239.

[188] G. Welper. « Transformed snapshot interpolation with high resolution transforms ». In: *SIAM J. Sci. Comput.* 42.4 (2020), A2037–A2061.

[189] H. Wendland. *Scattered data approximation*. Vol. 17. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2005, pp. x+336.

[190] K. Willcox and J. Peraire. « Balanced model reduction via the proper orthogonal decomposition ». In: *AIAA journal* 40.11 (2002), pp. 2323–2330.

[191] Y. Xu. « Asymptotics for orthogonal polynomials and Christoffel functions on a ball ». In: *Methods Appl. Anal.* 3.2 (1996), pp. 257–272.

[192] A. Ženíšek. « Extensions from the Sobolev spaces H^1 satisfying prescribed Dirichlet boundary conditions ». In: *Appl. Math.* 49.5 (2004), pp. 405–413.

[193] K. Zhang, D. Tao, X. Gao, X. Li, and J. Li. « Coarse-to-fine learning for single-image super-resolution ». In: *IEEE transactions on neural networks and learning systems* 28.5 (2016), pp. 1109–1122.

[194] Z. Zou, D. Kouri, and W. Aquino. « An adaptive local reduced basis method for solving PDEs with uncertain inputs and evaluating risk ». In: *Computer Methods in Applied Mechanics and Engineering* 345 (2019), pp. 302–322.