RESTRICTED ISOMETRY AND INFORMATION-BASED NUMERICAL ANALYSIS

by

Wu, Haoning

A Thesis Submitted in Partial Fulfilment of the Requirements for
the Degree of Doctor of Philosophy
at the University of Hong Kong.

July, 2023

Intentionally blank page.

Abstract of thesis entitled

RESTRICTED ISOMETRY AND INFORMATION-BASED NUMERICAL ANALYSIS

Submitted by

Wu, Haoning

for the degree of Doctor of Philosophy

at The University of Hong Kong

in July, 2023

This thesis aims to advance numerical analysis towards practical scenarios in which the available information for numerical algorithms is partial, contaminated, and priced. These scenarios represent information-based situations crucial in practical applications, particularly when the available information is predetermined or obtaining sufficient information can be difficult and costly. The resulting paradigm is referred to as information-based numerical analysis, and the main motivation is to develop numerical algorithms that achieve a reasonable level of accuracy with limited and possibly predetermined information. The thesis explores this new paradigm with concrete numerical algorithms in three numerical analysis topics.

The first part of the thesis focuses on numerical approximation, specifically on hyperinterpolation. Hyperinterpolation is a quadrature-based discretization of the orthogonal projection, and constructing a hyperinterpolant of degree n requires a positive-weight quadrature rule with exactness degree $2n$. Using the Marcinkiewicz–Zygmund (MZ) system of quadrature rules, which can be regarded as a restricted isometry of the quadrature rules in the numerical integration of polynomials of certain degrees, the thesis shows that hyperinterpolation can be constructed with a reasonable error bound in the presence of the quadrature exactness assumption, suggesting that it is reliable in regard of the information-based situation. The thesis also proposes a variant of this scheme in approximating singular and oscillatory functions, leveraging the product-integration method and attaining the desired accuracy with fewer quadrature points than the original hyperinterpolation.

The second part of the thesis proposes and analyzes a quadrature-based spectral method for solving the Allen–Cahn equation on spheres based on the approximation results in the previous part. This method employs hyperinterpolation and the MZ system of quadrature rules, achieving the theoretical benefits of the Galerkin method

at a computational cost comparable to the collocation method. This method does not necessarily rely on the quadrature exactness assumption, which confronts more practical simulations situations and distinguishes it from previous quadrature-based methods. Moreover, this method also possesses an effective maximum principle, which allows the numerical solutions to deviate from the sharp bound by a controllable discretization error.

The third part of the thesis focuses on compressed sensing and imaging, investigating models that reconstruct unknown signals and images from their incomplete and inaccurate measurements. These models can be formulated as solving an underdetermined linear system. The thesis introduces the springback penalty and the enhanced total variation (TV) regularization and establishes the exact and stable reconstruction theory for these models under the restricted isometry property (RIP) framework. The thesis shows that the springback and enhanced TV models have tighter reconstruction error bounds than various convex and non-convex models for scenarios where the amount of measurements is limited and the level of noise is significant.

Overall, this thesis contributes to the development of numerical algorithms that require less information while achieving the desired level of accuracy. The proposed algorithms in this thesis demonstrate their effectiveness and theoretical superiority over existing algorithms in various numerical analysis topics regarding information-based situations. The investigation in this thesis may enlighten the development of numerical algorithms for other problems that involve information-based situations.

RESTRICTED ISOMETRY AND INFORMATION-BASED NUMERICAL ANALYSIS

by

Wu, Haoning
B.Sc., *Jinan*

A Thesis Submitted in Partial Fulfilment of the Requirements for
the Degree of Doctor of Philosophy
at the University of Hong Kong.

July, 2023

Intentionally blank page.

Declaration

I hereby declare that this thesis represents my own work, except where due acknowledgement is made, and that it has not been previously included in a thesis, dissertation or report submitted to the University of Hong Kong or to any other institution for a degree, diploma or other qualifications.

Signed: 

Date: July 20, 2023

Intentionally blank page.

In memory of my grandfather

Intentionally blank page.

Acknowledgements

This thesis would not have been possible without the generous support, guidance, and encouragement of many individuals. Foremost among them is my advisor, Prof. Xiaoming Yuan, to whom I owe countless thanks for his exceptional mentorship and unwavering support throughout my academic journey. His academic and personal guidance have been pivotal in shaping my research taste and developing my own scientific outlook. His senses of mission and responsibility towards the society have always been an inspiration to me, and I strive to emulate his dedication to making a positive impact in the world. Moreover, I am deeply appreciative of his open-mindedness and encouragement to pursue my own research interests. It has been an absolute privilege to be his student.

Especially I wish to express my thanks to Dr. Congpei An of the Southwestern University of Finance and Economics for his encouragement, support, and accompany. He brought me into the wonder land of computational mathematics during my undergraduate years in Guangzhou, and my research has greatly benefited from our regular conversation and continued collaboration since then.

I am deeply grateful to Prof. Ding-Xuan Zhou for his hospitality during my visit to the University of Sydney in January 2023. I learned a great deal about approximation theory and learning theory from Prof. Zhou. I would also like to express my sincere gratitude to Prof. Ian H. Sloan of the University of New South Wales for offering me constructive advice on the first two parts of this thesis during our talk in his office. Prof. Sloan has undoubtedly made a profound impact on me in the fields of numerical analysis and approximation theory, which is evident throughout this thesis. My special thanks also go to Prof. Xiaobing Feng of the University of Tennessee for his feedback on my research and for kindly sharing his recent achievements with me during his short stay in Hong Kong in December 2022. Moreover, I would like to express my particular appreciation to Prof. Dr. Gitta Kutyniok of Ludwig-Maximilians-Universität München and Dr. Xiaosheng Zhuang of City University of Hong Kong for their generous support on my academic voyage.

Spending four years as a research postgraduate student in the Department of Mathematics was an incredible experience. I would like to express my heartfelt appreciation to Dr. Zhiwen Zhang for his warm support and attention on my research. I am also very grateful for Dr. Zheng Qu for her help in my academic journey. Additionally, I wish to extend my sincere thanks to our administrative staff in RR 408 and Sis. Ming for their excellent service and kind help. Furthermore, I am very

appreciative of the scholarships and financial support provided by the University and the Department, which enabled me to pursue my academic goals.

Upon completion of this thesis, I am especially grateful to Prof. Feng, Prof. Dr. Kutyniok, Dr. Zhang, and Dr. Zhuang for their invaluable time and dedication in my thesis examination.

Last but far from least, I am deeply indebted to my family, including my parents, grandparents, and all other family members, for their unconditional love and support. I am particularly grateful to my late grandfather, to whom this thesis is dedicated, for his encouragement and inspiration. Their belief in me has been a constant source of motivation from my childhood to the embarking of an academic journey.

Contents

Notations

General

$\mathbb{N}$	set of positive integers
$\mathbb{N}_0$	set of non-negative integers
$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\mathbb{R}^d$	set of the d-tuples of real numbers
$\mathbb{S}^d$	unit d-sphere $\{x \in \mathbb{R}^{d+1} : \|x\|_2 \leq 1\}$ in $\mathbb{R}^{d+1}$ $(d \geq 2)$
Ω	closure of a connected open domain or a smooth closed lower-dimensional manifold
(a, b)	open interval $\{x \in \mathbb{R} : a < x < b\}$ in $\mathbb{R}$
$[a, b]$	closed interval $\{x \in \mathbb{R} : a \leq x \leq b\}$ in $\mathbb{R}$

$\mathrm{dist}(x, y)$	geodesic distance between $x, y \in \mathbb{S}^d$
$\lfloor x \rfloor$	integer part of $x \in \mathbb{R}$
$\mathrm{sgn}(x)$	sign function of $x \in \mathbb{R}$
$\mathrm{supp}(x)$	support $\{1 \leq i \leq n : x_i \neq 0\}$ of $x \in \mathbb{R}^n$
$x \circ y$	Hadamard (entry-wise) product between x and y
δ_{ij}	Kronecker delta
$\Gamma(\cdot)$	Gamma function
$(\cdot)_n$	Pochhammer symbol
I	identity matrix
i.i.d.	independent and identically distributed

Asymptotics

$\mathcal{O}(\cdot)$	Big-Oh notation
$\sim$	$f(x) \sim g(x)$ denotes $f(x)/g(x) \to 1$ as $x \to \infty$
$\lesssim$	$a \lesssim b$ denotes there exists $c > 0$ such that $a \leq cb$
$\gtrsim$	$a \gtrsim b$ denotes there exists $c > 0$ such that $a \geq cb$
$\asymp$	$a_n \asymp b_n$ denotes there exists $c_1, c_2 > 0$ independent of n such that $c_1 a_n \leq b_n \leq c_2 a_n$

$a \ll b$	a is much less than b
$a \gg b$	a is much greater than b

Functional Analysis

$C(\Omega)$	space of continuous functions over Ω
$L^p(\Omega)$	space of all measurable functions from Ω to $\mathbb{R}$ or $\mathbb{C}$ whose absolute value raised to the p-th power has a finite integral
$W^{k,p}(\Omega)$	Sobolev space of order k for $1 \leq p \leq \infty$ over Ω
$H^k(\Omega)$	Sobolev space of order k with $p = 2$ over Ω
ℓ^p	collection of sequences of numbers for which the p-th power of each term has a finite sum.
$\|\cdot\|_S$	norm equipping the space S
$\|\cdot\|_p$	L^p or ℓ^p norm
$\langle\cdot,\cdot\rangle$	inner product
$\langle\cdot,\cdot\rangle_m$	discrete inner product by an m-point quadrature
$\partial(f(x))$	subdifferential of convex function f at x

Approximation Theory

$\mathbb{P}_n(\Omega)$	space of polynomials over Ω of total degree at most n
$\dim \mathbb{P}_n$	dimension of $\mathbb{P}_n(\Omega)$
d_n	dimension of $\mathbb{P}_n(\Omega)$
$\{p_\ell\}_{\ell=1}^{\dim \mathbb{P}_n}$	an L^2-orthonormal basis for $\mathbb{P}_n(\Omega)$
χ	generic polynomial
$\mathcal{P}_n$	orthogonal projection operator of degree n
$\mathcal{L}_n$	hyperinterpolation operator of degree n
$\mathcal{S}_n$	efficient hyperinterpolation operator of degree n
$\mathcal{U}_n$	unfettered hyperinterpolation operator of degree n
$\mathcal{Q}_n$	QMC hyperinterpolation operator of degree n
$E_n(f)$	best uniform approximation error of $f \in C(\Omega)$ by polynomials in $\mathbb{P}_n(\Omega)$
$Y_{\ell,k}$	real-valued spherical harmonic of degree ℓ and index k
$Z(d,\ell)$	number of mutually orthonormal real-valued spherical harmonics of degree ℓ

Differential Equations

∇	gradient or first-order Beltrami operator
Δ	Laplace or Laplace–Beltrami operator
$\mathbf{L}$	constant-coefficient linear differential operator

$\mathbf{N}$	constant-coefficient nonlinear differential (or non-differential) operator of lower order
$\mathcal{E}(u)$	energy functional of solution u
$\tilde{\mathcal{E}}(u)$	discrete energy functional of solution u

Compressed sensing and imaging

x	signal
$\bar{x}$	ground-truth signal
x^{opt}	recovered signal by solving some minimization problems (in theory)
x^*	global or local minimum point by solving some minimization problems numerically
A	sensing matrix
b	measurements of $\bar{x}$ (possibly contaminated by noise)

X	image
$\bar{X}$	ground-truth image
X^{opt}	recovered signal by solving some minimization problems (in theory)
X^*	global or local minimum point by solving some minimization problems numerically
$\mathcal{M}$	sensing operator
y	measurements of $\bar{X}$ (possibly contaminated by noise)
$\mathcal{H}$	bivariate Haar transform
$\mathcal{F}$	bivariate discrete Fourier transform

Λ	index set as a subset of $\{1, 2, \ldots, n\}$ (or with some super/subscripts)
$\lvert\Lambda\rvert$	cardinality of Λ
x_Λ	vector with the same entries as $x \in \mathbb{R}^n$ on indices Λ and zero entries on indices Λ^c
A_Λ	submatrix of A in $\mathbb{R}^{m\times\lvert\Lambda\rvert}$ with column indices Λ

Intentionally blank page.

Chapter 1

Introduction

This thesis delves into the practical aspects of numerical analysis, which is the field of study that deals with numerical methods that aim to find approximate solutions to problems rather than exact ones. The scope of numerical analysis encompasses various subareas, including but not limited to approximation and interpolation, numerical differentiation and integration, numerical linear algebra, numerical optimization, finding roots of nonlinear equations, numerical methods for ordinary and partial differential equations, and numerical methods for integral equations. Each of these subareas corresponds to a topic in mathematical analysis. In his essay [218] entitled "The definition of numerical analysis," Lloyd N. Trefethen presented his definition of numerical analysis, stating that

> *Numerical analysis is the study of algorithms for the problems of continuous mathematics.*

This definition has helped to shift the perception of numerical analysis from the study of rounding errors to the study of algorithms. As included in an appendix to Trefethen's popular book [223] on numerical linear algebra, this definition might have become one of the most familiar ones to numerical analysts of our generation.

In order to implement the developed numerical algorithms and further apply them to real-world problems, it is often necessary to approximate the continuous problem with a discrete one. This process, known as *discretization*, is a fundamental part of numerical analysis due to the floating point arithmetic of digital computers. Digital computers use a finite number of bits to represent a real number, resulting in a representation of only a finite subset of the real numbers. Consequently, there are gaps between the represented numbers, and we must take discrete samples of continuous objects for numerical computation. To analyze the accuracy, stability, and conditioning of algorithms in the context of floating point arithmetic, we refer to [113].

1.1 Information-based numerical analysis

Each subarea of numerical analysis emphasizes where to take the discrete samples. For example, in the case of univariate numerical approximation, the choice of sampling points depends on the nature of the function to be approximated. For periodic functions, equispaced points are a viable option, while for non-periodic functions, equispaced points lead to the undesirable Runge's phenomenon. Instead, Chebyshev technology can be used, which involves sampling the function at Chebyshev points. The superiority of Chebyshev technology in practice has been extensively documented in [221, Chapter 1], and this technology has also been extended to spectral methods such as pseudospectral methods [219] and spectral Galerkin methods [190].

The development of these numerical algorithms assumes full access to the function f, which is sufficient for computer implementation. However, there are still gaps between numerical analysis and real-world application, where the main motivation of this thesis arises. In practice, we may not be fortunate enough to be granted full access to the function f but only a set of samples of it (or more generally, a set of linear functional of f). Furthermore, obtaining these samples can be difficult and costly, and the sampling process may be contaminated. In addition, the distribution of samples may be predetermined by the problem at hand or by engineers, limiting our control over the sampling locations.

Despite these challenges, we may have some knowledge of the global properties of f, such as its belonging to a class of smooth, convex, or periodic functions. Even if not granted full access to f, we are still intrigued to develop numerical algorithms that can produce *reliable* results with limited information. This is what we call *information-based numerical analysis* in this thesis:

> *Information-based numerical analysis is the study of algorithms for the problems of continuous mathematics without full access to the concerned objects but only partial, contaminated, and priced information.*

In other words, this thesis aims to develop numerical methods that are robust to imperfect and limited sampling, with the number of needed samples reduced. The term "information-based," borrowed from the field of information-based complexity [217, 236], refers to the situation where

- Information is *partial*. That is, having *a priori* knowledge and a finite set of samples, we cannot, in general, solve the continuous mathematics problem exactly and uniquely.

- Information is *contaminated*. That is, it is computed with errors. Examples include sampling noise and round-off errors.
- Information is *priced*. That is, we are charged for each sample.

Here information refers not to Claude Shannon and information theory but to what we know about the problem to be solved.

The issue of information-based complexity has also sparked the interest of numerical analysts, as seen in Ivo Babuška's survey [16] on quadrature and finite element methods and Mike Powell's works [165, 166, 167, 168, 169, 170] on derivative-free optimization. However, our focus differs from the field of information-based complexity in the sense that we are not concerned with estimating the total cost of acquiring information and computation to obtain an approximate solution. Instead, we aim to develop algorithms that provide reliable solutions for problems of interest with a reasonable error bound.

Numerical analysis is a well-established field, and our methodology in this thesis enhances existing numerical methods by reducing the required number of samples or by weakening and even dismissing the original requirements and replacing them with milder ones on samples. A noteworthy feature of these mild conditions in this thesis is that they all represent the *restricted isometry* of the samples. Our investigation in this thesis focuses on numerical approximation, numerical methods for partial differential equations (PDEs), and compressed sensing (problems formulated as solving underdetermined systems of linear equations) among various subareas of numerical analysis, and we hope to extend our approach to other subareas that involve information-based situations in the future. We specifically focus on problems with domains of more than one dimension.

The paper is thus divided into three parts: polynomial approximation, numerical solutions to PDEs, and compressed sensing and imaging. Our goal is to develop algorithms that can reliably solve these problems with limited and imperfect information.

1.2 Polynomial approximation

In **Chapters 2–4**, we investigate polynomial approximation of functions f from a set of discrete samples $\{f(x_j)\}_{j=1}^m$ at m points $\{x_j\}_{j=1}^m$. In numerical analysis, two conventional methods are interpolation and approximation. The univariate polynomial approximation is a must-have chapter in almost every numerical analysis textbook. However, in the multivariate setting, the Mairhuber–Curtis Theorem states that there are no Haar spaces [232, Chapter 2]; that is, it is impossible to interpolate all

kinds of data at any set of samples with size equaling the dimension of the space of polynomials of degree at most n by a specific polynomial of degree n. If one still wants to interpolate function samples, he may consider radial basis function interpolation via solving some large-scale interpolation equations, which is beyond the scope of this thesis. For radial basis function approximation, we refer to [232].

For approximation, the discrete least squares approximation might be a simple but powerful method, which aims to find a polynomial p of degree n that minimizes the loss function $\sum_{j=1}^{m}[f(x_j) - p(x_j)]^2$. However, it is sometimes hard to analyze the approximation error of f by p unless we have an explicit expression of p or its coefficients with respect to some orthonormal basis.

Besides, an important tool in numerical analysis is the orthogonal projection. Let Ω be a bounded region of $\mathbb{R}^d$ with measure $\mathrm{d}\omega$, which is either the closure of a connected open domain or a smooth closed lower-dimensional manifold in $\mathbb{R}^d$. This region is assumed to have finite measure with respect to $\mathrm{d}\omega$. We denote by $\mathbb{P}_n \subset L^2(\Omega)$ the linear space of polynomials on Ω of degree at most n, equipped with the L^2 inner product

$$\langle v, z\rangle = \int_{\Omega} vz\mathrm{d}\omega. \tag{1.2.1}$$

Let $\{p_1, p_2 \ldots, p_{d_n}\} \subset \mathbb{P}_n$ be an orthonormal basis of $\mathbb{P}_n$ in the sense of

$$\langle p_\ell, p_{\ell'}\rangle = \delta_{\ell\ell'}$$

for $1 \le \ell, \ell' \le d_n$, where $d_n = \dim \mathbb{P}_n$ is the dimension of $\mathbb{P}_n$. Given an L^2 function f, the *orthogonal projection* of f onto the space $\mathbb{P}_n$ is defined as

$$\mathcal{P}_n f := \sum_{\ell=1}^{d_n} \langle f, p_\ell\rangle\, p_\ell \in \mathbb{P}_n. \tag{1.2.2}$$

However, the orthogonal projection cannot be implemented on computers because coefficients as inner products cannot be evaluated exactly.

In the early 1990s, Ian H. Sloan became intrigued by the conundrum of whether the interpolation of a periodic function on an interval (or equivalently, for a function on a circle) has properties as good as those of the more famous orthogonal projection (1.2.2). Though for spheres of dimension more than one and many other multidimensional regions, this is not the case and interpolation on spheres remains very problematic [234], a discrete approximation with the right properties is available, albeit using more points than interpolation. This approximation, now known as

hyperinterpolation, appeared in the paper [196], of which Sloan was rather proud[1]. Constructing hyperinterpolants requires an m-point quadrature rule of the form

$$\sum_{j=1}^{m} w_j g(x_j) \approx \int_{\Omega} g \mathrm{d}\omega, \tag{1.2.3}$$

where the quadrature points x_j belong to Ω and weights w_j are all positive for $j = 1, 2, \ldots, m$. For a comprehensive introduction to numerical integration, we refer to the classic book [68]. Assuming that the quadrature rule (1.2.3) has exactness degree $2n$, i.e.,

$$\sum_{j=1}^{m} w_j g(x_j) = \int_{\Omega} g \mathrm{d}\omega \quad \forall g \in \mathbb{P}_{2n}, \tag{1.2.4}$$

the hyperinterpolation operator $\mathcal{L}_n : C(\Omega) \to \mathbb{P}_n$ maps a continuous function $f \in C(\Omega)$ on Ω to

$$\mathcal{L}_n f := \sum_{\ell=1}^{d_n} \langle f, p_\ell \rangle_m p_\ell, \tag{1.2.5}$$

where

$$\langle v, z \rangle_m := \sum_{j=1}^{m} w_j v(x_j) z(x_j)$$

is a "discrete version" of the L^2 inner product (1.2.1). Thus, hyperinterpolation can be regarded as a discrete version of the orthogonal projection from $C(\Omega)$ onto $\mathbb{P}_n$ with respect to (1.2.1). Moreover, hyperinterpolation can be reduced to interpolation if a minimal quadrature exists and is applied — an m-point quadrature rule (1.2.3) is said to be *minimal* if it is exact for all polynomials of degree at most $2n$ and $m = \dim \mathbb{P}_n$. Furthermore, hyperinterpolation is a minimizer of the following discrete, weighted least squares approximation

$$\min_{p \in \mathbb{P}_n} \sum_{j=1}^{m} w_j [f(x_j) - p(x_j)]^2.$$

The subsequent development of hyperinterpolation has mainly focused on its application to the sphere, as documented in a number of works, including [67, 110, 128, 137, 177, 178, 202, 234]. Meanwhile, hyperinterpolation has also been explored on other regions, such as the disk [106], the square [39], the cube [40, 224], and spherical triangles [206]. While hyperinterpolation has been shown to be a powerful tool for approximating functions of multiple variables in all of these works, it should be noted that the exactness degree $2n$ of the quadrature rule (1.2.3) is

[1] See "A Fortunate Scientific Life" by Ian H. Sloan, included in the book entitled *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan.*

a central assumption in the construction of hyperinterpolants. This assumption is also maintained in some variants of hyperinterpolation, such as filtered hyperinterpolation [205] (which requires even more degrees) and some regularized versions of hyperinterpolation [8, 9, 10].

Given the highly restrictive nature of the quadrature exactness assumption, it is often impractical or impossible to obtain data on the desired quadrature points in practice. Therefore, one aim of this part is to relax and even bypass this assumption. We provide a recipe based on the Marcinkiewicz–Zygmund property, which assumes the existence of an $\eta \in [0, 1)$ such that

$$\left| \sum_{j=1}^{m} w_j \chi(x_j)^2 - \int_{\Omega} \chi^2 \mathrm{d}\omega \right| \leq \eta \int_{\Omega} \chi^2 \mathrm{d}\omega \quad \forall \chi \in \mathbb{P}_n.$$

If the quadrature exactness assumption (1.2.4) holds, then η is equal to 0. This property can be equivalently expressed as the Marcinkiewicz–Zygmund inequality [89, 143, 146] applied to polynomials of degree at most $2n$, and is referred to as the Marcinkiewicz-Zygmund property in this thesis. Specifically, it can be written as

$$(1-\eta) \int_{\Omega} \chi^2 \mathrm{d}\omega_d \leq \sum_{j=1}^{m} w_j \chi(x_j)^2 \leq (1+\eta) \int_{\Omega} \chi^2 \mathrm{d}\omega_d \quad \forall \chi \in \mathbb{P}_n.$$

This inequality can be interpreted as a *restricted isometry* of $\{\chi(x_j)\}_{j=1}^m$ in numerically evaluating the integral of χ^2 for any $\chi \in \mathbb{P}_n$. The motivation for introducing the Marcinkiewicz–Zygmund property is explained on page 21. With the property, it is shown that even if the quadrature exactness assumption (1.2.4) is weakened or dismissed, reasonable error bounds for the approximation by hyperinterpolation can still be obtained. Moreover, special attention is paid to examining the behavior of hyperinterpolation in approximating singular and oscillatory functions.

1.3 Numerical solutions to PDEs

In the second part of this thesis, consisting of **Chapter 5** only, we aim to apply our approximation scheme from the first part to compute smooth solutions of stiff, semilinear PDEs on the unit sphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\} \subset \mathbb{R}^d$ with dimension $d \geq 3$ of the form

$$u_t = \mathbf{L}u + \mathbf{N}(u), \quad u(0, x) = u_0(x),$$

where $u = u(t, x)$ with $(t, x) \in [0, \infty) \times \mathbb{S}^{d-1}$ is a function of time t and spatial variable $x \in \mathbb{S}^{d-1}$, $\mathbf{L}$ is a constant-coefficient linear differential operator, and $\mathbf{N}$ is a constant-coefficient nonlinear differential (or non-differential) operator of lower order.

In applications, equations of this kind typically arise when two or more different physical processes are combined, and many equations in science and engineering take this form (5.1.1).

PDEs on the sphere are often used to describe geological, meteorological, and oceanic problems, since the sphere can be regarded as a rough model of the earth. Moreover, solving PDEs on the sphere, which is the simplest version of smooth, compact manifolds, provides some insights for solving PDEs on an arbitrary, smooth, compact manifold.

As a model equation, we focus on the Allen–Cahn equation

$$u_t = \nu^2 \Delta u - F'(u), \quad u(0, x) = u_0(x),$$

where Δ is the Laplace–Beltrami operator on $\mathbb{S}^{d-1}$. This equation with linear diffusion $\nu^2 \Delta u$ and a nonlinear reaction term $F'(u)$ was introduced by Allen and Cahn in the 1970s to describe the process of phase separation in iron alloys [5]. In this reaction-diffusion equation, $u = u(t, x)$ is a scalar function typically representing the concentration of one of the two metallic components of the alloy. The nonlinear term has the usual double well form of

$$F'(u) = f(u) = u^3 - u$$

with

$$F(u) = \frac{1}{4}(u^2 - 1)^2.$$

We will also focus on the stiff case of $\nu \ll 1$, where numerical methods for solving the Allen–Cahn equation may be numerically unstable unless the time stepping size, which depends on ν, is extremely small. Thus, we aim to develop numerical methods that are stable with large time stepping sizes since long-time simulations of the Allen–Cahn equation and many other phase-field models are necessary for stable solutions. It is worth noting that the techniques presented in this chapter can be applied to more general PDEs.

The Allen–Cahn equation possesses two intrinsic properties, namely, energy stability and the maximum principle. In the literature, numerical methods for this equation have been designed to preserve these properties. As seen in [57, 87, 133, 192, 193, 241] and references therein, these methods aim to maintain energy stability. Additionally, many methods, such as [114, 136, 191], also aim to preserve the maximum principle of the numerical solutions.

However, in some cases, only modified energy stability can be analyzed, and

some undesired, stringent conditions on the numerical schemes are introduced for stability or accuracy, such as small time stepping sizes that depend on ν. These conditions can increase simulation time. Furthermore, these methods often rely on quadrature rules with exactness, which may not be practical when facing information-based situations, as previously discussed.

In this part, we present a quadrature-based spectral method for solving the Allen–Cahn equation using the approximation results from the previous part. Our motivation for this method is presented as follows. On the one hand, many existing numerical schemes for this equation assume small time stepping sizes that depend on ν, introducing stringent conditions on the numerical scheme and increasing the simulation time. We aim to lift these conditions and instead impose mild conditions solely on the polynomial degree of numerical solutions, independent of the time stepping size. This idea is motivated by a recent work [131], which proposes an effective maximum principle was proposed. This principle is an almost sharp maximum principle that allows the numerical solutions to deviate from the sharp bound by a controllable discretization error without introducing stringent conditions on the numerical scheme. Such an approach sounds practical and reasonable for numerical analysts. On the other hand, our method is designed to confront the information-based situation. It does not necessarily rely on the assumption of quadrature exactness for preserving the effective maximum principle and the L^∞ stability of the numerical solution. Even if a set of random samples of the initial condition $u(0,x) = u_0(x)$ is available, our numerical scheme still preserves the effective maximum principle. If the quadrature rule has sufficient exactness degrees, our method is energy stable and equivalent to the discrete Galerkin method.

1.4 Compressed sensing and imaging

In **Chapters 6–7**, we investigate the reconstruction of signals and images from their subsampled measurements, which can be modeled as solving an underdetermined linear system. Mathematically, a signal reconstruction problem can be expressed as estimating an unknown $\bar{x} \in \mathbb{R}^n$ from an underdetermined linear system

$$b = A\bar{x} + e, \tag{1.4.1}$$

where $A \in \mathbb{R}^{m\times n}$ is a full row-rank sensing matrix such as a projection or transformation matrix (see, e.g., [37, 43, 44]) with $m \ll n$, $b \in \mathbb{R}^m \setminus \{0\}$ is a vector of measurements, and $e \in \mathbb{R}^m$ is some unknown but bounded noise perturbation with $\|e\|_2 \leq \tau$. Physically, a signal of interest, or its coefficients under certain transformation, is often sparse (see, e.g., [37]). Hence, it is natural to seek a sparse solution to

the underdetermined linear system (1.4.1), though it has infinitely many solutions. We say that $x \in \mathbb{R}^n$ is *s-sparse* if $\|x\|_0 \leq s$, where $\|x\|_0$ counts the number of nonzero entries of x. To find the sparsest solution to (1.4.1), one may consider solving the following minimization problem:

$$\min_{x\in\mathbb{R}^n} \|x\|_0 \quad \text{s.t.} \quad \|Ax-b\|_2 \leq \tau, \tag{1.4.2}$$

in which $\|x\|_0$ serves as a penalty term of the sparsity, and it is referred to as the ℓ_0 penalty for convenience. Due to the discrete and discontinuous nature of the ℓ_0 penalty, the model (1.4.2) is NP-hard [37]. This means the model (1.4.2) is computationally intractable, and this difficulty has inspired many alternatives to the ℓ_0 penalty in the literature. A fundamental proxy of the model (1.4.2) is the basis pursuit (BP) problem proposed in [58]:

$$\min_{x\in\mathbb{R}^n} \|x\|_1 \quad \text{s.t.} \quad \|Ax-b\|_2 \leq \tau. \tag{1.4.3}$$

In this convex model, $\|x\|_1 := \sum_{i=1}^n |x_i|$ and it is called the ℓ_1 penalty hereafter. Recall that $\|x\|_1$ is the convex envelope of $\|x\|_0$ (see, e.g., [180]), and it induces sparsity most efficiently among all convex penalties (see [37]). The BP problem (1.4.3) has been intensively studied in voluminous papers since the seminal works [42, 43, 74], in which various conditions have been comprehensively explored for the exact reconstruction via the convex model (1.4.3).

The BP problem (1.4.3) is crucial for signal reconstruction, but its solution can suffer from over-penalization because the ℓ_1 penalty tends to underestimate high-amplitude components of the solution, as discussed in [83]. Therefore, it is reasonable to consider non-convex alternatives to the ℓ_1 penalty and upgrade the model (1.4.3) to achieve a more accurate reconstruction. In the literature, some non-convex penalties have been well studied, such as the smoothly clipped absolute deviation (SCAD) [83], the capped ℓ_1 penalty [244], the transformed ℓ_1 penalty [142, 243], and the ℓ_p penalty with $0 < p < 1$ [54, 55, 124]. Besides, one particular penalty is the minimax concave penalty (MCP) proposed in [240], and it has been widely shown to be effective in reducing the bias from the ℓ_1 penalty [240]. Moreover, the so-called ℓ_{1-2} penalty has been studied in the literature, e.g. [82, 238, 239], to mention a few. In summary, convex penalties are more tractable in both senses of theoretical analysis and numerical computation, while they are less effective for achieving the desired sparsity (i.e., the approximation to the ℓ_0 penalty is less accurate). Non-convex penalties are generally the opposite.

In the seminal compressed sensing papers [41, 74], reconstruction conditions have

been established for the BP model (1.4.3). These conditions rely on the restricted isometry property (RIP) of the *sensing matrix* A, as proposed in [44]. For an index set $T \subset \{1, 2, \ldots, n\}$ and an integer s with $|T| \leq s$, the *s-restricted isometry constant* (RIC) of $A \in \mathbb{R}^{m\times n}$ is the smallest $\delta_s \in (0, 1)$ such that

$$(1 - \delta_s)\|x\|_2^2 \leq \|A_T x\|_2^2 \leq (1 + \delta_s)\|x\|_2^2$$

for all subsets T with $|T| \leq s$ and all $x \in \mathbb{R}^{|T|}$. The matrix A is said to satisfy the *s-restricted isometry property* (RIP) with δ_s. If A satisfies the s-RIP, it means that the sensing matrix preserves the geometry of s-sparse signals up to a certain error level. This property forms the foundation of compressed sensing. As the name suggests, the RIP also represents a restricted isometry of the sensed measurements.

In this thesis, we aim to find a penalty that combines the advantages of both the ℓ_1 penalty and its non-convex alternatives while avoiding their drawbacks. Specifically, we propose the *springback* penalty

$$\mathcal{R}_\alpha^{\mathrm{SPB}}(x) = \|x\|_1 - \frac{\alpha}{2}\|x\|_2^2,$$

where $\alpha > 0$ is a carefully chosen model parameter. We establish the exact and stable reconstruction theory for the compressed sensing model using the springback penalty under the RIP framework. Furthermore, we theoretically demonstrate the superiority of the springback model over existing models in certain information-based situations where measurement noise is high and the number of measurements is limited, providing a sharper recovery bound. Overall, the springback penalty offers an improved model for signal reconstruction, with benefits in both theoretical analysis and numerical computation.

The compressed sensing theory is based on the assumption of the sparsity of the (vector) signal of interest or its coefficients under certain transformations. This assumption can also be extended to image reconstruction, as natural images X typically have (approximately) sparse gradients ∇X. An extension of the ℓ_1 penalty to imaging is the anisotropic total variation (TV) if the image gradient ∇X is considered as a vector.

Given linear measurements $y \in \mathbb{C}^m$ observed via

$$y = \mathcal{M}\bar{X} + e$$

from an unknown image $\bar{X} \in \mathbb{C}^{N\times N}$, where $\mathcal{M} : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ is a linear operator defined component-wisely by

$$[\mathcal{M}(\bar{X})]_j := \langle M_j, \bar{X}\rangle = \mathrm{tr}(M_j \bar{X}^*),$$

for suitable matrices M_j with m considerably smaller than N^2, and $e \in \mathbb{C}^m$ is a noise term bounded by $\|e\|_2 \le \tau$ with level $\tau \ge 0$, the reconstruction of the unknown $\bar{X}$ can be modeled as the following TV minimization problem:

$$\min_{X\in\mathbb{C}^{N\times N}} \|X\|_{\mathrm{TV}} \quad \text{s.t.} \quad \|\mathcal{M}X - y\|_2 \le \tau, \tag{1.4.4}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the TV semi-norm.

It is important to note that while the compressed sensing theory is applicable to sparse signals or signals that are sparse after an orthonormal transform, it cannot be applied to the TV model (1.4.4) because the gradient transform $\nabla : X \to \nabla X$ is not orthonormal, as mentioned in [157]. The first compressed sensing theory tailored for imaging and the TV model (1.4.4) was established in [157], also under the RIP framework. In the realm of images, we say that a linear operator $\mathcal{A} : \mathbb{C}^{n_1\times n_2} \to \mathbb{C}^m$ has the RIP of order s and level $\delta \in (0, 1)$ if

$$(1-\delta)\|X\|_2^2 \le \|\mathcal{A}X\|_2^2 \le (1+\delta)\|X\|_2^2 \quad \forall s\text{-sparse } X \in \mathbb{C}^{n_1\times n_2}, \tag{1.4.5}$$

and the smallest δ for (1.4.5) is said to be the *restricted isometry constant* (RIC) associated with $\mathcal{A}$.

We propose an extension of the compressed sensing theory with the springback penalty to image reconstruction by introducing the *enhanced TV* regularization:

$$\mathcal{R}\alpha(X) := \|X\|_{\mathrm{TV}} - \frac{\alpha}{2}\|\nabla X\|_2^2,$$

where $\alpha > 0$ is a carefully chosen parameter to ensure the positiveness or the well-definedness of (7.1.6), and $\|\nabla X\|_2^2$ is the sum of the squared magnitudes of ∇X. In the information-based situation, under certain weaker restricted isometry property conditions, the enhanced TV minimization model is shown to have tighter reconstruction error bounds than various TV-based models, especially when the noise level is significant and the number of measurements is limited.

1.5 Contributions

Specifically, we briefly summarize our results and contributions as follows. All our results are verified by adequate convincing numerical experiments.

In **Chapter 2**, we provide a recipe — the Marcinkiewicz–Zygmund property — for weakening the quadrature exactness assumption of hyperinterpolation. Specifically, we examine the behavior of hyperinterpolation when the required exactness degree $2n$ is relaxed to $n + k$ with $0 < k \leq n$. Using the Marcinkiewicz–Zygmund property, we show that the L^2 norm of the exactness-relaxing hyperinterpolation operator is bounded by a constant independent of n, and this approximation scheme is convergent as $n \to \infty$ if k is positively correlated to n. These results demonstrate that hyperinterpolation is a reliable approximation scheme, even when the original quadrature exactness assumption is partly ruined. Besides, the family of candidate quadrature rules for constructing hyperinterpolants can be significantly enriched, and the number of quadrature points can be considerably reduced. These facts suggest that hyperinterpolation is a good method for approximating functions regarding the information-based situation. As a potential cost, this relaxation may slow the convergence rate of hyperinterpolation in terms of the reduced degrees of quadrature exactness.

In **Chapter 3**, we focus on examining the efficiency of hyperinterpolation in approximating singular and oscillatory functions, which are commonly encountered in applied mathematics and physics. By efficiency, we mean achieving satisfactory accuracy with a considerably small number of sampling points, which is crucial in information-based situations. Singular and oscillatory functions typically require more sampling points to attain satisfactory accuracy, which can be computationally expensive. In this chapter, we propose a new approximation scheme called efficient hyperinterpolation, which employs the product-integration method to achieve the desired accuracy with fewer quadrature points than the original method. Building on our results in Chapter 2, we provide theorems that establish the superiority of efficient hyperinterpolation over the original method in approximating functions belonging to $L^1(\Omega)$, $L^2(\Omega)$, and $C(\Omega)$ spaces, respectively. This study can be regarded as an application of results in Chapter 2.

In **Chapter 4**, we further explore the idea of relaxing the quadrature exactness assumption introduced in Chapter 2. The goal is to completely replace this assumption with the Marcinkiewicz-Zygmund property, enabling the construction of hyperinterpolation using positive-weight quadrature rules that do not require exactness. This approach is referred to as unfettered hyperinterpolation. We provide a

reasonable error estimate for this new method, which consists of two terms: one representing the error estimate of the original hyperinterpolation with full quadrature exactness, and another term that compensates for the loss of exactness degrees. We offer a guide to controlling the newly introduced term in practice. Furthermore, if the quadrature points form a quasi-Monte Carlo (QMC) design, a refined error estimate is available. These findings confirm that hyperinterpolation is a dependable approximation scheme in the information-based situation.

In **Chapter 5**, we present a novel quadrature-based spectral method for solving the Allen-Cahn equation on spheres. Our method utilizes hyperinterpolation and the Marcinkiewicz-Zygmund system of quadrature rules to achieve the theoretical advantages of the Galerkin method while maintaining a computational cost comparable to the collocation method. This method eliminates the stringent requirements on the time step size and imposes mild conditions on the polynomial degree of numerical solutions. Additionally, it includes an effective maximum principle, which allows the numerical solutions to deviate from the sharp bound by a controllable discretization error. Our method does not rely on the quadrature exactness assumption to maintain this principle in the information-based situation. If the quadrature rule has sufficient exactness degrees, our method is guaranteed to be energy stable.

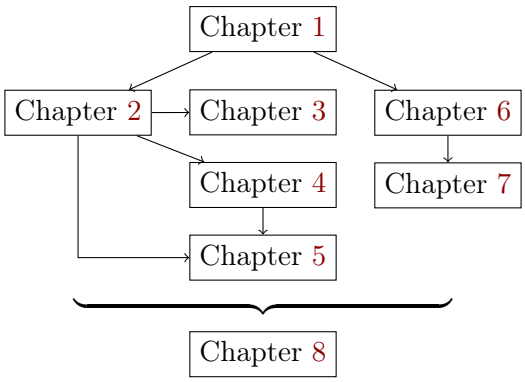
In **Chapter 6**, we introduce a novel penalty, known as the springback penalty, for creating models that can reconstruct an unknown signal from incomplete and inaccurate measurements. We establish exact and stable reconstruction theories for the reconstruction model using the springback penalty, both for sparse and nearly sparse signals, respectively, under the RIP framework. Furthermore, we derive an easily implementable difference-of-convex algorithm. Our model possesses theoretical superiority to some existing models, with a sharper reconstruction bound for certain scenarios where the number of measurements is limited, and the measurement noise level is high. In regard to the information-based situation, our model addresses the challenge of incomplete and inaccurate measurements and enables accurate signal reconstruction with limited data.

In **Chapter 7**, we extend our newly proposed penalty and the corresponding compressed sensing theory from Chapter 6 to image reconstruction. We base our approach on the observation that natural images typically have (approximately) sparse gradients. We propose the enhanced TV regularization, and we explain, from the perspective of PDEs, that the enhanced term corresponds to introducing a backward diffusion term for deblurring. As the gradient transform is not orthonormal, the

compressed sensing approach used in Chapter 6 cannot be applied to our image reconstruction model. However, since images are also sparse under some wavelet transforms, we establish a reconstruction theory for the enhanced TV model. In regard to the information-based situation, we show that under weaker restricted isometry property conditions, the enhanced TV minimization model has tighter reconstruction error bounds than various TV-based models when the amount of measurements is limited, and the level of noise is significant. Particularly, for variable-density sampled Fourier measurements, we show that the minimum number of measurements required for the enhanced TV model is approximately 30.86% of that established in [123] for the TV model.

In **Chapter 8**, we provide a chapter-wise conclusion on this thesis and discuss how deep learning can be engaged in exploring high-dimensional numerical analysis.

The overall organization of these chapters as well as their logical dependency is illustrated in the following chart. Chapters 2–5 and Chapters 7–8 serve as two independent part. The reader interested in approximation theory can consult Chapters 2–4. For the PDE results in Chapter 5, Chapters 2 and 4 provide approximation backgrounds. The reader interested in compressed sensing and imaging can directly start from Chapter 6.



Chapter 2

On the quadrature exactness of hyperinterpolation

This chapter investigates the role of quadrature exactness in the approximation scheme of hyperinterpolation. Constructing a hyperinterpolant of degree n requires a positive-weight quadrature rule with exactness degree $2n$. We examine the behavior of such approximation when the required exactness degree $2n$ is relaxed to $n + k$ with $0 < k \leq n$. Aided by the Marcinkiewicz–Zygmund inequality, we affirm that the L^2 norm of the exactness-relaxing hyperinterpolation operator is bounded by a constant independent of n, and this approximation scheme is convergent as $n \to \infty$ if k is positively correlated to n. Thus, the family of candidate quadrature rules for constructing hyperinterpolants can be significantly enriched, and the number of quadrature points can be considerably reduced. As a potential cost, this relaxation may slow the convergence rate of hyperinterpolation in terms of the reduced degrees of quadrature exactness. Our theoretical results are asserted by numerical experiments on three of the best-known quadrature rules: the Gauss quadrature, the Clenshaw–Curtis quadrature, and the spherical t-designs.

2.1 Introduction

Let Ω be a bounded region of $\mathbb{R}^d$ with measure $\mathrm{d}\omega$, which is either the closure of a connected open domain, or a smooth closed lower-dimensional manifold in $\mathbb{R}^d$. This region is assumed to have finite measure with respect to $\mathrm{d}\omega$, that is,

$$\int_\Omega \mathrm{d}\omega = V < \infty.$$

Let the space $L^p(\Omega)$ be equipped with the usual L^p norm $\|\cdot\|_p$ for $1 \leq p \leq \infty$, that is, for $g \in L^p(\Omega)$,

$$\|g\|_p := \begin{cases} \left(\int_\Omega |g|^p \mathrm{d}\omega\right)^{1/p}, & 1 \leq p < \infty, \\ \operatorname{ess\,sup}_{x\in\Omega} |g(x)|, & p = \infty. \end{cases}$$

The space $C(\Omega)$ of continuous functions is also equipped with the L^∞ norm. In particular, $L^p(\Omega)$ is a Hilbert space when $p = 2$, with the L^2 inner product defined as (1.2.1). This inner product also induces the L^2 norm, that is, $\|g\|_2 = \sqrt{\langle g, g\rangle}$ for $g \in L^2(\Omega)$.

Recall that a hyperinterpolation of degree n, defined as (1.2.5), is a discrete approximation of the L^2 orthogonal projection (1.2.2) that approximates continuous functions with polynomials in the space $\mathbb{P}_n$ of polynomials of degree at most n, introduced by Ian H. Sloan in [196] for his curiosity on whether interpolation has properties as good as those of the L^2 orthogonal projection. The discretization is achieved by evaluating the orthogonal coefficients (in terms of inner products) by some quadrature rules (1.2.3). The exactness degree $2n$ of the quadrature rule (1.2.3) is a central assumption in constructing hyperinterpolants. Moreover, if one considers hyperinterpolation on some regions where quadrature theory has not been well established, this exactness assumption has also potentially spurred the development of quadrature theory and orthogonal polynomials on these regions. Indeed, quadrature exactness contributes to the standard principle for designing quadrature rules: they should be exact for a certain class of integrands, e.g., polynomials under a fixed degree. This exactness principle is the departing point of most discussions on quadrature. Still, there has been growing concern recently about whether this principle is reliable in designing quadrature rules, as discussed by Trefethen in [222]. The main message of [222] is that the exactness principle proves to be an unreliable guide to actual accuracy. According to Trefethen, the exactness principle is a matter of algebra, concerned with whether or not certain quantities are exactly zero; however, quadrature is a problem of analysis, focusing on whether or not certain quantities are small. Thus, we are intrigued to know whether the required exactness degree $2n$ in constructing hyperinterpolants of degree n is superfluous.

This question is answered as the main results of this chapter: When $2n$ is relaxed to $n + k$, where $0 < k \leq n$, i.e., reduced at least to $n + 1$, the norm of $\mathcal{L}_n$ as an operator from $C(\Omega)$ to $L^2(\Omega)$ is bounded by some constant, and the error estimate $\|\mathcal{L}_n f - f\|_2$ is bounded in terms of $E_k(f)$, which is the best uniform error of f by a polynomial in $\mathbb{P}_k$. In addition, if k is positively correlated to n, then the scheme of hyperinterpolation is convergent as $n \to \infty$. This relaxation helps

hyperinterpolation to get rid of the disadvantage that, remarked by Hesse and Sloan in [110], it needs function values at the given points of the positive-weight quadrature rule with exactness degree $2n$. In real-world applications, data sampling may be expensive. This relaxation may enlighten us to develop hyperinterpolation-based methods for problems that are in favor of a high-order approximation but against extensive data sampling. When data sampling is cheap, this relaxation may also help to speed up our computation.

We note that the generalized hyperinterpolation [67, 178], defined on the sphere, only requires a positive-weight quadrature rule with exactness degree $n+1$ rather than $2n$. However, the definition of this scheme is different from that of the original hyperinterpolation. In this chapter, we focus on the original hyperinterpolation and investigate the effects of relaxing the quadrature exactness. Moreover, our investigation pertains to a general region Ω, while the generalized hyperinterpolation is only studied on the sphere.

In the next section, we present the main theoretical results on the exactness-relaxing hyperinterpolation, with the proof of our main Theorem 2.2.8 given in Section 2.3. To verify our theory, we conduct some numerical experiments on the interval $[-1,1]$ and the unit sphere $\mathbb{S}^2$ in Section 2.4.

2.2 Hyperinterpolation with exactness-relaxing quadrature rules

The hyperinterpolation of degree n with an exactness-relaxing quadrature rule is defined as follows.

Assumption 2.2.1 *The m-point quadrature rule (1.2.3), with nodes $x_j \in \Omega$ and weights $w_j > 0$ for $j = 1, 2, \ldots, m$, has exactness degree $n+k$ with $0 < k \leq n$, where $n, k \in \mathbb{N}$.*

Definition 2.2.2 *Let $\langle \cdot, \cdot \rangle_m$ be an m-point quadrature rule fulfilling Assumption 2.2.1 and $\{p_\ell\}_{\ell=1}^{d_n} \subset \mathbb{P}_n$ be an orthonormal basis of $\mathbb{P}_n$. Given $f \in C(\Omega)$, the hyperinterpolant of degree n to f is defined as*

$$\mathcal{L}_n f := \sum_{\ell=1}^{d_n} \langle f, p_\ell \rangle_m p_\ell. \tag{2.2.1}$$

This scheme (2.2.1) is essentially the hyperinterpolation scheme (1.2.5), except that the degree of quadrature exactness is relaxed. Thus the scheme (2.2.1) is also a discrete version of the orthogonal projection from $C(\Omega)$ onto $\mathbb{P}_n$ with respect to the

L^2 inner product (1.2.1). To tell the difference between schemes (1.2.5) and (2.2.1), we refer to Sloan's hyperinterpolation as the *original hyperinterpolation*. We denote by $\mathcal{L}_n^{\mathrm{S}}$ the original hyperinterpolation operator in the following texts, where S stands for Sloan.

What kind of benefits and costs does the relaxation of quadrature exactness bring to the analysis and implementation of hyperinterpolation? Here is an immediate benefit. We know that an m-point quadrature rule with exactness degree $2n$ requires $m \geq d_n$ quadrature points, see [196, Lemma 2], and such a quadrature rule is said to be *minimal* if $m = d_n$. This fact suggests that m should satisfy $m \geq d_n$ for $\mathcal{L}_n^{\mathrm{S}}$, and it also admits the following rather simple but interesting theorem.

Theorem 2.2.3 *The number of quadrature points for the hyperinterpolation* (2.2.1) *satisfies*

$$m \geq \begin{cases} d_{(n+k)/2} = d_{(n+k)/2}, & \text{when } n+k \text{ is even}, \\ d_{(n+k+1)/2} = d_{(n+k+1)/2}, & \text{when } n+k \text{ is odd}. \end{cases}$$

The benefit brought by the theorem is two-fold. On the one hand, for minimal quadrature rules used in constructing hyperinterpolants, the required amount of quadrature points can be considerably reduced from d_n to $d_{(n+k)/2}$ or $d_{(n+k+1)/2}$, depending on the parity of $n+k$. Such reduction is more pronounced in higher-dimensional regions. On the other hand, for quadrature rules demanding more nodes to achieve the exactness degree $2n$, which used to be deemed impractical, some of them can be added into the family of candidate quadrature rules to construct hyperinterpolants efficiently. For example, a typical choice of quadrature rules for hyperinterpolation on $[-1, 1]$ is the Gaussian quadrature, and now the Clenshaw–Curtis quadrature can also be considered a good choice; see more details in Section 2.4.

Obviously, such relaxation is not cost-free. The original hyperinterpolant (1.2.5) is a projection for $f \in \mathbb{P}_n$, that is, $\mathcal{L}_n^{\mathrm{S}} f = f$ for all $f \in \mathbb{P}_n$; see [196, Lemma 4]. However, due to the loss of some exactness degrees, this property is preserved only for polynomials of degree at most k, asserted by the following lemma.

Lemma 2.2.4 *If $f \in \mathbb{P}_k$, then $\mathcal{L}_n$ defined in Definition 2.2.2 admits $\mathcal{L}_n f = f$.*

Proof. For $f \in \mathbb{P}_k$, it may be expressed as

$$f = \sum_{\ell=1}^{d_k} a_\ell p_\ell,$$

where $a_\ell = \int_\Omega f p_\ell \mathrm{d}\omega$ and $d_k = \dim \mathbb{P}_k$. The exactness degree $n+k$ admits $\langle p_{\ell'}, p_\ell \rangle_m = \delta_{\ell\ell'}$ for $1 \le \ell' \le d_k$ and $1 \le \ell \le d_n$. Thus,

$$\mathcal{L}_n f = \sum_{\ell=1}^{d_n} \left\langle \sum_{\ell'=1}^{d_k} a_{\ell'} p_{\ell'}, p_\ell \right\rangle_m p_\ell = \sum_{\ell=1}^{d_n} \left(\sum_{\ell'=1}^{d_k} a_{\ell'} \langle p_{\ell'}, p_\ell \rangle_m \right) p_\ell = \sum_{\ell=1}^{d_k} a_\ell p_\ell,$$

leading to $\mathcal{L}_n f = f$. □

Corollary 2.2.5 *For $f \in C(\Omega)$, we have*

$$\mathcal{L}_n(\mathcal{L}_k f) = \mathcal{L}_k(\mathcal{L}_n f) = \mathcal{L}_k(\mathcal{L}_k f) = \mathcal{L}_k f.$$

Proof. As $\mathcal{L}_k f \in \mathbb{P}_k$, Lemma 2.2.4 immediately implies $\mathcal{L}_n(\mathcal{L}_k f) = \mathcal{L}_k f$. Similar to the proof of Lemma 2.2.4, we have

$$\begin{aligned} \mathcal{L}_k(\mathcal{L}_n f) &= \sum_{\ell=1}^{d_k} \left\langle \sum_{\ell'=1}^{d_n} \langle f, p_{\ell'} \rangle_m p_{\ell'}, p_\ell \right\rangle_m p_\ell = \sum_{\ell=1}^{d_k} \left(\sum_{\ell'=1}^{d_n} \langle f, p_{\ell'} \rangle_m \langle p_{\ell'}, p_\ell \rangle_m \right) p_\ell \\ &= \sum_{\ell=1}^{d_k} \langle f, p_\ell \rangle_m p_\ell = \mathcal{L}_k f, \end{aligned}$$

and similarly,

$$\mathcal{L}_k(\mathcal{L}_k f) = \sum_{\ell=1}^{d_k} \left(\sum_{\ell'=1}^{d_k} \langle f, p_{\ell'} \rangle_m \langle p_{\ell'}, p_\ell \rangle_m \right) p_\ell = \sum_{\ell=1}^{d_k} \langle f, p_\ell \rangle_m p_\ell = \mathcal{L}_k f.$$

Thus, the corollary is completely proved. □

Remark 2.2.6 *Lemma 2.2.4 indicates that the exactness degree $2n$ can be relaxed at least to $n+1$; otherwise, the projection property $\mathcal{L}_n f = f$ for all $f \in \mathbb{P}_k$ does not maintain for any non-trivial polynomial spaces.*

Remark 2.2.7 *There may be an illusion that for the exactness-relaxing hyperinterpolation (2.2.1), there holds $\mathcal{L}_n f = f$ for $f \in \mathbb{P}_{\lfloor (n+k)/2 \rfloor}$, induced from the fact that for $\mathcal{L}_n^{\mathrm{S}}$ with exactness degree $2n$, $\mathcal{L}_n^{\mathrm{S}} f = f$ for all $f \in \mathbb{P}_n$. However, according to the proof of Lemma 2.2.4, this is not true. Indeed, $\langle p_{\ell'}, p_\ell \rangle_m$ with exactness degree $n+k$ may not be the Kronecker $\delta_{\ell\ell'}$ for $p_{\ell'} \in \mathbb{P}_{\lfloor (n+k)/2 \rfloor}$ and $p_\ell \in \mathbb{P}_n$.*

This decay of projection-maintaining degrees is followed by Theorem 2.2.8 below, indicating that the convergence rate of $\mathcal{L}_n^{\mathrm{S}}$ is slowed from $E_n(f)$ to $E_k(f)$. It was

proved in [196] that

$$\|\mathcal{L}_n^{\mathrm{S}} f\|_2 \leq V^{1/2}\|f\|_\infty \tag{2.2.2}$$

and

$$\|\mathcal{L}_n^{\mathrm{S}} f - f\|_2 \leq 2V^{1/2}E_n(f). \tag{2.2.3}$$

To tell the difference between the stability result (2.2.2) of $\mathcal{L}_n^{\mathrm{S}}$ and that of $\mathcal{L}_n$, we note that the stability result (2.2.2) stems from

$$\|\mathcal{L}_n^{\mathrm{S}} f\|_2^2 + \langle f - \mathcal{L}_n^{\mathrm{S}} f, f - \mathcal{L}_n^{\mathrm{S}} f\rangle_{m'} = \langle f, f\rangle_{m'} = \sum_{j=1}^{m} w_j f(x_j)^2 \leq V\|f\|_\infty^2$$

and the non-negativeness of $\langle f - \mathcal{L}_n^{\mathrm{S}} f, f - \mathcal{L}_n^{\mathrm{S}} f\rangle_{m'}$, where $\langle\cdot,\cdot\rangle_{m'}$ denotes an m-point quadrature rule (1.2.3) with exactness degree $2n$ and this notation is only used here; see the proof in [196]. However, due to the relaxation of exactness degrees, we can only claim

$$\|\mathcal{L}_n f\|_2^2 + \langle f - \mathcal{L}_n f, f - \mathcal{L}_n f\rangle_m + \sigma_{n,k,f} = \langle f, f\rangle_m,$$

where

$$\sigma_{n,k,f} = \langle \mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle - \langle \mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle_m \tag{2.2.4}$$

stands for the error in evaluating the integral of $(\mathcal{L}_n f - \mathcal{L}_k f)^2$ over Ω by the quadrature rule (1.2.3) with exactness degree $n+k$; see the equation (2.3.8) in our proof in the next section. Even though it is possible (and often occurs) that

$$\langle f - \mathcal{L}_n f, f - \mathcal{L}_n f\rangle_m + \sigma_{n,k,f} \geq 0$$

if the quadrature rule (1.2.3) converges fast enough, we cannot make such a claim rigorously in general. Therefore, it is natural to endow the quadrature rule (1.2.3) with some convergence property.

We assume that there exists an $\eta \in [0,1)$, which is independent of n and χ, such that

$$\left|\sum_{j=1}^{m} w_j \chi(x_j)^2 - \int_\Omega \chi^2 \mathrm{d}\omega\right| \leq \eta \int_\Omega \chi^2 \mathrm{d}\omega \quad \forall \chi \in \mathbb{P}_n. \tag{2.2.5}$$

If $k = n$, i.e., the quadrature exactness is not relaxed, then $\eta = 0$. This convergence property (2.2.5) can be regarded as the Marcinkiewicz–Zygmund inequality [89, 143, 146] applied to polynomials of degree at most $2n$, and we refer to it as the *Marcinkiewicz–Zygmund property* below. From the expression (2.2.4) of $\sigma_{n,k,f}$ we

immediately observe that the Marcinkiewicz–Zygmund property (2.2.5) suffices to bound $|\sigma_{n,k,f}|$.

Theorem 2.2.8 *Given $f \in C(\Omega)$, let $\mathcal{L}_n f \in \mathbb{P}_n$ be defined by (2.2.1), where the m-point quadrature rule (1.2.3) not only fulfills Assumption 2.2.1 with $0 < k \leq n$ but also has the Marcinkiewicz–Zygmund property (2.2.5) with $\eta \in [0, 1)$. Then*

$$\|\mathcal{L}_n f\|_2 \leq \frac{V^{1/2}}{\sqrt{1-\eta}} \|f\|_\infty \tag{2.2.6}$$

and

$$\|\mathcal{L}_n f - f\|_2 \leq \left(\frac{1}{\sqrt{1-\eta}} + 1 \right) V^{1/2} E_k(f). \tag{2.2.7}$$

The hyperinterpolant $\mathcal{L}_n f$ may not converge to f as $n \to \infty$ if k is fixed. If k is additionally positively correlated to n, then

$$\|\mathcal{L}_n f - f\|_2 \to 0 \quad \text{as} \quad n \to \infty. \tag{2.2.8}$$

Remark 2.2.9 *By "k is additionally positively correlated to n," we mean that $n \to \infty$ implies $k \to \infty$. This condition ensures the convergence result (2.2.8) as $n \to \infty$. The converse statement that $k \to \infty$ implies $n \to \infty$ automatically holds because $k \leq n$.*

Remark 2.2.10 *If $k = n$, i.e., the degree of quadrature exactness is not relaxed, then the stability result (2.2.6), the error estimate (2.2.7), and the convergence result (2.2.8) are the same as those for $\mathcal{L}_n^{\mathrm{S}}$ in [196]. If $0 < k < n$, then as a cost of the relaxation of exactness, the error estimation (2.2.7) is now controlled by $E_k(f)$ rather than $E_n(f)$. Since $E_k(f) \geq E_n(f)$ if $k < n$, this estimation (2.2.7) reveals an effect of relaxing the quadrature exactness. That is, we can use fewer quadrature points than the original hyperinterpolation, but the corresponding error estimation will be somewhat amplified. Moreover, if $k \leq 0$, i.e., the degree of quadrature exactness is relaxed to n or even less, then no convergence information can be offered by Theorem 2.2.8.*

An immediate application of Theorem 2.2.8 is to a generalization of the method of "product integration", see discussions in [196]. In this method, the integral over Ω of the form $\int_\Omega h f \mathrm{d}\omega$, where f is smooth and h contains any singularities in the product integrand, is approximated by

$$\int_\Omega h f \mathrm{d}\omega \approx \int_\Omega h(\mathcal{L}_n f)\mathrm{d}\omega = \sum_{\ell=1}^{d_n} \langle f, p_\ell \rangle_m \int_\Omega h p_\ell \mathrm{d}\omega = \sum_{j=1}^{m} W_j f(x_j), \tag{2.2.9}$$

where

$$W_j = w_j \sum_{\ell=1}^{d_n} p_\ell(x_j) \int_\Omega h p_\ell \mathrm{d}\omega, \quad j = 1, 2, \ldots, m. \tag{2.2.10}$$

Applying the Cauchy–Schwarz inequality over Ω to $\int_\Omega h(\mathcal{L}_n f - f)\mathrm{d}\omega$, Theorem 2.2.8 immediately implies the following result.

Corollary 2.2.11 *Let h be measurable on Ω with respect to $\mathrm{d}\omega$ and satisfy $\|h\|_2 < \infty$, and let $\{W_j\}_{j=1}^m$ be given by (2.2.10). Under the conditions of Theorem 2.2.8, the approximation error of $\int_\Omega hf\mathrm{d}\omega$ in terms of (2.2.9) is estimated by*

$$\left| \sum_{j=1}^m W_j f(x_j) - \int_\Omega hf\mathrm{d}\omega \right| \le \left(\frac{1}{\sqrt{1-\eta}} + 1 \right) \|h\|_2 V^{1/2} E_k(f).$$

A further discussion on hyperinterpolation and the product-integration method will be given in Chapter 3.

Remark 2.2.12 *In the light of Theorem 2.2.8, we expect that the required exactness degree in constructing other variants of hyperinterpolants, such as filtered hyperinterpolants [205] and Lasso hyperinterpolants [9], can also be reduced, and corresponding theory can be developed.*

2.3 Proof of the main theorem

2.3.1 Preparation

The hyperinterpolant $\mathcal{L}_n f$ can be decomposed into

$$\mathcal{L}_n f := \mathcal{L}_k f + (\mathcal{L}_n - \mathcal{L}_k) f, \tag{2.3.1}$$

where $\mathcal{L}_n - \mathcal{L}_k : C(\Omega) \to \mathbb{P}_n$ is a linear operator mapping $f \in C(\Omega)$ to

$$(\mathcal{L}_n - \mathcal{L}_k) f := \sum_{\ell=d_k+1}^{d_n} \langle f, p_\ell \rangle_m p_\ell.$$

In the following proof of Theorem 2.2.8, we shall treat $\mathcal{L}_k f$ and $(\mathcal{L}_n - \mathcal{L}_k) f$ separately. For the former component, the degree $n + k \ge 2k$ of quadrature exactness leads to

$$\langle \mathcal{L}_k f, \mathcal{L}_k f \rangle = \langle \mathcal{L}_k f, \mathcal{L}_k f \rangle_m. \tag{2.3.2}$$

For the latter component, the orthogonality of $\{p_\ell\}$ renders

$$\langle(\mathcal{L}_n - \mathcal{L}_k)f, (\mathcal{L}_n - \mathcal{L}_k)f\rangle = \sum_{\ell=d_k+1}^{d_n} \langle f, p_\ell\rangle_m^2 = \langle f, (\mathcal{L}_n - \mathcal{L}_k)f\rangle_m. \tag{2.3.3}$$

Before proving Theorem 2.2.8, we present a lemma involving $\langle\mathcal{L}_k f, \mathcal{L}_k f\rangle_m$ and $\langle f, (\mathcal{L}_n - \mathcal{L}_k)f\rangle_m$.

Lemma 2.3.1 *Adopt the conditions of Theorem 2.2.8. Let $\mathcal{L}_k : C(\Omega) \to \mathbb{P}_k$ be the hyperinterpolation operator of degree k, defined with an m-point quadrature with exactness degree $n + k$. Then*

(a) $\langle f - \mathcal{L}_k f, \chi\rangle_m = 0$ *and* $\langle f - \mathcal{L}_n f, \chi\rangle_m = 0$ *for all* $\chi \in \mathbb{P}_k$,

(b) $\langle\mathcal{L}_k f, \mathcal{L}_k f\rangle_m + \langle f - \mathcal{L}_k f, f - \mathcal{L}_k f\rangle_m = \langle f, f\rangle_m$,

(c) $\langle\mathcal{L}_k f, \mathcal{L}_k f\rangle_m + \langle\mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle_m = \langle\mathcal{L}_n f, \mathcal{L}_n f\rangle_m$,

(d) $\langle f - \mathcal{L}_n f, f - \mathcal{L}_n f\rangle_m + 2\langle f, \mathcal{L}_n f - \mathcal{L}_k f\rangle_m = \langle f - \mathcal{L}_k f, f - \mathcal{L}_k f\rangle_m + \langle\mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle_m$.

Proof. (a) Note that any $\chi \in \mathbb{P}_k$ can be expressed as $\chi = \sum_{\ell=1}^{d_k} a_\ell p_\ell$, where $a_\ell = \int_\Omega \chi p_\ell \mathrm{d}\omega$. The first equation holds since

$$\begin{aligned}\langle f - \mathcal{L}_k f, \chi\rangle_m &= \sum_{\ell=1}^{d_k} a_\ell \left\langle f - \sum_{\ell'=1}^{d_k} \langle f, p_{\ell'}\rangle_m p_{\ell'}, p_\ell \right\rangle_m \\ &= \sum_{\ell=1}^{d_k} a_\ell \left(\langle f, p_\ell\rangle_m - \sum_{\ell'=1}^{d_k} \langle f, p_{\ell'}\rangle_m \langle p_{\ell'}, p_\ell\rangle_m \right) = 0.\end{aligned}$$

Similarly,

$$\langle f - \mathcal{L}_n f, \chi\rangle_m = \sum_{\ell=1}^{d_k} a_\ell \left(\langle f, p_\ell\rangle_m - \sum_{\ell'=1}^{d_n} \langle f, p_{\ell'}\rangle_m \langle p_{\ell'}, p_\ell\rangle_m \right) = 0.$$

(b) Letting $\chi = \mathcal{L}_k f$, the first equation in statement (a) implies $\langle\mathcal{L}_k f, \mathcal{L}_k f\rangle_m = \langle f, \mathcal{L}_k f\rangle_m$. Thus

$$\begin{aligned}&\langle\mathcal{L}_k f, \mathcal{L}_k f\rangle_m + \langle f - \mathcal{L}_k f, f - \mathcal{L}_k f\rangle_m \\ &\quad = 2\langle\mathcal{L}_k f, \mathcal{L}_k f\rangle_m - 2\langle f, \mathcal{L}_k f\rangle_m + \langle f, f\rangle_m \\ &\quad = \langle f, f\rangle_m.\end{aligned}$$

(c) Letting $\chi = \mathcal{L}_k f$ in both equations in statement (a), we have

$$\langle \mathcal{L}_k f, \mathcal{L}_k f\rangle_m = \langle f, \mathcal{L}_k f\rangle_m = \langle \mathcal{L}_n f, \mathcal{L}_k f\rangle_m.$$

Thus

$$\begin{aligned}&\langle \mathcal{L}_k f, \mathcal{L}_k f\rangle_m + \langle \mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle_m\\ &\quad =2\langle \mathcal{L}_k f, \mathcal{L}_k f\rangle_m - 2\langle \mathcal{L}_n f, \mathcal{L}_k f\rangle_m + \langle \mathcal{L}_n f, \mathcal{L}_n f\rangle_m\\ &\quad =\langle \mathcal{L}_n f, \mathcal{L}_n f\rangle_m.\end{aligned}$$

(d) It is immediate that

$$\langle g - \mathcal{L}_n g, g - \mathcal{L}_n g\rangle_m = \langle g, g\rangle_m - 2\langle g, \mathcal{L}_n g\rangle_m + \langle \mathcal{L}_n g, \mathcal{L}_n g\rangle_m \tag{2.3.4}$$

holds for any $g \in C(\Omega)$. Lemma 2.2.4 implies $\mathcal{L}_n(\mathcal{L}_k f) = \mathcal{L}_k f$. Then replacing g by $f - \mathcal{L}_k f$, the left-hand side of (2.3.4) becomes

$$\begin{aligned}&\langle f - \mathcal{L}_k f - \mathcal{L}_n(f - \mathcal{L}_k f), f - \mathcal{L}_k f - \mathcal{L}_n(f - \mathcal{L}_k f)\rangle_m\\ &\quad =\langle f - \mathcal{L}_k f - \mathcal{L}_n f + \mathcal{L}_k f, f - \mathcal{L}_k f - \mathcal{L}_n f + \mathcal{L}_k f\rangle_m\\ &\quad =\langle f - \mathcal{L}_n f, f - \mathcal{L}_n f\rangle_m,\end{aligned}$$

and three terms on the right-hand side becomes $\langle g, g\rangle_m = \langle f - \mathcal{L}_k f, f - \mathcal{L}_k f\rangle_m$,

$$\begin{aligned}-2\langle g, \mathcal{L}_n g\rangle_m &= -2\langle f - \mathcal{L}_k f, \mathcal{L}_n(f - \mathcal{L}_k f)\rangle_m\\ &= -2\langle f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle_m\\ &= -2\langle f, \mathcal{L}_n f - \mathcal{L}_k f\rangle_m,\end{aligned} \tag{2.3.5}$$

and

$$\begin{aligned}\langle \mathcal{L}_n g, \mathcal{L}_n g\rangle_m &=\langle \mathcal{L}_n(f - \mathcal{L}_k f), \mathcal{L}_n(f - \mathcal{L}_k f)\rangle_m\\ &=\langle \mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle_m,\end{aligned}$$

respectively, where the last step in (2.3.5) holds since the orthogonality of $\{p_\ell\}_{\ell=1}^{d_n}$ and the quadrature exactness degree $n + k$ imply $\langle p_\ell, p_{\ell'}\rangle_m = 0$ for $\ell = 1, 2, \ldots, d_k$ and $\ell' = d_k + 1, \ldots, d_n$, and then

$$\langle \mathcal{L}_k f, (\mathcal{L}_n - \mathcal{L}_k) f\rangle_m = \left\langle \sum_{\ell=1}^{d_k} \langle f, p_\ell\rangle_m p_\ell, \sum_{\ell'=d_k+1}^{d_n} \langle f, p_{\ell'}\rangle_m p_{\ell'} \right\rangle_m = 0. \tag{2.3.6}$$

Hence, the equality (2.3.4) suggests the proof of statement (d). □

2.3.2 Proof of Theorem 2.2.8

Now we are prepared to prove Theorem 2.2.8.

Proof of Theorem 2.2.8. According to the decomposition (2.3.1), we have

$$\begin{aligned}\|\mathcal{L}_n f\|_2^2 =&\langle \mathcal{L}_n f, \mathcal{L}_n f\rangle = \langle \mathcal{L}_k f + (\mathcal{L}_n - \mathcal{L}_k)f, \mathcal{L}_k f + (\mathcal{L}_n - \mathcal{L}_k)f\rangle \\ =&\langle \mathcal{L}_k f, \mathcal{L}_k f\rangle + \langle (\mathcal{L}_n - \mathcal{L}_k)f, (\mathcal{L}_n - \mathcal{L}_k)f\rangle,\end{aligned}$$

where the last step holds since $\langle \mathcal{L}_k f, (\mathcal{L}_n - \mathcal{L}_k)f\rangle = 0$, which can be proved similarly to (2.3.6) and using the fact that $\langle p_\ell, p_{\ell'}\rangle = 0$ for $\ell = 1, 2, \ldots, d_k$ and $\ell' = d_k + 1, \ldots, d_n$. The observations (2.3.2) and (2.3.3) then lead to

$$\|\mathcal{L}_n f\|_2^2 = \langle \mathcal{L}_k f, \mathcal{L}_k f\rangle_m + \langle f, (\mathcal{L}_n - \mathcal{L}_k)f\rangle_m.$$

To derive the stability result (2.2.6), summing up the equations in Lemma 2.3.1(b,c,d), after easy computations, we have

$$\begin{aligned}2\langle \mathcal{L}_k f, \mathcal{L}_k f\rangle_m + 2\langle f, (\mathcal{L}_n - \mathcal{L}_k)f\rangle_m + \langle f - \mathcal{L}_n f, f - \mathcal{L}_n f\rangle_m \\ =\langle f, f\rangle_m + \langle \mathcal{L}_n f, \mathcal{L}_n f\rangle_m.\end{aligned} \tag{2.3.7}$$

Recalling the expression (2.2.4) of

$$\sigma_{n,k,f} = \langle \mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle - \langle \mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle_m$$

and the observation (2.3.3), we have

$$\begin{aligned}\langle f, (\mathcal{L}_n - \mathcal{L}_k)f\rangle_m =&\langle \mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle \\ =&\langle \mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle_m + \sigma_{n,k,f}.\end{aligned}$$

Together with statement (c) of Lemma 2.3.1, we have

$$\begin{aligned}\langle \mathcal{L}_n f, \mathcal{L}_n f\rangle_m &= \langle \mathcal{L}_k f, \mathcal{L}_k f\rangle_m + \langle \mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f\rangle_m \\ &= \langle \mathcal{L}_k f, \mathcal{L}_k f\rangle_m + \langle f, (\mathcal{L}_n - \mathcal{L}_k)f\rangle_m - \sigma_{n,k,f}.\end{aligned}$$

Thus, replacing a sum of $\langle \mathcal{L}_k f, \mathcal{L}_k f\rangle_m + \langle f, (\mathcal{L}_n - \mathcal{L}_k)f\rangle_m$ on the left-hand side of (2.3.7) with $\langle \mathcal{L}_n f, \mathcal{L}_n f\rangle_m + \sigma_{n,k,f}$ gives

$$\begin{aligned}\langle \mathcal{L}_k f, \mathcal{L}_k f\rangle_m + \langle f, (\mathcal{L}_n - \mathcal{L}_k)f\rangle_m + \sigma_{n,k,f} + \langle f - \mathcal{L}_n f, f - \mathcal{L}_n f\rangle_m \\ = \langle f, f\rangle_m.\end{aligned} \tag{2.3.8}$$

As $\sigma_{n,k,f}$ stands for the error in evaluating the integral of $(\mathcal{L}_n f - \mathcal{L}_k f)^2$ over Ω by

the quadrature rule (1.2.3) with exactness degree $n+k$, the Marcinkiewicz–Zygmund property (2.2.5) implies

$$|\sigma_{n,k,f}| \leq \eta \langle \mathcal{L}_n f - \mathcal{L}_k f, \mathcal{L}_n f - \mathcal{L}_k f \rangle = \eta \langle f, (\mathcal{L}_n - \mathcal{L}_k) f \rangle_m.$$

Thus, together with the non-negativeness of $\langle f - \mathcal{L}_n f, f - \mathcal{L}_n f \rangle_m$, the expression (2.3.8) leads to

$$\langle \mathcal{L}_k f, \mathcal{L}_k f \rangle_m + (1-\eta) \langle f, (\mathcal{L}_n - \mathcal{L}_k) f \rangle_m \leq \langle f, f \rangle_m,$$

that is,

$$\langle f, (\mathcal{L}_n - \mathcal{L}_k) f \rangle_m \leq \frac{1}{1-\eta} \left(\langle f, f \rangle_m - \langle \mathcal{L}_k f, \mathcal{L}_k f \rangle_m \right).$$

Hence, we have

$$\begin{aligned} \|\mathcal{L}_n f\|_2^2 =& \langle \mathcal{L}_k f, \mathcal{L}_k f \rangle_m + \langle f, (\mathcal{L}_n - \mathcal{L}_k) f \rangle_m \\ \leq& \frac{1}{1-\eta} \langle f, f \rangle_m - \frac{\eta}{1-\eta} \langle \mathcal{L}_k f, \mathcal{L}_k f \rangle_m \\ \leq& \frac{1}{1-\eta} \langle f, f \rangle_m, \end{aligned}$$

and the stability result (2.2.6) follows from

$$\langle f, f \rangle_m = \sum_{j=1}^m w_j f(x_j)^2 \leq \sum_{j=1}^m w_j \|f\|_\infty^2 = V \|f\|_\infty^2.$$

The error bound (2.2.7) can be derived from a standard argument. For any $\chi \in \mathbb{P}_k$, with the aid of Lemma 2.2.4, there holds

$$\mathcal{L}_n f - f = \mathcal{L}_n (f - \chi) - (f - \chi).$$

Using the stability result (2.2.6), we have

$$\begin{aligned} \|\mathcal{L}_n f - f\|_2 = \|\mathcal{L}_n (f-\chi) - (f-\chi)\|_2 &\leq \|\mathcal{L}_n (f-\chi)\|_2 + \|f - \chi\|_2 \\ &\leq \frac{V^{1/2}}{\sqrt{1-\eta}} \|f-\chi\|_\infty + V^{1/2} \|f-\chi\|_\infty \\ &= \left(\frac{1}{\sqrt{1-\eta}} + 1 \right) V^{1/2} \|f-\chi\|_\infty. \end{aligned}$$

This estimate implies, as it holds for all $\chi \in \mathbb{P}_k$, that

$$\begin{aligned}\|\mathcal{L}_n f - f\|_2 &\leq \left(\frac{1}{\sqrt{1-\eta}} + 1\right) V^{1/2} \inf_{\chi \in \mathbb{P}_k} \|f - \chi\|_\infty \\ &= \left(\frac{1}{\sqrt{1-\eta}} + 1\right) V^{1/2} E_k(f).\end{aligned}$$

If k is fixed, then $E_k(f)$ is fixed, suggesting that no convergence result of $\mathcal{L}_n f$ as $n \to \infty$ can be concluded. On the other hand, if k is positively correlated to n, then $E_k(f) \to 0$ and hence $\|\mathcal{L}_n f - f\|_2 \to 0$ as $n \to \infty$. □

2.4 Examples and numerical experiments

We now apply Theorem 2.2.8 to two regions: the interval $[-1, 1] \subset \mathbb{R}$ and the 2-sphere $\mathbb{S}^2 \subset \mathbb{R}^3$. For the simplicity of the narrative, we assume that the following mentioned quadrature rules have the Marcinkiewicz–Zygmund property (2.2.5) with $\eta = 3/4$, a quite loose assumption for $\eta \in [0, 1)$. All codes were written by MATLAB R2022a, and all numerical experiments were conducted on a laptop (16 GB RAM, Intel® CoreTM i7-9750H Processor) with macOS Monterey 12.4.

2.4.1 On the interval

Let $\Omega = [-1, 1]$ with $\mathrm{d}\omega = \omega(x)\mathrm{d}x$, where $\omega(x) \geq 0$ is a weight function on $[-1, 1]$ and different $\omega(x)$ leads to different value of $V = \int_{-1}^{1} \omega(x)\mathrm{d}x$. The space $\mathbb{P}_n$ is a linear space of polynomials of degree at most n on $[-1, 1]$, hence $d_n = n + 1$.

In the following example, we consider $\omega(x) = 1$ (thus $V = 2$), and quadrature rules with such weight function include the Gauss–Legendre quadrature and the Clenshaw–Curtis quadrature. We refer the reader to [222] for background information about quadrature rules on $[-1, 1]$. The Gauss–Legendre quadrature rule is a typical choice of quadrature rules for the original hyperinterpolation $\mathcal{L}_n^{\mathrm{S}}$, as an m-point Gauss–Legendre quadrature has exactness degree $2m - 1$. For effective testing of Gaussian quadrature rules, we refer the reader to [96]. Thus, an $(n + 1)$-point Gauss–Legendre quadrature can fulfill the exactness requirement $2n$ of $\mathcal{L}_n^{\mathrm{S}}$. Meanwhile, the Clenshaw–Curtis quadrature [62] in the Chebyshev points, which has exactness degree $m - 1$ if m quadrature points are adopted, is not considered practical in constructing the original hyperinterpolants. Indeed, one needs a $(2n+1)$-point Clenshaw–Curtis quadrature to construct an original hyperinterpolant $\mathcal{L}_n^{\mathrm{S}} f$. However, in the light of Theorem 2.2.8, we have the following corollary.

Corollary 2.4.1 *Let $\langle\cdot,\cdot\rangle_m$ used in Definition 2.2.2 be an m-point Gauss-Legendre quadrature with $(n+2)/2 \le m \le (2n+1)/2$, or an m-point Clenshaw–Curtis quadrature with $n+2 \le m \le 2n+1$. Under the conditions of Theorem 2.2.8 with $\eta = 3/4$, the exactness-relaxing hyperinterpolant $\mathcal{L}_n f$ satisfies*

$$\|\mathcal{L}_n f - f\|_2 \le \begin{cases} 3V^{1/2} E_{2m-1-n}(f) & \text{when using the Gauss–Legendre quadrature,} \\ 3V^{1/2} E_{m-1-n}(f) & \text{when using the Clenshaw–Curtis quadrature.} \end{cases}$$

It is worth noting that the m-point Newton–Cotes quadrature in the equispaced points with $n+2 \le m \le 2n+1$, though having exactness degree exceeding $n+1$, fails to fulfill the assumption of positive weights, as the Newton–Cotes weights have alternating signs. However, this does not suggest the impossibility of constructing hyperinterpolants in the equispaced points. Quadrature rules with exactness $n+k$ in the equispaced points, even in the scattered points, can be designed in the spirit of optimal recovery rather than the exactness principle. As suggested in [70], given m distinct points $\{x_j\}_{j=1}^m$, one can design a quadrature with exactness degree $n+k$ by obtaining its quadrature weights $\{w_j\}_{j=1}^m$ from solving

$$\min_{w_1,w_2,\ldots,w_m} \sum_{j=1}^m |w_j| \quad \text{s.t.} \quad \sum_{j=1}^m w_j v(x_j) = \int_{-1}^1 v \quad \forall v \in \mathbb{P}_{n+k}. \tag{2.4.1}$$

In general, the number m of quadrature points in the rule (2.4.1) should be much larger than the exactness-oriented quadrature rules to achieve the exactness degree $n+k$. For example, to design an m-equispaced-point quadrature with exactness degree $n+k$ in the spirit of (2.4.1), m, n, and k shall satisfy $n+k = \mathcal{O}(\sqrt{m \ln m})$, see [70, Theorem 3.6]. Thus, we have the following result.

Corollary 2.4.2 *Let $\langle\cdot,\cdot\rangle_m$ used in Definition 2.2.2 be an m-point quadrature designed by (2.4.1), where the quadrature points are equispaced points on $[-1,1]$, and the weights should be positive. Under the conditions of Theorem 2.2.8 with $\eta = 3/4$, the error of the exactness-relaxing hyperinterpolant $\mathcal{L}_n f$ is controlled by*

$$\|\mathcal{L}_n f - f\|_2 \le 3V^{1/2} E_k(f).$$

We present a toy example on the interval $[-1,1]$ to illustrate Theorem 2.2.8 on $\Omega = [-1,1]$. We are interested in a 40-degree hyperinterpolant $\mathcal{L}_{40} f$ of $f = \exp(-x^2)$ and $f = |x|^{5/2}$, with $\{p_\ell\}_{\ell=1}^{41}$ chosen as normalized Legendre polynomials $\{P_\ell\}_{\ell=0}^{40}$. The former test function $f = \exp(-x^2)$ is an analytic function (so smooth enough) and the latter $f = |x|^{5/2}$ is only continuous (not even differentiable).

Constructing $\mathcal{L}_{40}^{\mathrm{S}} f$ requires a quadrature rule with exactness degree 80, thus one may consider a 41-point Gauss quadrature with exactness degree 81. Besides, we also construct $\mathcal{L}_{40} f$ using a 25-point Gauss-Legendre quadrature, a 50-point Clenshaw–Curtis quadrature, and a 186-point quadrature (2.4.1) in equispaced points with exactness degree 49. These quadrature rules all have the exactness degree 49, which is far from the required degree 80 for $\mathcal{L}_{40}^{\mathrm{S}} f$, but they also enable us to obtain hyperinterpolants with considerably small errors. On the other hand, the relaxation of quadrature exactness, suggested in Theorem 2.2.8, slows the convergence rates of hyperinterpolants. That is, the L^2 error estimation of $\mathcal{L}_{40}^{\mathrm{S}} f$ is controlled by $E_{40}(f)$, suggested by the estimation (2.2.3) derived in Sloan's original work [196], while that of $\mathcal{L}_{40} f$ is controlled by $E_9(f)$, according to our error estimation (2.2.7).

The performance of $\mathcal{L}_{40} f$ in the approximation of both functions is displayed in Figures 2.1 and 2.2, respectively. Our theoretical analysis of the effects of the relaxing quadrature exactness is also verified in both figures. Besides, the numerical results suggest that such effects may also be related to the smoothness of functions to be approximated. That is, the error of $\mathcal{L}_{40}^{\mathrm{S}} f$ is much smaller than the errors of $\mathcal{L}_{40} f$ using three different quadrature rules for the analytic function $f = \exp(-x^2)$, but just slightly smaller than those for the non-differentiable function $f = |x|^{5/2}$. Moreover, it is pretty interesting that the hyperinterpolant $\mathcal{L}_{40} f$ with the 50-point Clenshaw–Curtis quadrature performs better than that using the 25-point Gauss–Legendre quadrature and the 186-point quadrature (2.4.1) in equispaced points, though three quadrature rules have the same exactness degree 49. This finding is worthy of further study. To the authors' best knowledge, the connection between the Clenshaw–Curtis quadrature and the performance of hyperinterpolation has not been established. Some possibly useful results that help us to establish such a connection can be found in Trefethen's famous paper [220].

2.4.2 On the sphere

Let $\Omega = \mathbb{S}^2 \subset \mathbb{R}^3$ with $\mathrm{d}\omega = \omega(x)\mathrm{d}x$, where $\omega(x)$ is an area measure on $\mathbb{S}^2$. Thus $V = \int_{\mathbb{S}^2} \mathrm{d}\omega = 4\pi$ denotes the surface area of $\mathbb{S}^2$. In this example, $\mathbb{P}_n$ can be regarded as the space of spherical polynomials of degree at most n. Let the basis $\{p_\ell\}_{\ell=1}^{d_n}$ be a set of orthonormal spherical harmonics $\{Y_{\ell,k} : \ell = 0, 1, \ldots, n, k = 1, \ldots, 2\ell + 1\}$, and the dimension of $\mathbb{P}_n$ is $d_n = (n+1)^2$. Many positive-weight quadrature rules can achieve the desired exactness degree, such as rules using spherical t-designs [69] and tensor-product quadrature rules from rules on the interval [202], which are both designed on structural quadrature points. Thanks to the work of Mhaskar, Narcowich, and Ward [146], it was also proved that positive-weight quadrature rules with desired polynomial exactness could be designed from scattered data. All of

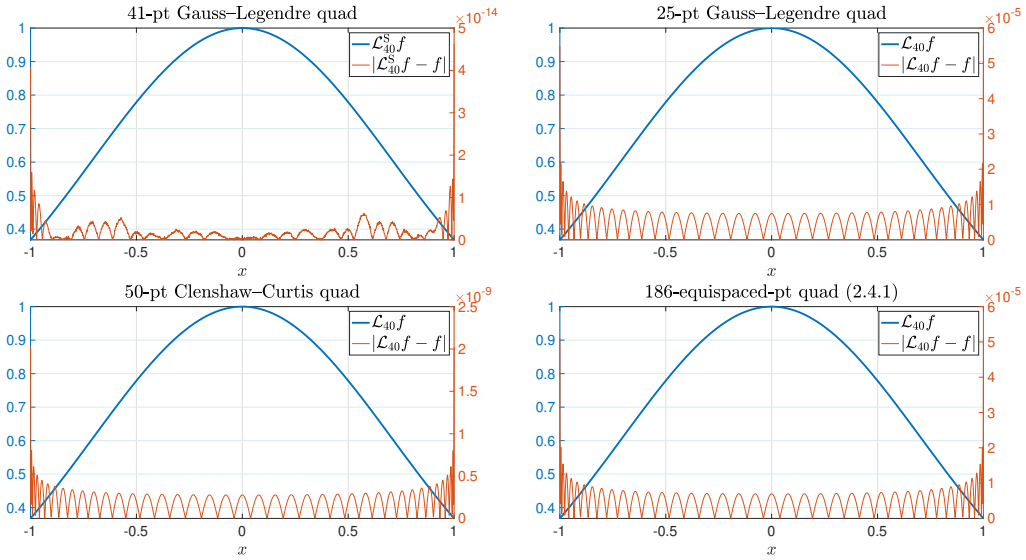


Figure 2.1: Hyperinterpolants $\mathcal{L}_{40}^{\mathrm{S}} f$ and $\mathcal{L}_{40} f$ of $f = \exp(-x^2)$, constructed by various quadrature rules. The estimation of $\|\mathcal{L}_{40}^{\mathrm{S}} f - f\|_2$ is controlled by $E_{40}(f)$, while that of $\|\mathcal{L}_{40} f - f\|_2$ by $E_9(f)$.

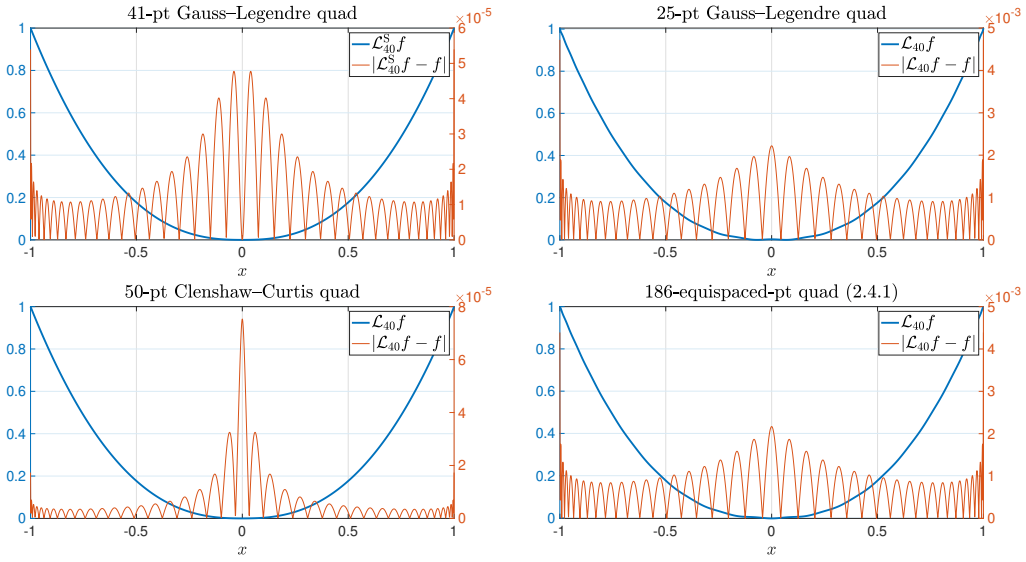


Figure 2.2: Hyperinterpolants $\mathcal{L}_{40}^{\mathrm{S}} f$ and $\mathcal{L}_{40} f$ of $f = |x|^{5/2}$, constructed by various quadrature rules. The estimation of $\|\mathcal{L}_{40}^{\mathrm{S}} f - f\|_2$ is controlled by $E_{40}(f)$, while that of $\|\mathcal{L}_{40} f - f\|_2$ by $E_9(f)$.

these rules requires $m = \mathcal{O}(k^2)$ points to achieve the exactness degree k. Thus roughly speaking, to construct an original hyperinterpolant requires $4cn^2$ points, where $c > 0$ is some constant, while in the light of Theorem 2.2.8, only $c(n+k)^2$ points with $0 < k \leq n$ are needed.

For the sake of easy implementation, we discuss Theorem 2.2.8 with quadrature rules using spherical t-designs, which can be implemented easily and efficiently. A point set $\{x_1, x_2, \ldots, x_m\} \subset \mathbb{S}^2$ is said to be a *spherical t-design* [69] if it satisfies

$$\frac{1}{m}\sum_{j=1}^{m} v(x_j) = \frac{1}{4\pi}\int_{\mathbb{S}^2} v \mathrm{d}\omega \quad \forall v \in \mathbb{P}_t. \tag{2.4.2}$$

It can be seen that spherical t-design is a set of points on the sphere such that an equal-weight quadrature rule in these points integrates all (spherical) polynomials up to degree t exactly. In this chapter, we employ well conditioned spherical t-designs [7], which are suitable for numerical integration and interpolation. The study in [8] revealed that well conditioned spherical t-designs can be used to realize hyperinterpolation and regularization approximation successfully. Well conditioned spherical t-designs require at least $(t+1)^2$ quadrature points to achieve the exactness degree t [7]. Thus, it requires at least $(2n+1)^2$ points to construct an original hyperinterpolant of degree n. However, thanks to Theorem 2.2.8, we have the following result.

Corollary 2.4.3 *Let $\langle \cdot, \cdot \rangle_m$ used in Definition 2.2.2 be the quadrature rule (2.4.2) using a spherical $(n+k)$-design with $0 < k \leq n$. The number m of quadrature points should satisfy $m \geq (n+k+1)^2$. Under the conditions of Theorem 2.2.8 with $\eta = 3/4$, the exactness-relaxing hyperinterpolant $\mathcal{L}_n f$ satisfies*

$$\|\mathcal{L}_n f - f\|_2 \leq 6\pi^{1/2} E_k(f).$$

In particular, if the spherical $(n+k)$-design with $m = (n+k+1)^2$ is used, then

$$\|\mathcal{L}_n f - f\|_2 \leq 6\pi^{1/2} E_{\sqrt{m}-n-1}(f).$$

We present a toy illustration on the sphere, making use of the well conditioned spherical t-designs [7] with $m = (t+1)^2$. We are interested in a 25-degree hyperinterpolant $\mathcal{L}_{25} f$ of a Wendland function f: Let $\mathbf{z}_1 = [1,0,0]^{\mathrm{T}}$, $\mathbf{z}_2 = [-1,0,0]^{\mathrm{T}}$, $\mathbf{z}_3 = [0,1,0]^{\mathrm{T}}$, $\mathbf{z}_4 = [0,-1,0]^{\mathrm{T}}$, $\mathbf{z}_5 = [0,0,1]^{\mathrm{T}}$, and $\mathbf{z}_6 = [0,0,-1]^{\mathrm{T}}$, the testing function f is defined as

$$f(\mathbf{x}) = \sum_{i=1}^{6} \phi_2(\|\mathbf{z}_i - \mathbf{x}\|_2), \tag{2.4.3}$$

where $\phi_2(r) := \tilde{\phi}_2(r/\delta_2)$ is a normalized Wendland function [60], with

$$\tilde{\phi}_2(r) := (\max\{1-r, 0\})^6 (35r^2 + 18r + 3)/3$$

been an original Wendland function [231] and $\delta_2 = (9\Gamma(5/2))/(2\Gamma(3))$. According to the original definition of hyperinterpolation (1.2.5), one shall use a spherical 50-design and its corresponding quadrature rule to construct $\mathcal{L}_{25}^{\mathbb{S}} f$. To tell the difference between $\mathcal{L}_{25}^{\mathbb{S}} f$ and $\mathcal{L}_{25} f$, we also use a sphere 30-design and its corresponding quadrature rule to construct $\mathcal{L}_{25} f$. Both designs are displayed in Figure 2.3.

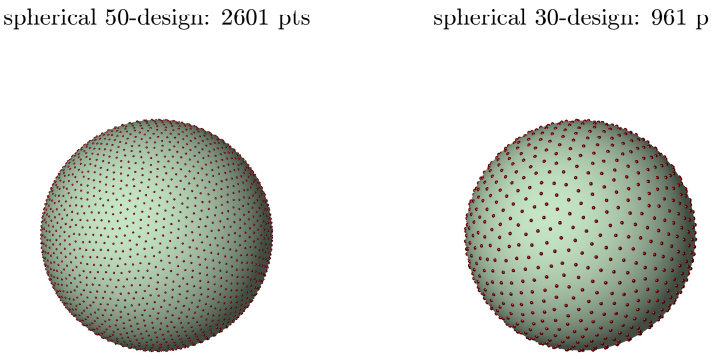


Figure 2.3: Spherical 50- and 30-designs, generated by the method proposed in [7].

The original hyperinterpolant $\mathcal{L}_{25}^{\mathbb{S}} f$ of the Wendland-type function (2.4.3) and the corresponding error are plotted in the upper row of Figure 2.4. According to Sloan [196], the L^2 error estimation of $\mathcal{L}_{25}^{\mathbb{S}} f$ is controlled by $E_{25}(f)$. Corollary 2.4.3 indicates that $\mathcal{L}_{25} f$ can be obtained using an exactness-relaxing quadrature rule. This is shown in the lower row in Figure 2.4, where a sphere 30-design and its corresponding quadrature rule are used. Corollary 2.4.3 also suggests that the L^2 error estimation of $\mathcal{L}_{25} f$ is thus controlled by $E_5(f)$.

Along with the Wendland-type function (2.4.3), we additionally test the function $f(\mathbf{x}) = f(x, y, z) = |x + y + z|$ with $\mathbf{x} = [x, y, z]^{\mathrm{T}} \in \mathbb{S}^2$. Similar to the above test, the original hyperinterpolant $\mathcal{L}_{25}^{\mathbb{S}} f$ and the corresponding error are plotted in the upper row of Figure 2.5, and the hyperinterpolant $\mathcal{L}_{25} f$ and its error are shown in the lower row of Figure 2.5. This test also validates our theory on the effects of the relaxing quadrature exactness. Moreover, as the function $f(x, y, z) = |x + y + z|$ is not differentiable, similar to the non-differentiable function $f(x) = |x|^{5/2}$ on $[-1, 1]$, we see than the error of $\mathcal{L}_{25}^{\mathbb{S}} f$ is just slightly smaller than that of $\mathcal{L}_{25} f$.

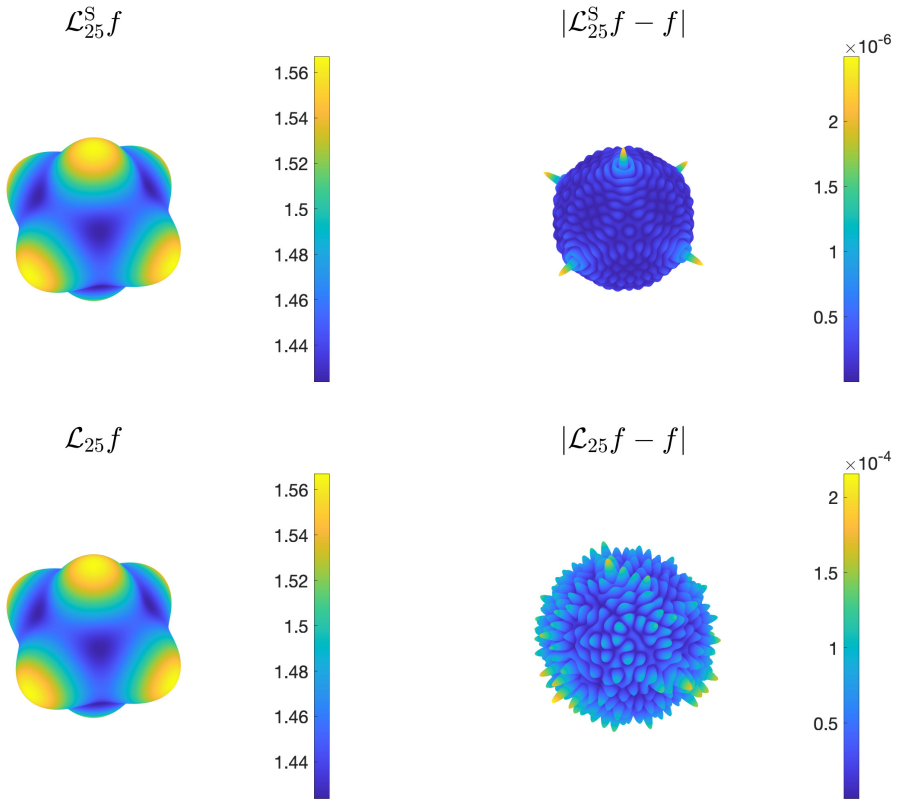


Figure 2.4: Hyperinterpolants $\mathcal{L}_{25}^{\mathrm{S}} f$ and $\mathcal{L}_{25} f$ of a Wendland-type function (2.4.3), constructed by spherical t-designs with $t = 50$ (upper row) and 30 (lower row), respectively. The estimation of $\|\mathcal{L}_{25}^{\mathrm{S}} f - f\|_2$ is controlled by $E_{25}(f)$, while that of $\|\mathcal{L}_{25} f - f\|_2$ by $E_5(f)$.

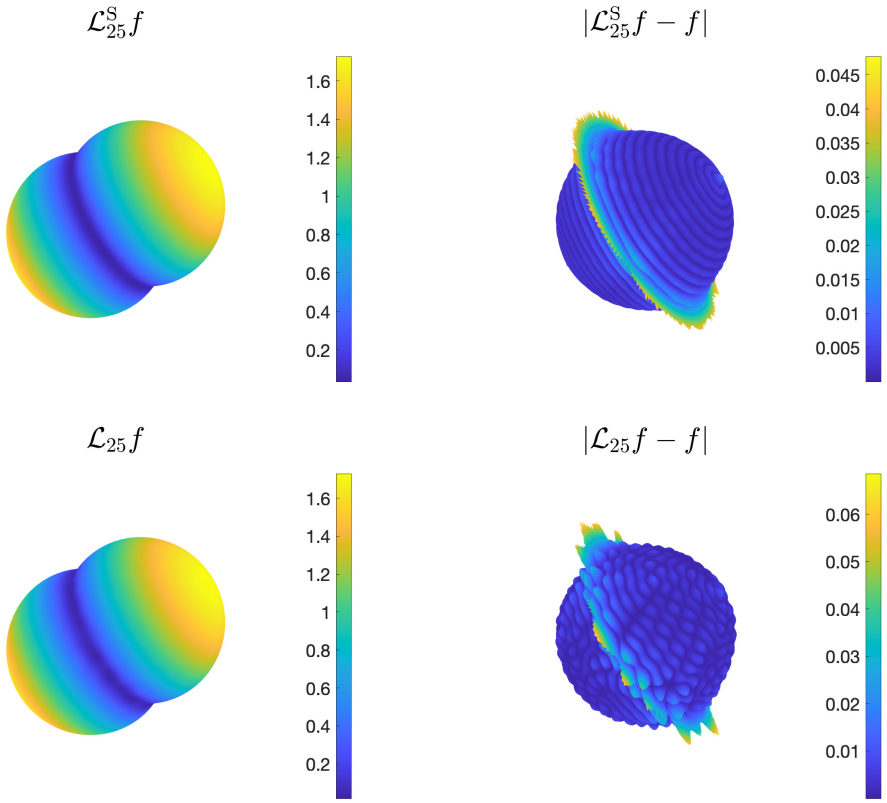


Figure 2.5: Hyperinterpolants $\mathcal{L}_{25}^{\mathrm{S}} f$ and $\mathcal{L}_{25} f$ of $f(\mathbf{x}) = f(x, y, z) = |x + y + z|$, constructed by spherical t-designs with $t = 50$ (upper row) and 30 (lower row), respectively. The estimation of $\|\mathcal{L}_{25}^{\mathrm{S}} f - f\|_2$ is controlled by $E_{25}(f)$, while that of $\|\mathcal{L}_{25} f - f\|_2$ by $E_5(f)$.

Chapter 3

Hyperinterpolation of singular and oscillatory functions

Note: *We denote by $n + n'$ the relaxed quadrature exactness degree in this chapter instead of $n + k$ in Chapter 2.*

Singular and oscillatory functions occupy pivotal positions in a wide array of applications, and their approximation is crucial for solving applied mathematics problems efficiently. As discussed in Chapter 2, hyperinterpolation is a discrete projection method approximating functions with the L^2 orthogonal projection coefficients obtained by numerical integration. However, this approach may be inefficient for approximating singular and oscillatory functions, requiring a large number of integration points to achieve satisfactory accuracy. To address this issue, we propose a new approximation scheme in this chapter, called efficient hyperinterpolation, which leverages the product-integration methods to attain the desired accuracy with fewer numerical integration points than the original scheme. We provide theorems that explain the superiority of efficient hyperinterpolation over the original method in approximating such functions belonging to $L^1(\Omega)$, $L^2(\Omega)$, and $C(\Omega)$ spaces, respectively, and demonstrate through numerical experiments on the interval and the sphere that our approach outperforms the original method in terms of accuracy when using a limited number of integration points.

3.1 Introduction

Recall that Ω is a bounded region of $\mathbb{R}^d$, either the closure of a connected open domain or a smooth closed lower-dimensional manifold in $\mathbb{R}^d$. The region is assumed to have finite measure with respect to a given measure $\mathrm{d}\omega$, that is, $\int_\Omega \mathrm{d}\omega = V < \infty$.

We are interested in the efficient numerical approximation of functions in the form of

$$F(x) = K(x)f(x) \tag{3.1.1}$$

by some polynomials on Ω, where $K \in L^1(\Omega)$ is a real- or complex-valued absolutely integrable function, which needs not be continuous or of one sign, and $f \in C(\Omega)$ is a continuous (and preferably smooth) function. By *efficient*, we mean that a considerably small amount of sampling points is enough for such approximation with satisfactory accuracy. We also investigate scenarios of $K \in L^2(\Omega)$ and $C(\Omega)$ to refine the general (but rough) analysis for the case of $K \in L^1(\Omega)$.

3.1.1 Sources of singular and oscillatory functions

Functions in the form of (3.1.1) frequently feature in mathematical physics and applied mathematics [63, 101]. Some differential equations naturally have solutions with oscillatory behaviors and singularities. For example, the fundamental solutions of the Helmholtz equation are given by

$$G(x,y) = \begin{cases} \dfrac{i}{4} H_0^{(1)}(\kappa|x-y|) & \text{for } x,y \in \mathbb{R}^2 \\ \dfrac{1}{4\pi} \dfrac{e^{i\kappa|x-y|}}{|x-y|} & \text{for } x,y \in \mathbb{R}^3, \end{cases}$$

where $|x-y|$ denotes the usual Euclidean distance between x and y, $H_0^{(1)}(z)$ is the Hankel function of the first kind and of order zero, and κ is known as the wave number when the equation is applied to waves. The fundamental solution of the biharmonic differential equation in $\mathbb{R}^2$ is given by

$$G(x,y) = \frac{1}{8\pi}|x-y|^2 \log|x-y| \quad \text{for } x,y \in \mathbb{R}^2.$$

Another important source of singular and oscillatory functions can be found in the study of

$$\frac{Y_{\ell,k}(y)}{|x-y|} \quad \text{for } x,y \in \mathbb{S}^2$$

for the electromagnetic field and wave computation [15, 61, 101], where $Y_{\ell,k}$ is the spherical harmonic of degree ℓ and order k.

As we can see, many fundamental solutions are functions with singularities and oscillatory behaviors. The approximation of such functions helps us develop approximation methods to solve related mathematical physics problems. Thus, designing

an efficient method for such approximation is a fascinating area of computational mathematics.

3.1.2 The approximation basics

A typical constructive approximation scheme of degree n for $F = Kf$ consists of two stages: evaluating the integrals

$$\int_\Omega (Kf) p_\ell \mathrm{d}\omega, \quad \ell = 1, 2, \ldots, d_n, \tag{3.1.2}$$

and then approximating F by

$$\mathcal{P}_n F := \sum_{\ell=1}^{d_n} \left(\int_\Omega (Kf) p_\ell \mathrm{d}\omega \right) p_\ell. \tag{3.1.3}$$

This scheme (3.1.3) is the L^2 orthogonal projection (1.2.2) of F onto $\mathbb{P}_n$. To link the orthogonal projection to applications immediately, we make use of some quadrature rules (1.2.3). With the assumption (1.2.4) of quadrature exactness

$$\sum_{j=1}^{m} w_j g(x_j) = \int_\Omega g \mathrm{d}\omega \quad \forall g \in \mathbb{P}_{2n},$$

the hyperinterpolant of degree n, constructed for the approximation of $F \in C(\Omega)$, is defined as

$$\mathcal{L}_n F := \sum_{\ell=1}^{d_n} \langle Kf, p_\ell \rangle_m p_\ell. \tag{3.1.4}$$

We refer the reader to [101, 110, 128, 177, 197, 202, 203, 234] for some follow-up works on the general analysis of hyperinterpolation and [9, 137, 149, 205] for some variants of classical hyperinterpolation. The approximation of the form (3.1.4) using rotationally invariant quadrature rules on the 2-sphere $\mathbb{S}^2$ was also investigated in [4].

However, it is well known that if K is singular and highly oscillatory, it is inefficient to evaluate the integrals (3.1.2) directly using some classical numerical integration rules. Instead, one shall evaluate them in a semi-analytical way: for the evaluation of an integral of the form

$$\int_\Omega K(x) f(x) \mathrm{d}\omega(x),$$

one shall replace f by its polynomial interpolant or approximant of degree n, expressed as

$$f \approx \sum_{\ell=1}^{d_n} c_\ell p_\ell,$$

and evaluate the integral by

$$\int_\Omega K(x) f(x) \mathrm{d}\omega(x) \approx \sum_{\ell=1}^{d_n} c_\ell \int_\Omega K(x) p_\ell(x) \mathrm{d}\omega(x).$$

This idea for numerical integration may be referred to as the *product-integration rule* in the classical literature [198, 199, 200]. This rule was initially designed on $[-1,1]$ for $K \in L^1[-1,1]$ and $f \in C[-1,1]$, and it converges to the exact integral as the number of quadrature points approaches the infinity if $K \in L^p[-1,1]$ for some $p > 1$ is additionally assumed. In the context of highly oscillatory integrals with an oscillatory $K \in C(\Omega)$, this approach is also known as the *Filon-type method* [90, 118]. In most of these references, f is approximated by its interpolant, and it is generally assumed that the *modified moments*

$$\int_\Omega K(x) p_\ell(x) \mathrm{d}\omega(x), \quad \ell = 1, 2, \ldots, d_n \tag{3.1.5}$$

can be computed accurately by using special functions or efficiently by invoking some stable iterative procedures. Besides, f may also be replaced by its approximant. For example, the idea of replacing f with its hyperinterpolant has emerged in the first paper [196] on hyperinterpolation. It may be better to replace f with its hyperinterpolant rather than the interpolant: The L^2 operator norm of hyperinterpolation is bounded if the regional area/volume V is finite [196], but there is no guarantee of the boundedness of the L^2 operator norm of polynomial interpolation over general regions; see a piece of evidence from [196].

In this spirit, we propose *efficient hyperinterpolation*, a general scheme for approximating functions in the form of (3.1.1), provided that the modified moment (3.1.5) can be readily obtained. We approximate f by its hyperinterpolant and the resulting scheme is defined as

$$\mathcal{S}_n F := \sum_{\ell=1}^{d_n} \left(\int_\Omega K(\mathcal{L}_n f) p_\ell \mathrm{d}\omega \right) p_\ell. \tag{3.1.6}$$

Along with the classical hyperinterpolation (3.1.4), this scheme can be regarded as another discrete approximation of the L^2 orthogonal projection (3.1.3). The main theoretical results of this chapter are the stability and error analysis for this scheme,

and this scheme is verified to be efficient when the amount of quadrature points is considerably small.

Although singular and oscillatory integration has been well studied in the classical literature, we found these studies were not widely linked to hyperinterpolation. Here is a possible explanation for this gap. The required quadrature exactness degree $2n$ for a hyperinterpolant of degree n *de facto* ensures a sufficient amount of numerical integration points when n is relatively large. Thus, directly evaluating the integrals (3.1.2) by the classical numerical integration methods may also lead to relatively satisfactory accuracy.

In Chapter 2, we discussed what if the required exactness $2n$ is relaxed to $n+n'$, where $0 < n' \leq n$. This discussion provides a regime where efficient hyperinterpolation may perform much more accurately than classical hyperinterpolation. In particular, if K is continuous, we show that for the classical hyperinterpolation of degree n, the approximation error is bounded as

$$\|\mathcal{L}_n F - F\|_2 \lesssim E_{n'}(Kf),$$

where

$$E_{n'}(g) = \inf_{\chi \in \mathbb{P}_{n'}} \|g - \chi\|_\infty$$

for $g \in C(\Omega)$; while for efficient hyperinterpolation of degree n, there holds

$$\|\mathcal{S}_n F - F\|_2 \lesssim E_{n'}(f) + E_n(K\chi^*),$$

where $\chi^* \in \mathbb{P}_{n'}$ is the best uniform approximation of f in $\mathbb{P}_{n'}$, that is,

$$\|f - \chi^*\|_\infty = E_{n'}(f).$$

Thus, the controlling term $E_{n'}(Kf)$ is considerably greater than $E_{n'}(f)$ and $E_n(K\chi^*)$ when $n' < n$, f is smooth enough, and K is awkward enough to be approximated by lower degree polynomials, asserting the outperformance of efficient hyperinterpolation in this scenario.

The rest of this chapter is organized as follows. In the next section, we review some results of the classical hyperinterpolation and discuss some properties of the efficient modification. The implementation of efficient hyperinterpolation is elaborated in Section 3.3. In Section 3.4, we analyze the stability and the error bound for efficient hyperinterpolation when $K \in L^1(\Omega)$. This analysis is refined in Section 3.5 with the assumptions that $K \in L^2(\Omega)$ and $K \in C(\Omega)$. In particular, we discuss in Section 3.5.3 why the classical hyperinterpolation may be inefficient when

approximating functions in the form of (3.1.1). In Section 3.6, we investigate efficient hyperinterpolation on the interval and the sphere, respectively, and give some numerical results.

3.2 Hyperinterpolation and efficient hyperinterpolation

Hyperinterpolation (3.1.4) uses classical numerical integration methods to evaluate the L^2 orthogonal projection coefficients (3.1.2). However, the classical methods prove to be inefficient in the presence of a singular or an oscillatory K. Thus, we propose efficient hyperinterpolation (3.1.6) to achieve satisfactory approximation accuracy by using a considerably small amount of quadrature points. In this section, we review some results of (3.1.4) and discuss some properties of (3.1.6).

3.2.1 Hyperinterpolation

As introduced, the original definition (3.1.4) of hyperinterpolants of degree n requires an m-point quadrature rule (1.2.3) with polynomial exactness $2n$ [196], and this requirement on quadrature exactness has been relaxed to $n + n'$ with $0 < n' \leq n$ in Chapter 2.

The definition (3.1.4) is also restricted to the approximation of continuous functions. Thus, if K is additionally assumed to be continuous, then it was derived in [196] that $\mathcal{L}_n F$ defined by (3.1.4) with quadrature exactness $2n$ shall satisfy

$$\|\mathcal{L}_n F\|_2 \leq V^{1/2} \|F\|_\infty \tag{3.2.1}$$

and

$$\|\mathcal{L}_n F - F\|_2 \leq 2V^{1/2} E_n(F), \tag{3.2.2}$$

where

$$E_n(g) = \inf_{\chi \in \mathbb{P}_n} \|g - \chi\|_\infty$$

denotes the best uniform approximation error of $g \in C(\Omega)$ by a polynomial in $\mathbb{P}_n$.

Let the quadrature rule (1.2.3) have exactness degree $n + n'$ with $0 < n' \leq n$, and let it satisfy the *Marcinkiewicz–Zygmund* property that there exists an $\eta \in [0, 1)$ such that

$$\left| \sum_{j=1}^{m} w_j \chi(x_j)^2 - \int_\Omega \chi^2 \mathrm{d}\omega \right| \leq \eta \int_\Omega \chi^2 \mathrm{d}\omega \quad \forall \chi \in \mathbb{P}_n, \tag{3.2.3}$$

and $\eta = 0$ if $n' = n$. The property (3.2.3) is referred to as the Marcinkiewicz–Zygmund property as it can be regarded as the Marcinkiewicz–Zygmund inequality

[89, 126, 143, 146] applied to polynomials of degree at most $2n$; see [12] for more details. If the quadrature rule (1.2.3) with exactness degree $n + n'$ satisfies the Marcinkiewicz–Zygmund property (3.2.3) with $\eta \in [0, 1)$, then it was derived in [12] that

$$\|\mathcal{L}_n F\|_2 \leq \frac{V^{1/2}}{\sqrt{1-\eta}} \|F\|_\infty \tag{3.2.4}$$

and

$$\|\mathcal{L}_n F - F\|_2 \leq \left(\frac{1}{\sqrt{1-\eta}} + 1 \right) V^{1/2} E_{n'}(F). \tag{3.2.5}$$

For the sake of generality, we have the following assumption for the rest of this chapter.

Assumption 3.2.1 *The quadrature rule (1.2.3) has exactness degree $n+n'$ with $0 < n' \leq n$, and it satisfies the Marcinkiewicz–Zygmund property (3.2.3) with $\eta \in [0, 1)$.*

3.2.2 Properties of efficient hyperinterpolation

We then make a short discussion on the relations among the L^2 orthogonal projection $\mathcal{P}_n$, hyperinterpolation $\mathcal{L}_n$, and efficient hyperinterpolation $\mathcal{S}_n$. Note that $\mathcal{P}_n \chi = \chi$ for all $\chi \in \mathbb{P}_n$, while for $\mathcal{L}_n$ with quadrature exactness $n + n'$ $(0 < n' \leq n)$, there only holds $\mathcal{L}_n \chi = \chi$ for all $\chi \in \mathbb{P}_{n'}$; see [12].

It is immediately observed that the efficient hyperinterpolation (3.1.6) can be represented in terms of the L^2 orthogonal projection $\mathcal{P}_n$ and hyperinterpolation $\mathcal{L}_n$.

Lemma 3.2.2 *Let $K \in L^1(\Omega)$ and $f \in C(\Omega)$. Then $\mathcal{S}_n F = \mathcal{P}_n(K \mathcal{L}_n f)$.*

Remark 3.2.3 *This observation in Lemma 3.2.2 may simplify our proofs below, but it cannot explain the computational benefits of the efficient hyperinterpolation and using modified moments. The aim of this chapter is to demonstrate the latter issue.*

We have the following lemma on the relation between $\mathcal{S}_n$ and $\mathcal{P}_n$.

Lemma 3.2.4 *Let $K \in L^1(\Omega)$. Then $\mathcal{S}_n(K\chi) = \mathcal{P}_n(K\chi)$ for all $\chi \in \mathbb{P}_{n'}$.*

Proof. Note that $\mathcal{L}_n \chi = \chi$ for all $\chi \in \mathbb{P}_{n'}$. Thus

$$\mathcal{S}_n(K\chi) = \mathcal{P}_n(K \mathcal{L}_n \chi) = \mathcal{P}_n(K\chi),$$

which proves this lemma. □

We then discuss the relation between $\mathcal{S}_n$ and $\mathcal{L}_n$. We can see that if $K = 1$, i.e., $F = f$, then $\mathcal{S}_n F = \mathcal{L}_n F$. Indeed, with the property that $\mathcal{P}_n \chi = \chi$ for all $\chi \in \mathbb{P}_n$,

we have

$$\mathcal{S}_n f = \mathcal{P}_n(\mathcal{L}_n f) = \mathcal{L}_n f.$$

If $K \neq 1$, we have the following lemma.

Lemma 3.2.5 *Let $K \in L^2(\Omega)$. Then $\langle K\mathcal{L}_n f - \mathcal{S}_n F, \chi\rangle = 0$ for all $\chi \in \mathbb{P}_n$.*

Proof. As $\mathcal{S}_n F = \mathcal{P}_n(K\mathcal{L}_n f)$, this lemma is proved by the projection property of the orthogonal projection $\mathcal{P}_n$: For any $f \in L^2(\Omega)$,

$$\langle \mathcal{P}_n f - f, \chi\rangle = 0$$

for all $\chi \in \mathbb{P}_n$. □

Lemma 3.2.5 suggests that $\mathcal{S}_n F$ is the orthogonal projection of $K\mathcal{L}_n f$ onto $\mathbb{P}_n$ as well as the following least squares property.

Theorem 3.2.6 *Let $K \in L^2(\Omega)$. Then*

$$\langle K\mathcal{L}_n f - \mathcal{S}_n F, K\mathcal{L}_n f - \mathcal{S}_n F\rangle = \min_{\chi\in\mathbb{P}_n} \langle K\mathcal{L}_n f - \chi, K\mathcal{L}_n f - \chi\rangle.$$

Proof. For any $\chi \in \mathbb{P}_n$, we have

$$K\mathcal{L}_n f - \chi = K\mathcal{L}_n f - \mathcal{S}_n F + \mathcal{S}_n F - \chi,$$

and by Lemma 3.2.5, we have

$$\langle K\mathcal{L}_n f - \mathcal{S}_n F, \mathcal{S}_n F - \chi\rangle = 0.$$

Thus, the Pythagorean theorem suggests

$$\|K\mathcal{L}_n f - \mathcal{S}_n F\|_2^2 + \|\mathcal{S}_n F - \chi\|_2^2 = \|K\mathcal{L}_n f - \chi\|_2^2,$$

which implies

$$\|K\mathcal{L}_n f - \mathcal{S}_n F\|_2^2 \leq \|K\mathcal{L}_n f - \chi\|_2^2$$

for all $\chi \in \mathbb{P}_n$ and

$$\|K\mathcal{L}_n f - \mathcal{S}_n F\|_2^2 = \|K\mathcal{L}_n f - \chi\|_2^2$$

if $\chi = \mathcal{S}_n F$. Hence the theorem is proved. □

3.3 Implementation of efficient hyperinterpolation

To implement efficient hyperinterpolation (3.1.6), the key step is to evaluate its coefficients. Note that for $\ell = 1, 2, \ldots, d_n$, each coefficient

$$\begin{aligned}\int_\Omega K(\mathcal{L}_n f)p_\ell \mathrm{d}\omega &= \int_\Omega K\left[\sum_{\ell'=1}^{d_n}\left(\sum_{j=1}^{m} w_j f(x_j)p_{\ell'}(x_j)\right)p_{\ell'}\right]p_\ell \mathrm{d}\omega \\ &= \sum_{j=1}^{m} w_j \left(\sum_{\ell'=1}^{d_n} p_{\ell'}(x_j)\int_\Omega K p_{\ell'}p_\ell \mathrm{d}\omega\right) f(x_j) \\ &= \sum_{j=1}^{m} W_{j\ell} f(x_j),\end{aligned}$$

where

$$W_{j\ell} := w_j \left(\sum_{\ell'=1}^{d_n} p_{\ell'}(x_j)\int_\Omega K p_{\ell'}p_\ell \mathrm{d}\omega\right), \quad j = 1, 2, \ldots, m.$$

Thus, the weights $\{W_{j\ell}\}$ can be computed analytically or stably if one can evaluate

$$\alpha_{\ell'\ell} := \int_\Omega K p_{\ell'}p_\ell \mathrm{d}\omega, \quad 1 \le \ell', \ell \le d_n \tag{3.3.1}$$

in the same manner. Note that $p_{\ell'}p_\ell$ is another polynomial of degree $n_1 + n_2$, where $n_1 := \deg p_{\ell'}$ and $n_2 := \deg p_\ell$. Thus, it can be expanded as

$$p_{\ell'}p_\ell = \sum_{r=1}^{d_{n_1+n_2}} c_r q_r,$$

where $\{q_r\}_{r=1}^{d_{2n}}$ is an orthonormal basis of $\mathbb{P}_{2n}$, which could be chosen from the same orthogonal family of $\{p_\ell\}$ or not, and the coefficients

$$c_r := \int_\Omega p_{\ell'}p_\ell q_r \mathrm{d}\mu, \quad r = 1, 2, \ldots, d_{n_1+n_2}. \tag{3.3.2}$$

In the expression (3.3.2), $\mathrm{d}\mu$ is the Lebesgue–Stieltjes measure associated with μ. Sometimes we may have

$$\mathrm{d}\mu(x) = \mu(x)\mathrm{d}x,$$

and $\mu(x)$ is referred to as the weight function of the orthogonal family $\{q_r\}$.

As introduced, it is generally assumed that the modified moments

$$\beta_r := \int_\Omega K q_r \mathrm{d}\omega \tag{3.3.3}$$

can be computed by using special functions or invoking some stable iterative procedures. In the implementation of efficient hyperinterpolation, we adopt this assumption for $r = 1, 2, \ldots, d_{2n}$. Thus, the weights

$$W_{j\ell} = w_j \sum_{\ell'=1}^{d_n} p_{\ell'}(x_j)\alpha_{\ell'\ell} = w_j \left[\sum_{\ell'=1}^{d_n} p_{\ell'}(x_j) \left(\sum_{r=1}^{d_{n_1+n_2}} c_r \beta_r \right) \right] \tag{3.3.4}$$

can be computed analytically or stably for $j = 1, 2, \ldots, m$ and $\ell = 1, 2, \ldots, d_n$.

The above discussion suggests how to implement efficient hyperinterpolation (3.1.6) in the form of

$$\mathcal{S}_n F = \sum_{\ell=1}^{d_n} \left(\sum_{j=1}^{m} W_{j\ell} f(x_j) \right) p_\ell. \tag{3.3.5}$$

Here is a pseudocode describing the whole procedure, which is easy to be implemented.

Algorithm. Efficient hyperinterpolant (3.1.6) for the approximation of $F = Kf$

Compute the modified moments (3.3.3) for $r = 1, 2, \ldots, d_{2n}$, save as $\{\beta_r\}_{r=1}^{d_{2n}}$;
for $\ell = 1 : d_n$
  for $\ell' = 1 : d_n$
    for $r = 1 : d_{n_1+n_2}$ % n_1+n_2 = degree of $p_{\ell'} p_\ell$
      $c_r = \langle p_{\ell'} p_\ell, q_r \rangle$;
    end
    $\alpha_{\ell'\ell} = \sum_{r=1}^{d_{n_1+n_2}} c_r \beta_r$;
  end
  for $j = 1 : m$
    $W_{j\ell} = w_j \sum_{\ell'=1}^{d_n} p_{\ell'}(x_j)\alpha_{\ell'\ell}$
  end
end
$\mathcal{S}_n F = \sum_{\ell=1}^{d_n} \left(\sum_{j=1}^{m} W_{j\ell} f(x_j) \right) p_\ell$.

3.4 Exploratory estimate: absolutely integrable kernels

We now analyze efficient hyperinterpolation for the approximation of

$$F = Kf$$

when $K \in L^1(\Omega)$. This case is the most general one among $K \in L^1(\Omega)$, $L^2(\Omega)$, and $C(\Omega)$, as there holds

$$C(\Omega) \subset L^2(\Omega) \subset L^1(\Omega)$$

for a bounded and closed subset Ω of $\mathbb{R}^d$. As $L^1(\Omega)$ does not carry any inner products, we can only give a general but rough analysis. What's more, since $F = Kf \in L^1(\Omega)$, we can only give an L^1 error analysis. We shall refine our analysis in the next section by assuming $K \in L^2(\Omega)$ and $C(\Omega)$.

Theorem 3.4.1 *Given $K \in L^1(\Omega)$ and $f \in C(\Omega)$, let $F = Kf$ and let $\mathcal{S}_n F$ be defined as (3.1.6), where the m-point quadrature rule (1.2.3) fulfills the Assumption 3.2.1. Then*

$$\|\mathcal{S}_n F\|_2 \leq \frac{V^{1/2} A_n}{\sqrt{1-\eta}} \|f\|_\infty, \tag{3.4.1}$$

where

$$A_n = \sqrt{\sum_{\ell=1}^{d_n} \sum_{\ell'=1}^{d_n} \alpha_{\ell'\ell}^2}$$

with $\alpha_{\ell'\ell}$ defined as (3.3.1), and

$$\begin{aligned} &\|\mathcal{S}_n F - F\|_1 \leq \\ &\qquad \left(\frac{V A_n}{\sqrt{1-\eta}} + \|K\|_1 \right) E_{n'}(f) + \left(V^{1/2} \sum_{\ell=1}^{d_n} \|p_\ell\|_\infty + 1 \right) E_n^{(1)}(K\chi^*), \end{aligned} \tag{3.4.2}$$

where $E_n^{(1)}(g) := \inf_{\chi \in \mathbb{P}_n} \|g - \chi\|_1$ and $\chi^* \in \mathbb{P}_{n'}$ is the best uniform approximation of f in $\mathbb{P}_{n'}$.

Proof. By Parseval's identity, we have

$$\|\mathcal{S}_n F\|_2^2 = \sum_{\ell=1}^{d_n} \left(\int_\Omega K(\mathcal{L}_n f) p_\ell \mathrm{d}\omega \right)^2 = \sum_{\ell=1}^{d_n} \left(\sum_{\ell'=1}^{d_n} \langle f, p_{\ell'} \rangle_m \int_\Omega K p_{\ell'} p_\ell \mathrm{d}\omega \right)^2.$$

By applying the Cauchy–Schwarz inequality and Parseval's identity again, we have

$$\|\mathcal{S}_n F\|_2^2 \leq \sum_{\ell=1}^{d_n} \left(\sum_{\ell'=1}^{d_n} \langle f, p_{\ell'} \rangle_m^2 \right) \left(\sum_{\ell'=1}^{d_n} \alpha_{\ell'\ell}^2 \right) = \|\mathcal{L}_n f\|_2^2 \sum_{\ell=1}^{d_n} \sum_{\ell'=1}^{d_n} \alpha_{\ell'\ell}^2,$$

which leads to

$$\|\mathcal{S}_n F\|_2 \leq A_n \|\mathcal{L}_n f\|_2.$$

By the stability result (3.2.4) with F changed to f, we have the stability result (3.4.1). For any $\chi \in \mathbb{P}_{n'}$, we have

$$\begin{aligned}\|\mathcal{S}_n F - F\|_1 &= \|\mathcal{S}_n(F - K\chi) - (F - K\chi) + (\mathcal{S}_n(K\chi) - K\chi)\|_1 \\ &\le V^{1/2}\|\mathcal{S}_n(F - K\chi)\|_2 + \|F - K\chi\|_1 + \|\mathcal{S}_n(K\chi) - K\chi\|_1 \\ &\le \frac{VA_n}{\sqrt{1-\eta}}\|f - \chi\|_\infty + \|K\|_1\|f - \chi\|_\infty + \|\mathcal{S}_n(K\chi) - K\chi\|_1,\end{aligned}$$

where the last inequality is obtained by applying the stability result (3.4.1) and Hölder's inequality to

$$F - K\chi = K(f - \chi),$$

respectively. As the above estimate applied to an arbitrary $\chi \in \mathbb{P}_{n'}$, letting $\chi = \chi^*$ gives

$$\|\mathcal{S}_n F - F\|_1 \le \left(\frac{VA_n}{\sqrt{1-\eta}} + \|K\|_1\right) E_{n'}(f) + \|\mathcal{S}_n(K\chi^*) - K\chi^*\|_1. \tag{3.4.3}$$

By Lemma 3.2.4, the term

$$\|\mathcal{S}_n(K\chi^*) - K\chi^*\|_1 = \|\mathcal{P}_n(K\chi^*) - K\chi^*\|_1.$$

Thus for any $\chi \in \mathbb{P}_n$, we have

$$\mathcal{P}_n(K\chi^*) - K\chi^* = \mathcal{P}_n(K\chi^* - \chi) - (K\chi^* - \chi)$$

and

$$\|\mathcal{P}_n(K\chi^*) - K\chi^*\|_1 \le V^{1/2}\|\mathcal{P}_n(K\chi^* - \chi)\|_2 + \|K\chi^* - \chi\|_1.$$

As for any $g \in L^1(\Omega)$, there holds

$$\begin{aligned}\|\mathcal{P}_n g\|_2 = \left(\sum_{\ell=1}^{d_n}\left(\int_\Omega g p_\ell \mathrm{d}\omega\right)^2\right)^{1/2} &\le \sum_{\ell=1}^{d_n}\left|\int_\Omega g p_\ell \mathrm{d}\omega\right| \\ \le \sum_{\ell=1}^{d_n}\|g p_\ell\|_1 \le \|g\|_1 \sum_{\ell=1}^{d_n}\|p_\ell\|_\infty,\end{aligned}$$

we have

$$\|\mathcal{P}_n(K\chi^*) - K\chi^*\|_1 \le \left(V^{1/2}\sum_{\ell=1}^{d_n}\|p_\ell\|_\infty + 1\right)\|K\chi^* - \chi\|_1.$$

Since this estimate applied to an arbitrary $\chi \in \mathbb{P}_n$, we have

$$\|\mathcal{P}_n(K\chi^*) - K\chi^*\|_1 \leq \left(V^{1/2} \sum_{\ell=1}^{d_n} \|p_\ell\|_\infty + 1 \right) E_n^{(1)}(K\chi^*).$$

Together with (3.4.3), we have the error bound (3.4.2). □

3.5 Refined estimates: square-integrable and continuous kernels

We then refine our general analysis in Section 3.4 by assuming $K \in L^2(\Omega)$ and $C(\Omega)$, respectively. Inner products emerge as a powerful tool in such refinement. For example, we used the estimate

$$\|\mathcal{P}_n g\|_2 \leq \|g\|_1 \sum_{\ell=1}^{d_n} \|p_\ell\|_\infty$$

for $g \in L^1(\Omega)$ in the proof of Theorem 3.4.1, but we have

$$\|\mathcal{P}_n g\|_2 \leq \|g\|_2 \quad \forall g \in L^2(\Omega), \tag{3.5.1}$$

and

$$\|\mathcal{P}_n g\|_2 \leq V^{1/2}\|g\|_\infty \quad \forall g \in C(\Omega) \tag{3.5.2}$$

with the aid of inner products. Indeed, the inequality (3.5.1) is none other than Bessel's inequality. By generalized Hölder's inequality,

$$\|g\|_2 \leq V^{1/2}\|g\|_\infty$$

for $g \in C(\Omega)$, thus

$$\|\mathcal{P}_n g\|_2 \leq V^{1/2}\|g\|_\infty$$

for $g \in C(\Omega)$.

3.5.1 Analysis with square-integrable kernels

When $K \in L^2(\Omega)$, we have the following theorem.

Theorem 3.5.1 *Let $K \in L^2(\Omega)$ and adopt the rest conditions of Theorem 3.4.1. Then*

$$\|\mathcal{S}_n F\|_2 \leq \|K\|_2 \|\mathcal{L}_n\|_\infty \|f\|_\infty, \tag{3.5.3}$$

where $\|\mathcal{L}_n\|_\infty$ *denotes the norm of* $\mathcal{L}_n$ *as an operator from* $C(\Omega)$ *to* $C(\Omega)$*, and*

$$\|\mathcal{S}_n F - F\|_2 \leq (1 + \|\mathcal{L}_n\|_\infty)\|K\|_2 E_{n'}(f) + 2E_n^{(2)}(K\chi^*), \tag{3.5.4}$$

where $E_n^{(2)}(g) := \inf_{\chi\in\mathbb{P}_n} \|g - \chi\|_2$ *and* $\chi^* \in \mathbb{P}_{n'}$ *is the best uniform approximation of* f *in* $\mathbb{P}_{n'}$*.*

Proof. Recall that $\mathcal{S}_n F = \mathcal{P}_n(K\mathcal{L}_n f)$. Thus, Bessel's inequality (3.5.1) suggests

$$\|\mathcal{S}_n f\|_2 \leq \|K\mathcal{L}_n f\|_2.$$

By the generalized Hölder's inequality,

$$\|\mathcal{S}_n F\|_2 \leq \|K\|_2\|\mathcal{L}_n f\|_\infty \leq \|K\|_2\|\mathcal{L}_n\|_\infty\|f\|_\infty.$$

For any $\chi \in \mathbb{P}_{n'}$, we have

$$\begin{aligned}\|\mathcal{S}_n F - F\|_2 &= \|\mathcal{S}_n(F - K\chi) - (F - K\chi) + (\mathcal{S}_n(K\chi) - K\chi)\|_2\\ &\leq \|\mathcal{S}_n(F - K\chi)\|_2 + \|F - K\chi\|_2 + \|\mathcal{S}_n(K\chi) - K\chi\|_2\\ &\leq \|K\|_2\|\mathcal{L}_n\|_\infty\|f - \chi\|_\infty + \|K\|_2\|f - \chi\|_\infty + \|\mathcal{S}_n(K\chi) - K\chi\|_2,\end{aligned}$$

where the last inequality is obtained by applying the stability (3.5.3) and generalized Hölder's inequality to

$$F - K\chi = K(f - \chi),$$

respectively. Letting $\chi = \chi^*$ gives

$$\|\mathcal{S}_n F - F\|_2 \leq (\|\mathcal{L}\|_\infty + 1)\,\|K\|_2 E_{n'}(f) + \|\mathcal{S}_n(K\chi^*) - K\chi^*\|_2. \tag{3.5.5}$$

Similar to the proof of Theorem 3.4.1, Lemma 3.2.4 implies

$$\|\mathcal{S}_n(K\chi^*) - K\chi^*\|_2 = \|\mathcal{P}_n(K\chi^*) - K\chi^*\|_2.$$

By the estimate (3.5.1), for any $\chi \in \mathbb{P}_n$, we have

$$\begin{aligned}\|\mathcal{P}_n(K\chi^*) - K\chi^*\|_2 &\leq \|\mathcal{P}_n(K\chi^* - \chi)\|_2 + \|K\chi^* - \chi\|_2\\ &\leq 2\|K\chi^* - \chi\|_2.\end{aligned}$$

Since this estimate applied to an arbitrary $\chi \in \mathbb{P}_n$, we have

$$\|\mathcal{P}_n(K\chi^*) - K\chi^*\|_2 \leq 2E_n^{(2)}(K\chi^*).$$

Together with (3.5.5), we have the error bound (3.5.4). □

Remark 3.5.2 *For $\mathcal{L}_n$ with quadrature exactness $2n$, some studies on $\|\mathcal{L}_n\|_\infty$ in various regions can be found in [39, 40, 106, 128, 202, 203, 224]. This operator norm awaits further investigation for $\mathcal{L}_n$ with quadrature exactness $n+n'$ $(0<n'<n)$. Nevertheless, the norm $\|\mathcal{L}_n\|_\infty$ cannot be uniformly bounded by any constant in general.*

Remark 3.5.3 *The fact that $\|\mathcal{L}_n\|_\infty$ is not uniformly bounded has spurred the development of* filtered hyperinterpolation *on the sphere and then on general regions [137, 149, 205]. The filtered hyperinterpolation operator, as an operator from $C(\Omega)\to C(\Omega)$, has a uniformly bounded norm. Thus, a possible future work may be the combination of efficient and filtered hyperinterpolation so that a better result of the approximation of $F=Kf$ with $K\in L^2(\Omega)$ can be expected.*

3.5.2 Analysis with continuous kernels

If K is continuous, then we have the following analysis.

Theorem 3.5.4 *Let $K\in C(\Omega)$ and adopt the rest conditions of Theorem 3.4.1. Then*

$$\|\mathcal{S}_nF\|_2\le\frac{V^{1/2}}{\sqrt{1-\eta}}\|K\|_\infty\|f\|_\infty, \tag{3.5.6}$$

where $\|\mathcal{L}_n\|_\infty$ denotes the norm of $\mathcal{L}_n$ as an operator from $C(\Omega)$ to $C(\Omega)$, and

$$\|\mathcal{S}_nF-F\|_2\le\left(\frac{V^{1/2}}{\sqrt{1-\eta}}\|K\|_\infty+\|K\|_2\right)E_{n'}(f)+2V^{1/2}E_n(K\chi^*), \tag{3.5.7}$$

where $\chi^*\in\mathbb{P}_{n'}$ is the best uniform approximation of f in $\mathbb{P}_{n'}$.

Proof. In the proof of Theorem 3.5.1, we have obtained

$$\|\mathcal{S}_nF\|_2\le\|K\mathcal{L}_nf\|_2.$$

Thus, for $K\in C(\Omega)$, by generalized Hölder's inequality, we have

$$\|\mathcal{S}_nF\|_2\le\|K\|_\infty\|\mathcal{L}_nf\|_2,$$

and by the stability result (3.2.4) of $\mathcal{L}_n$, we have the stability (3.5.6) of S_n.

Similar to the case of $K \in L^2(\Omega)$, for any $\chi \in \mathbb{P}_{n'}$, we have

$$\begin{aligned}\|\mathcal{S}_n F - F\|_2 &\leq \|\mathcal{S}_n(F - K\chi)\|_2 + \|K(f-\chi)\|_2 + \|\mathcal{S}_n(K\chi) - K\chi\|_2 \\ &\leq \frac{V^{1/2}}{\sqrt{1-\eta}}\|K\|_\infty \|f-\chi\|_\infty + \|K\|_2\|f-\chi\|_\infty + \|\mathcal{S}_n(K\chi) - K\chi\|_2.\end{aligned}$$

Letting $\chi = \chi^*$ leads to

$$\|\mathcal{S}_n F - F\|_2 = \left(\frac{V^{1/2}}{\sqrt{1-\eta}}\|K\|_\infty + \|K\|_2\right) E_{n'}(f) + \|\mathcal{S}_n(K\chi^*) - K\chi^*\|_2. \tag{3.5.8}$$

By Lemma 3.2.4 and the estimate (3.5.2), for any $\chi \in \mathbb{P}_n$, we have

$$\begin{aligned}\|\mathcal{S}_n(K\chi^*) - K\chi^*\|_2 = \|\mathcal{P}_n(K\chi^*) - K\chi^*\|_2 &\leq \|\mathcal{P}_n(K\chi^* - \chi)\|_2 + \|K\chi^* - \chi\|_2 \\ &\leq V^{1/2}\|K\chi^* - \chi\|_\infty + V^{1/2}\|K\chi^* - \chi\|_\infty \\ &= 2V^{1/2}\|K\chi^* - \chi\|_\infty.\end{aligned}$$

Thus $\|\mathcal{S}_n(K\chi^*) - K\chi^*\|_2 \leq 2V^{1/2}E_n(K\chi^*)$. Together with (3.5.8), we have the estimate (3.5.7). □

3.5.3 The potential inefficiency of classical hyperinterpolation

The classical hyperinterpolation (3.1.4) is defined to approximate continuous functions. The approximation of $F = Kf \in C(\Omega)$ by efficient hyperinterpolation is described by Theorem 3.5.4. Thus, if we let $K = 1$, then both the stability result (3.5.6) and the error bound (3.5.7) of efficient hyperinterpolation reduce to (3.2.4) and (3.2.5) of the classical hyperinterpolation, respectively, derived in [12]. Furthermore, if the quadrature rule (1.2.3) has exactness degree $2n$, that is, $\eta = 0$, then they reduce to the original results (3.2.1) and (3.2.2) derived by Sloan in [196].

But what if $K \neq 1$ and K is awkward enough to be approximated? In this case,

$$\|\mathcal{S}_n F - F\|_2 \lesssim E_{n'}(f) + E_n(K\chi^*),$$

where χ^* is the best uniform approximation of f in $\mathbb{P}_{n'}$. However, for the classical hyperinterpolation there holds

$$\|\mathcal{L}_n F - F\|_2 \lesssim E_{n'}(Kf).$$

Thus, if f is smooth enough so that $E_n(K\chi^*)$ dominates the bound of $\|\mathcal{S}_n F - F\|_2$, and if $n' < n$ and K is awkward enough so that $E_{n'}(Kf)$ is considerably greater

than $E_n(K\chi^*)$, efficient hyperinterpolation shall give a better approximation than the classical one in the sense of estimated error bounds.

On the other hand, it is inappropriate to claim that efficient hyperinterpolation is always better than the classical hyperinterpolation in the approximation of $F = Kf$. If the singularity of K is relatively weak (for a singular K), or if K oscillates slowly (for an oscillatory K), then the classical hyperinterpolation may generate a comparable or even better approximation of F than efficient hyperinterpolation.

3.6 Examples and numerical experiments

We now numerically investigate efficient hyperinterpolation (3.1.6) on two specific regions, the interval $[-1,1] \subset \mathbb{R}$ and the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$. On each region, we test oscillatory and singular terms K, respectively. A key issue is how to evaluate the modified moments (3.3.3). We shall discuss the computational issues of the moments respectively on each region and for each K. All numerical results are carried out by using MATLAB R2022a on a laptop (16 GB RAM, Intel CoreTM i7-9750H Processor) with macOS Monterey 12.3.

3.6.1 On the interval

Let $\Omega = [-1,1]$. In this case, $d_n = n+1$. There is merit in adopting orthogonal polynomials as the basis [97, 210] for the approximation of functions defined on $[-1,1]$. In our experiments, we let $\{p_\ell\}_{\ell=1}^{d_n}$ be normalized Legendre polynomials $\{\tilde{P}_\ell\}_{\ell=0}^{n}$, and let $\{q_r\}_{r=1}^{d_{2n}}$ be Chebyshev polynomials $\{T_r\}_{r=0}^{2n}$. Thus for any $\tilde{P}_{\ell'}\tilde{P}_\ell$ with $0 \le \ell', \ell \le n$, it can be expressed as

$$\tilde{P}_{\ell'}\tilde{P}_\ell = \sum_{r=0}^{2n} c_r T_r,$$

where the coefficients are given for $r \ge 1$ by

$$c_r = \frac{2}{\pi}\int_{-1}^{1} \frac{\tilde{P}_{\ell'}(x)\tilde{P}_\ell(x)T_r(x)}{\sqrt{1-x^2}}\mathrm{d}x, \quad r = 1,\dots,2n,$$

and for $r = 0$ by the same formula with the factor $\pi/2$ changed to $1/\pi$ for $r = 0$ [221]. In the expression of c_r, $(1-x^2)^{-0.5}$ is the weight function associated to the Chebyshev polynomials, and $\langle \tilde{P}_{\ell'}\tilde{P}_\ell, T_r\rangle$ is divided by the factor $\langle T_r, T_r\rangle$ since $\{T_r\}_{r=0}^{2n}$ are not orthonormal. In our experiments, these coefficients $\{c_r\}$ are obtained by the `chebcoeffs` command included in Chebfun [75]. For the quadrature rule (1.2.3),

we use the Gauss–Legendre quadrature. It is well-known that the m-point Gauss–Legendre quadrature has exactness degree $2m-1$.

Oscillatory functions. We first test

$$K(x) = e^{i\kappa x}$$

with $\kappa > 0$, which is an oscillatory term regularly appearing in applications. For the evaluation of

$$\beta_r = \int_{-1}^{1} e^{i\kappa x} T_r(x) \mathrm{d}x, \quad r = 0, 1, \ldots, 2n, \tag{3.6.1}$$

we invoke the stable algorithm proposed in [73] for implementing the Filon–Clenshaw–Curtis rule [72, 73]. One may also investigates other oscillatory terms in this spirit. For example, one could study Bessel functions with the aid of Clenshaw–Curtis–Filon method [237]. For the function $f \in C[-1,1]$, we let

$$f(x) = (1.2 - x^2)^{-1}.$$

For $\kappa = 100$, we let $n = 120$ and $m = 70$; that is, the theoretical error of classical hyperinterpolation is controlled by $E_{19}(e^{i100x} f)$, while that of efficient hyperinterpolation is controlled by $E_{19}(f)$ and $E_{120}(e^{i100x}\chi^*)$, where $\chi^* \in \mathbb{P}_{19}$ is the best uniform approximation of f in $\mathbb{P}_{19}$. The approximation results are displayed in Figure 3.1, in which we see that efficient hyperinterpolation generates a good approximation, but the classical one fails to do so. Moreover, for $\kappa = 160$, we let $n = 180$ and $m = 100$; that is, the theoretical error of classical hyperinterpolation is controlled by $E_{19}(e^{i160x} f)$, while that of efficient hyperinterpolation is controlled by $E_{19}(f)$ and $E_{180}(e^{i160x}\chi^*)$, where $\chi^* \in \mathbb{P}_{19}$ is the best uniform approximation of f in $\mathbb{P}_{19}$. The approximation results are displayed in Figure 3.2, which convey the same message as the case of $\kappa = 100$.

We continue with a more detailed investigation on the approximation of

$$F(x) = e^{i\kappa x}(1.2 - x^2)^{-1}$$

with $\kappa = 100$ and 160. For $\kappa = 100$, we test $n = 100$, 120, and 150; for $\kappa = 160$, we consider $n = 160$, 180, and 210. For each (κ, n), we test several numbers m of quadrature points. The L^2 errors of each hyperinterpolant are listed in Table 3.1; these errors are evaluated by the command `norm` in Chebfun, with the function F and its approximants treated as Chebfun objects. In each setting, the error of efficient hyperinterpolation is always less than that of classical hyperinterpolation. Apart from this, Table 3.1 conveys some other interesting messages.

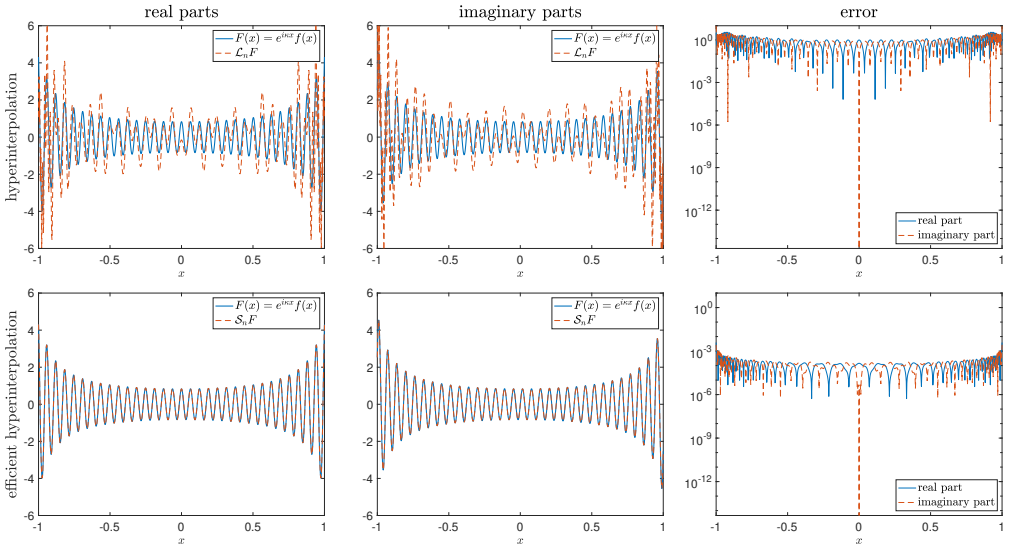


Figure 3.1: Approximation of $F(x) = e^{i\kappa x}(1.2 - x^2)^{-1}$ by $\mathcal{L}_n$ and $\mathcal{S}_n$ with $(\kappa, n, m) = (100, 120, 70)$.

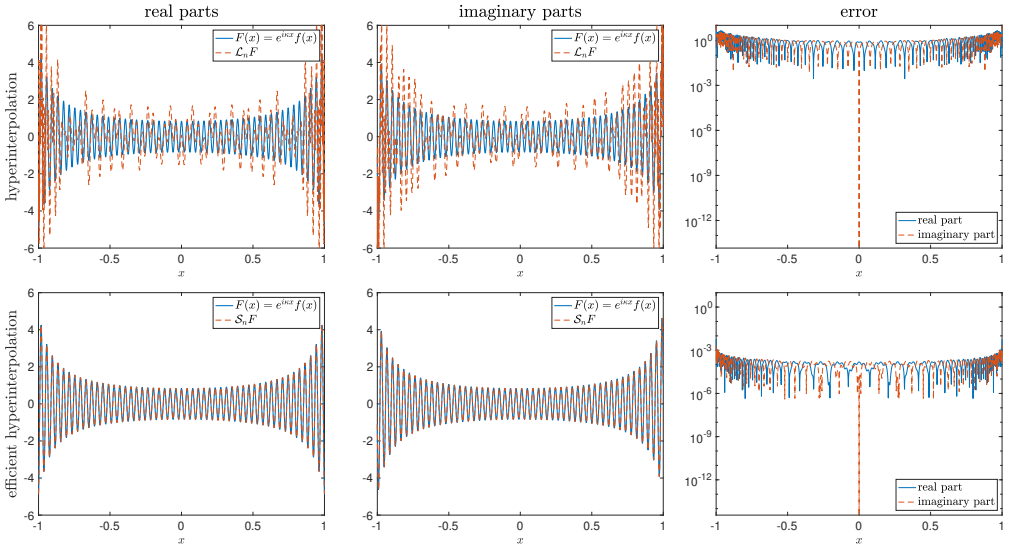


Figure 3.2: Approximation of $F(x) = e^{i\kappa x}(1.2 - x^2)^{-1}$ by $\mathcal{L}_n$ and $\mathcal{S}_n$ with $(\kappa, n, m) = (160, 180, 100)$.

Table 3.1: Errors of hyperinterpolation and efficient hyperinterpolation with different (n, m) for the approximation of $F(x) = K(x)f(x)$ with $K(x) = e^{i\kappa x}$ and $f(x) = (1.2 - x^2)^{-1}$, with $\kappa = 100$ and 160.

	$n = 100$ $K(x) = e^{i100x}$		$n = 120$ $K(x) = e^{i100x}$		$n = 150$ $K(x) = e^{i100x}$	
m	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$
60	2.1437	0.2064	2.3310	2.1556	2.7565	2.6291
70	1.7667	0.2064	2.1339	3.7060e-04	2.3565	2.3635
80	1.3929	0.2064	1.7547	8.2733e-06	2.2603	0.02830
100	0.3428	0.2064	1.0354	8.2730e-06	1.5477	8.3481e-10
120	0.2064	0.2064	1.8091e-05	8.2730e-06	0.7998	1.4644e-13
150	0.2064	0.2064	8.2730e-06	8.2730e-06	9.6996e-14	9.2094e-14
180	0.2064	0.2064	8.2730e-06	8.2730e-06	7.6783e-14	6.9940e-14
	$n = 160$ $K(x) = e^{i160x}$		$n = 180$ $K(x) = e^{i160x}$		$n = 210$ $K(x) = e^{i160x}$	
m	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$
70	2.6149	2.3755	2.8822	2.5556	3.2681	2.8106
100	2.0502	0.2014	2.2357	3.7455e-04	2.3368	2.3372
120	1.5749	0.2014	1.7994	5.8491e-05	2.1408	4.8505e-06
150	0.8957	0.2014	1.1128	5.8491e-05	1.4421	1.6188e-13
180	0.2014	0.2014	1.1543e-04	5.8491e-05	0.7253	1.4140e-13
210	0.2014	0.2014	5.8491e-05	5.8491e-05	2.9417e-13	2.0787e-13
240	0.2014	0.2014	5.8491e-05	5.8491e-05	1.2553e-13	1.2040e-13

Let n be fixed. When the exactness degree of the quadrature rule is less than $2n$, i.e., $2m-1<2n$, the limited number of quadrature points slow the convergence of classical hyperinterpolation, as its error bound is controlled by $E_{n'}(Kf)=E_{2m-1-n}(Kf)$. Meanwhile, efficient hyperinterpolation may work well because its error bound is controlled by $E_{n'}(f)=E_{2m-1-n}(f)$ and $E_n(K\chi^*)$. When $2m-1\geq 2n$, by our analysis, the accuracy of both schemes only depends on n.

On the other hand, let m be fixed. When $2m-1<2n$, increasing n may not help in improving the accuracy of classical hyperinterpolation; on the contrary, it may slow its convergence, as $E_{2m-1-n}(Kf)$ is enlarged as n increases. However, if $E_n(K\chi^*)$ dominates the error bound of efficient hyperinterpolation, then increasing n shall improve the accuracy of efficient hyperinterpolation.

Singular functions. We then test three singular terms K, which are

$$K(x)=\begin{cases}(1+x)^{-1/3},\\ |x-1|^{-0.2},\\ (1-x^2)^{-0.5}.\end{cases}$$

For the first two cases, we compute

$$\beta_r=\int_{-1}^{1}K(x)T_r(x)\mathrm{d}x,\quad r=0,1,\dots,2n,$$

by the built-in command `quadgk` in MATLAB, which is a stable procedure developed in [189]. For the third case, as $(1-x^2)^{-0.5}$ is the weight function associated to the Chebyshev polynomials, we have $\beta_0=\pi$ and $\beta_r=0$ for all $r\geq 1$. For the continuous function $f\in C[-1,1]$, we let

$$f(x)=e^{-x^2}.$$

For each K, we report the L^1 errors of classical and efficient hyperinterpolation with $n=6,9,12,\dots,120$, and $m=\lceil 1.1n/2\rceil$, $\lceil 1.2n/2\rceil$, and $\lceil 1.5n/2\rceil$. These errors are evaluated numerically by the MATLAB built-in command `quadgk` ,and they are plotted in Figure 3.3. We can summarize from these errors that when the available data (the number of quadrature points) is limited, then the error of efficient hyperinterpolation is generally less than that of classical hyperinterpolation. It is also interesting to see that classical hyperinterpolation may perform better than efficient hyperinterpolation as the amount of quadrature points increases. For example, see the subplots on the bottom left and bottom right of Figure 3.3. An interesting related fact is that the functions $K(x)=(1+x)^{-1/3}$ and $K(x)=(1-x^2)^{-0.5}$ is smoother than $K(x)=|x-1|^{-0.2}$ in the sense of differentiability. Hence, it is interesting

to identify the critical number of quadrature points that the outperformance of the classical and efficient hyperinterpolation switches as future work. In particular, this critical number may be related to the smoothness of F.

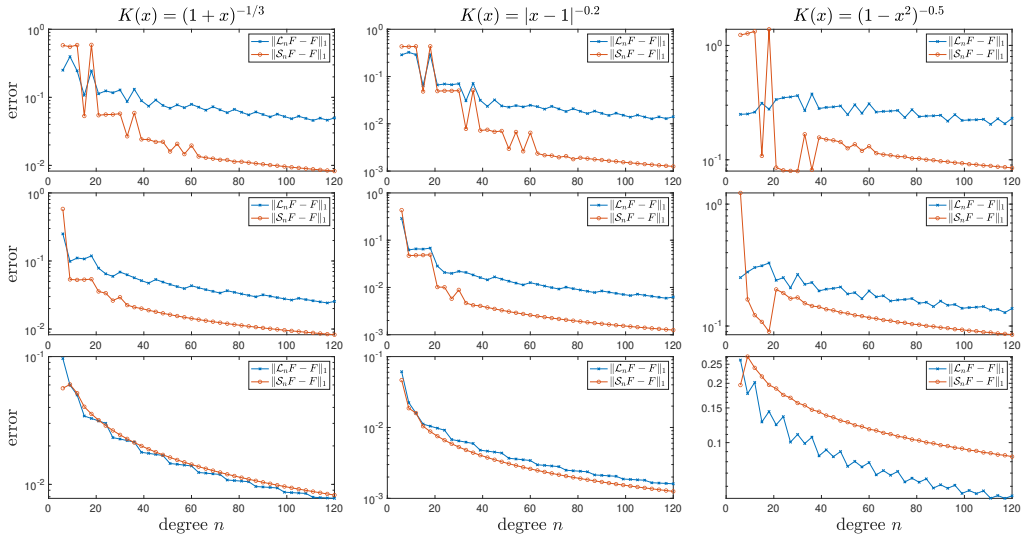


Figure 3.3: Errors of hyperinterpolation and efficient hyperinterpolation with different (n, m) for the approximation of $F(x) = K(x)f(x)$ with three singular K's and $f(x) = e^{-x^2}$. From top row to bottom row: $m = \lceil 1.1n/2 \rceil$, $\lceil 1.2n/2 \rceil$, and $\lceil 1.5n/2 \rceil$, respectively.

3.6.2 On the sphere

Let $\Omega = \mathbb{S}^2 \subset \mathbb{R}^3$ with $\mathrm{d}\omega(x) = \omega(x)\mathrm{d}x$, where $\omega(x)$ is an area measure on $\mathbb{S}^2$. Thus

$$V = \int_{\mathbb{S}^2} \mathrm{d}\omega = 4\pi$$

denotes the surface area of $\mathbb{S}^2$. In this example, $\mathbb{P}_n$ can be regarded as the space of spherical polynomials of degree at most n. Let the basis $\{p_\ell\}_{\ell=1}^{d_n}$ be a set of orthonormal real spherical harmonics $\{Y_{\ell,k} : \ell = 0, 1, \ldots, n, \text{ and } k = -\ell, -\ell+1, \ldots, \ell-1, \ell\}$, and the dimension of $\mathbb{P}_n$ is $d_n = (n+1)^2$. Let $\{q_r\}_{r=1}^{d_{2n}}$ also be the set of orthonormal real spherical harmonics $\{Y_{\ell,k} : \ell = 0, 1, \ldots, 2n, \text{ and } k = -\ell, -\ell+1, \ldots, \ell-1, \ell\}$.

For the quadrature rule (1.2.3), we use the rule based on spherical t-designs, which can be implemented easily and efficiently. A point set $\{x_1, x_2, \ldots, x_m\} \subset \mathbb{S}^2$ is said to be a *spherical t-design* [69] if it satisfies

$$\frac{1}{m}\sum_{j=1}^{m} v(x_j) = \frac{1}{4\pi}\int_{\mathbb{S}^2} v\mathrm{d}\omega \quad \forall v \in \mathbb{P}_t. \tag{3.6.2}$$

In other words, it is a set of points on the sphere such that an equal-weight quadrature rule in these points integrates all (spherical) polynomials up to degree t exactly. Spherical t-designs require at least $(t+1)^2$ quadrature points to achieve the exactness degree t. For generating spherical t-designs, we make use of the well-conditioned spherical t-designs [7] with $m = (t+1)^2$.

For any $Y_{\ell',k'}Y_{\ell,k}$ with $0 \le \ell', \ell \le n$, $-\ell' \le k' \le \ell'$, and $-\ell \le k \le \ell$, it can be expressed as

$$Y_{\ell',k'}Y_{\ell,k} = \sum_{\ell''=0}^{2n} \sum_{k''=-\ell''}^{\ell''} c_{\ell''k''} Y_{\ell'',k''},$$

where the coefficients

$$c_{\ell''k''} = \int_{\mathbb{S}^2} (Y_{\ell',k'}Y_{\ell,k}) Y_{\ell'',k''} \mathrm{d}\omega, \quad \ell'' = 0, \ldots, 2n, \quad k'' = -\ell'', \ldots, \ell''$$

are evaluated by a quadrature rule using spherical $(\ell + \ell' + \ell'')$-designs.

We may use boldface letters to denote a point on $\mathbb{S}^2$, say $\boldsymbol{x} = [x, y, z]^{\mathrm{T}}$, in order to avoid any potential ambiguity. The Euclidean distance between two points $\boldsymbol{\xi}$ and $\boldsymbol{x}$ on the sphere $\mathbb{S}^2$ is defined as

$$|\boldsymbol{\xi} - \boldsymbol{x}| := \sqrt{2(1 - \boldsymbol{\xi} \cdot \boldsymbol{x})},$$

where $\cdot$ denotes the inner product in $\mathbb{R}^3$.

Oscillatory functions. The spherical harmonics themselves are highly oscillatory when their degrees become large. Thus we let $K = Y_{\bar{\ell},\bar{k}}$ for some $\bar{\ell}, \bar{k} \in \mathbb{N}$. In this case, the modified moments can be evaluated by

$$\beta_r := \beta_{\ell''k''} = \int_{\mathbb{S}^2} Y_{\bar{\ell},\bar{k}} Y_{\ell'',k''} \mathrm{d}\omega = \delta_{\bar{\ell},\ell''} \delta_{\bar{k},k''}.$$

For the continuous function $f \in C(\mathbb{S}^2)$, we let

$$f(\boldsymbol{x}) = f(x, y, z) = \cos(\cosh(xz) - 2y).$$

We investigate two kinds of oscillatory terms, $(\bar{\ell}, \bar{k}) = (12, 8)$ and $(32, -24)$. For $K = Y_{12,8}$, we let $n = 20$ and $m = 625$, that is, $t = 24$, the theoretical error of classical hyperinterpolation is controlled by $E_4(Y_{12,8} f)$, while that of efficient hyperinterpolation is controlled by $E_4(f)$ and $E_{20}(Y_{12,8}\chi^*)$, where $\chi^* \in \mathbb{P}_4$ is the best uniform approximation of f in $\mathbb{P}_4$. The approximation results are displayed in Figure 3.4, in which we see that efficient hyperinterpolation generates a good approximation, but the classical one does not. For $K = Y_{32,-24}$, we let $n = 40$ and

$m = 2209$, that is, $t = 46$, the theoretical error of classical hyperinterpolation is controlled by $E_6(Y_{32,-24}f)$, while that of efficient hyperinterpolation is controlled by $E_6(f)$ and $E_{40}(Y_{32,-24}\chi^*)$, where $\chi^* \in \mathbb{P}_6$ is the best uniform approximation of f in $\mathbb{P}_6$. The approximation results are displayed in Figure 3.5, which convey the same message as the case of $K = Y_{12,8}$.

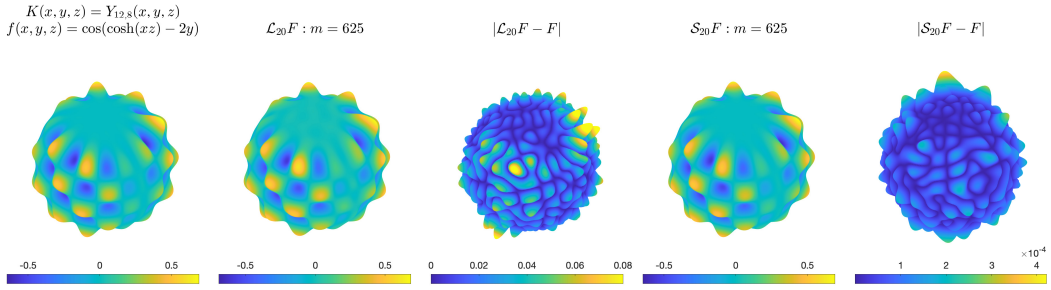


Figure 3.4: Approximation of $F = Y_{12,8}f$ with $f(x,y,z) = \cos(\cosh(xz) - 2y)$ by hyperinterpolation $\mathcal{L}_n$ and efficient hyperinterpolation $\mathcal{S}_n$.

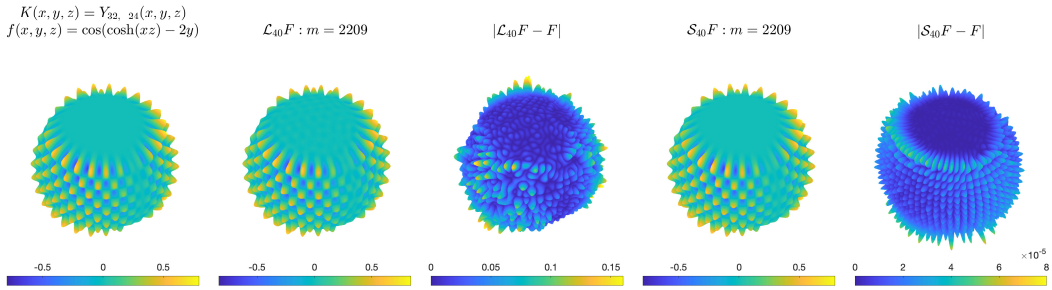


Figure 3.5: Approximation of $F = Y_{32,-24}f$ with $f(x,y,z) = \cos(\cosh(xz) - 2y)$ by hyperinterpolation $\mathcal{L}_n$ and efficient hyperinterpolation $\mathcal{S}_n$.

Similar to Table 3.1, we list the L^2 errors of the classical and efficient hyperinterpolation in different settings in Table 3.2. These errors are evaluated by the command `norm` in Spherefun [216], a part of Chebfun for computing with functions defined on the surface of the unit sphere, and the functions and their approximants are regarded as Spherefun objects. We see that the error of efficient hyperinterpolation is always less than (or eventually equal to) that of the classical hyperinterpolation.

Singular functions. For singular functions, we test three different singular terms. Their forms and the evaluation of modified moments

$$\beta_r := \beta_{\ell'' k''} = \int_{\mathbb{S}^2} K(\boldsymbol{x}) Y_{\ell'',k''}(\boldsymbol{x}) \mathrm{d}\omega(\boldsymbol{x})$$

for $\ell'' = 0, \ldots, 2n$ and $k'' = -\ell'', \ldots, \ell''$ are elaborated as follows.

Table 3.2: Errors of hyperinterpolation and efficient hyperinterpolation with different (n, m) for the approximation of $F(x, y, z) = K(x, y, z)f(x, y, z)$ with two K's and $f(x, y, z) = \cos(\cosh(xz) - 2y)$.

	$n = 16$ $K(\boldsymbol{x}) = Y_{12,8}(\boldsymbol{x})$		$n = 18$ $K(\boldsymbol{x}) = Y_{12,8}(\boldsymbol{x})$		$n = 20$ $K(\boldsymbol{x}) = Y_{12,8}(\boldsymbol{x})$	
m	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$
484	0.1427	0.0116	0.1359	0.0092	0.1233	0.0082
529	0.1271	0.0097	0.1160	0.0031	0.1181	0.0044
576	0.1090	0.0086	0.0993	9.1533e-04	0.0932	8.0575e-04
625	0.0910	0.0086	0.0973	7.0753e-04	0.0861	2.9376e-04
841	0.0530	0.0086	0.0425	5.9738e-04	0.0439	5.9812e-05
1089	0.0285	0.0086	0.0189	5.9737e-04	0.0112	5.9767e-05
1369	0.0098	0.0086	6.4698e-04	5.9737e-04	1.6743e-04	5.9767e-05
1681	0.0086	0.0086	5.9749e-04	5.9737e-04	5.9776e-05	5.9767e-05
2025	0.0086	0.0086	5.9737e-04	5.9737e-04	5.9767e-05	5.9767e-05
	$n = 36$ $K(\boldsymbol{x}) = Y_{32,-24}(\boldsymbol{x})$		$n = 38$ $K(\boldsymbol{x}) = Y_{32,-24}(\boldsymbol{x})$		$n = 40$ $K(\boldsymbol{x}) = Y_{32,-24}(\boldsymbol{x})$	
m	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$	$\|\mathcal{L}_n F - F\|_2$	$\|\mathcal{S}_n F - F\|_2$
1849	0.2092	0.0086	0.1868	0.0031	0.1674	0.0028
2025	0.1622	0.0083	0.1469	6.3438e-04	0.1433	2.7101e-04
2209	0.1327	0.0083	0.1295	5.9311e-04	0.1252	4.7438e-05
2401	0.1180	0.0083	0.1160	5.9286e-04	0.1167	4.4752e-05
3249	0.0736	0.0083	0.0689	5.9286e-04	0.0673	4.4728e-05
4225	0.0432	0.0083	0.0391	5.9286e-04	0.0350	4.4728e-05
5329	0.0174	0.0083	0.0091	5.9286e-04	0.0053	4.4728e-05
6561	0.0083	0.0083	5.9286e-04	5.9286e-04	4.4731e-05	4.4728e-05
7921	0.0083	0.0083	5.9282e-04	5.9286e-04	4.4728e-05	4.4728e-05

- Let
$$K(\boldsymbol{x}) = |\boldsymbol{\xi} - \boldsymbol{x}|^\nu,$$
where $\nu > -1$, and $\boldsymbol{\xi}$ is an algebraic type singularity if $\nu < 0$. Then
$$\beta_{\ell'' k''} = 2^{\nu+2}\pi \left(-\frac{\nu}{2}\right)_{\ell''} \frac{\Gamma(\frac{\nu+2}{2})}{\Gamma(\ell'' + \nu/2 + 2)} Y_{\ell'',k''}(\boldsymbol{\xi}),$$
where $\Gamma(\cdot)$ is the Gamma function, and $(\cdot)_n = \Gamma(\cdot+n)/\Gamma(\cdot)$ is the Pochhammer symbol [15].

- Let
$$K(\boldsymbol{x}) = \log|\boldsymbol{\xi} - \boldsymbol{x}|,$$
where $\boldsymbol{\xi}$ is a logarithmic type singularity. Then
$$\beta_{\ell'' k''} = \frac{|\mathbb{S}^1|}{2} \left(\int_{-1}^{1} \log(2(1-t)) P_{\ell''}(t) \mathrm{d}t \right) Y_{\ell'',k''}(\boldsymbol{\xi}), \tag{3.6.3}$$
where $|\mathbb{S}^1| = 2\pi$ is the length of the unit circle $\mathbb{S}^1$, and P_ℓ denote the Legendre polynomials of degree ℓ (without normalization).

- Let
$$K(\boldsymbol{x}) = |\boldsymbol{\xi} - \boldsymbol{x}|^{\nu_1} |\boldsymbol{\xi} + \boldsymbol{x}|^{\nu_2},$$
where $\nu_1, \nu_2 > -1$, and $\boldsymbol{\xi}$ and $-\boldsymbol{\xi}$ are two algebraic type singularities if $\nu_1, \nu_2 < 0$. Then
$$\begin{aligned}\beta_{\ell'' k''} =& (-1)^{\ell''} 2^{(\nu_1+\nu_2)/2} |\mathbb{S}^1| R_{\ell,3} \\ & \left(\int_{-1}^{1} (1-t)^{\nu_1/2} (1+t)^{\nu_2/2} \left[\left(\frac{\mathrm{d}}{\mathrm{d}t} \right)^{\ell''} (1-t^2)^{\ell''} \right] \mathrm{d}t \right) Y_{\ell'',k''}(\boldsymbol{\xi}),\end{aligned}$$
where
$$R_{n,s} = \frac{\Gamma(\frac{s-1}{2})}{2^n \Gamma(n + \frac{s-1}{2})}.$$

There results can be found in [15, Chapter 3]. In particular, the modified moments of the third term can be evaluated by
$$\beta_{\ell'' k''} = 2^{(\nu_1+\nu_2)/2} |\mathbb{S}^1| \left(\int_{-1}^{1} (1-t)^{\nu_1/2} (1+t)^{\nu_2/2} P_{\ell''}(t) \mathrm{d}t \right) Y_{\ell'',k''}(\boldsymbol{\xi}), \tag{3.6.4}$$
with the aid the Rodrigues' formula
$$P_n(x) = \frac{1}{2^n n!} \left(\frac{\mathrm{d}}{\mathrm{d}x} \right)^n \left[(x^2-1)^n \right] = (-1)^n \frac{1}{2^n n!} \left(\frac{\mathrm{d}}{\mathrm{d}x} \right)^n (1-x^2)^n$$

for Legendre polynomials[1]. For the continuous function $f \in C(\mathbb{S}^2)$, we consider

$$\begin{aligned} f_1(\boldsymbol{x}) = f_1(x,y,z) =& 0.75\exp(-(9x-2)^2/4 - (9y-2)^2/4 - (9z-2)^2/4) \\ &+ 0.75\exp(-(9x+1)^2/49 - (9y+1)/10 - (9z+1)/10) \\ &+ 0.5\exp(-(9x-7)^2/4 - (9y-3)^2/4 - (9z-5)^2/4) \\ &- 0.2\exp(-(9x-4)^2 - (9y-7)^2 - (9z-5)^2), \end{aligned} \tag{3.6.5}$$

which is analytic on the sphere, and

$$f_2(\boldsymbol{x}) = f_2(x,y,z) = \exp(x+y+z). \tag{3.6.6}$$

The integrals in modified moments (3.6.3) and (3.6.4) are evaluated by the MATLAB built-in command `quadgk`. For each K, we report the L^1 errors of classical and efficient hyperinterpolation with $n = 2,3,4,\ldots,40$, and $m = (\lceil 1.1n\rceil + 1)^2$, $(\lceil 1.2n\rceil + 1)^2$, and $(\lceil 1.5n\rceil + 1)^2$. The singularity $\boldsymbol{\xi}$ in the definitions of $K(\boldsymbol{x})$ is set as $\boldsymbol{\xi} = [\sqrt{2}/2, \sqrt{2}/2, 0]^{\mathrm{T}}$. These errors of approximating f_1 and f_2 are numerically evaluated by a 50,000-point equal-weight quadrature rule, and they are plotted in Figures 3.6 and 3.7, respectively. Unlike the experiments on the singular functions with endpoint singularities on $[-1,1]$, all singularities on the sphere are interior. Thus, the numerical integration of spherical singular functions becomes extremely unstable: the actual performance of numerical integration depends on the point distribution around the singularities. This technical issue is also reflected in the approximation of singular functions by numerically integrating the L^2 projection coefficients, i.e., the approximation by classical hyperinterpolation. We see from Figures 3.6 and 3.7 that it seems impossible to predict the actual accuracy of classical hyperinterpolation in the approximation of $F(x,y,z) = K(x,y,z)f(x,y,z)$ with f defined as (3.6.5), with three kinds of singular K listed above. Indeed, the stability and error bounds of classical hyperinterpolation in [12, 196] are only valid for the approximation of continuous functions. On the other hand, we see that the actual accuracy of efficient hyperinterpolation is stable and predictable: the point distribution around singularities does not affect the performance of efficient hyperinterpolation, and the approximation error decays as n increases.

[1]It may be unstable to evaluate the integral $\int_{-1}^{1}(1-t)^{\nu_1/2}(1+t)^{\nu_2/2}(\frac{\mathrm{d}}{\mathrm{d}t})^n(1-t^2)^n\mathrm{d}t$ by taking the n-th derivative and then evaluating the resulting integral, as the factor accumulated as $n!$ after differentiation may be huge. Thus the error of representing numbers by double-precision floating-point numbers, according to IEEE Standard 754, may be inaccurate.

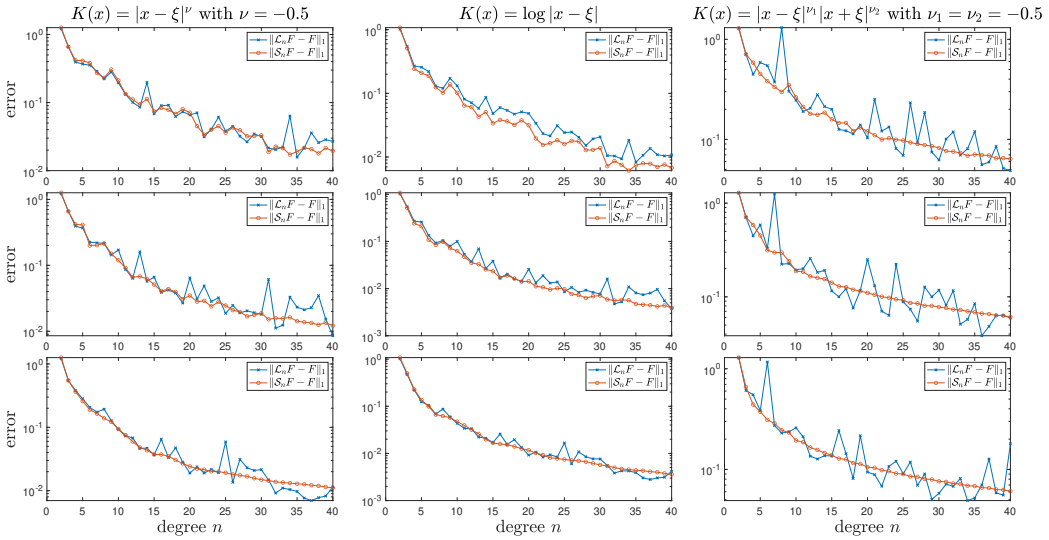


Figure 3.6: Errors of hyperinterpolation and efficient hyperinterpolation with different (n, m) for the approximation of $F(x, y, z) = K(x, y, z)f_1(x, y, z)$ with three singular K's and $f_1(x, y, z)$ defined as (3.6.5). The singularity $\boldsymbol{\xi}$ in the definitions of $K(\boldsymbol{x})$ is set as $\boldsymbol{\xi} = [\sqrt{2}/2, \sqrt{2}/2, 0]^{\mathrm{T}}$. From top row to bottom row: $m = (\lceil 1.1n \rceil+1)^2$, $(\lceil 1.2n \rceil + 1)^2$, and $(\lceil 1.5n \rceil + 1)^2$, respectively.

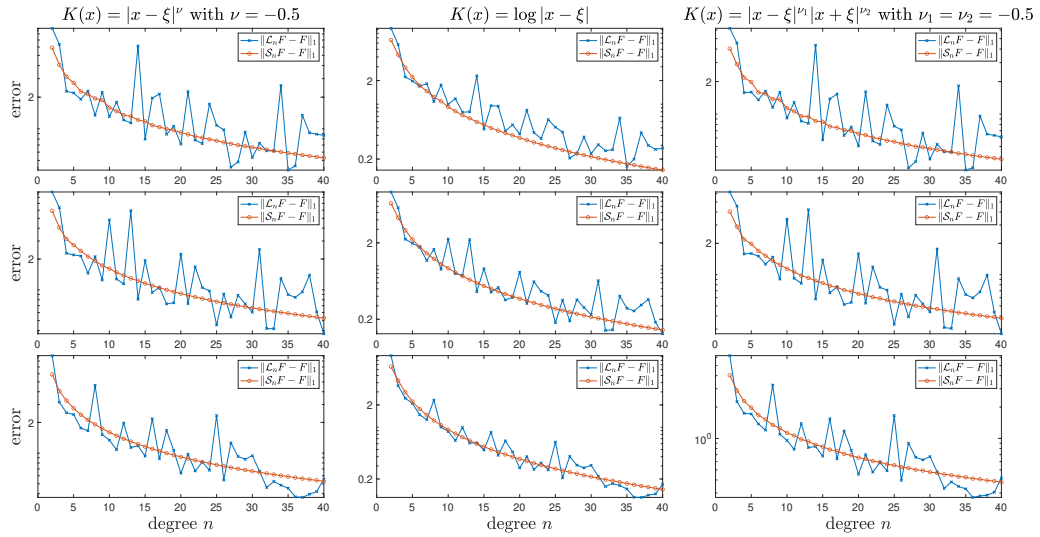


Figure 3.7: Errors of hyperinterpolation and efficient hyperinterpolation with different (n, m) for the approximation of $F(x, y, z) = K(x, y, z)f_2(x, y, z)$ with three singular K's and $f_2(x, y, z)$ defined as (3.6.6). The singularity $\boldsymbol{\xi}$ in the definitions of $K(\boldsymbol{x})$ is set as $\boldsymbol{\xi} = [\sqrt{2}/2, \sqrt{2}/2, 0]^{\mathrm{T}}$. From top row to bottom row: $m = (\lceil 1.1n \rceil+1)^2$, $(\lceil 1.2n \rceil + 1)^2$, and $(\lceil 1.5n \rceil + 1)^2$, respectively.

Chapter 4

Bypassing the quadrature exactness of hyperinterpolation

Let us continue the discussion on relaxing the quadrature exactness assumption (1.2.4) in Chapter 2. Recall that in Sloan's original manuscript [196], hyperinterpolation (1.2.5) of degree n is a discrete approximation of the L^2-orthogonal projection (1.2.2) of degree n with its Fourier coefficients evaluated by a positive-weight quadrature rule (1.2.3) that exactly integrates all spherical polynomials of degree at most $2n$. This chapter aims to bypass this quadrature exactness assumption (1.2.4) by replacing it with the Marcinkiewicz–Zygmund property (2.2.5) proposed in Chapter 2. Consequently, hyperinterpolation can be constructed by a positive-weight quadrature rule (not necessarily with quadrature exactness). This scheme is referred to as *unfettered hyperinterpolation*. This chapter provides a reasonable error estimate for unfettered hyperinterpolation. The error estimate generally consists of two terms: a term representing the error estimate of the original hyperinterpolation of full quadrature exactness and another introduced as compensation for the loss of exactness degrees. A guide to controlling the newly introduced term in practice is provided. In particular, if the quadrature points form a quasi-Monte Carlo (QMC) design, then there is a refined error estimate. Numerical experiments verify the error estimates and the practical guide.

In particular, we focus on the approximation of spherical functions in this chapter, utilizing some tools from spherical harmonic analysis.

4.1 Introduction

Let $\mathbb{S}^d := \{x \in \mathbb{R}^{d+1} : \|x\|_2 = 1\}$ be the unit sphere in the Euclidean space $\mathbb{R}^{d+1}$ for $d \geq 2$, endowed with the surface measure $\mathrm{d}\omega_d$; that is, $|\mathbb{S}^d| := \int_{\mathbb{S}^d} \mathrm{d}\omega_d$ denotes the surface area of the unit sphere $\mathbb{S}^d$. Many real-world applications can be modeled as spherical problems. A critical task of spherical modeling is to find an effective data

fitting strategy to approximate the underlying mapping between input and output data. Hyperinterpolation is a simple yet powerful method for fitting spherical data, and it has received a great deal of interest since its birth, see references listed in Chapter 1. Given sampled data $\{(x_j, y_j)\}_{j=1}^m \subset \mathbb{S}^d \times \mathbb{R}$, the underlying mapping can be modeled as a spherical hyperinterpolant of degree n in the form of

$$x \in \mathbb{S}^d \mapsto \sum_{j=1}^m w_j y_j G_n(x, x_j) \in \mathbb{R}, \tag{4.1.1}$$

where $w_j > 0$, $j = 1, 2, \ldots, m$, are some prescribed weights,

$$G_n(x, y) = \sum_{\ell=0}^n \sum_{k=1}^{Z(d,\ell)} Y_{\ell,k}(x) Y_{\ell,k}(y)$$

is a kernel generated by the spherical harmonics $\{Y_{\ell,k}\}$ of degree at most n, and the precise number $Z(d, \ell)$ of spherical harmonics of exact degree ℓ is given in (4.2.1).

The simplicity of spherical hyperinterpolation is manifested in the modeled mapping (4.1.1). Unlike many other fitting techniques that usually need to solve a system of linear equations to obtain the modeled mapping, e.g., the least squares, the spherical hyperinterpolation (4.1.1) can be directly written down and immediately generates the output from any input $x \in \mathbb{S}^d$ without any mathematical manipulations but only addition and multiplication. Moreover, adding a new data pair or withdrawing an existing one can be directly achieved without a new computation from scratch.

However, the construction of hyperinterpolation of degree n requires a positive-weight quadrature rule (4.1.2)

$$\sum_{j=1}^m w_j f(x_j) \approx \int_{\mathbb{S}^d} f \mathrm{d}\omega_d \tag{4.1.2}$$

to be exact for polynomials up to degree $2n$, that is,

$$\sum_{j=1}^m w_j f(x_j) = \int_{\mathbb{S}^d} f \mathrm{d}\omega_d \quad \forall f \in \mathbb{P}_{2n}(\mathbb{S}^d), \tag{4.1.3}$$

where $\mathbb{P}_n(\mathbb{S}^d)$ be the space of spherical polynomials of degree at most n. A convenient L^2-orthonormal basis (with respect to $\mathrm{d}\omega_d$) for $\mathbb{P}_n$ is provided by the spherical harmonics $\{Y_{\ell,k} : k = 1, 2, \ldots Z(d, \ell); \ell = 0, 1, 2, \ldots, n\}$. The spherical hyperinterpolation operator $\mathcal{L}_n : C(\mathbb{S}^d) \to \mathbb{P}_n(\mathbb{S}^d)$ maps a continuous function $f \in C(\mathbb{S}^d)$ on the

sphere $\mathbb{S}^d$ to

$$\mathcal{L}_n f := \sum_{\ell=0}^{n} \sum_{k=1}^{Z(d,\ell)} \langle f, Y_{\ell,k} \rangle_m Y_{\ell,k} \in \mathbb{P}_n(\mathbb{S}^d), \tag{4.1.4}$$

where

$$\langle f, g \rangle_m := \sum_{j=1}^{m} w_j f(x_j) g(x_j)$$

is the numerical evaluation of the inner product

$$\langle f, g \rangle := \int_{\mathbb{S}^d} f(x) g(x) \mathrm{d}\omega_d$$

by the quadrature rule (4.1.2) with the exactness assumption (4.1.3). In other words, the hyperinterpolation (4.1.4) of $f \in C(\mathbb{S}^d)$ can be regarded as a discrete version of the famous L^2-orthogonal projection

$$\mathcal{P}_n f := \sum_{\ell=0}^{n} \sum_{k=1}^{Z(d,\ell)} \langle f, Y_{\ell,k} \rangle \, Y_{\ell,k} \in \mathbb{P}_n(\mathbb{S}^d) \tag{4.1.5}$$

of f from $C(\mathbb{S}^d)$ onto $\mathbb{P}_n(\mathbb{S}^d)$. Sometimes we may consider equal-weight quadrature rules of the form

$$\frac{|\mathbb{S}^d|}{m} \sum_{j=1}^{m} f(x_j) \approx \int_{\mathbb{S}^d} f \mathrm{d}\omega_d. \tag{4.1.6}$$

Regarding this very restrictive nature of (4.1.3) that it is impractical and sometimes impossible to obtain data on the desired quadrature points in practice, our aim in this chapter is to bypass this quadrature exactness assumption by replacing it with the *Marcinkiewicz–Zygmund property*; namely, we assume that there exists an $\eta \in [0, 1)$ such that (2.2.5) holds. If $n' = n$, i.e., the quadrature exactness is not relaxed, then the exactness (4.1.3) implies $\eta = 0$. Then the construction of hyperinterpolation is feasible with many more quadrature rules outside the traditional candidates. Traditionally, quadrature rules using spherical t-designs are used to construct hyperinterpolation. As we can see in this chapter, quadrature rules using scattered points, equal area points, minimal energy points, maximal determinant points, and many other kinds of points are also feasible for constructing hyperinterpolation. The Marcinkiewicz–Zygmund property (2.2.5) is equivalent to

$$(1-\eta) \int_{\mathbb{S}^d} \chi^2 \mathrm{d}\omega_d \le \sum_{j=1}^{m} w_j \chi(x_j)^2 \le (1+\eta) \int_{\mathbb{S}^d} \chi^2 \mathrm{d}\omega_d \quad \forall \chi \in \mathbb{P}_n(\mathbb{S}^d),$$

which can be regarded as the Marcinkiewicz–Zygmund inequality [89, 143, 146] applied to polynomials χ^2 of degree at most $2n$ with $\chi \in \mathbb{P}_n(\mathbb{S}^d)$, and it has been utilized in Chapter 2 that quadrature rules are assumed to have exactness degree $n + n'$ with $0 < n' \leq n$ for the construction of hyperinterpolation.

To tell the difference between the original hyperinterpolation $\mathcal{L}_n$ and the hyperinterpolation relying *only* on the Marcinkiewicz–Zygmund property (2.2.5), we refer to the latter as the *unfettered hyperinterpolation*, indicating that the application of hyperinterpolation is no longer limited by the quadrature exactness assumption, and denote it by

$$\mathcal{U}_n f := \sum_{\ell=0}^{n} \sum_{k=1}^{Z(d,\ell)} \langle f, Y_{\ell,k} \rangle_m Y_{\ell,k} \in \mathbb{P}_n(\mathbb{S}^d), \tag{4.1.7}$$

where the quadrature rule (4.1.2) for evaluating $\langle f, Y_{\ell,k} \rangle_m$ is only assumed to satisfy the property (2.2.5).

We derive in this chapter that

$$\|\mathcal{U}_n f - f\|_{L^2} \leq \left(\sqrt{1+\eta} \left(\sum_{j=1}^{m} w_j \right)^{1/2} + |\mathbb{S}^d|^{1/2} \right) E_n(f) + \sqrt{\eta^2 + 4\eta} \|\chi^*\|_{L^2}, \tag{4.1.8}$$

where $E_n(f)$ denotes the best uniform error of $f \in C(\mathbb{S}^d)$ by a polynomial in $\mathbb{P}_n(\mathbb{S}^d)$, that is,

$$E_n(f) := \inf_{\chi \in \mathbb{P}_n(\mathbb{S}^d)} \|f - \chi\|_\infty,$$

and $\chi^* \in \mathbb{P}_n(\mathbb{S}^d)$ is the best approximation polynomial of f in $\mathbb{P}_n(\mathbb{S}^d)$ in the sense of

$$\|f - \chi^*\|_\infty = E_n(f).$$

Thus, no matter what kind of point distributions is adopted, it is sufficient for a reasonable approximation error bound to control the numerical integration error so that the constant η in the Marcinkiewicz–Zygmund property (2.2.5) is reasonably small.

The L^2 error estimate (4.1.8) reduces to the classical result

$$\|\mathcal{L}_n f - f\|_{L^2} \leq 2|\mathbb{S}^d|^{1/2} E_n(f)$$

of hyperinterpolation derived in [196] when the quadrature exactness degree is assumed to be $2n$, because such an assumption leads to $\eta = 0$ and

$$\sum_{j=1}^{m} w_j = \int_{\mathbb{S}^d} \mathrm{d}\omega_d = |\mathbb{S}^d|.$$

If the quadrature exactness degree is assumed to be $n + n'$ with $0 < n' \leq n$, then the estimate (4.1.8) can be refined as

$$\|\mathcal{U}_n f - f\|_{L^2} \leq \left(\sqrt{1+\eta} + 1\right) |\mathbb{S}^d|^{1/2} E_{n'}(f),$$

and this convergence rate in terms of $E_{n'}(f)$ coincides with the result in Chapter 2 that

$$\|\mathcal{L}_n f - f\|_{L^2} \leq \left(\frac{1}{\sqrt{1-\eta}} + 1\right) |\mathbb{S}^d|^{1/2} E_{n'}(f) \tag{4.1.9}$$

under the same assumption. A Sobolev analog to the error estimate (4.1.8), i.e., the error measured by a Sobolev norm, is also established in this chapter.

We also highlight the connection between the unfettered hyperinterpolation and QMC designs. Historically, quadrature exactness is often a starting point in designing quadrature rules. Nevertheless, this trend has recently received growing concerns regarding whether exactness is a reliable designing principle, see, e.g., [222]. The concept of QMC designs, introduced by Brauchart, Saff, Sloan, and Womersley in [35], is an important quadrature-designing principle against this historical trend. QMC designs include many points distributions that are easy to obtain numerically, and quadrature rules using QMC designs provide the same asymptotic order of convergence as rules with quadrature exactness when the integrand belongs to the Sobolev space $H^s(\mathbb{S}^d)$ with $s > d/2$. Moreover, quadrature exactness is not a necessary assumption for QMC designs. If the quadrature points form a QMC design, then we show quadrature rules using them also satisfy the Marcinkiewicz–Zygmund property (2.2.5). Hence hyperinterpolation using QMC designs is a special case in the general framework of unfettered hyperinterpolation. However, the general error estimate (4.1.8) may not be sharp for hyperinterpolation using QMC designs, and we can refine them. Regarding the particularity of QMC designs, we may refer to the hyperinterpolation of $f \in H^s(\mathbb{S}^d)$ using QMC designs, though a special case of unfettered hyperinterpolation, as the *QMC hyperinterpolation*, and denote it by

$$\mathcal{Q}_n f := \sum_{\ell=0}^{n} \sum_{k=1}^{Z(d,\ell)} \langle f, Y_{\ell,k} \rangle_m \, Y_{\ell,k} \in \mathbb{P}_n(\mathbb{S}^d), \tag{4.1.10}$$

where the quadrature rule (4.1.2) for evaluating $\langle f, Y_{\ell,k}\rangle_m$ adopt a QMC design for $H^s(\mathbb{S}^d)$ as the set of quadrature points. We show in this chapter that for $f \in H^s(\mathbb{S}^d)$,

$$\|\mathcal{Q}_n f - f\|_{L^2} \le c''(s,d)\left(n^{-s} + \frac{1}{m^{s/d}}\sqrt{\frac{Z(d+1,n)}{a_n^{(s)}}}\right)\|f\|_{H^s},$$

where $c''(s,d) > 0$ is some constant depending only on c and s, and $a_n^{(s)}$ is of order $(1+n)^{-2s}$.

The rest of this chapter is organized as follows. Section 4.2 collects some technical facts regarding spherical harmonics, our Sobolev space setting, spherical t-designs, and QMC designs. Section 4.3 gives the approximation theory of the unfettered hyperinterpolation under the only assumption of the Marcinkiewicz–Zygmund property (2.2.5). Section 4.4 develops the approximation theory of the QMC hyperinterpolation under the only assumption that $\{x_j\}_{j=1}^m$ is a QMC design. Section 4.5 contains numerical experiments that validate our theory.

4.2 Spherical harmonics analysis and spherical designs

We are concerned with real-valued functions on the sphere $\mathbb{S}^d$ in the Euclidean space $\mathbb{R}^{d+1}$ for $d \ge 2$.

4.2.1 Spherical harmonics and hyperinterpolation

Let $L^2(\mathbb{S}^d)$ denote the Hilbert space of all square-integrable functions on $\mathbb{S}^d$ with the inner product

$$\langle f, g\rangle := \int_{\mathbb{S}^d} f(x)g(x)\mathrm{d}\omega_d(x)$$

and the induced norm

$$\|f\|_{L^2} := \sqrt{\langle f, f\rangle}.$$

By $C(\mathbb{S}^d)$ we denote the space of continuous functions on $\mathbb{S}^d$, endowed with the uniform norm

$$\|f\|_\infty := \operatorname{ess\,sup}_{x\in\mathbb{S}^d}|f(x)|.$$

The restriction to $\mathbb{S}^d$ of a homogeneous and harmonic polynomial of total degree ℓ defined on $\mathbb{R}^{d+1}$ is called a *spherical harmonic of degree* ℓ on $\mathbb{S}^d$. We denote, as usual, by $\{Y_{\ell,k} : k = 1, 2, \ldots, Z(d,\ell)\}$ a collection of L^2-orthonormal real-valued

spherical harmonics of exact degree ℓ, where

$$Z(d,0)=1,\ Z(d,\ell)=(2\ell+d-1)\frac{\Gamma(\ell+d-1)}{\Gamma(d)\Gamma(\ell+1)}\sim\frac{2}{\Gamma(d)}\ell^{d-1}\text{ as }\ell\to\infty, \tag{4.2.1}$$

where $\Gamma(z)$ is the Gamma function. The spherical harmonics of degree $\ell\in\{0,1,2,\ldots\}$ satisfy the addition theorem [151, Theorem 2], that is,

$$\sum_{k=1}^{Z(d,\ell)}Y_{\ell,k}(x)Y_{\ell,k}(y)=\frac{Z(d,\ell)}{|\mathbb{S}^d|}P_\ell^{(d)}(x\cdot y),$$

where $P_\ell^{(d)}$ is the normalized Gegenbauer polynomial on $[-1,1]$, orthogonal with respect to the weight function $(1-t^2)^{d/2-1}$, and normalized such that $P_\ell^{(d)}(1)=1$. As an immediate application of the addition theorem, we have

$$\|Y_{\ell,k}\|_\infty\le\left(\frac{Z(d,\ell)}{|\mathbb{S}^d|}\right)^{1/2} \tag{4.2.2}$$

for all $\ell=0,1,2,\ldots$ and $k=1,2,\ldots,Z(d,\ell)$. Indeed, for any spherical harmonic $Y_{\ell,k}$, suppose $|Y_{\ell,k}(x)|$ attains $\|Y_{\ell,k}\|_\infty$ at the point $x^*\in\mathbb{S}^d$, then

$$\begin{aligned}\|Y_{\ell,k}\|_\infty=|Y_{\ell,k}(x^*)|&\le\left(\sum_{k=1}^{Z(d,\ell)}|Y_{\ell,k}(x^*)|^2\right)^{1/2}\\&=\left(\frac{Z(d,\ell)}{|\mathbb{S}^d|}P_\ell^{(d)}(1)\right)^{1/2}=\left(\frac{Z(d,\ell)}{|\mathbb{S}^d|}\right)^{1/2}.\end{aligned}$$

Besides, it is well known (see, e.g., [151, pp. 38–39]) that each spherical harmonic $Y_{\ell,k}$ of exact degree ℓ is an eigenfunction of the negative Laplace–Beltrami operator $-\Delta_d^*$ for $\mathbb{S}^d$ with eigenvalue

$$\lambda_\ell:=\ell(\ell+d-1). \tag{4.2.3}$$

The family $\{Y_{\ell,k}:k=1,\ldots,Z(d,\ell);\ell=0,1,2,\ldots\}$ of spherical harmonics forms a complete L^2-orthonormal (with respect to $\mathrm{d}\omega_d$) system for the Hilbert space $L^2(\mathbb{S}^d)$. Thus, for any $f\in L^2(\mathbb{S}^d)$, it can be represented by a Laplace–Fourier series

$$f(x)=\sum_{\ell=0}^{\infty}\sum_{k=1}^{Z(d,\ell)}\hat{f}_{\ell,k}Y_{\ell,k}(x)$$

with coefficients

$$\hat{f}_{\ell,k}:=\langle f,Y_{\ell,k}\rangle=\int_{\mathbb{S}^d}f(x)Y_{\ell,k}(x)\mathrm{d}\omega_d(x) \tag{4.2.4}$$

for $\ell=0,1,2,\ldots$ and $k=1,2,\ldots,Z(d,\ell)$.

The space $\mathbb{P}_n(\mathbb{S}^d)$ of all spherical polynomials of degree at most n (i.e., the restriction to $\mathbb{S}^d$ of all polynomials in $\mathbb{R}^{d+1}$ of degree at most n) coincides with the span of all spherical harmonics up to (and including) degree n, and its dimension satisfies $\dim \mathbb{P}_n(\mathbb{S}^d) = Z(d+1, n)$. The space $\mathbb{P}_n(\mathbb{S}^d)$ is also a reproducing kernel Hilbert space with the reproducing kernel

$$G_n(x, y) = \sum_{\ell=0}^{n} \sum_{k=1}^{Z(d,\ell)} Y_{\ell,k}(x) Y_{\ell,k}(y) \tag{4.2.5}$$

in the sense that

$$\langle \chi, G(\cdot, x) \rangle = \chi(x) \quad \forall \chi \in \mathbb{P}_n(\mathbb{S}^d), \tag{4.2.6}$$

see, e.g., [176]. Given $f \in C(\mathbb{S}^d)$, it is often simpler in practice to express the hyperinterpolant $\mathcal{L}_n f$ using the reproducing kernel $G_n(\cdot, \cdot)$ defined by (4.2.5). By rearranging the summation,

$$\begin{aligned} \mathcal{L}_n f(x) &= \sum_{\ell=0}^{n} \sum_{k=1}^{Z(d,\ell)} \left(\sum_{j=1}^{m} w_j f(x_j) Y_{\ell,k}(x_j) \right) Y_{\ell,k}(x) \\ &= \sum_{j=1}^{m} w_j f(x_j) G_n(x, x_j). \end{aligned}$$

Since such a summation-rearranging procedure does not depend on the quadrature exactness, such an expression also applies to $\mathcal{U}_n f$ and $\mathcal{Q}_n f$. What makes the above three expressions different is the quadrature rules used for constructing different kinds of hyperinterpolants.

4.2.2 Sobolev spaces

The study of hyperinterpolation in a Sobolev space setting can be traced back to the work [110] by Hesse and Sloan. The Sobolev space $H^s(\mathbb{S}^d)$ on the sphere $\mathbb{S}^d$ may be defined for $s \geq 0$ as the set of all functions $f \in L^2(\mathbb{S}^d)$ whose Laplace–Fourier coefficients (4.2.4) satisfy

$$\sum_{\ell=0}^{\infty} \sum_{k=1}^{Z(d,\ell)} (1 + \lambda_\ell)^s |\hat{f}_{\ell,k}|^2 < \infty,$$

where λ_ℓ is given as (4.2.3). When $s = 0$, we have $H^0(\mathbb{S}^d) = L^2(\mathbb{S}^d)$. The norm in $H^s(\mathbb{S}^d)$ may be defined as the square root of the expression on the left-hand side of the last inequality; however, in this chapter, we shall take advantage of the freedom to define equivalent Sobolev space norms. Let $s > d/2$ be fixed and suppose we are

given a sequence of positive real numbers $(a_\ell^{(s)})_{\ell\geq 0}$ satisfying

$$a_\ell^{(s)} \asymp (1+\lambda_\ell)^{-s} \asymp (1+\ell)^{-2s}. \tag{4.2.7}$$

Then we can define a norm in $H^s(\mathbb{S}^d)$ by

$$\|f\|_{H^s} := \left(\sum_{\ell=0}^{\infty}\sum_{k=1}^{Z(d,\ell)} \frac{1}{a_\ell^{(s)}}|\hat{f}_{\ell,k}|^2\right)^{1/2}.$$

The norm $\|\cdot\|_{H^s}$ therefore depends on the particular choice of the sequence $(a_\ell^{(s)})_{\ell\geq 0}$, but a change to this sequence merely leads to an *equivalent* Sobolev norm.

The following lemmas are necessary for our analysis.

Lemma 4.2.1 *For any $f \in \mathbb{P}_n(\mathbb{S}^d)$,*

$$\|f\|_{H^s} \leq \tilde{c}\,(n+1)^s\,\|f\|_{L^2},$$

where $\tilde{c} > 0$ is a constant.

Proof. It is straightforward that for any $f \in \mathbb{P}_n(\mathbb{S}^d)$,

$$\begin{aligned}\|f\|_{H^s} &= \left(\sum_{\ell=0}^{n}\sum_{k=1}^{Z(d,\ell)} \frac{1}{a_\ell^{(s)}}|\hat{f}_{\ell,k}|^2\right)^{1/2} \leq \left(\frac{1}{a_n^{(s)}}\|f\|_{L^2}^2\right)^{1/2} \\ &\leq \tilde{c}\,(n+1)^s\,\|f\|_{L^2},\end{aligned}$$

where we used the order (4.2.7) of $(a_\ell^{(s)})_{\ell\geq 0}$. □

Lemma 4.2.2 *If $s > d/2$, then*

$$\|fg\|_{H^s} \leq \check{c}\|f\|_{H^s}\|g\|_{H^s},$$

where $\check{c} > 0$ is some constant.

Proof. For any Lipschitz domain Ω, let $W^{s,2}(\Omega)$ be the Sobolev space of those functions in $L^2(\Omega)$ whose distributional derivatives up to (and including) order s are in $L^2(\Omega)$. Note that the Sobolev spaces $H^s(\mathbb{S}^d)$ can also be defined with the help of charts (that is, the so-called Sobolev spaces over boundaries), giving the space $W^{s,2}(\mathbb{S}^d)$ with an *equivalent* norm, that is,

$$c_1\|f\|_{H^s} \leq \|f\|_{W^{s,2}(\mathbb{S}^d)} \leq c_2\|f\|_{H^s}, \tag{4.2.8}$$

where $c_1, c_2 > 0$ are some constants; see [138, Chapter 7.3] or [14, Chapter 7.2.3]. If $s > d/2$, then the Sobolev space $W^{s,2}(\mathbb{S}^d)$ is a Banach algebra, that is, for any $f, g \in W^{s,2}(\mathbb{S}^d)$,

$$\|fg\|_{W^{s,2}(\mathbb{S}^d)} \leq c_3 \|f\|_{W^{s,2}(\mathbb{S}^d)} \|g\|_{W^{s,2}(\mathbb{S}^d)}, \tag{4.2.9}$$

where $c_3 > 0$ is some constant; we refer to [1, Theorem 5.23] or [144, Section 6.1] for this result. Together with (4.2.8) and (4.2.9), we have the desired estimate. □

Remark 4.2.3 *The norm equivalence* (4.2.8) *is also identified and utilized in some other spherical approximation schemes, see, e.g., [112, 127].*

4.2.3 Spherical t-designs and QMC designs

As briefly mentioned earlier in Chapter 2, a spherical t-design, introduced in the remarkable paper [69] by Delsarte, Goethals, and Seidel, is a set of points $\{x_j\}_{j=1}^m \subset \mathbb{S}^d$ with the characterizing property that an equal-weight quadrature rule in these points exactly integrates all polynomials of degree at most t, that is,

$$\frac{|\mathbb{S}^d|}{m} \sum_{j=1}^{m} \chi(x_j) = \int_{\mathbb{S}^d} \chi(x) \mathrm{d}\omega_d(x) \quad \forall \chi \in \mathbb{P}_t. \tag{4.2.10}$$

A majority of studies in the literature on spherical designs care about the relation between m and t in (4.2.10). It was known by Seymour and Zaslavsky [187] that a spherical t-design always exists if m is sufficiently large, but no quantitative results on the size of m were established. In the original manuscript [69] of spherical t-designs, lower bounds on m of exact order t^d were derived in the sense that

$$m \geq \begin{cases} \binom{d + t/2}{d} + \binom{d + t/2 - 1}{d} & \text{for even } t, \\ 2\binom{d + \lfloor t/2 \rfloor}{d} & \text{for odd } t; \end{cases}$$

but according to Bannai and Damerell [17, 18], the number m of quadrature points could attain these lower bounds only for a few small values of t. Bondarenko, Radchenko, and Viazovska asserted in [28] that for each $m \geq ct^d$ with some positive but unknown constant $c > 0$, there exists a spherical t-design in $\mathbb{S}^d$ consisting of m points.

Quadrature rules (4.1.2) using spherical t-designs are known to have fast-convergence property when the integrand belongs to the Sobolev space H^s; namely, given $s > d/2$,

there exists $C(s,d) > 0$ depending only on s and d such that for every m-point spherical t-design $\{x_j\}_{j=1}^m$ on $\mathbb{S}^d$, there holds

$$\sup_{\substack{f\in H^s(\mathbb{S}^d),\\ \|f\|_{H^s}\le 1}} \left| \frac{|\mathbb{S}^d|}{m} \sum_{j=1}^m f(x_j) - \int_{\mathbb{S}^d} f(x) \mathrm{d}\omega_d \right| \le \frac{C(s,d)}{t^s}. \tag{4.2.11}$$

The estimate (4.2.11) was established gradually: It was first proved for the particular case $s = 3/2$ and $d = 2$ in [108], then extended to all $s > 1$ for $d = 2$ in [109], and finally extended to all $s > d/2$ and all $d \ge 2$ in [34]. The condition $s > d/2$ is a natural one because functions to be approximated in this chapter are assumed to be continuous, and by the Sobolev embedding theorem, $H^s(\mathbb{S}^d)$ is continuously embedded in $C(\mathbb{S}^d)$ if $s > d/2$.

If only spherical t-designs with $m \asymp t^d$ are concerned, then the upper bound on the error (4.2.11) is of order $m^{-s/d}$. Here comes the concept of QMC designs, introduced by Brauchart, Saff, Sloan, and Womersley in [35]: Given $s > d/2$, a sequence $\{x_j\}_{j=1}^m$ of m-point configurations on $\mathbb{S}^d$ with $m \to \infty$ is said to be a sequence of *QMC designs* for $H^s(\mathbb{S}^d)$ if there exists $c(s,d) > 0$ independent of m such that

$$\sup_{\substack{f\in H^s(\mathbb{S}^d),\\ \|f\|_{H^s}\le 1}} \left| \frac{|\mathbb{S}^d|}{m} \sum_{j=1}^m f(x_j) - \int_{\mathbb{S}^d} f(x) \mathrm{d}\omega_d \right| \le \frac{c(s,d)}{m^{s/d}}. \tag{4.2.12}$$

In a nutshell, quadrature rules using QMC designs provide the same asymptotic order of convergence as exact rules (e.g., rules using spherical t-designs) when the integrand belongs to the Sobolev space H^s, but are easier to obtain numerically. For more studies on the numerical integration on the sphere with the integrand belonging to a Sobolev space, we refer the reader to [31, 32, 111].

A substantial definition related to QMC designs $\{x_j\}_{j=1}^m$ is the *QMC strength*, denoted by s^*. For every sequence of QMC designs $\{x_j\}_{j=1}^m$, there is some number s^* such that $\{x_j\}_{j=1}^m$ is a sequence of QMC designs for all s satisfying $d/2 < s \le s^*$ and is not a QMC design for $s > s^*$. Even if the integrand f is infinitely differentiable, the convergence rate of the numerical integration error (4.2.12) using a QMC design with strength s^* is controlled by $m^{-s^*/d}$.

4.3 General framework of unfettered hyperinterpolation

With the aid of the reproducing property (4.2.6), the Marcinkiewicz–Zygmund property (2.2.5) implies the following lemma.

Lemma 4.3.1 *For any $\chi \in \mathbb{P}_n(\mathbb{S}^d)$, we have*

(a) $(1-\eta)\|\chi\|_{L^2}^2 \le \langle \mathcal{U}_n\chi, \chi\rangle \le (1+\eta)\|\chi\|_{L^2}^2$.

(b) $(1-\eta)\|\chi\|_{L^2} \le \|\mathcal{U}_n\chi\|_{L^2} \le (1+\eta)\|\chi\|_{L^2}$.

(c) $\|\mathcal{U}_n\chi - \chi\|_{L^2}^2 \le (\eta^2+4\eta)\|\chi\|_{L^2}^2$.

Proof. (a) The reproducing property (4.2.6) of $G_n(\cdot,\cdot)$ implies

$$\begin{aligned}\langle \mathcal{U}_n\chi, \chi\rangle &= \left\langle \sum_{j=1}^m w_j\chi(x_j)G_n(x,x_j), \chi(x)\right\rangle \\ &= \sum_{j=1}^m w_j\chi(x_j)\left\langle G_n(x,x_j), \chi(x)\right\rangle \\ &= \sum_{j=1}^n w_j\chi(x_j)^2.\end{aligned}$$

Thus by the Marcinkiewicz–Zygmund property (2.2.5),

$$\begin{aligned}(1-\eta)\|\chi\|_{L^2}^2 = (1-\eta)\int_{\mathbb{S}^d}\chi^2 \mathrm{d}\omega_d &\le \sum_{j=1}^n w_j\chi(x_j)^2 \\ &\le (1+\eta)\int_{\mathbb{S}^d}\chi^2\mathrm{d}\omega_d = (1+\eta)\|\chi\|_{L^2}^2.\end{aligned}$$

(b) By part (a), we have

$$(1-\eta)\|\chi\|_{L^2}^2 \le \langle \mathcal{U}_n\chi,\chi\rangle \le \|\mathcal{U}_n\chi\|_{L^2}\|\chi\|_{L^2},$$

leading to

$$(1-\eta)\|\chi\|_{L^2} \le \|\mathcal{U}_n\chi\|_{L^2}.$$

We also have

$$\begin{aligned}\|\mathcal{U}_n\chi\|_{L^2}^2 \le \langle \mathcal{U}_n\chi, \mathcal{U}_n\chi\rangle &= \left\langle \sum_{j=1}^m w_j\chi(x_j)G_n(x,x_j), \mathcal{U}_n\chi(x)\right\rangle \\ &= \sum_{j=1}^m w_j\chi(x_j)\mathcal{U}_n\chi(x_j) \\ &\le \left(\sum_{j=1}^m w_j\chi(x_j)^2\right)^{1/2}\left(\sum_{j=1}^m w_j\left(\mathcal{U}_n\chi(x_j)\right)^2\right)^{1/2} \\ &\le (1+\eta)\|\chi\|_{L^2}\|\mathcal{U}_n\chi\|_{L^2},\end{aligned}$$

where the first inequality is due to the Cauchy–Schwarz inequality, and the second one is ensured by the Marcinkiewicz–Zygmund property (2.2.5). Thus part (b) is proved.

(c) Using parts (a) and (b) above, it is straightforward that

$$\begin{aligned}\|\mathcal{U}_n\chi - \chi\|_{L^2}^2 &= \|\mathcal{U}_n\chi\|_{L^2}^2 - 2\langle \mathcal{U}_n\chi, \chi\rangle + \|\chi\|_{L^2}^2 \\ &\le (1+\eta)^2\|\chi\|_{L^2}^2 - 2(1-\eta)\|\chi\|_{L^2}^2 + \|\chi\|_{L^2}^2 \\ &= (\eta^2+4\eta)\|\chi\|_{L^2}^2.\end{aligned}$$

Hence this lemma is proved. □

We are now ready to state our main theorem.

Theorem 4.3.2 *Given $f \in C(\mathbb{S}^d)$, let $\mathcal{U}_n f \in \mathbb{P}_n$ be its unfettered hyperinterpolant defined by (4.1.7), where the m-point positive-weight quadrature rule (4.1.2) is only assumed to have the Marcinkiewicz–Zygmund property (2.2.5) with $\eta \in [0,1)$. Then*

$$\|\mathcal{U}_n f\|_{L^2} \le \sqrt{1+\eta}\left(\sum_{j=1}^m w_j\right)^{1/2}\|f\|_\infty, \tag{4.3.1}$$

and

$$\begin{aligned}\|\mathcal{U}_n f - f\|_{L^2} \le &\left(\sqrt{1+\eta}\left(\sum_{j=1}^m w_j\right)^{1/2} + |\mathbb{S}^d|^{1/2}\right) E_n(f) \\ &+ \sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2},\end{aligned} \tag{4.3.2}$$

where $E_n(f)$ denotes the best uniform error of f by a polynomial in $\mathbb{P}_n(\mathbb{S}^d)$ and $\chi^* \in \mathbb{P}_n(\mathbb{S}^d)$ denotes the best approximation polynomial of f in $\mathbb{P}_n(\mathbb{S}^d)$ in the sense of $\|f-\chi^*\|_\infty = E_n(f)$.

Proof. For any $f \in C(\mathbb{S}^d)$, we have $\mathcal{U}_n f \in \mathbb{P}_n$ and hence

$$\langle G_n(x, x_j), \mathcal{U}_n f(x)\rangle = \mathcal{U}_n f(x_j).$$

Thus,

$$\begin{aligned}\langle \mathcal{U}_n f, \mathcal{U}_n f\rangle &= \left\langle \sum_{j=1}^m w_j f(x_j) G_n(x, x_j), \mathcal{U}_n f(x)\right\rangle = \sum_{j=1}^m w_j f(x_j)\mathcal{U}_n f(x_j)\\ &\le \left(\sum_{j=1}^m w_j f(x_j)^2\right)^{1/2}\left(\sum_{j=1}^m w_j \left(\mathcal{U}_n\chi(x_j)\right)^2\right)^{1/2}\\ &\le \left(\sum_{j=1}^m w_j\right)^{1/2} \|f\|_\infty\sqrt{1+\eta}\|\mathcal{U}_n f\|_{L^2},\end{aligned}$$

where the first inequality is due to the Cauchy–Schwarz inequality and the second one holds by using

$$\sum_{j=1}^m w_j f(x_j)^2 \le \|f\|_\infty^2 \sum_{j=1}^m w_j$$

and the Marcinkiewicz–Zygmund property (2.2.5). This estimate immediately implies the stability result (4.3.1).

The error bound (4.3.2) is obtained by the following argument. For any $\chi \in \mathbb{P}_n$, we have

$$\begin{aligned}&\|\mathcal{U}_n f - f\|_{L^2}\\ =&\|\mathcal{U}_n(f-\chi) + (\chi - f) + (\mathcal{U}_n\chi - \chi)\|_{L^2}\\ \le&\|\mathcal{U}_n(f-\chi)\|_{L^2} + \|f-\chi\|_{L^2} + \|\mathcal{U}_n\chi - \chi\|_{L^2}\\ \le&\sqrt{1+\eta}\left(\sum_{j=1}^m w_j\right)^{1/2}\|f-\chi\|_\infty + |\mathbb{S}^d|^{1/2}\|f-\chi\|_\infty + \|\mathcal{U}_n\chi - \chi\|_{L^2}.\end{aligned}$$

It follows, since this estimate holds for all polynomials in $\mathbb{P}_n(\mathbb{S}^d)$, that

$$\|\mathcal{U}_n f - f\|_{L^2} \le \left(\sqrt{1+\eta}\left(\sum_{j=1}^m w_j\right)^{1/2} + |\mathbb{S}^d|^{1/2}\right) E_n(f) + \|\mathcal{U}_n\chi^* - \chi^*\|_{L^2}.$$

By part (c) of Lemma 4.3.1, we have $\|\mathcal{U}_n\chi^* - \chi^*\|_{L^2} \le \sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$. □

4.3.1 Connections in the literature

If the quadrature rule (4.1.2) is additionally assumed to integrate all constant functions (polynomials of degree zero) exactly, that is,

$$\sum_{j=1}^{m} w_j = |\mathbb{S}^d|,$$

then we have

$$\|\mathcal{U}_n f\|_{L^2} \leq \sqrt{1+\eta}|\mathbb{S}^d|^{1/2}\|f\|_\infty$$

and

$$\|\mathcal{U}_n f - f\|_{L^2} \leq \left(\sqrt{1+\eta}+1\right)|\mathbb{S}^d|^{1/2}E_n(f) + \sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}.$$

If the quadrature rule (4.1.2) exactly integrates all polynomials of degree at most $2n$, i.e., the constant η is zero, then the stability result (4.3.1) and error bound (4.3.2) reduce to the classical results of hyperinterpolation in [196]; namely,

$$\|\mathcal{U}_n f\|_{L^2} \leq |\mathbb{S}^d|^{1/2}\|f\|_\infty$$

and

$$\|\mathcal{U}_n f - f\|_{L^2} \leq 2|\mathbb{S}^d|^{1/2}E_n(f).$$

If the quadrature rule (4.1.2) has exactness degree $n+n'$ with $0 < n' \leq n$, then $\mathcal{U}_n\chi = \chi$ for all $\chi \in \mathbb{P}_{n'}(\mathbb{S}^d)$, see Lemma 2.2.4. By the stability result (4.3.1), we have for any $\chi \in \mathbb{P}_{n'}(\mathbb{S}^d)$,

$$\|\mathcal{U}_n f - f\|_{L^2} \leq \|\mathcal{U}_n(f-\chi) - (f-\chi)\|_{L^2} \leq \|\mathcal{U}_n(f-\chi)\|_{L^2} + \|f-\chi\|_{L^2}.$$

As this estimate holds for all $\chi \in \mathbb{P}_{n'}(\mathbb{S}^d)$, it is straightforward that

$$\|\mathcal{U}_n f - f\|_{L^2} \leq \left(\sqrt{1+\eta}+1\right)|\mathbb{S}^d|^{1/2}E_{n'}(f), \tag{4.3.3}$$

which has the same convergence rate in terms of $E_{n'}(f)$ as our previous estimate (4.1.9) in [12]. In [12], we make use of the discrete orthogonal projection property (see [12, Lemma 3.1]) to obtain the estimate (4.1.9), while in this chapter we utilize the reproducing property (4.2.6) for the estimate (4.3.3).

Moreover, in light of Theorem 4.3.2 and the study on spherical hyperinterpolation in a Sobolev space setting by Hesse and Sloan in [110], we have the following Sobolev estimates, which reduce to their results in [110] when the exactness degree $2n$

is assumed. For simplicity and without loss of generality, we assume $\sum_{j=1}^{m} w_j = |\mathbb{S}^d|$ in Corollary 4.3.3. Note that $H^s(\mathbb{S}^d) \subset L^2(\mathbb{S}^d)$.

Corollary 4.3.3 *Let $d \geq 2$, and let t and s be fixed real numbers with $s \geq t \geq 0$ and $s > d/2$. Under the conditions of Theorem 4.3.2, for any unfettered hyperinterpolation operator $\mathcal{U}_n : H^s(\mathbb{S}^d) \to H^t(\mathbb{S}^d)$, there hold*

$$\begin{aligned}\|\mathcal{U}_n f\|_{H^t} \leq & \tilde{c}\left[\left(\sqrt{1+\eta}|\mathbb{S}^d|^{1/2}+1\right)(n+1)^{d/2+t-s}\|f\|_{H^s}\right. \\ & \left.+(n+1)^t\sqrt{\eta^2+4\eta}\|f\|_{L^2}\right]+\|f\|_{H^s}\end{aligned} \tag{4.3.4}$$

and

$$\begin{aligned}\|\mathcal{U}_n f - f\|_{H^t} \leq & \tilde{c}\left[\left(\sqrt{1+\eta}|\mathbb{S}^d|^{1/2}+1\right)(n+1)^{d/2+t-s}E_n(f;H^s(\mathbb{S}^d))\right. \\ & \left.+\tilde{c}(n+1)^t\sqrt{\eta^2+4\eta}\|f\|_{L^2}\right],\end{aligned} \tag{4.3.5}$$

where $\tilde{c} > 0$ is some constant that may vary line to line, and $E_n(f;H^s(\mathbb{S}^d))$ is the best H^s approximation error of $f \in H^s(\mathbb{S}^d)$ by a polynomial in $\mathbb{P}_n(\mathbb{S}^d)$, that is,

$$E_n(f;H^s(\mathbb{S}^d)) := \inf_{\chi\in\mathbb{P}_n(\mathbb{S}^d)} \|f-\chi\|_{H^s}.$$

Remark 4.3.4 *When the exactness degree of the rule (4.1.2) is assumed to be $2n$, we have $\eta = 0$, and hence the results (4.3.4) and (4.3.5) reduce to the respective results of the original hyperinterpolation (some constants may be different) derived by Hesse and Sloan in [110].*

Proof. Similar to the decomposition of $\|\mathcal{U}_n f - f\|_{L^2}$ in the proof of Theorem 4.3.2, we have

$$\|\mathcal{U}_n f - f\|_{H^t} \leq \|\mathcal{U}_n(f-\mathcal{P}_n f)\|_{H^t} + \|\mathcal{P}_n f - f\|_{H^t} + \|\mathcal{U}_n(\mathcal{P}_n f) - \mathcal{P}_n f\|_{H^t}. \tag{4.3.6}$$

The first term on the right-hand side of (4.3.6) can be bounded by

$$\begin{aligned}\|\mathcal{U}_n(f-\mathcal{P}_n f)\|_{H^t} &\leq \tilde{c}(n+1)^t\|\mathcal{U}_n(f-\mathcal{P}_n f)\|_{L^2} \\ &\leq \tilde{c}(n+1)^t\sqrt{1+\eta}|\mathbb{S}^d|^{1/2}\|f-\mathcal{P}_n f\|_\infty \\ &\leq \tilde{c}(n+1)^t\sqrt{1+\eta}|\mathbb{S}^d|^{1/2}(n+1)^{d/2-s}\|f-\mathcal{P}_n f\|_{H^s},\end{aligned}$$

where the first inequality is due to Lemma 4.2.1, the second is due to the stability result (4.3.1), and the third is due to [110, Lemma 3.5]. This lemma also guarantees that

$$\|\mathcal{P}_n f - f\|_{H^t} \leq \tilde{c}(n+1)^{t-s}\|\mathcal{P}_n f - f\|_{H^s}.$$

The third term can be estimated as

$$\begin{aligned}\|\mathcal{U}_n(\mathcal{P}_n f) - \mathcal{P}_n f\|_{H^t} &\leq \tilde{c}(n+1)^t \|\mathcal{U}_n(\mathcal{P}_n f) - \mathcal{P}_n f\|_{L^2} \\ &\leq \tilde{c}(n+1)^t \sqrt{\eta^2+4\eta}\|\mathcal{P}_n f\|_{L^2} \\ &\leq \tilde{c}(n+1)^t \sqrt{\eta^2+4\eta}\|f\|_{L^2}\end{aligned}$$

where the first inequality is due to Lemma 4.2.1, the second is due to part (c) of Lemma 4.3.1, and the third is due to the fact that the norm of $\mathcal{P}_n$ as an operator from $L^2(\mathbb{S}^d)$ onto $L^2(\mathbb{S}^d)$ is 1. Thus we have

$$\begin{aligned}\|\mathcal{U}_n f - f\|_{H^t} \leq& \tilde{c}\left[\left(\sqrt{1+\eta}|\mathbb{S}^d|^{1/2}+1\right)(n+1)^{d/2+t-s}E_n(f;H^s(\mathbb{S}^d))\right. \\ &\left.+(n+1)^t\sqrt{\eta^2+4\eta}\|f\|_{L^2}\right],\end{aligned}$$

where $E_n(f;H^s(\mathbb{S}^d)) = \|f - \mathcal{P}_n f\|_{H^s}$ is verified by [110, Equ. (3.22)].

As $\|f - \mathcal{P}_n f\|_{H^s} \leq \|f\|_{H^s}$ and $\|f\|_{H^t} \lesssim \|f\|_{H^s}$, we have

$$\begin{aligned}\|\mathcal{U}_n f\|_{H^t} \leq& \|\mathcal{U}_n f - f\|_{H^t} + \|f\|_{H^t} \\ \leq& \tilde{c}\left[\left(\sqrt{1+\eta}|\mathbb{S}^d|^{1/2}+1\right)(n+1)^{d/2+t-s}\|f\|_{H^s}\right. \\ &\left.+(n+1)^t\sqrt{\eta^2+4\eta}\|f\|_{L^2}\right] + \|f\|_{H^s},\end{aligned}$$

which completes the proof of this corollary. □

4.3.2 Scattered data

In the work [126] of Le Gia and Mhaskar, the Marcinkiewicz–Zygmund inequality (2.2.5) was established for the case where quadrature points are randomly distributed:

Proposition 4.3.5 ([126, p. 463]) *Let $\gamma > 0$ and $\eta \in (0,1)$. For an equal-weight quadrature rule (4.1.6) with an independent random sample of m quadrature points drawn from the distribution ω_d, there exists a constant $\bar{c} := \bar{c}(\gamma)$ such that if $m \geq \bar{c}n^d \log n/\eta^2$, then the Marcinkiewicz–Zygmund property (2.2.5) holds with probability exceeding $1 - \bar{c}n^{-\gamma}$.*

With Proposition 4.3.5, we can obtain a probabilistic description of Theorem 2.2.8.

Corollary 4.3.6 *Adopt conditions of Theorem 2.2.8 and Proposition 4.3.5, where the quadrature rule for constructing $\mathcal{U}_n f$ takes the form of (4.1.6) and uses $m \geq$*

$\bar{c}(\gamma)n^d \log n/\eta^2$ *quadrature points. Then the stability result* (2.2.6) *and error bound* (2.2.7) *are valid with probability exceeding* $1 - \bar{c}n^{-\gamma}$.

As we can see, having bypassed the quadrature exactness assumption of the original hyperinterpolation, Theorem 2.2.8 provides a general framework of analyzing the behavior of the unfettered hyperinterpolation. What we need to do in practice is to control the constant η occurred in the Marcinkiewicz–Zygmund property (2.2.5). As a practical guide, if the quadrature points are independently random samples from the distribution ω_d, then Corollary 4.3.6 suggests a simple way to decrease η by increasing the number m of quadrature points.

4.4 Unfettered hyperinterpolation with QMC designs

Provided that $\{x_j\}_{j=1}^m$ forms a QMC design for $H^s(\mathbb{S}^d)$, it can be managed to satisfy the Marcinkiewicz–Zygmund property (2.2.5), as shown in Section 4.4.1. Hence the unfettered hyperinterpolation using QMC designs is a special case of the general framework analyzed in Theorem 4.3.2. Recall that we refer to such approximation as the QMC hyperinterpolation, denoted by $\mathcal{Q}_n$. However, the obtained error estimate may not be optimal due to the generality of Theorem 4.3.2, and we can find a sharper estimate customized for the unfettered hyperinterpolation using QMC designs.

4.4.1 QMC hyperinterpolation in the general framework of unfettered hyperinterpolation

It is critical to note that the numerical integration error (4.2.12) of the QMC design-based quadrature rule and the Marcinkiewicz–Zygmund property (2.2.5) are not interchangeable. The error (4.2.12) applies to all functions in $H^s(\mathbb{S}^d)$, but the property (2.2.5) only holds for polynomial χ^2 with $\chi \in \mathbb{P}_n(\mathbb{S}^d)$. On the other hand, if the integrand in the quadrature rule (4.1.6) is χ^2 with $\chi \in \mathbb{P}_n(\mathbb{S}^d)$, the error bound (4.2.12) suggests

$$\left| \frac{|\mathbb{S}^d|}{m} \sum_{j=1}^m \chi(x_j)^2 - \int_{\mathbb{S}^d} \chi^2 \mathrm{d}\omega_d \right| \leq \frac{c(s,d)}{m^{s/d}} \|\chi^2\|_{H^s} \tag{4.4.1}$$

with the controlling term $\|\chi^2\|_{H^s}$ instead of $\int_{\mathbb{S}^d} \chi^2 \mathrm{d}\omega_d$. Nevertheless, we can find an upper bound of $\|\chi^2\|_{H^s}$ in terms of $\int_{\mathbb{S}^d} \chi^2 \mathrm{d}\omega_d$ to transform the error (4.4.1) into a

Marcinkiewicz–Zygmund property (2.2.5). With the aid of Lemma 4.2.1, we have

$$\begin{aligned}\|\chi^2\|_{H^s} &\leq \tilde{c}(2n+1)^s\|\chi^2\|_{L^2}\\ &\leq \tilde{c}(2n+1)^s\|\chi\|_\infty\|\chi\|_{L^2}\\ &\leq \tilde{c}(2n+1)^s\frac{\|\chi\|_\infty}{\|\chi\|_{L^2}}\int_{\mathbb{S}^d}\chi^2 \mathrm{d}\omega_d.\end{aligned}$$

For any

$$\chi = \sum_{\ell=0}^{n}\sum_{k=1}^{Z(d,\ell)}\alpha_{\ell,k}Y_{\ell,k} \in \mathbb{P}_n(\mathbb{S}^d),$$

we have

$$\frac{\|\chi\|_\infty}{\|\chi\|_{L^2}} \leq \frac{\sum_{\ell=0}^{n}\sum_{k=1}^{Z(d,\ell)}|\alpha_{\ell,k}|\|Y_{\ell,k}\|_\infty}{\sqrt{\sum_{\ell=0}^{n}\sum_{k=1}^{Z(d,\ell)}|\alpha_{\ell,k}|^2}} \leq \sqrt{\frac{Z(d,n)}{|\mathbb{S}^d|}Z(d+1,n)},$$

where we used the estimate (4.2.2) on the uniform norm of $Y_{\ell,k}$ and regard $\{\alpha_{\ell,k}\}$ as a vector of size $Z(d+1,n)$. Then we can let

$$\eta = \frac{c(s,d)\tilde{c}}{m^{s/d}}(2n+1)^s\sqrt{\frac{Z(d,n)}{|\mathbb{S}^d|}Z(d+1,n)} \tag{4.4.2}$$

and enforce it to be in $(0,1)$. Thus in this case, with the asymptotic result (4.2.1) of the size of $Z(d,\ell)$, the number m should have a lower bound of order

$$n^{d+\frac{d^2}{s}-\frac{d}{2s}}$$

as $n \to \infty$. Moreover, note that for a fixed degree n, the convergence rate of the term $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$ in the error estimate (4.3.2) in Theorem 4.3.2 with respect to m is $m^{-s/(2d)}$.

4.4.2 Approximation theory of QMC hyperinterpolation

We then show that the QMC hyperinterpolation has a sharper error estimate than the general estimate (4.3.2) in Theorem 4.3.2.

Theorem 4.4.1 *Given $f \in H^s(\mathbb{S}^d) \subset L^2(\mathbb{S}^d)$, let $\mathcal{Q}_n f \in \mathbb{P}_n(\mathbb{S}^d)$ be its QMC hyperinterpolant defined by (4.1.10), where the m-point equal-weight quadrature rule (4.1.6) adopts a QMC design for $H^s(\mathbb{S}^d)$ as quadrature points. Then*

$$\|\mathcal{Q}_n f\|_{L^2} \leq \|f\|_{L^2} + \frac{c'(s,d)}{m^{s/d}}(n+1)^s\|f\|_{H^s}, \tag{4.4.3}$$

where $c'(s,d) > 0$ is some constant depending only on s and d, and

$$\|\mathcal{Q}_n f - f\|_{L^2} \le c''(s,d) \left(n^{-s} + \frac{1}{m^{s/d}} \sqrt{\frac{Z(d+1,n)}{a_n^{(s)}}} \right) \|f\|_{H^s}, \tag{4.4.4}$$

where $c''(s,d) > 0$ is some constant depending only on s and d.

Proof. For $f \in H^s(\mathbb{S}^d)$, we have

$$\begin{aligned}
\|\mathcal{Q}_n f\|_{L^2}^2 = \langle \mathcal{Q}_n f, \mathcal{Q}_n f \rangle &= \left\langle \sum_{j=1}^m w_j f(x_j) G_n(x, x_j), \mathcal{Q}_n f(x) \right\rangle \\
&= \sum_{j=1}^m w_j f(x_j) \mathcal{Q}_n f(x_j) \\
&\le \int_{\mathbb{S}^d} (\mathcal{Q}_n f) f \mathrm{d}\omega_d + \frac{c(s,d)}{m^{s/d}} \|(\mathcal{Q}_n f) f\|_{H^s} \\
&\le \|f\|_{L^2} \|\mathcal{Q}_n f\|_{L^2} + \frac{c(s,d)\check{c}}{m^{s/d}} \|f\|_{H^s} \|\mathcal{Q}_n f\|_{H^s} \\
&\le \|f\|_{L^2} \|\mathcal{Q}_n f\|_{L^2} + \frac{c(s,d)\check{c}}{m^{s/d}} \|f\|_{H^s} (n+1)^s \|\mathcal{Q}_n f\|_{L^2},
\end{aligned}$$

where the first inequality is due to the integration error (4.2.12) using QMC designs, the second one is due to the Cauchy–Schwarz inequality and Lemma 4.2.2 with $\check{c}$ given there, and the last one is due to Lemma 4.2.1. Hence we have the stability result (4.4.3).

For the error estimate (4.4.4), we have

$$\|\mathcal{Q}_n f - f\|_{L^2} \le \|\mathcal{Q}_n f - \mathcal{P}_n f\|_{L^2} + \|\mathcal{P}_n f - f\|_{L^2},$$

where $\mathcal{P}_n$ is the L^2-orthgonal projection operator (4.1.5). For the term $\|\mathcal{P}_n f - f\|_{L^2}$, we have

$$\begin{aligned}
\|\mathcal{P}_n f - f\|_{L^2}^2 &= \sum_{\ell=n+1}^{\infty} \sum_{k=1}^{Z(d,\ell)} |\langle f, Y_{\ell,k} \rangle|^2 \\
&= \sum_{\ell=n+1}^{\infty} \sum_{k=1}^{Z(d,\ell)} |\langle f, Y_{\ell,k} \rangle|^2 \frac{a_\ell^{(s)}}{a_\ell^{(s)}} \\
&\lesssim n^{-2s} \|f\|_{H^s}^2,
\end{aligned}$$

where we use the asymptotic relation $a_n^{(s)} \asymp (1+n)^{2s}$. For the term $\|\mathcal{Q}_n f - \mathcal{P}_n f\|_{L^2}$, we have

$$\|\mathcal{Q}_n f - \mathcal{P}_n f\|_{L^2}^2 = \sum_{\ell=0}^{n} \sum_{k=1}^{Z(d,\ell)} \left| \langle f, Y_{\ell,k} \rangle_m - \langle f, Y_{\ell,k} \rangle \right|^2$$

and

$$\begin{aligned} \left| \langle f, Y_{\ell,k} \rangle_m - \langle f, Y_{\ell,k} \rangle \right|^2 &\le \left(\frac{c(s,d)}{m^{s/d}} \|f Y_{\ell,k}\|_{H^s} \right)^2 \\ &\le \left(\frac{c(s,d)\check{c}}{m^{s/d}} \|f\|_{H^s} \|Y_{\ell,k}\|_{H^s} \right)^2, \end{aligned}$$

where the first inequality is described by the integration error (4.2.12) using QMC designs, and the second is due to Lemma 4.2.2. Note that

$$\|Y_{\ell,k}\|_{H^s}^2 = \sum_{\ell'=0}^{n} \sum_{k'=1}^{Z(d,\ell)} \frac{1}{a_\ell^{(s)}} |\langle Y_{\ell,k}, Y_{\ell',k'} \rangle|^2 = \frac{1}{a_\ell^{(s)}}.$$

Thus

$$\begin{aligned} \|\mathcal{Q}_n f - \mathcal{P}_n f\|_{L^2}^2 &\le \left(\frac{c(s,d)\check{c}}{m^{s/d}} \|f\|_{H^s} \right)^2 \frac{1}{a_n^{(s)}} \sum_{\ell=0}^{n} \sum_{k=1}^{Z(d,\ell)} 1 \\ &= \left(\frac{c(s,d)\check{c}}{m^{s/d}} \|f\|_{H^s} \right)^2 \frac{Z(d+1,n)}{a_n^{(s)}}, \end{aligned}$$

leading to the error estimate (4.4.4). □

The estimate (4.4.4) consists of two terms, one represents the error of the original hyperinterpolation, and the other is newly introduced in terms of m. In addition to hyperinterpolation, the fully discrete needlet approximation [225] using spherical needlets [152, 153] and using quadrature rules without exactness assumption also has error estimates of this type, see a recent contribution in [33].

Corollary 4.4.2 *If $f \in \mathbb{P}_n(\mathbb{S}^d) \subset H^s(\mathbb{S}^d)$, then $\|\mathcal{P}_n f - f\|_{L^2} = 0$ and*

$$\|\mathcal{Q}_n f - f\|_{L^2} \le \frac{c(s,d)\check{c}}{m^{s/d}} \sqrt{\frac{Z(d+1,n)}{a_n^{(s)}}} \|f\|_{H^s}.$$

Remark 4.4.3 *If the number m of quadrature points has a lower bound of order $(n+1)^d$, then $\|\mathcal{Q}_n f\|_{L^2}$ is stable in the sense of*

$$\|\mathcal{Q}_n f\|_{L^2} \le \|f\|_{L^2} + c'(s,d)\|f\|_{H^s}.$$

Recall from (4.2.7) *that* $a_n^{(s)} \asymp (1+n)^{-2s}$ *and from* (4.2.1) *that*

$$Z(d+1,n) \sim \frac{2}{\Gamma(d+1)} n^d$$

as $n \to \infty$. *Thus if* m *has a lower bound of order*

$$n^{d+\frac{d^2}{2s}},$$

then $\|\mathcal{Q}_n f - f\|_{L^2}$ *does not blow up as* $n \to \infty$. *Moreover, if* m *has a lower bound of order*

$$(n+1)^{d+\varepsilon_1} n^{\frac{d^2}{2s}+\varepsilon_2} \tag{4.4.5}$$

where $\varepsilon_1, \varepsilon_2 > 0$, *then* $\|\mathcal{Q}_n f - f\|_{L^2} \to 0$ *as* $n \to \infty$.

If the QMC hyperinterpolation is regarded as a special case of the unfettered hyperinterpolation, then the expression (4.4.2) on η requires m to have a lower bound of order

$$(2n+1)^{d+\varepsilon_1} n^{\frac{2d^2-d}{2s}+\varepsilon_2} \tag{4.4.6}$$

so that $\eta \to 0$ and hence $\|\mathcal{Q}_n f - f\|_{L^2} \to 0$ as $n \to \infty$. For the same values of ε_1 and ε_2, the order (4.4.6) derived from regarding the QMC hyperinterpolation as a special case of the unfettered hyperinterpolation is unconditionally greater than the order (4.4.5) derived from Theorem 4.4.1, as

$$\frac{d^2}{2s} < \frac{d^2}{s} - \frac{d}{2s}$$

holds for any $d \geq 1$. Moreover, as the term $E_n(f)$ in the estimate (4.3.2) in Theorem 4.3.2 also has convergence rate of n^{-s}, what essentially varies the general estimate (4.3.2) and the refined estimate (4.4.4) is the other term in both estimates: the term $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$ in the estimate (4.3.2) and the term $\frac{1}{m^{s/d}}\sqrt{\frac{Z(d+1,n)}{a_n^{(s)}}}\|f\|_{H^s}$ in the refined estimate (4.4.4). For a fixed degree n, we have demonstrated in Section 4.4.1 that the convergence rate of the term in (4.3.2) with respect to m is $m^{-s/(2d)}$, and we can see the convergence rate of the term in (4.4.4) is $m^{-s/d}$.

Corollary 4.4.4 *With the aid of Remark 4.4.3, we know that if* $E_n(f) \lesssim n^{-s}$, *then letting*

$$m \gtrsim (n+1)^d n^{\frac{d^2}{2s}} n^d$$

gives

$$\|\mathcal{Q}_n f - f\|_{L^2} \lesssim n^{-s}.$$

Remark 4.4.5 *For the above results, we assume* $f \in H^s(\mathbb{S}^d)$ *and* $\{x_j\}_{j=1}^m$ *is a QMC*

design for $H^s(\mathbb{S}^d)$. Recall the concept of QMC strength. Suppose $f \in H^{s'}$ and $\{x_j\}_{j=1}^m$ is a QMC design with strength s^*, then s in the above results should be $s = \min\{s', s^*\}$.

4.5 Numerical experiments

4.5.1 Point sets and test functions

Many different sequences of point sets on the sphere have been introduced in the literature. In the following experiments, we use points sets including

- Random scattered points generated by the following MATLAB commands:

```
rvals = 2*rand(m,1)-1;
elevation = asin(rvals);
azimuth = 2*pi*rand(m,1);
% convert to Cartesian coordinates
[x1,x2,x3] = sph2cart(azimuth,elevation,ones(m,1));
```

- Equal area points [173] based on an algorithm given in [130];
- Fekete points which maximize the determinant for polynomial interpolation [204];
- Coulomb energy points, which minimize

$$\sum_{i,j=1}^{m} \frac{1}{\|x_i - x_j\|_2};$$

- Spherical t-designs.

Random scattered points are directly generated in MATLAB, equal area points are generated based on the Recursive Zonal Equal Area (EQ) Sphere Partitioning Toolbox by Leopardi, Fekete points and Coulomb energy points are computed by Womersley in advance and are available on his website[1], and spherical t-designs are generated as the so-called well conditioned spherical t-designs in [7].

Moreover, we consider four kinds of test functions, including

- A polynomial $f_1(x) = (x_1 + x_2 + x_3)^2 \in \mathbb{P}_6(\mathbb{S}^2)$;
- $f_2(x_1, x_2, x_3) := |x_1 + x_2 + x_3| + \sin^2(1 + |x_1 + x_2 + x_3|)$, which is continuous but non-smooth;

[1] Robert Womersley, *Interpolation and Cubature on the Sphere*, http://www.maths.unsw.edu.au/~rsw/Sphere/; accessed in August, 2022.

- The Franke function for the sphere [179, p. 146]

$$\begin{aligned} f_3(x_1,x_2,x_3) := \\ & 0.75\exp(-((9x_1-2)^2)/4-((9x_2-2)^2)/4-((9x_3-2)^2)/4) \\ +&0.75\exp(-((9x_1+1)^2)/49-((9x_2+1))/10-((9x_3+1))/10) \\ +&0.5\exp(-((9x_1-7)^2)/4-((9x_2-3)^2)/4-((9x_3-5)^2)/4) \\ -&0.2\exp(-((9x_1-4)^2)-((9x_2-7)^2)-((9x_3-5)^2)), \end{aligned}$$

  which is in $C^\infty(\mathbb{S}^d)$;

- The sums of six compactly supported Wendland radial basis function [225]

$$f_{4,\sigma} := \sum_{i=1}^{6} \phi_\sigma(z_i - x), \quad \sigma \geq 0,$$

  where $z_1 = [1,0,0]^{\mathrm{T}}$, $z_2 = [-1,0,0]^{\mathrm{T}}$, $z_3 = [0,1,0]^{\mathrm{T}}$, $z_4 = [0,-1,0]^{\mathrm{T}}$, $z_5 = [0,0,1]^{\mathrm{T}}$, and $z_6 = [0,0,-1]^{\mathrm{T}}$. The original Wendland functions

$$\tilde{\phi}_\sigma(r) := \begin{cases} (1-r)_+^2, & \sigma = 0, \\ (1-r)_+^4(4r+1), & \sigma = 1, \\ (1-r)_+^6(35r^2+18r+3)/3, & \sigma = 2, \\ (1-r)_+^8(32r^3+25r^2+8r+1), & \sigma = 3, \\ (1-r)_+^{10}(429r^4+450r^3+210r^2+50r+5)/5, & \sigma = 4, \end{cases}$$

  are defined in [231], where $(r)_+ := \max\{r, 0\}$ for $r \in \mathbb{R}$, and the normalized Wendland functions (test functions below) as defined in [30] are

$$\phi_\sigma(r) := \tilde{\phi}_\sigma\left(\frac{r}{\delta_\sigma}\right), \quad \delta_\sigma := \frac{3(\sigma+1)\Gamma(\sigma+1/2)}{2\Gamma(\sigma+1)}, \quad \sigma \geq 0.$$

  The normalized Wendland functions converge pointwise to a Gaussian as $\sigma \to \infty$, see [60]; moreover, $f_{4,\sigma} \in H^{\sigma+3/2}(\mathbb{S}^d)$, see [129, 154].

4.5.2 Unfettered hyperinterpolation and scattered data

We start with a very interesting example of the unfettered hyperinterpolation with scattered data. As we have discussed in Theorem 4.3.2 and Corollary 4.3.6, the performance (i.e., the L^2 error) of the unfettered hyperinterpolation is heavily dependent on the constant η, and what we need to do is to control this constant. In particular, if the degree n and the number m of quadrature points are fixed, Corollary 4.3.6 suggests that η has a lower bound of order $\sqrt{n^2 \log n/m}$. It is immediate

to see that η is positively correlated to n and negatively to m. Moreover, the term $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$ in the error bound (4.3.2) has a lower bound of order

$$\sqrt{\frac{n^2\log n}{m}+4\sqrt{\frac{n^2\log n}{m}}}.$$

That is, for a given n, the term $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$ has a lower bound of order $m^{-1/4}$.

We first solely investigate the term $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$ that arises as an artifact when the quadrature exactness assumption is discarded and leads to the divergence of the unfettered hyperinterpolation by examining the test function $f_1 \in \mathbb{P}_6(\mathbb{S}^d)$. As $E_n(f_1) = 0$ for all $n \geq 6$, we can focus on this term $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$ by letting $n \geq 6$. The L^2 errors are depicted in Figure 4.1. For each pair of (n, m), we test ten

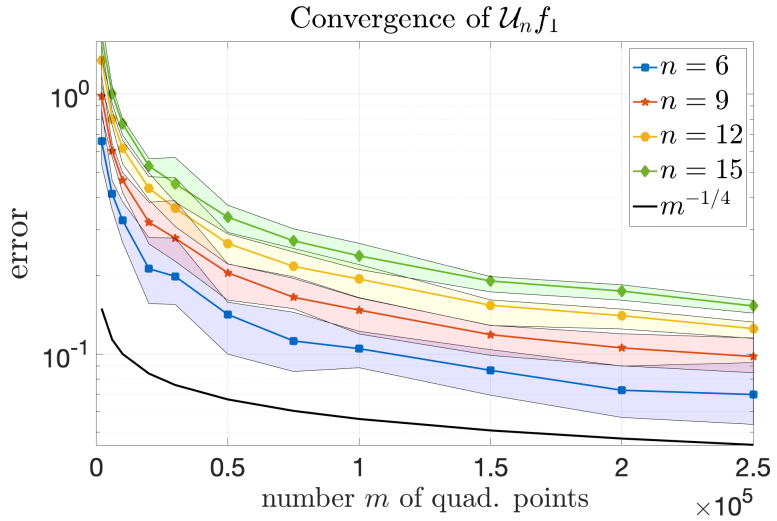


Figure 4.1: Convergence of the unfettered hyperinterpolation in the approximation of f_1.

times and report the average in terms of solid lines with markers; the maximal and minimal errors among these ten tests contribute to the upper and lower bounds of the filled region. We have at least three observations. Firstly, a larger degree n of the unfettered hyperinterpolation, counterintuitively but rigorously asserted by our theory, leads to a larger value of $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$, because Corollary 4.3.6 suggests that η is negatively related to n. Secondly, as n increases, the unfettered hyperinterpolation becomes more stable in the sense that the gap between the maximal and minimal errors among the ten tests for each pair of (n, m) shrinks. This is also asserted by Corollary 4.3.6 that the error bound (4.3.2) is valid with probability exceeding $1 - \bar{c}n^{-\gamma}$. Thirdly, as m increases, the decaying rate of the unfettered hyperinterpolation with respect to m for each n coincides with the rate of $m^{-1/4}$.

This observation is partially covered by our theory that the term $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$ has a lower bound of order $m^{-1/4}$, see discussions in the previous paragraph, and a reasonable conjecture is that there may hold $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2} \asymp m^{-1/4}$.

After characterizing the behavior of the term $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$, we then consider the L^2 error of the unfettered hyperinterpolation. If $E_n(f)$ is not zero, then error estimate (4.3.2) is controlled by two terms, $E_n(f)$ and $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$. We repeat the above procedure for non-polynomial functions f_2 and f_3, and the L^2 errors are displayed in Figure 4.2, in which we only report the average errors. We see that when m is relatively small, the term $\sqrt{\eta^2+4\eta}\|\chi^*\|_{L^2}$ dominates the error bound, so a smaller n leads to a smaller η and hence a smaller error bound; when m is relatively large, η becomes tiny, and the term $E_n(f)$ dominates the error bound, so a larger n leads to a smaller error bound.

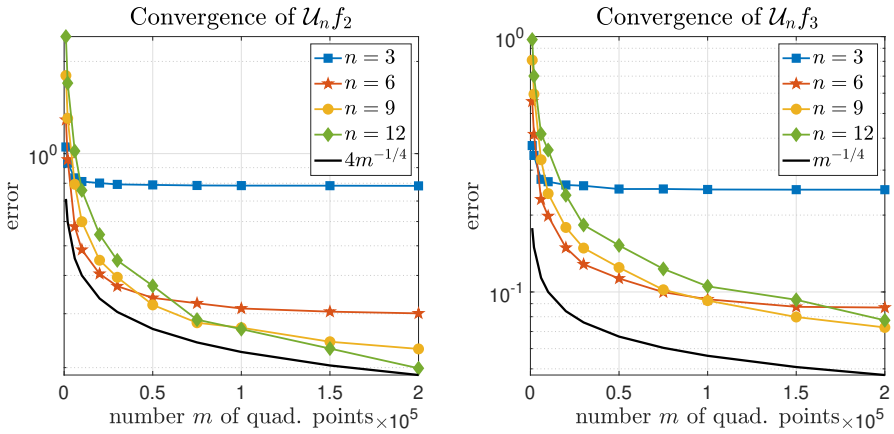


Figure 4.2: Convergence of the unfettered hyperinterpolation in the approximation of f_2 and f_3.

Thus, we may conclude a *rule of thumb* for determining the degree n of the unfettered hyperinterpolation in real-world applications: If the number of samples is limited, then choose a small n; on the other hand, if the samples are relatively sufficient, then choose a large n.

4.5.3 QMC hyperinterpolation and QMC designs

We then investigate the QMC hyperinterpolation, using equal area points, Coulomb energy points, Fekete points, and spherical t-designs. We first consider the approximation of $f_1 \in \mathbb{P}_6$ by the QMC hyperinterpolation using equal area points, and we show that the refined error estimate (4.4.4) in Theorem 4.4.1 is indeed sharper than the estimate (4.3.2) in Theorem 4.3.2. A convergence result of quadrature rules using

equal area points can be found in [111, Section 6.1]. For any $n \geq 6$, we have

$$\|\mathcal{Q}_n f_1 - f_1\|_{L^2} \leq \frac{c''(s,d)}{m^{s/d}} \sqrt{\frac{Z(d+1,n)}{a_n^{(s)}}} \|f_1\|_{H^s}, \tag{4.5.1}$$

in the light of Corollary 4.4.2. As the QMC strength s^* of equal area points is conjectured in [35] to be 2, we may expect the decaying rate of $\|\mathcal{Q}_n f_1 - f_1\|_{L^2}$ with respect to m to be m^{-1} on the 2-sphere $\mathbb{S}^2$. However, from the general framework of the unfettered hyperinterpolation, we can only expect the decaying rate to be $m^{-1/2}$; see discussions at the end of Section 4.4.1. The L^2 errors are depicted in Figure 4.3, which perfectly coincide with these deductions from our theory. We see

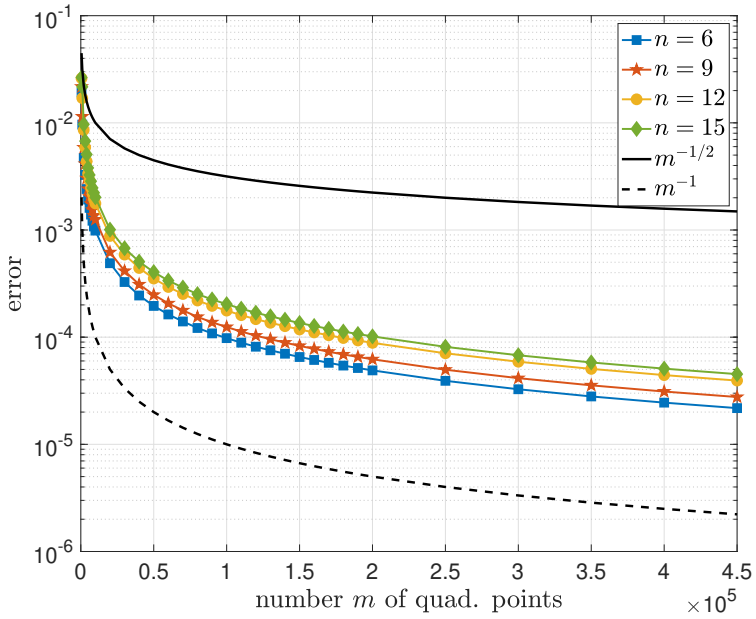


Figure 4.3: Convergence of the QMC hyperinterpolation in the approximation of f_1 using equal area points.

that although the QMC hyperinterpolation can be regarded as a special case in the general framework of unfettered hyperinterpolation, the general estimate may not be sharp. Moreover, we find that a smaller n leads to a smaller error, suggested by the error bound (4.5.1).

We then consider the approximation of the normalized Wendland function $f_{4,2}$ by QMC hyperinterpolation, in which the term $n^{-s}\|f_{4,2}\|_{H^s}$ cannot be ignored. Thus, the terms n^{-s} and $m^{-s/2}$ jointly determine the convergence rate of $\|\mathcal{Q}_n f_{4,2} - f_{4,2}\|_{L^2}$. It is conjectured in [35] that the strength of Fekete points, equal area points, and

Coulomb energy points is 1.5, 2, and 2, respectively. The L^2 errors are depicted in Figure 4.4.

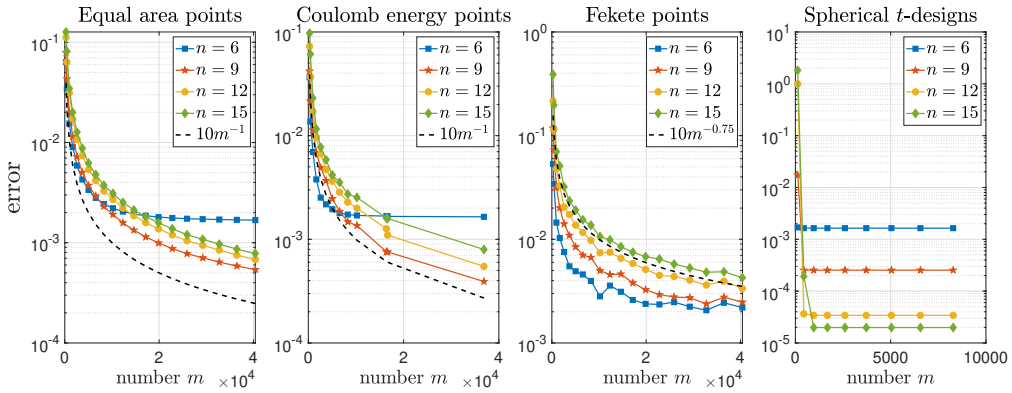


Figure 4.4: Convergence of the QMC hyperinterpolation in the approximation of $f_{4,2}$ using different kinds of point sets.

Similarly to the unfettered hyperinterpolation using scattered data, we see that

$$\frac{1}{m^{s/d}}\sqrt{\frac{Z(d+1,n)}{a_n^{(s)}}}\|f\|_{H^s}$$

dominates the error bound when m is relatively small, so a smaller n leads to a smaller error; and the term $n^{-s}\|f\|_{H^s}$ dominates the error bound when m is relatively large. We observe that each error curve flattens as m increases, and the curve of $n = 6$ is higher than others when m is large enough. Note that each curve corresponds to a fixed degree n. Thus the rule of thumb for determining the degree n of the unfettered hyperinterpolation also applies to the QMC hyperinterpolation. The error curves of the QMC hyperinterpolation using spherical t-designs quickly flatten once the number m of spherical t-designs renders the required quadrature exactness degrees. The convergence of the QMC hyperinterpolation using Fekete points is not monotonic. In light of Womersley's caveat on his website, the non-monotonic convergence is possibly caused by the fact that all computed Fekete points are only approximate local maximizers of the determinant for polynomial interpolation.

We then study the performance of the QMC hyperinterpolation in the approximation of functions with different levels of smoothness. As we mentioned, the normalized Wendland function $f_{4,\sigma}$ belongs to $H^{\sigma+3/2}(\mathbb{S}^d)$. The L^2 errors of the QMC hyperinterpolation of degree $n = 5$ in the approximation of $f_{4,\sigma}$ with $\sigma = 0, 1, \ldots, 4$ are displayed in Figure 4.5, and the degree is intentionally set so small that error curves corresponding to different σ can be distinguished. As we expect, the QMC hyperinterpolation is better in terms of L^2 errors if the function to be approximated

is smoother.

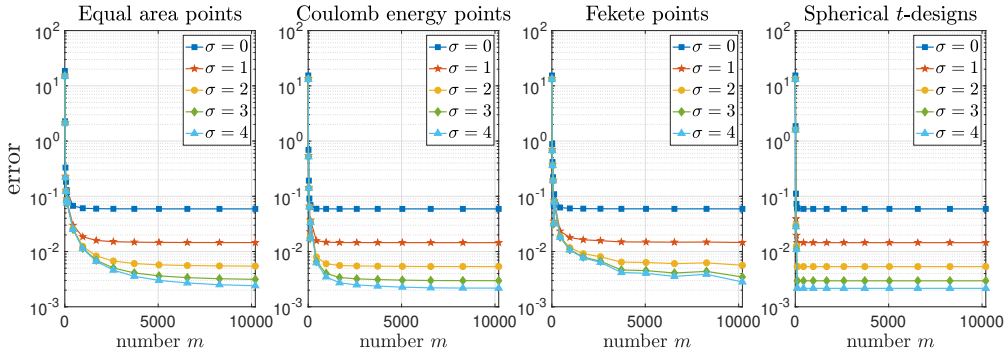


Figure 4.5: Convergence of the QMC hyperinterpolation in the approximation of $f_{4,\sigma}$ with $\sigma = 0, 1, 2, 3, 4$.

Finally, we give a numerical example related to Remark 4.4.3 and Corollary 4.4.4 by considering the approximation of $f_{4,\sigma}$. As we mentioned in Section 4.2.3, to form a spherical t-design, m should satisfy $m \asymp t^d$. Thus, to construct an original hyperinterpolant $\mathcal{L}_n f$ of degree n on the 2-sphere $\mathbb{S}^2$ requires m to be of order n^2, and we have $\|\mathcal{L}_n f - f\|_{L^2} \to 0$ as $n \to \infty$. According to Remark 4.4.3, m should have a lower bound of order

$$(n+1)^{d+\varepsilon_1} n^{\frac{d^2}{2s}+\varepsilon_2}$$

for any $\varepsilon_1, \varepsilon_2 > 0$ to imply $\|\mathcal{Q}_n f - f\|_{L^2} \to 0$ as $n \to \infty$. The L^2 errors with respect to the degree n are depicted in Figure 4.6, and we let $m = (n+1)^2$ and $\lceil (n+1)^2 n^{\frac{2}{\sigma+3/2}} \rceil$.

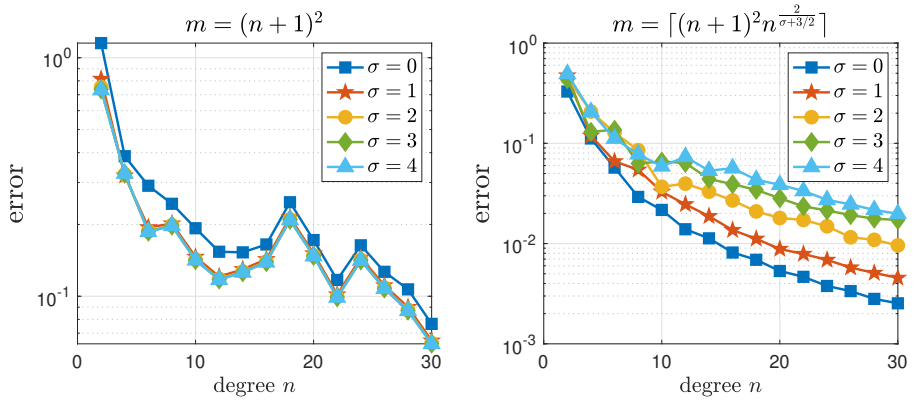


Figure 4.6: Performance of the QMC hyperinterpolation in the approximation of $f_{4,\sigma}$ with $m = (n+1)^2$ and $m = \lceil (n+1)^2 n^{\frac{2}{\sigma+3/2}} \rceil$.

The choice of $m = (n+1)^2$, which suffices to ensure the convergence of the original hyperinterpolation as $n \to \infty$, fails to imply the *monotonic* convergence of the QMC

hyperinterpolation. The choice of

$$m = \lceil (n+1)^2 n^{\frac{2}{\sigma+3/2}} \rceil, \tag{4.5.2}$$

according to our theory, can ensure the convergence of $\mathcal{Q}_n f$ as $n \to \infty$, as shown in Figure 4.6. It may be strange to find that a larger σ leads to a larger error level; this is due to the choice (4.5.2) of m: a larger σ implies a smaller m.

By Corollary 4.4.4, if we let $m \gtrsim (n+1)^2 n^{\frac{2}{s}+2}$, then we can expect $\|\mathcal{Q}_n f - f\|_{L^2} \lesssim n^{-s}$. This corollary is asserted by Figure 4.7, in which we investigate the approximation of $f_{4,2}$ using equal area points. We know that $f_{4,2} \in H^{2+3/2}(\mathbb{S}^d)$, thus we test on five choices of the number m, namely,

$$m = \beta \lceil (n+1)^2 n^{2+\frac{2}{2+3/2}} \rceil$$

with $\beta = 1, 2, 3, 4, 5$. We see that the decaying rates of five choices all coincide with $n^{-(2+3/2)}$. This observation suggests

$$\|\mathcal{Q}_n f_{4,2} - f_{4,2}\|_{L^2} \lesssim n^{-(2+3/2)},$$

and more importantly, successfully verifies our theory on the QMC hyperinterpolation.

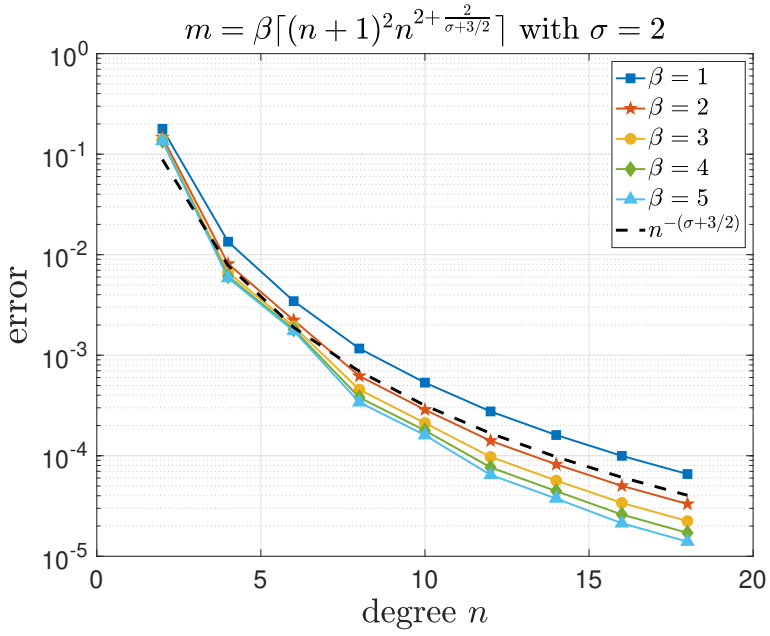


Figure 4.7: Convergence of the QMC hyperinterpolation in the approximation of $f_{4,2}$ with m been a multiple β of $\lceil (n+1)^2 n^{2+\frac{2}{\sigma+3/2}} \rceil$ for $\beta = 1, 2, 3, 4, 5$.

Chapter 5

A spectral method for the Allen–Cahn equation on spheres

Note: *To respect the tradition of numerical PDEs that u^n always denotes the numerical solution at time $t = n\tau$, where τ is the time stepping size, we denote by N the degree of orthogonal projection and hyperinterpolation in this chapter. Also, we focus on $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ instead of $\mathbb{S}^d$ in this chapter. Some preliminaries on $\mathbb{S}^{d-1}$ will be summarized again.*

In this chapter, we propose a novel quadrature-based spectral method for solving the Allen–Cahn equation on spheres, without quadrature exactness. Instead of assumptions on certain exactness degrees, we employ a restricted isometry relation based on the Marcinkiewicz–Zygmund system of quadrature rules to quantify the quadrature error of polynomial integrands. The new method only assumes some conditions on the polynomial degree of numerical solutions to derive the maximum principle and energy stability, and thus it is substantially different from methods in the literature which usually require stringent conditions on the time stepping size, a prior assumption on the Lipschitz continuity of the nonlinear term in the equation, or the L^∞ boundedness of the numerical solutions. Hence, the new method is practically suitable for long-time simulations. Further, we develop an almost sharp maximum principle that allows controllable deviation of numerical solutions from the sharp bound, and show that the new method is energy stable and equivalent to the Galerkin method if the quadrature rule exhibits sufficient exactness degrees. In addition, we propose an energy-stable mixed-quadrature scheme which works well even with randomly sampled initial condition data. We validate the theoretical results about the energy stability and the almost sharp maximum principle by numerical experiments on the 2-sphere $\mathbb{S}^2$.

5.1 Introduction

We are interested in computing smooth solutions of stiff, semi-linear partial differential equations (PDEs) on the unit sphere $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\} \subset \mathbb{R}^d$ with dimension $d \geq 3$ of the form

$$u_t = \mathbf{L}u + \mathbf{N}(u), \quad u(0, x) = u_0(x), \tag{5.1.1}$$

where $u = u(t, x)$ with $(t, x) \in [0, \infty) \times \mathbb{S}^{d-1}$ is a function of time t and spatial variable $x \in \mathbb{S}^{d-1}$, $\mathbf{L}$ is a constant-coefficient linear differential operator, and $\mathbf{N}$ is a constant-coefficient nonlinear differential (or non-differential) operator of lower order. Many applications in science and engineering, especially simulations of the combination of two or more different physical processes, requires smooth solutions of specific cases of (5.1.1). In this chapter, we focus on the Allen–Cahn equation

$$u_t = \nu^2 \Delta u - F'(u), \quad u(0, x) = u_0(x), \tag{5.1.2}$$

where Δ is the Laplace–Beltrami operator, but the main techniques to be developed can be extended to more general cases. Introduced by Allen and Cahn in [5] for describing the process of phase separation in iron alloys, the Allen–Cahn equation (5.1.2) is a reaction-diffusion equation with a linear diffusion term $\nu^2 \Delta u$ and a nonlinear reaction term $F'(u)$. In (5.1.2), $u = u(t, x)$ is a scalar function typically representing the concentration of one of the two metallic components of the alloy. The nonlinear term has the usual double well form of

$$F'(u) = f(u) = u^3 - u$$

with

$$F(u) = \frac{1}{4}(u^2 - 1)^2.$$

We will also focus on the stiff case of $\nu \ll 1$; namely, numerical methods for solving the equation (5.1.2) may be numerically unstable unless the time step size depending on ν is taken to be extremely small. Since stable solutions are always necessary for long-time simulations of the Allen–Cahn equation (5.1.2) as well as many other phase-field models, we aim at developing numerical methods which are stable with large time step sizes.

The Allen–Cahn equation (5.1.2) possesses two intrinsic properties, namely, energy stability and the maximum principle. We consider the energy functional

$$\mathcal{E}(u) := \int_{\mathbb{S}^{d-1}} \left(\frac{1}{2}\nu^2 |\nabla u|^2 + F(u) \right) \mathrm{d}\omega_d, \tag{5.1.3}$$

where $\mathrm{d}\omega_d$ is the surface measure on $\mathbb{S}^{d-1}$, i.e., $\int_{\mathbb{S}^{d-1}} \mathrm{d}\omega_d = |\mathbb{S}^{d-1}|$ denotes the surface area of $\mathbb{S}^{d-1}$. As an L^2-gradient flow, $\mathcal{E}$ is a decreasing function of the time t in the sense of

$$\frac{\mathrm{d}\mathcal{E}(u(t))}{\mathrm{d}t} = -\|u_t\|_{L^2}^2 \leq 0.$$

Therefore, for smooth solutions of the equation (5.1.2), it holds that the energy decay

$$\mathcal{E}(u(t,\cdot)) \leq \mathcal{E}(u(s,\cdot))$$

for any $0 \leq s \leq t < \infty$. Moreover, due to the particular structure of the Allen–Cahn system (5.1.2), we also have the L^∞ maximum principle for the solution to (5.1.2). That is, if the L^∞ norm of u_0 is bounded by some constant, then that of the entire solution should also be bounded by this constant.

For the Allen–Cahn equation (5.1.2) and many related phase-field models, various numerical methods have been proposed to preserve energy stability and the sharp maximum principle. For the literature on ensuring the (modified) energy stability and preserving the maximum principle in numerical simulations of the Allen–Cahn equation (5.1.2) and related phase field models, we refer to, e.g., [19, 25, 29, 57, 76, 77, 78, 80, 85, 86, 87, 88, 132, 193, 207, 241] and references therein. Although preserving both properties is highly desirable for numerical simulations, sometimes only modified energy stability can be analyzed, and some unwanted stringent conditions on the numerical schemes are always introduced, such as small time stepping sizes that depend on ν. These conditions increase the simulation time unfavorably. For finite difference schemes, it was analyzed in [211] that discrete energy stability holds for $0 < \tau \leq 1/2$ and it was further extended to spectral methods for $0 < \tau \leq 0.86$ in [131].

5.1.1 Motivation

In this chapter, we propose a quadrature-based spectral method for the Allen–Cahn equation (5.1.2) on $\mathbb{S}^{d-1}$ by treating all numerical solutions as spherical polynomials of degree N. These polynomials of degree N are allowed to deviate from the sharp bound of the initial data by a controllable discretization error. Our three-fold motivation arise from the practical simulation of the Allen–Cahn equation (5.1.2).

As introduced, many numerical schemes assume a small time stepping size depending on ν. Meanwhile, though there are many methods (see, e.g., [23, 184, 192]) provide unconditional stability results for any time stepping size, their analysis explicitly relies on a Lipschitz assumption on the nonlinearity term or certain *a priori* L^∞ bounds on the numerical solutions. Thus our first motivation for considering

spectral methods is to remove all these stringent and technical conditions and establish a more reasonable analysis by imposing conditions onto the degree N solely. The degree N is independent of the time stepping size. This idea is motivated by a recent work [131], in which an *effective maximum principle* was proposed. This principle is an almost sharp maximum principle allowing the numerical solutions to deviate from the sharp bound by a controllable discretization error without introducing stringent conditions on the numerical scheme. Such an approach sounds practical and reasonable for numerical analysis.

Our second motivation arises from the sampling process in fully discrete practical simulation, which involves the usage of projection operators and numerical integration. There are several methods for discretizing the spatial part of the Allen–Cahn equation (5.1.2) with spectral accuracy. For equations on $\mathbb{S}^{d-1}$ investigated in this chapter, we consider spherical harmonics [15, 151]. One of the main reasons is that spherical harmonics are eigenfunctions of the negative Laplace–Beltrami operator on the sphere, and thus we can avoid the discretization of differential operators.

In [131], the following implicit-explicit spectral scheme

$$\begin{cases} \dfrac{u^{n+1}-u^n}{\tau} = \nu^2 \Delta u^{n+1} - \mathcal{P}_N\left((u^n)^3 - u^n\right), \quad n \geq 0, \\ u^0 = \mathcal{P}_N u_0 \end{cases} \tag{5.1.4}$$

was proposed for the Allen–Cahn equation (5.1.2) with periodic boundary conditions, where $\tau > 0$ is the size of time step, u^n denotes the numerical solution at time $t = n\tau$, and the operator $\mathcal{P}_N$ projects any periodic function to its first N-modes. For the cases of no boundary (such as on compact manifolds) and non-periodic boundary conditions, we can project an L^2 function onto the space $\mathbb{P}_N$ of polynomials of degree at most N. In particular, on $\mathbb{S}^{d-1}$, a convenient L^2-orthonormal basis (with respect to $\mathrm{d}\omega_d$) for $\mathbb{P}_N := \mathbb{P}_N(\mathbb{S}^{d-1})$ is provided by the spherical harmonics $\{Y_{\ell,k} : k = 1, 2, \ldots Z(d,\ell); \ell = 0, 1, 2, \ldots, N\}$ with dimension $\dim \mathbb{P}_N = Z(d+1, N)$, where

$$Z(d,0) = 1, \;\; Z(d,\ell) = (2\ell + d - 2)\frac{\Gamma(\ell + d - 2)}{\Gamma(d-1)\Gamma(\ell+1)} \sim \frac{2}{\Gamma(d-1)}\ell^{d-2} \text{ as } \ell \to \infty.$$

The orthogonal projection is defined as

$$\mathcal{P}_N f = \sum_{\ell=0}^{N} \sum_{k=1}^{Z(d,\ell)} \langle f, Y_{\ell,k}\rangle Y_{\ell,k}, \tag{5.1.5}$$

with the inner product defined as

$$\langle v, z\rangle := \int_{\mathbb{S}^{d-1}} vz \mathrm{d}\omega_d. \tag{5.1.6}$$

The scheme (5.1.4) is equivalent to the *Galerkin* scheme

$$\left\langle \frac{u^{n+1}-u^n}{\tau}, \chi \right\rangle = \left\langle \nu^2 \Delta u^{n+1}, \chi \right\rangle - \left\langle (u^n)^3 - u^n, \chi \right\rangle \quad \forall \chi \in \mathbb{P}_N. \tag{5.1.7}$$

However, for practical simulations, the inner products in either the Galerkin scheme (5.1.7) or the orthogonal projection operator (5.1.5) occurred in the scheme (5.1.4) should be evaluated by some quadrature rules. For example, an m-point positive-weight spherical quadrature rule takes the form of

$$\sum_{j=1}^{m} w_j g(x_j) \approx \int_{\mathbb{S}^{d-1}} g \mathrm{d}\omega_d, \tag{5.1.8}$$

where quadrature points $x_j \in \mathbb{S}^{d-1}$ and weights $w_j > 0$ for $j = 1, 2, \ldots, m$. For numerical integration on the sphere, we refer the reader to [111]. Thus, we are intrigued to investigate how the effective maximum principle behaves after the evaluation of inner products by some quadrature rules. Such investigation is also critical because the analysis of the scheme (5.1.4) may not properly quantify the actual behavior of the numerical solutions. In the practice of spectral methods, the quadrature rules are always chosen to have the exactness degree of $2N$, that is,

$$\sum_{j=1}^{m} w_j g(x_j) = \int_{\mathbb{S}^{d-1}} g \mathrm{d}\omega_d \quad \forall g \in \mathbb{P}_{2N}. \tag{5.1.9}$$

Our third motivation arises from the following question: What if we do not have full access to the initial data u_0 but only a set of samples $\{u_0(x_j)\}_{j=1}^m$ whose data sites $\{x_j\}_{j=1}^m$ cannot be determined by us? In this case, the quadrature rule (5.1.8) with points $\{x_j\}_{j=1}^m$ might not have the desired exactness (5.1.9), but we need to investigate the behavior of the numerical solutions. On the one hand, such a consideration comes in line with the trend of interest in the numerical analysis community that the necessity of quadrature exactness should be re-accessed, because what matters in practice is the accuracy for integrating non-polynomial functions, see, e.g., [12, 222]. On the other hand, even when the quadrature rule (5.1.8) with exactness can be used, this investigation is still necessary, as explained below. Numerical integration on surfaces is quite different from the Euclidean space. On the one-dimensional Euclidean interval $[-1, 1]$, for example, we have Gauss quadrature rules. When the dimension

increases, we can at least consider tensor products to construct quadrature rules with high exactness degrees. However, the situation is different on surfaces. Let us take the sphere as an example. A spherical t-design, introduced in [69], is a set of points $\{x_j\}_{j=1}^m \subset \mathbb{S}^{d-1}$ with the characterizing property that an equal-weight quadrature rule in these points exactly integrates all polynomials of degree at most t, that is,

$$\frac{|\mathbb{S}^{d-1}|}{m}\sum_{j=1}^{m}\chi(x_j) = \int_{\mathbb{S}^{d-1}}\chi(x)\mathrm{d}\omega_d(x) \quad \forall \chi \in \mathbb{P}_t.$$

Therefore, the quadrature rule (5.1.8) with quadrature points as a spherical $2N$-design satisfies the quadrature exactness assumption. It was even verified in [28] that for each $m \geq ct^{d-1}$ with some positive but unknown constant $c > 0$, there exists a spherical t-design in $\mathbb{S}^{d-1}$ consisting of m points. However, the distribution of a spherical t-designs is still unknown. In order to link spherical t-designs to numerical analysis, the distribution of spherical t-designs is obtained by solving some equivalent optimization problems, and the table of computed spherical t-designs are reserved for further applications in approximation theory and numerical analysis. Some equivalent optimization problems are given in [7, 59, 233]. However, this table only contains the distributions of spherical t-designs of a limited range of t; for large t, it is very time-consuming to solve the corresponding optimization problems. This fact also explains our third motivation: if we do not have the points distribution leading to high-degree quadrature exactness at hand, we should use point distributions without quadrature exactness.

5.1.2 Our Scheme

Consider discretizing the orthogonal projection operator $\mathcal{P}_N$ directly in the scheme (5.1.4) as

$$\mathcal{L}_N f = \sum_{\ell=0}^{N}\sum_{k=1}^{Z(d,\ell)}\langle f, Y_{\ell,k}\rangle_m Y_{\ell,k}, \tag{5.1.10}$$

where

$$\langle v, z\rangle_m := \sum_{j=1}^{m} w_j v(x_j) z(x_j) \tag{5.1.11}$$

is a "discrete version" of the L^2 inner product (5.1.6). This is a fully discrete scheme. Note that the operator (5.1.10) is now always referred to as the hyperinterpolation operator, which was originally introduced by Sloan in [196] (cf. Chapters 2–4). Hence, for the Allen–Cahn equation (5.1.2) on the sphere $\mathbb{S}^{d-1}$, we propose the

following scheme:

$$\begin{cases} \dfrac{u^{n+1}-u^n}{\tau} = \nu^2 \Delta u^{n+1} - \mathcal{L}_N\left((u^n)^3 - u^n\right), \quad n \geq 0, \\ u^0 = \mathcal{L}_N u_0. \end{cases} \tag{5.1.12}$$

Recall that spherical harmonics are eigenfunctions of the negative Laplace–Beltrami operator. The scheme (5.1.12) is already fully discrete because there is no need to discretize the Laplace–Beltrami operator. Moreover, implementing the scheme (5.1.12) involves updating the coefficients of the numerical solution u^n, and it requires only vector-matrix multiplications. During each time evolution from n to $n+1$, we need to evaluate the coefficients of the hyperinterpolant $\mathcal{L}_N((u^n)^3-(u^n))$, which can be accomplished in $\dim \mathbb{P}_N + 2m$ floating point operations (flops). We also need to update the coefficients of u^{n+1}, which can be done in $3\dim \mathbb{P}_N$ flops. Therefore, each time evolution of the scheme (5.1.12) can be achieved in $2\left((m+1)(\dim \mathbb{P}_N)+m\right)$ flops, allowing us to achieve the theoretical benefits of the Galerkin method (5.1.7) at a computational cost comparable to the collocation method.

In [196], the construction of hyperinterpolation relies on the quadrature exactness (5.1.9). However, recent works in [11, 12] has relaxed and even bypassed this assumption. In this chapter, the quadrature exactness (5.1.9) is not a necessary assumption on our scheme; we only make the following three natural and simple assumptions:

Assumption 5.1.1 *For the quadrature rule (5.1.8), we assume that*

(I) *it integrates all constants exactly; namely,* $\sum_{j=1}^m w_j = \int_{\mathbb{S}^{d-1}} \mathrm{d}\omega_d = |\mathbb{S}^{d-1}|$;

(II) $\{(x_j, w_j)\}_{j=1}^m$ *forms a Marcinkiewicz–Zygmund (MZ) system of order 2 with respect to* $\mathbb{P}_N$*; namely, for every* $N \geq 0$ *and* $\chi \in \mathbb{P}_N$*, there exists a constant* $\eta < 1$*, independent of* χ *and* N*, such that*

$$\left| \sum_{j=1}^m w_j \chi(x_j)^2 - \int_{\mathbb{S}^{d-1}} \chi^2 d\omega_d \right| \leq \eta \int_{\mathbb{S}^{d-1}} \chi^2 d\omega_d \quad \forall \chi \in \mathbb{P}_N; \tag{5.1.13}$$

(III) *it converges to* $\int_{\mathbb{S}^{d-1}} g \mathrm{d}\omega_d$ *as* $m \to \infty$ *for all* $g \in C(\mathbb{S}^{d-1})$.

Assumption (I) holds if the quadrature rule (5.1.8) is equal-weight, that is, $w_j = |\mathbb{S}^{d-1}|/m$ for all $j = 1, 2, \ldots, m$, or of the quadrature rule (5.1.8) has exactness degree at least one. If this assumption does not hold, we only need to replace the term $|\mathbb{S}^{d-1}|$ in our theoretical results to $\sum_{j=1}^m w_j$. Assumption (II) is equivalent to the Marcinkiewicz–Zygmund inequality, which has been heavily investigated in

[89, 143, 146]. From a numerical perspective, Assumption (II) merely indicates that the relative error of evaluating the integral of χ^2 via the rule (5.1.8) should be less than one for any $\chi \in \mathbb{P}_N$. Moreover, it should be noted that Assumption (II) implies $m \to \infty$ as $N \to \infty$. Assumption (III) is a natural assumption regarding the performance of quadrature rules.

5.1.3 Outline of the chapter

In the chapter, we investigate the L^∞ stability and energy stability for the scheme (5.1.12), and also explore the effective maximum principle for this scheme. In the next section, we introduce some preliminaries on spherical harmonics and the Sobolev space on the sphere. In Section 5.3, for the scheme (5.1.12) with quadrature rules (5.1.8) only fulfilling Assumption 5.1.1, we establish the L^∞ stability for $0 < \tau < 2$ and effective maximum principle for $0 < \tau \leq 1/2$. In Section 5.4, we demonstrate that if the quadrature exactness (5.1.9) is assumed, then our scheme (5.1.12) is equivalent to a fully discrete Galerkin method and it has discrete energy stability for $0 < \tau \leq 0.86$. Moreover, if the quadrature rule (5.1.8) is assumed to have exactness degree of $4N$, we demonstrate the stability of the original energy (5.1.3). Our theoretical assertions are verified by some numerical experiments on the unit sphere $\mathbb{S}^2$ in Section 5.5.

5.2 Preliminaries

We are concerned with real-valued functions on the sphere $\mathbb{S}^{d-1}$ in the Euclidean space $\mathbb{R}^d$ for $d \geq 3$. For the case of $d = 2$, since $\mathbb{S}^1$ can be regarded as a special case of the one-dimensional torus, we refer the reader to the case of tori in [131].

5.2.1 Geometric properties of point distributions

A critical assumption in Assumption 5.1.1 is that the set of $\{(x_j, w_j)\}_{j=1}^m$ is assumed to form an MZ system of order 2 with respect to $\mathbb{P}_N$. A natural concern is under what conditions can the assumption holds. This assumption is related to the quality of distribution of quadrature points $\mathcal{X}_m := \{x_j\}_{j=1}^m$. We define the *mesh norm* $h_{\mathcal{X}_m}$ of the quadrature point set $\mathcal{X}_m \subset \mathbb{S}^{d-1}$ as

$$h_{\mathcal{X}_m} := \max_{x \in \mathbb{S}^{d-1}} \min_{x_j \in \mathcal{X}_m} \operatorname{dist}(x, x_j),$$

where $\operatorname{dist}(x, y) := \cos^{-1}(x \cdot y)$ is the geodesic distance between $x, y \in \mathbb{S}^{d-1}$. In other words, the mesh norm can be regarded as the geodesic radius of the largest hole in the mesh $\mathcal{X}_m$. Thus, it was investigated in [89, 146] that the Assumption (II) in

Assumption 5.1.1 holds if

$$N \lesssim \frac{\eta}{2h_{\mathcal{X}_m}}. \tag{5.2.1}$$

This assumption holds even when $\mathcal{X}_m$ consists of random points. When the quadrature rule (5.1.8) is equal-weight, it was shown in [126] that if an independent random sample of m points drawn from the distribution ω_d, then there exists a constant $\bar{c} := \bar{c}(\gamma)$ such that the MZ inequality (5.1.13) holds with probability exceeding $1 - \bar{c}N^{-\gamma}$ on the condition of

$$m \geq \bar{c}\frac{N^{d-1}\log N}{\eta^2}.$$

5.2.2 Spherical harmonics and hyperinterpolation

The restriction to $\mathbb{S}^{d-1}$ of a homogeneous and harmonic polynomial of total degree ℓ defined on $\mathbb{R}^d$ is called a *spherical harmonic of degree* ℓ on $\mathbb{S}^{d-1}$. We denote, as usual, by $\{Y_{\ell,k} : k = 1, 2, \ldots, Z(d,\ell)\}$ a collection of L^2-orthonormal real-valued spherical harmonics of exact degree ℓ. Besides, it is well known (see, e.g., [151, pp. 38–39]) that each spherical harmonic $Y_{\ell,k}$ of exact degree ℓ is an eigenfunction of the negative Laplace–Beltrami operator $-\Delta$ for $\mathbb{S}^{d-1}$ with eigenvalue

$$\lambda_\ell := \ell(\ell + d - 2). \tag{5.2.2}$$

The family $\{Y_{\ell,k}\}$ of spherical harmonics forms a complete L^2-orthonormal (with respect to ω_d) system for the Hilbert space $L^2(\mathbb{S}^{d-1})$. Thus, for any $f \in L^2(\mathbb{S}^{d-1})$, it can be represented by a Laplace–Fourier series

$$f(x) = \sum_{\ell=0}^{\infty}\sum_{k=1}^{Z(d,\ell)} \hat{f}_{\ell,k} Y_{\ell,k}(x)$$

with coefficients

$$\hat{f}_{\ell,k} := \langle f, Y_{\ell,k}\rangle = \int_{\mathbb{S}^{d-1}} f(x)Y_{\ell,k}(x)\mathrm{d}\omega_d(x)$$

for $\ell = 0, 1, 2, \ldots$ and $k = 1, 2, \ldots, Z(d,\ell)$.

The space $\mathbb{P}_N := \mathbb{P}_N(\mathbb{S}^{d-1})$ of all spherical polynomials of degree at most N (i.e., the restriction to $\mathbb{S}^{d-1}$ of all polynomials in $\mathbb{R}^d$ of degree at most N) coincides with the span of all spherical harmonics up to (and including) degree N, and its dimension satisfies

$$\dim \mathbb{P}_N = Z(d+1, N) = \mathcal{O}(N^{d-1}).$$

The space $\mathbb{P}_N$ is also a reproducing kernel Hilbert space with the reproducing kernel

$$G_N(x,y) = \sum_{\ell=0}^{N} \sum_{k=1}^{Z(d,\ell)} Y_{\ell,k}(x) Y_{\ell,k}(y) \tag{5.2.3}$$

in the sense that

$$\langle \chi, G_N(\cdot, x) \rangle = \chi(x) \quad \forall \chi \in \mathbb{P}_N(\mathbb{S}^{d-1});$$

see, e.g., [176]. The following lemma, occurred in the proof of Theorem 5.5.2 in [203], plays a critical role in our following analysis.

Lemma 5.2.1 ([203]) *For any given point $x_0 \in \mathbb{S}^{d-1}$, there holds*

$$\|G_N(x_0,\cdot)\|_{L^2}^2 = \frac{Z(d+1,N)}{|\mathbb{S}^{d-1}|}.$$

Given $f \in C(\mathbb{S}^{d-1})$, it is often simpler in practice to express the hyperinterpolant $\mathcal{L}_N f$ using the reproducing kernel $G_N(\cdot,\cdot)$ defined by (5.2.3). By rearranging the summation,

$$\mathcal{L}_N f(x) = \sum_{\ell=0}^{N} \sum_{k=1}^{Z(d,\ell)} \left(\sum_{j=1}^{m} w_j f(x_j) Y_{\ell,k}(x_j) \right) Y_{\ell,k}(x) = \sum_{j=1}^{m} w_j f(x_j) G_N(x, x_j).$$

Lemma 5.2.2 *The norm of the hyperinterpolation operator constructed using quadrature rules (5.1.8) fulfilling Assumption 5.1.1 in the setting of $C(\mathbb{S}^{d-1})$ to $C(\mathbb{S}^{d-1})$ is bounded by*

$$\|\mathcal{L}_N\|_\infty := \sup_{f \in C(\Omega)} \frac{\|\mathcal{L}_N f\|_\infty}{\|f\|_\infty} = \mathcal{O}\left(\sqrt{1+\eta} N^{\frac{d-1}{2}} \right). \tag{5.2.4}$$

Proof. It was derived in [203] that

$$\|\mathcal{L}_N\|_\infty \le |\mathbb{S}^{d-1}|^{1/2} \left(\sum_{j=1}^{m} w_j G_N(x_0, x_j)^2 \right)^{1/2},$$

where $x_0 \in \mathbb{S}^{d-1}$ is a certain point. Since $\{(x_j, w_j)\}_{j=1}^m$ forms an MZ system of order 2 (Assumption II), we have

$$\begin{aligned} \|\mathcal{L}_N\|_\infty \le & |\mathbb{S}^{d-1}|^{1/2} \left((1+\eta) \int_{\mathbb{S}^{d-1}} G_N(x_0, x)^2 \mathrm{d}\omega_d(x) \right)^{1/2} \\ \le & |\mathbb{S}^{d-1}|^{1/2} \sqrt{1+\eta} \|G_N(x_0,\cdot)\|_{L^2} \\ \le & \sqrt{1+\eta} (\dim \mathbb{P}_N)^{1/2} = \mathcal{O}\left(\sqrt{1+\eta} N^{\frac{d-1}{2}} \right), \end{aligned}$$

where in the last inequality we use Lemma 5.2.1. □

Remark 5.2.3 *The following historical note partly explains the impact of discretizing the inner products (5.1.6) via some quadrature rules (5.1.8). The uniform operator norm of $\mathcal{P}_N$ satisfies*

$$\|\mathcal{P}_N\|_\infty \asymp N^{\frac{d-2}{2}},$$

and the case of $\mathbb{S}^2$ ($d = 3$) can be dated back to Gronwall [103]. However, the uniform norm $\|\mathcal{L}_N\|_\infty$ of the hyperinterpolation operator constructed using quadrature rules (5.1.8) with quadrature exactness (5.1.9) is bounded as

$$\|\mathcal{L}_N\|_\infty = \mathcal{O}(n^{\frac{d-1}{2}}).$$

That is, the rate of growth of the uniform norm $\|\mathcal{L}_N\|_\infty$ of the hyperinterpolation operator with quadrature exactness (5.1.9) , as shown in [203] that is worse by a factor of $n^{1/2}$ than the optimal result for $\mathcal{P}_N$. Only for the special case of $d = 3$ and under a mild additional assumption on the quadrature rule (5.1.8), the improved result of $\|\mathcal{L}_N\|_\infty \asymp n^{1/2}$ was achieved in [203].

5.2.3 Sobolev spaces

The study of hyperinterpolation in a Sobolev space setting can be traced back to the work [110] by Hesse and Sloan. We define the the Sobolev space for $s \geq 0$ as the set of all functions $f \in L^2(\mathbb{S}^{d-1})$ whose Laplace–Fourier coefficients satisfy

$$\sum_{\ell=0}^{\infty} \sum_{k=1}^{Z(d,\ell)} (1+\lambda_\ell)^s |\hat{f}_{\ell,k}|^2 < \infty,$$

where λ_ℓ is given as (5.2.2). When $s = 0$, we have $H^0(\mathbb{S}^{d-1}) = L^2(\mathbb{S}^{d-1})$. The norm in $H^s(\mathbb{S}^{d-1})$ is therefore defined as

$$\|f\|_{H^s} := \left(\sum_{\ell=0}^{\infty} \sum_{k=1}^{Z(d,\ell)} (1+\lambda_\ell)^s |\hat{f}_{\ell,k}|^2 \right)^{1/2}.$$

The following lemma is necessary for our analysis, which was first presented in [110].

Lemma 5.2.4 *For any $f \in \mathbb{P}_N$,*

$$\|f\|_{H^s} \leq c_1 N^s \|f\|_{L^2},$$

where $c_1 > 0$ is a constant.

Denote $\mathcal{L}_{>N} := I - \mathcal{L}_N$. The behavior of $\mathcal{L}_{>N}$ in the $\|\cdot\|_\infty$ sense is described as follows.

Lemma 5.2.5 *Given $f \in C(\mathbb{S}^{d-1})$ and $t > \frac{d-1}{2}$, the stability of $\mathcal{L}_{>N}$ as an operator from $C(\mathbb{S}^{d-1})$ to $C(\mathbb{S}^{d-1})$ can be controlled by*

$$\|\mathcal{L}_{>N} f\|_\infty \le (1 + \|\mathcal{L}_N\|_\infty)\, E_N(f) + c_2 \eta N^t \|\chi^*\|_{L^2},$$

where $c_2 > 0$ is a constant only depending on η, $E_N(f) = \inf_{\chi \in \mathbb{P}_N} \|f - \chi\|_\infty$ denotes the best uniform approximation error of f in $\mathbb{P}_N$, $\chi^* \in \mathbb{P}_N$ is the best approximation of f in $\mathbb{P}_N$ such that $\|f - \chi^*\|_\infty = E_N(f)$. Furthermore, we have

$$\|\mathcal{L}_{>N} f\|_\infty \le \left(1 + \|\mathcal{L}_N\|_\infty + c_2 \eta N^t\right) E_N(f) + c_2 \eta N^t \|f\|_\infty. \tag{5.2.5}$$

Proof. For any $\chi \in \mathbb{P}_N$, we have

$$\mathcal{L}_{>N} f = f - \mathcal{L}_N f = f - \chi - \mathcal{L}_N(f - \chi) - (\mathcal{L}_N \chi - \chi),$$

and hence

$$\|f - \mathcal{L}_N f\|_\infty \le \|f - \chi\|_\infty + \|\mathcal{L}_N\|_\infty \|f - \chi\|_\infty + \|\mathcal{L}_N \chi - \chi\|_\infty,$$

Since this holds for arbitrary $\chi \in \mathbb{P}_N$, we have

$$\|\mathcal{L}_N \chi - f\|_\infty \le (1 + \|\mathcal{L}_N\|_\infty)\, E_N(f) + \|\mathcal{L}_N \chi^* - \chi^*\|_\infty.$$

Then we control the term $\|\mathcal{L}_N \chi^* - \chi^*\|_\infty$ with the aid of the Sobolev embedding of $H^t(\mathbb{S}^{d-1})$ into $C(\mathbb{S}^{d-1})$ for any $t > \frac{d-1}{2}$. Thus, by Lemma 5.2.4,

$$\begin{aligned}
\|\mathcal{L}_N \chi^* - \chi^*\|_\infty \lesssim & \|\mathcal{L}_N \chi^* - \chi^*\|_{H^t} \\
\le & c_1 N^t \|\mathcal{L}_N \chi^* - \chi^*\|_{L^2} \\
\le & c_1 N^t \sqrt{\eta^2 + 4\eta} \|\chi^*\|_{L^2} \\
\le & c_2 \eta N^t,
\end{aligned}$$

where we use the fact that

$$\|\mathcal{L}_N \chi - \chi\|_{L^2}^2 \le (\eta^2 + 4\eta) \|\chi\|_{L^2}^2$$

for any $\chi \in \mathbb{P}_N$, which was proved in Lemma 4.3.1 The estimate (5.2.5) is immediately obtained by noting that $\|\chi^*\|_\infty \le \|f\|_\infty + E_N(f)$. □

5.3 L^∞ stability and effective maximum principle

We now study the L^∞ stability and effective maximum principle of the spectral scheme (5.1.12) with quadrature rules (5.1.8) fulfilling Assumption 5.1.1 for the Allen–Cahn equation (5.1.2) on $\mathbb{S}^{d-1} \subset \mathbb{R}^d$. A key observation is that for $f \in H^s(\mathbb{S}^{d-1})$ with $s > \frac{d-1}{2}$, the best approximation error $E_N(f)$ in $\mathbb{P}_N$ can be bounded as

$$E_N(f) \le \frac{c_3(f)}{N^{s-\frac{d-1}{2}}} \|f\|_{H^s},$$

where $c_3(f) > 0$ is some constant depending on f; such an error rate can be obtained by [171] together with the Sobolev embedding into Hölder spaces.

5.3.1 The case of $0 < \tau \le 1/2$

We first consider the case of $0 < \tau \le 1/2$.

Theorem 5.3.1 (L^∞ stability for $0 < \tau \le 1/2$) *Let $0 < \alpha_0 \le 1$, $0 < \tau \le 1/2$, and s_0 be a constant marginally larger than $(d-1)/2$. Assume $u_0 \in H^s(\mathbb{S}^{d-1})$ with $s > d-1$ and $\|u_0\|_\infty \le 1$. If $\eta = \tilde{c}N^{-\varepsilon}$ for any $\tilde{c} \ge 0$ and $\varepsilon > s_0$ and $N \ge N_1 := N_1(\alpha_0, \nu, s, d, u_0, \varepsilon)$, then*

$$\sup_{n \ge 0} \|u^n\|_\infty \le 1 + \alpha_0.$$

Proof. This theorem is proved by induction.

Step 1: Initial data. As $\mathcal{L}_{>N} = I - \mathcal{L}_N$, by Lemmas 5.2.2 and 5.2.5 we have

$$\begin{aligned}
\|\mathcal{L}_N u_0\|_\infty \le& \|u_0\|_\infty + \|\mathcal{L}_{>N} u_0\|_\infty \\
\le& 1 + (1 + \|\mathcal{L}_N\|_\infty) E_N(u_0) + c_2 \eta N^{s_0} \|\chi^*\|_{L^2} \\
\le& 1 + c_3 \left(1 + \sqrt{1+\eta}(\dim \mathbb{P}_N)^{1/2}\right) N^{-(s-\frac{d-1}{2})} \|u_0\|_{H^s} + c_2 \eta N^{s_0} \|\chi^*\|_{L^2} \\
\le& 1 + \tilde{c}_3 (1 + \sqrt{2} N^{\frac{d-1}{2}}) N^{-(s-\frac{d-1}{2})} \|u_0\|_{H^s} + c_2 \tilde{c} N^{-\varepsilon + s_0} \|\chi^*\|_{L^2} \\
\le& 1 + \alpha_0
\end{aligned}$$

if $N \ge N_1'(\alpha_0, s, d, \|u_0\|_{H^s}, \varepsilon)$ is large enough such that

$$\tilde{c}_3 (1 + \sqrt{2} N^{\frac{d-1}{2}}) N^{-(s-\frac{d-1}{2})} \|u_0\|_{H^s} + c_2 \tilde{c} N^{-\varepsilon + s_0} \|\chi^*\|_{L^2} \le \alpha_0, \tag{5.3.1}$$

where $\tilde{c}_3 > 0$ is a constant stemming from $\dim \mathbb{P}_N = \mathcal{O}(N^{d-1})$.

Step 2: Induction. The inductive assumption is

$$\|u^n\|_\infty \leq 1 + \alpha_0.$$

We intend to show

$$u^{n+1} \leq 1 + \alpha_0.$$

Afterwards, repeating the argument for $-u^{n+1}$ gives

$$-(1 + \alpha_0) \leq u^{n+1}.$$

Thus, we have

$$\|u^{n+1}\|_\infty \leq 1 + \alpha_0.$$

Note that the scheme (5.1.12) is equivalent to

$$(1 - \tau\nu^2\Delta)u^{n+1} = u^n + \tau\mathcal{L}_N\left(u^n - (u^n)^3\right).$$

Denote $u^n := 1 + \zeta^n$. Then the inductive assumption implies

$$-(2 + \alpha_0) \leq \zeta^n \leq \alpha_0.$$

For $\zeta^{n+1} := u^{n+1} - 1$, we have

$$\begin{aligned}
(1 - \tau\nu^2\Delta)\zeta^{n+1} &= \zeta^n + \tau\mathcal{L}_N\left((1 + \zeta^n) - (1 + \zeta^n)^3\right) \\
&= \zeta^n + \tau\mathcal{L}_N\left(-2\zeta^n - 3(\zeta^n)^2 - (\zeta^n)^3\right) \\
&= \zeta^n + \tau\left(-2\zeta^n - 3(\zeta^n)^2 - (\zeta^n)^3 - \mathcal{L}_{>N}(-2\zeta^n - 3(\zeta^n)^2 - (\zeta^n)^3)\right) \\
&= (1 - 2\tau)\zeta^n - \tau(\zeta^n)^2(3 + \zeta^n) + \tau\mathcal{L}_{>N}(2\zeta^n + 3(\zeta^n)^2 + (\zeta^n)^3).
\end{aligned}$$

Note that $3 + \zeta^n \geq 1 - \alpha_0 \geq 0$. Since

$$\sup_{1 \leq j \leq n} \|u^j\|_\infty \leq 1 + \alpha_0,$$

we can use the discrete smoothing estimate (cf. [134]) to the following iterated scheme

$$\begin{aligned}
u^{n+1} &= (I - \tau\nu^2\Delta)^{-1}u^n - (I - \tau\nu^2\Delta)^{-1}\tau\mathcal{L}_N(f(u^n)) \\
&=: T_0 u^n - \tau T_0 \mathcal{L}_N(f(u^n)) \\
&= T_0^{J+1}u^{n-J} - \tau\sum_{j=1}^{J} T_0^{j+1}\mathcal{L}_N(f(u^{n-j})) - \tau T_0\mathcal{L}_N(f(u^n)),
\end{aligned}$$

where $T_0 := (I - \tau\nu^2\Delta)^{-1}$, to show that

$$\sup_{1\le j\le n} \|u^j\|_{H^s} \le c_{\nu,u_0,s,d},$$

where $c_{\nu,u_0,s,d}$ is some constant depending only on μ, u_0, s and d. As the following analysis ensures $\sup_{1\le j\le n+1} \|u^j\|_\infty \le 1+\alpha_0$, we also have $\|u^{n+1}\|_{H^s} \le c_{\nu,u_0,s,d}$ in the next iteration. By the maximum principle and Lemmas 5.2.2 and 5.2.5, we have

$$\begin{aligned}
&\max \zeta^{n+1}\\
&\le (1-2\tau)\alpha_0 + \tau \left\|\mathcal{L}_{>N}(2\zeta^n + 3(\zeta^n)^2 + (\zeta^n)^3)\right\|_\infty\\
&\le (1-2\tau)\alpha_0 + \tau \Big[\left(1+\sqrt{1+\eta}N^{\frac{d-1}{2}} + c_2\eta N^{s_0}\right) N^{-(s-\frac{d-1}{2})}\cdot\\
&\quad \|2\zeta^n + 3(\zeta^n)^2 + (\zeta^n)^3\|_{H^s} + c_2\eta N^{s_0}\|2\zeta^n + 3(\zeta^n)^2 + (\zeta^n)^3\|_\infty\Big]\\
&\le (1-2\tau)\alpha_0 + \tau\left(N^{\frac{d-1}{2}-s} + N^{d-1-s} + N^{\frac{d-1}{2}+s_0-s-\varepsilon} + N^{s_0-\varepsilon}\right)\mathrm{const}(\nu,u_0,s,d)\\
&\le \alpha_0
\end{aligned}$$

if $N \ge N_1''(\alpha_0,\nu,s,d,\varepsilon)$ is large enough, which leads to $\max u^{n+1} \le 1+\alpha_0$. Thus this theorem is proved by letting $N_1 = \max\{N_1', N_1''\}$. □

Remark 5.3.2 *The situation* (5.3.1) *in our proof requires* $\alpha_0 > 0$. *This requirement stems from the fact that the hyperinterpolation operator does not preserve the sharp uniform bound. That is,* $\|u_0\|_\infty \le 1$ *does not* necessarily *imply* $\|\mathcal{L}_N u_0\|_\infty \le 1$; *see Remark 5.2.3. In the following theorems, we may directly assume that* $\|\mathcal{L}_N u_0\|_\infty \le 1+\alpha_0$ *for* $0 \le \alpha_0 \le 1$. *Then* $\alpha_0 = 0$ *is possible because spectral error term brought by hyperinterpolation has been explicitly recorded in such an assumption.*

Theorem 5.3.3 (Effective maximum principle for $0 < \tau \le 1/2$**)** *Let* $0 < \tau \le 1/2$ *and* s_0 *be a constant marginally larger than* $(d-1)/2$. *Assume* $u_0 \in H^s(\Omega)$ *with* $s > d-1$ *and* $\|u^0\|_\infty \le 1+\alpha_0$ *for some* $0 \le \alpha_0 \le 1$. *If* $\eta = \tilde{c}N^{-\varepsilon}$ *for any* $\tilde{c} \ge 0$ *and* $\varepsilon > s_0$ *and* $N \ge N_2 := N_2(\nu,s,d,u_0,\varepsilon)$, *then for any* $n \ge 1$,

$$\|u^n\|_\infty \le 1 + \theta^n\alpha_0 + \frac{1-\theta^n}{1-\theta}\tau C_{\nu,u_0,s,d}\left(\sqrt{1+\eta}N^{d-1-s} + \eta N^{s_0+\frac{d-1}{2}-s} + \eta N^{s_0}\right), \tag{5.3.2}$$

where $\theta = 1-2\tau$, *and* $C_{\nu,u_0,s,d} > 0$ *is a constant depending on* ν, u_0, s, *and* d. *Consequently,*

$$\limsup_{n\to\infty} \|u^n\|_\infty \le 1 + \frac{1}{2}C_{\nu,u_0,s,d}\left(\sqrt{1+\eta}N^{d-1-s} + \eta N^{s_0+\frac{d-1}{2}-s} + \eta N^{s_0}\right)$$

and

$$\limsup_{N\to\infty} \|u^n\|_\infty \le 1 + \theta^n \alpha_0.$$

Proof. By the inductive step in the proof of Theorem 5.3.1, there exists $N_2 := N_2(\nu, s, d, u_0, \varepsilon) > 0$ such that for any $N \ge N_2$, we have the weakest estimate $\sup_{n\ge 0} \|u^n\|_\infty \le 2$. Denote $u^n := 1 + \zeta^n$ and define $\alpha_n := \max \zeta^n$. Then by repeating the procedure in the proof of Theorem 5.3.1, we have

$$\alpha_{n+1} \le (1-2\tau)\alpha_n + \tau C_{\nu,u_0,s,d}\left(\sqrt{1+\eta} N^{d-1-s} + \eta N^{s_0+\frac{d-1}{2}-s} + \eta N^{s_0}\right)$$

where the constant $C_{\nu,u_0,s,d} > 0$ depends on ν, u_0, s, and d. Similar estimate also holds for $\tilde{\alpha}_n := \max(-1 - u^n)$. Thus for $\theta = 1 - 2\tau$, iterating in n then gives the effective maximum principle (5.3.2). Letting $n \to \infty$ and $N \to \infty$ leads to both limit cases, respectively. □

5.3.2 When the step size τ exceeds $1/2$

We now consider the case of $1/2 < \tau < 2$, with the aid of a prototype iterative system investigated in [131, Lemma 3.3].

Lemma 5.3.4 (Prototype iterative system for the maximum principle) *Let $0 < \tau < 2$ and $p(x) = (1+\tau)x - \tau x^3$. Consider the recurrent relation*

$$\alpha_{n+1} := \max_{|x|\le\alpha_n} |p(x)| + \zeta, \quad n \ge 0,$$

where $\zeta > 0$.

1. *Case $0 < \tau \le 1/2$. Let $\alpha_0 = 2$. There exists an absolute constant $\zeta_0 > 0$ sufficiently small such that for all $0 \le \zeta \le \zeta_0$, we have $1 \le \alpha_n \le 2$ for all n.*
2. *Case $1/2 < \tau \le 2 - \epsilon_0$ for some $0 < \epsilon_0 \le 1$. Let*

$$\alpha_0 = \frac{1}{2}\left(\frac{(1+\tau)^{3/2}}{\sqrt{3\tau}} \cdot \frac{2}{3} + \sqrt{\frac{2+\tau}{\tau}}\right).$$

Then there exists a constant $\zeta_0 > 0$ depending only on ϵ_0 such that if $0 < \zeta \le \zeta_0$, then for all $n \ge 1$, we have

$$\frac{(1+\tau)^{3/2}}{\sqrt{3\tau}} \cdot \frac{2}{3} + \zeta \le \alpha_n \le \alpha_0.$$

Remark 5.3.5 *For $\tau \geq 2$, such a stability result does not hold; see counterexamples provided in Remark 3.7 and Corollary 3.1 in [131].*

Theorem 5.3.6 (L^∞ stability for $1/2 < \tau < 2$) *Let $1/2 < \tau \leq 2 - \epsilon_0$ for some $0 < \epsilon_0 \leq 1$,*

$$M_0 = \frac{1}{2}\left(\frac{(1+\tau)^{3/2}}{\sqrt{3\tau}} \cdot \frac{2}{3} + \sqrt{\frac{2+\tau}{\tau}}\right),$$

and s_0 be a constant marginally larger than $(d-1)/2$. Assume $u_0 \in H^s(\mathbb{S}^{d-1})$ with $s > d-1$ and $\|u^0\|_\infty \leq M_0$. If $\eta = \tilde{c}N^{-\varepsilon}$ for any $\tilde{c} \geq 0$ and $\varepsilon > s_0$ and $N \geq N_3 := N_3(\epsilon_0, \nu, s, d, u_0, \varepsilon)$, then

$$\sup_{n\geq 0} \|u^n\|_\infty \leq M_0.$$

Remark 5.3.7 *As suggested in [131], the bound M_0 can be replaced with any number*

$$\tilde{M}_0 \in \left(\frac{(1+\tau)^{3/2}}{\sqrt{3\tau}}, \sqrt{\frac{2+\tau}{\tau}}\right). \tag{5.3.3}$$

Correspondingly, N_3 in Theorem 5.3.6 should also depend on $\tilde{M}_0$; or more precisely, on its distance to the end points of the interval in (5.3.3).

Proof. We adopt the same induction setting in the proof of Theorem 5.3.1. Note that

$$\begin{aligned}
&(1-\tau\nu^2\Delta)u^{n+1} \\
&\quad = u^n - \mathcal{L}_N u^n + \mathcal{L}_N\left((1+\tau)u^n - \tau(u^n)^3\right) \\
&\quad = (u^n - \mathcal{L}_N u^n) + \left((1+\tau)u^n - \tau(u^n)^3\right) - \mathcal{L}_{>N}\left((1+\tau)u^n - \tau(u^n)^3\right).
\end{aligned}$$

Recall $p(x) = (1+\tau)x - \tau x^3$. Then by the maximum principle,

$$\begin{aligned}
\|u^{n+1}\|_\infty \leq& \|u^n - \mathcal{L}_N u^n\|_\infty + \|p(u^n)\|_\infty + \|\mathcal{L}_{>N}\left((1+\tau)u^n - \tau(u^n)^3\right)\|_\infty \\
\leq& \|u^n - \mathcal{L}_N u^n\|_\infty + M_0 - \zeta + \|\mathcal{L}_{>N}\left((1+\tau)u^n - \tau(u^n)^3\right)\|_\infty,
\end{aligned}$$

where the estimates for $\|u^n - \mathcal{L}_N u^n\|_\infty$ and $\|\mathcal{L}_{>N}\left((1+\tau)u^n - \tau(u^n)^3\right)\|_\infty$ are similar to that in the proof of Theorem 5.3.3. Then the theorem then follows from Lemma 5.3.4 and induction. □

5.4 Refined results with quadrature exactness and related schemes

In this section, we demonstrate that our scheme (5.1.12) is equivalent to the discrete Galerkin scheme if the quadrature exactness (5.1.9) is assumed. With such an assumption, we can also investigate the energy stability of our scheme (5.1.12), which is not mentioned in Section 5.3.

5.4.1 Discrete Galerkin method

It should be noted that though this chapter only focuses on the Allen–Cahn equation (5.1.2), such equivalence also holds for other semi-linear partial differential equations (5.1.1), namely, $u_t = \mathbf{L}u + \mathbf{N}(u)$. In the spirit of our scheme (5.1.12), we consider the following semi-discrete scheme

$$\frac{u^{n+1} - u^n}{\tau} = \mathbf{L}u^{n+1} + \mathcal{L}_N(\mathbf{N}(u^n)) \tag{5.4.1}$$

for the semi-linear PDE (5.1.1), where $\tau > 0$ is the size of time stepping, and $u^n \in \mathbb{P}_N$ denotes the numerical solution at $t = n\tau$. If the quadrature exactness (5.1.9) is assumed, then the hyperinterpolation operator $\mathcal{L}_N$ is a discrete projection operator in the sense of

$$\langle f - \mathcal{L}_N f, \chi \rangle_m = 0 \quad \forall \chi \in \mathbb{P}_N, \tag{5.4.2}$$

and

$$\mathcal{L}_N \chi = \chi \quad \forall \chi \in \mathbb{P}_N; \tag{5.4.3}$$

both properties were shown in [196]. The scheme (5.4.1) is equivalent to

$$\mathcal{L}_N \left(\frac{u^{n+1} - u^n}{\tau} - \mathbf{L}u^{n+1} - \mathbf{N}(u^n) \right) = \frac{u^{n+1} - u^n}{\tau} - \mathbf{L}u^{n+1} - \mathcal{L}_N(\mathbf{N}(u^n)) = 0,$$

obtained using the linearity of $\mathcal{L}_N$ and the property (5.4.3). Then with the property (5.4.2), we know

$$\left\langle \frac{u^{n+1} - u^n}{\tau} - \mathbf{L}u^{n+1} - \mathbf{N}(u^n), \chi \right\rangle_m = 0 \quad \forall \chi \in \mathbb{P}_N,$$

which is further equivalent to

$$\frac{1}{\tau} \langle u^{n+1} - u^n, \chi \rangle_m = \langle \mathbf{L}u^{n+1}, \chi \rangle_m + \langle \mathbf{N}(u^n), \chi \rangle_m \quad \forall \chi \in \mathbb{P}_N, \tag{5.4.4}$$

the *discrete* Galerkin method for the scheme (5.4.1) on the quadrature points $\mathcal{X}_m$.

Focusing on the Allen–Cahn equation (5.1.2), the above discussion suggests that if the quadrature exactness (5.1.9) is assumed, then our scheme (5.1.12) is equivalent to

$$\frac{1}{\tau}\left\langle u^{n+1}-u^n,\chi\right\rangle_m=\left\langle \nu^2\Delta u^{n+1},\chi\right\rangle_m-\left\langle (u^n)^3-u^n,\chi\right\rangle_m\quad\forall\chi\in\mathbb{P}_N,\tag{5.4.5}$$

with $u^0=\mathcal{L}_N u_0\in\mathbb{P}_N$. The scheme (5.4.5) describes a quadrature-based Galerkin method, and it may be also known as the *qualocation method*, or more precisely, *quadrature-modified collocation method*, firstly investigated by Sloan and Wendland in [195, 201]. The motivation of the qualocation method is to design numerical schemes achieving the theoretical benefits of the Galerkin method at a computational cost comparable to the collocation method.

5.4.2 Refined results

An immediate consequence of the quadrature exactness (5.1.9) is $\eta=0$. Thus we have the following corollary of the theorems in Section 5.3. Note that if $\eta=0$, then N_1, N_2, and N_3 do not necessarily depend on ε.

Corollary 5.4.1 *Consider the scheme (5.1.12) for the Allen–Cahn equation (5.1.2) on $\mathbb{S}^{d-1}$, where the quadrature rule (5.1.8) has exactness degree at least $2N$. Assume $u_0\in H^s(\mathbb{S}^{d-1})$ with $s>d-1$. Then the following holds:*

1. *L^∞ stability for $0<\tau\le 1/2$. Let $0<\alpha_0\le 1$ and $0<\tau\le 1/2$. Assume $\|u_0\|_\infty\le 1$. If $N\ge N_4:=N_4(\alpha_0,\nu,s,d,u_0)$, then*

$$\sup_{n\ge 0}\|u^n\|_\infty\le 1+\alpha_0.$$

2. Effective maximum principle for $0<\tau\le 1/2$. *Let $0<\tau\le 1/2$. Assume $\|u^0\|_\infty\le 1+\alpha_0$ for some $0<\alpha_0\le 1$. If $N\ge N_4':=N_4'(\nu,s,d,u_0)$, then for any $n\ge 1$,*

$$\|u^n\|_\infty\le 1+\theta^n\alpha_0+\frac{1-\theta^n}{1-\theta}\tau C_{\nu,u_0,s,d}N^{d-1-s},$$

where $\theta=1-2\tau$, and $C_{\nu,u_0,s,d}>0$ is a constant depending on ν, u_0, s, and d.

3. *L^∞-stability for $1/2<\tau<2$. Let $1/2<\tau<2-\epsilon_0$ for some $0<\epsilon_0\le 1$, and let*

$$M_0=\frac{1}{2}\left(\frac{(1+\tau)^{3/2}}{\sqrt{3\tau}}\cdot\frac{2}{3}+\sqrt{\frac{2+\tau}{\tau}}\right).$$

Assume $\|u^0\|_\infty \le M_0$. *If* $N \ge N_4'' := N_4''(\epsilon_0, \nu, s, d, u_0)$, *then*

$$\sup_{n\ge 0} \|u^n\|_\infty \le M_0.$$

Remark 5.4.2 *Recall the historical note in Remark 5.2.3. If we consider the Allen–Cahn equation (5.1.2) on* $\mathbb{S}^2$, *then the order of* $\|\mathcal{L}_N\|_\infty$ *can be reduced by* $n^{1/2}$, *and the results in Corollary 5.4.1 can be improved correspondingly.*

We then consider the energy stability of the numerical solutions in the presence of quadrature exactness. Recall that the energy functional $\mathcal{E}(u)$ of u is defined as (5.1.3) and its discrete version can be defined as

$$\tilde{\mathcal{E}}(u) := \sum_{j=1}^{m} w_j \left(\frac{\nu^2}{2} (\nabla u(x_j) \cdot \nabla u(x_j)) + F(u(x_j)) \right), \tag{5.4.6}$$

which is discretized by the quadrature rule (5.1.8).

Lemma 5.4.3 (Energy estimate) *For any* $n \ge 0$, *if the quadrature rule (5.1.8) has exactness degree* $2N$, *then the sequence* $\{u^n\}_{n\ge 0}$ *generated by the scheme (5.1.12) satisfies*

$$\begin{aligned} &\tilde{\mathcal{E}}(u^{n+1}) - \tilde{\mathcal{E}}(u^n) + \left(\frac{1}{\tau} + \frac{1}{2}\right) \sum_{j=1}^{m} w_j (u^{n+1}(x_j) - u^n(x_j))^2 \\ &\quad \le \frac{3}{2} \max\left\{ \|u^n\|_\infty^2, \|u^{n+1}\|_\infty^2 \right\} \sum_{j=1}^{m} w_j (u^{n+1}(x_j) - u^n(x_j))^2, \end{aligned} \tag{5.4.7}$$

where the discrete energy $\tilde{\mathcal{E}}(u)$ *of* u *is given by (5.4.6). Furthermore, if the quadrature rule (5.1.8) has exactness degree* $4N$, *then the sequence* $\{u^n\}_{n\ge 0}$ *generated by the scheme (5.1.12) satisfies*

$$\begin{aligned} &\mathcal{E}(u^{n+1}) - \mathcal{E}(u^n) + \left(\frac{1}{\tau} + \frac{1}{2}\right) \int_{\mathbb{S}^{d-1}} (u^{n+1} - u^n)^2 \mathrm{d}\omega_d \\ &\quad \le \frac{3}{2} \max\left\{ \|u^n\|_\infty^2, \|u^{n+1}\|_\infty^2 \right\} \int_{\mathbb{S}^{d-1}} (u^{n+1} - u^n)^2 \mathrm{d}\omega_d, \end{aligned} \tag{5.4.8}$$

where the energy $\mathcal{E}(u)$ *of* u *is given by (5.1.3).*

Proof. Note that

$$\begin{aligned} \frac{1}{\tau} \int_{\mathbb{S}^{d-1}} (u^{n+1} - u^n)^2 \mathrm{d}\omega_d &= \left\langle \frac{u^{n+1} - u^n}{\tau}, u^{n+1} - u^n \right\rangle \\ &= \left\langle \nu^2 \Delta u^{n+1} - \mathcal{L}_N \left((u^n)^3 - u^n \right), u^{n+1} - u^n \right\rangle \\ &= \nu^2 \left\langle \Delta u^{n+1}, u^{n+1} - u^n \right\rangle - \left\langle \mathcal{L}_N \left(f(u^n) \right), u^{n+1} - u^n \right\rangle. \end{aligned} \tag{5.4.9}$$

For the first term on the right-hand side of (5.4.9), the Green–Beltrami identity suggests

$$\begin{aligned}\nu^2 \left\langle \Delta u^{n+1}, u^{n+1} - u^n \right\rangle &= -\nu^2 \int_{\mathbb{S}^{d-1}} \nabla u^{n+1} \cdot \nabla (u^{n+1} - u^n) \mathrm{d}\omega_d \\ = -\frac{\nu^2}{2} \left(\int_{\mathbb{S}^{d-1}} |\nabla u^{n+1}|^2 \mathrm{d}\omega_d - \int_{\mathbb{S}^{d-1}} |\nabla u^n|^2 \mathrm{d}\omega_d \int_{\mathbb{S}^{d-1}} |\nabla (u^{n+1} - u^n)|^2 \mathrm{d}\omega_d \right).\end{aligned} \tag{5.4.10}$$

Observing that all the integrands in the integrals and inner products (regarded as integrals) in the above expressions (5.4.9) and (5.4.10) are polynomials of degree at most $2N$, these integrals and inner products can be replaced by their discrete versions (5.1.8) and (5.1.11), respectively, with the assumption that the quadrature exactness degree is assumed to be $2N$ or $4N$.

Meanwhile, as

$$F(u^{n+1}) = F(u^n) + f(u^n)(u^{n+1} - u^n) + \frac{1}{2} f'(\xi)(u^{n+1} - u^n)^2,$$

where ξ lies between u^n and u^{n+1}, we then have

$$\begin{aligned}\sum_{j=1}^{m} w_j F(u^{n+1}(x_j)) \le \sum_{j=1}^{m} F(u^n(x_j)) + \left\langle f(u^n), u^{n+1} - u^n \right\rangle_m \\ + \left(\frac{3}{2} \max \left\{ \|u^n\|_\infty^2, \|u^{n+1}\|_\infty^2 \right\} - \frac{1}{2} \right) \sum_{j=1}^{m} w_j (u^{n+1}(x_j) - u^n(x_j))^2\end{aligned} \tag{5.4.11}$$

and

$$\begin{aligned}\int_{\mathbb{S}^{d-1}} F(u^{n+1}) \mathrm{d}\omega_d \le \int_{\mathbb{S}^{d-1}} F(u^n) \mathrm{d}\omega_d + \left\langle f(u^n), u^{n+1} - u^n \right\rangle \\ + \left(\frac{3}{2} \max \left\{ \|u^n\|_\infty^2, \|u^{n+1}\|_\infty^2 \right\} - \frac{1}{2} \right) \int_{\mathbb{S}^{d-1}} (u^{n+1} - u^n)^2 \mathrm{d}\omega_d.\end{aligned} \tag{5.4.12}$$

When the quadrature exactness degree is $2N$, we know from (5.4.11) and the discrete versions of equations (5.4.9) and (5.4.10) that

$$
\begin{aligned}
&\tilde{\mathcal{E}}(u^{n+1}) - \tilde{\mathcal{E}}(u^n) + \frac{\nu^2}{2}\sum_{j=1}^{m} w_j |\nabla(u^{n+1}(x_j) - u^n(x_j))|^2 \\
&+ \left(\frac{1}{\tau} + \frac{1}{2}\right)\sum_{j=1}^{m} w_j (u^{n+1}(x_j) - u^n(x_j))^2 \\
\leq &\frac{3}{2}\max\left\{\|u^n\|_\infty^2, \|u^{n+1}\|_\infty^2\right\}\sum_{j=1}^{m} w_j (u^{n+1}(x_j) - u^n(x_j))^2 \\
&+ \left\langle f(u^n) - \mathcal{L}_N(f(u^n)), u^{n+1} - u^n\right\rangle_m .
\end{aligned}
$$

The property (5.4.2) suggests

$$
\left\langle f(u^n) - \mathcal{L}_N(f(u^n)), u^{n+1} - u^n\right\rangle_m = 0.
$$

Hence the estimate (5.4.7) holds.

When the quadrature exactness degree is $4N$, we know from (5.4.11), (5.4.9), and (5.4.10) that

$$
\begin{aligned}
&\mathcal{E}(u^{n+1}) - \mathcal{E}(u^n) + \frac{\nu^2}{2}\int_{\mathbb{S}^{d-1}} |\nabla(u^{n+1} - u^n)|^2 \mathrm{d}\omega_d \\
&+ \left(\frac{1}{\tau} + \frac{1}{2}\right)\int_{\mathbb{S}^{d-1}} (u^{n+1} - u^n)^2 \mathrm{d}\omega_d \\
\leq &\frac{3}{2}\max\left\{\|u^n\|_\infty^2, \|u^{n+1}\|_\infty^2\right\}\int_{\mathbb{S}^{d-1}} (u^{n+1} - u^n)^2 \mathrm{d}\omega_d \\
&+ \left\langle f(u^n) - \mathcal{L}_N(f(u^n)), u^{n+1} - u^n\right\rangle .
\end{aligned}
\tag{5.4.13}
$$

We have

$$
\left\langle f(u^n) - \mathcal{L}_N(f(u^n)), u^{n+1} - u^n\right\rangle = \left\langle f(u^n) - \mathcal{L}_N(f(u^n)), u^{n+1} - u^n\right\rangle_m ,
$$

because the quadrature exactness degree is $4N$, and by the property (5.4.2) again, we have the estimate (5.4.8). □

Remark 5.4.4 *Lemma 5.4.3 immediately suggests that if*

$$
\frac{1}{\tau} + \frac{1}{2} \geq \frac{3}{2}\sup_{n\geq 0}\|u^n\|_\infty^2,
$$

then

$$
\tilde{\mathcal{E}}(u^{n+1}) \leq \tilde{\mathcal{E}}(u^n)
$$

when the quadrature exactness degree is $2N$*, and*

$$\mathcal{E}(u^{n+1}) \leq \mathcal{E}(u^n)$$

when the quadrature exactness degree is $4N$*.*

Remark 5.4.5 *From the proof of Lemma 5.4.3, we can see that if we do not make the quadrature exactness assumption, the terms* $\left\langle f(u^n) - \mathcal{L}_N(f(u^n)), u^{n+1} - u^n \right\rangle_m$ *and* $\left\langle f(u^n) - \mathcal{L}_N(f(u^n)), u^{n+1} - u^n \right\rangle$ *cannot be guaranteed zero or negative. Thus we cannot claim on the (discrete) energy decay of the numerical solutions generated by (5.1.12). However, in practice, the estimate (5.4.13) may suggest that if we consider a sufficiently large number* m *(depending on* ν*) of quadrature points to construct* $\mathcal{L}_N$ *such that*

$$\frac{\nu^2}{2} \int_{\mathbb{S}^{d-1}} |\nabla(u^{n+1} - u^n)|^2 \mathrm{d}\omega_d \geq \left\langle f(u^n) - \mathcal{L}_N(f(u^n)), u^{n+1} - u^n \right\rangle,$$

one may still have energy stability; we do not rigorously investigate this numerical issue in this chapter.

Theorem 5.4.6 (Energy stability for $0 < \tau \leq 1/2$**)** *Let* $0 < \tau \leq 1/2$*. Assume* $u_0 \in H^s(\mathbb{S}^{d-1})$ *with* $s > d-1$ *and* $\|u_0\|_\infty \leq 1$*. Then there exists* $N_5 := N_5(\nu, s, d, u_0)$ *such that for* $N \geq N_5$*, we have the discrete energy decay*

$$\tilde{\mathcal{E}}(u^{n+1}) \leq \tilde{\mathcal{E}}(u^n), \quad n \geq 0$$

if the quadrature rule (5.1.8) has exactness degree $2N$*, and the energy decay*

$$\mathcal{E}(u^{n+1}) \leq \mathcal{E}(u^n), \quad n \geq 0$$

if the quadrature rule (5.1.8) has exactness degree $4N$*.*

Proof. Let $\alpha_0 = \sqrt{\frac{5}{3}} - 1$ in Corollary 5.4.1. Then there exists $N_5(\nu, s, d, u_0)$ such that for $N \geq N_5$,

$$\sup_{n \geq 0} \|u^n\|_\infty \leq \sqrt{\frac{5}{3}}.$$

Then there clearly holds

$$\frac{1}{\tau} + \frac{1}{2} \geq \frac{5}{2} \geq \frac{3}{2} \sup_{n \geq 0} \|u^n\|_\infty^2,$$

and hence, by Lemma 5.4.3, we have both energy decaying estimates. □

With the aid of Theorem 5.3.6 and Remark 5.3.7, we now derive the energy stability result for $\tau \geq 1/2$. This result is only valid for $1/2 < \tau < \tau_1 \approx 0.86$. Consider the equation

$$\frac{1}{2} + \frac{1}{x} = \frac{3}{2} \cdot \left(\frac{2}{3} \cdot \frac{(1+x)^{3/2}}{\sqrt{3x}} \right)^2.$$

It is easy to check that

$$x = \tau_1 = \frac{1}{2}\left(-2 + (9 - 3\sqrt{6})^{1/3} + (9 + 3\sqrt{6})^{1/3} \right) \approx 0.860018$$

is the unique real-valued solution to this equation. Thus, if

$$1/2 < \tau \leq \tau_1 - \epsilon_0,$$

where $0 < \epsilon_0 \leq 0.1$, then

$$\frac{1}{2} + \frac{1}{\tau} \geq \frac{3}{2} \left(\frac{(1+\tau)^{3/2}}{\sqrt{3\tau}} \cdot \frac{2}{3} + \zeta(\epsilon_0) \right)^2, \tag{5.4.14}$$

where $\zeta(\epsilon_0) > 0$ only depends on ϵ_0. Thus, we have the following theorem.

Theorem 5.4.7 (Energy stability for $1/2 < \tau < \tau_1$) *Let $1/2 < \tau \leq \tau_1 - \epsilon_0$ for some $0 < \epsilon_0 \leq 0.1$, and let*

$$M_1 = \frac{(1+\tau)^{3/2}}{\sqrt{3\tau}} \cdot \frac{2}{3} + \zeta(\epsilon_0),$$

where $\zeta(\epsilon_0)$ is the same as the one in (5.4.14). Assume $u_0 \in H^s(\mathbb{S}^{d-1})$ with $s > d-1$ and $\|u^0\|_\infty \leq M_1$. If $N \geq N_6 := N_6(\tau, \epsilon_0, \nu, s, d, u_0)$, then we have the discrete energy decay

$$\tilde{\mathcal{E}}(u^{n+1}) \leq \tilde{\mathcal{E}}(u^n), \quad n \geq 0$$

if the quadrature rule (5.1.8) has exactness degree $2N$, and the energy decay

$$\mathcal{E}(u^{n+1}) \leq \mathcal{E}(u^n), \quad n \geq 0$$

if the quadrature rule (5.1.8) has exactness degree $4N$.

Proof. With $\eta = 0$, Theorem 5.3.6 and Remark 5.3.7 immediately suggest the L^∞ stability of

$$\sup_{n \geq 0} \|u^n\|_\infty \leq M_1.$$

In the light of the energy estimates in Lemma 5.4.3, it suffices to ensure

$$\frac{1}{2}+\frac{1}{\tau}\geq\frac{3}{2}M_1^2=\frac{3}{2}\left(\frac{(1+\tau)^{3/2}}{\sqrt{3\tau}}\cdot\frac{2}{3}+\eta(\epsilon_0)\right)^2,$$

which is exactly (5.4.14). □

5.4.3 An mixed quadrature-based scheme

Theoretical results in Section 5.3 suggest that our scheme (5.1.12) may not have energy stability if the exactness of the quadrature (5.1.8) is not assumed. Recall that our third motivation for proposing and investigating the scheme (5.1.12) is that we may not have the luxury of obtaining samples of the initial condition from quadrature points that we desire. However, we can consider the following mixed quadrature-based scheme

$$\begin{cases}\dfrac{u^{n+1}-u^n}{\tau}=\nu^2\Delta u^{n+1}-\tilde{\mathcal{L}}_N\left((u^n)^3-u^n\right), & n\geq 0,\\ u^0=\mathcal{L}_N u_0,\end{cases}\tag{5.4.15}$$

where $\mathcal{L}_N$ is constructed by quadrature rules (5.1.8) satisfying Assumption 5.1.1 only and $\tilde{\mathcal{L}}_N$ is the hyperinterpolation operator constructed by quadrature rules with quadrature exactness degree of $2N$ or $4N$. Thus if $u_0\in H^s(\mathbb{S}^{d-1})$ with $s>d-1$, $s_0>\frac{d-1}{2}$, and $\eta=\tilde{c}N^{-\varepsilon}$ for any $\tilde{c}\geq 0$ and $\varepsilon>s_0$, then the performance of the mixed quadrature-based scheme (5.4.15) can also be characterized by Corollary 5.4.1, Theorem 5.4.6, and Theorem 5.3.6. The imposed assumptions only aim to guarantee (5.3.1). Thus, with this mixed quadrature-based scheme, even for a set of scattered data of u_0, it is still possible to generate a sequence of numerical solutions quantified by Corollary 5.4.1.

5.5 Numerical experiments

In this section we present some numerical experiments on the 2-sphere $\mathbb{S}^2\subset\mathbb{R}^3$ to verify the analysis presented in the previous sections. It is worth noting that $|\mathbb{S}^2|=4\pi$. For simplicity, we consider quadrature rules (5.1.8) with equal-weight weights

$$w_j=\frac{4\pi}{m}\quad,j=1,2,\ldots,m.$$

Numerous point sets on the sphere have been introduced in the literature. In our experiments, we use the following points sets including randomly scattered points, equal

area points, Fekete points, Coulomb energy points, and well-conditioned spherical t-designs, with descriptions on these point set provided in Chapter 4. Fekete points and Coulomb energy points are precomputed by R. Womersley and are available on his website[1]. All codes were written by MATLAB R2022a, and all numerical experiments were conducted on a laptop (16 GB RAM, Intel® CoreTM i7–9750H Processor) with macOS Monterey 12.5.

We begin with an experiment that illustrates how the phases are separated using the above-mentioned five different types of quadrature points. We set $\nu = 10^{-1}$ and

$$u(0, x, y, z) = \cos(\cosh(5xz) - 10y) \tag{5.5.1}$$

and solve for u up to time $t = 70$. The numerical solution at times $t = 0, 5, 10, 15, 70$ are shown in Figure 5.1. The initial condition quickly converges to a metastable state $u \approx \pm 1$ (yellow area indicates $u \approx 1$ whereas blue area indicates $u \approx -1$) at time around $t = 10$ (for equal area points, Coulomb energy points, and spherical t-designs) and around $t = 15$ (for random points and Fekete points) and eventually to the stable solution $u = 1$ at around $t = 70$. We note that random points may perform slightly worse than points with certain properties, and the inferior performance of Fekete points, as cautioned by Womersley on his website, may be due to the fact that all computed Fekete points are only approximate local maximizers of the determinant for polynomial interpolation. Nevertheless, this experiment suggests that our proposed practical scheme (5.1.12) is a viable method, and we are confident in verifying our theoretical analysis from the previous sections.

In our second experiment, we aim to test the effective maximum principle and the L^∞ stability of the numerical solutions generated by our scheme (5.1.12) using quadrature rules without exactness. Namely, we verify our analysis in Section 5.3 using random points, equal area points, Fekete points, and Coulomb energy points. The uniform norms $\|u^n\|_\infty$ of the numerical solution u^n to the Allen–Cahn equation (5.1.2) with $\nu = 0.1$ and initial condition (5.5.1) are documented in Figure 5.2. We theoretically demonstrate that if $\tau \leq 1/2$, then the effective maximum principle holds, that is, for any fixed N, the upper bound of $\|u^n\|_\infty$ decreases as time advances. This principle suggests that although $\|u^n\|_\infty$ may backtrack, it eventually decreases. This is verified by the first column of Figure 5.2, in which $\tau = 0.5$ ensures the effective maximum principle. If $1/2 < \tau < 2$, then we know $\|u^n\|_\infty$ is bounded by $\|u^0\|_\infty$, which is illustrated by the second and third columns of Figure 5.2.

[1]Robert Womersley, *Interpolation and Cubature on the Sphere*, http://www.maths.unsw.edu.au/~rsw/Sphere/; accessed in March, 2023.

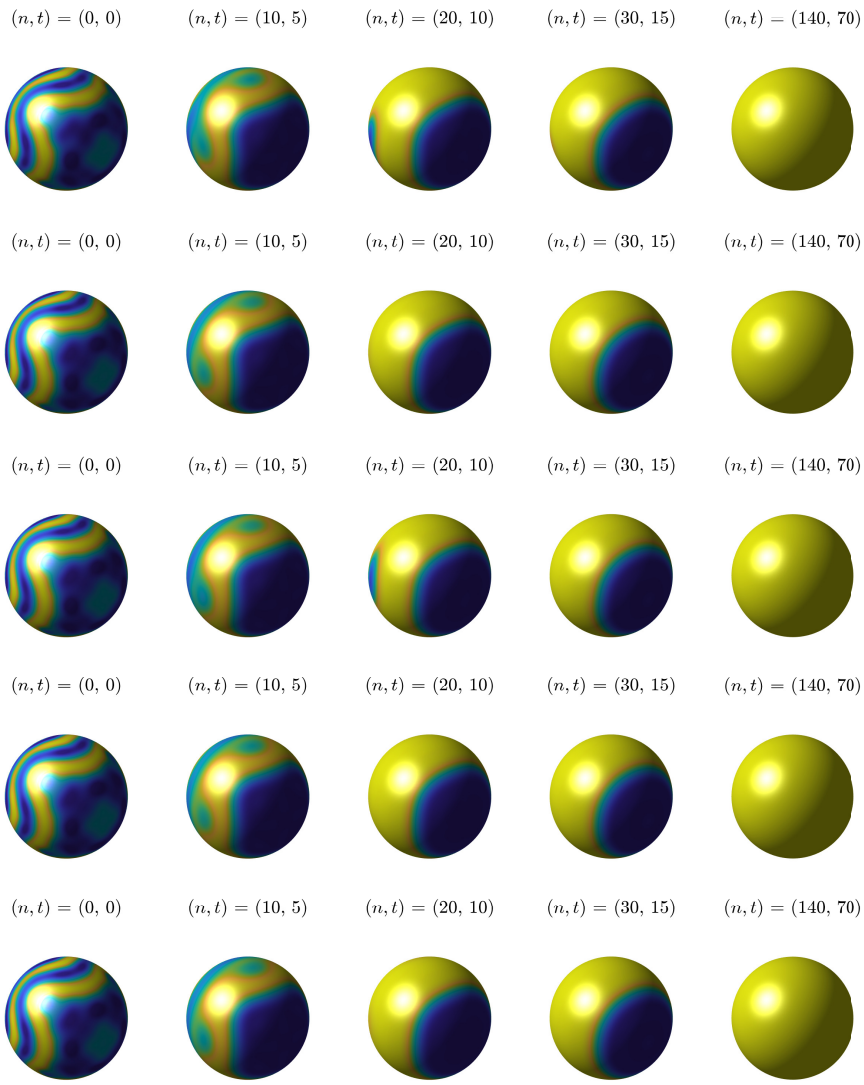


Figure 5.1: Numerical solution to the Allen–Cahn equation (5.1.2) with $\nu = 0.1$ and initial condition (5.5.1) using our scheme (5.1.12) with $\tau = 0.5$, $N = 15$, and different quadrature points. From top row to bottom row: $m = \lfloor 120N^2 \ln N \rfloor = 73,117$ random points; $m = (2N+1)^2 = 961$ equal area points; $m = 961$ Fekete points; $m = 961$ Coulomb energy points; and $m = 961$ spherical $2N$-designs.

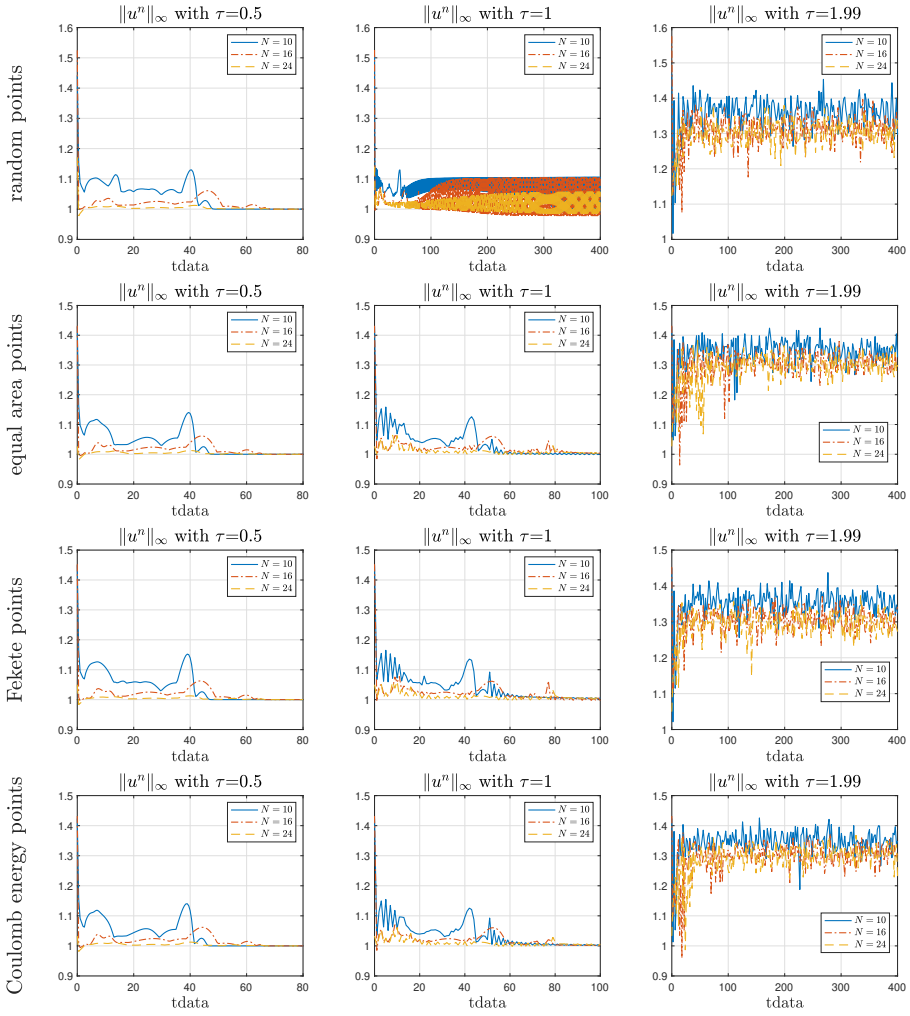


Figure 5.2: Uniform norms of the numerical solution to the Allen–Cahn equation (5.1.2) with $\nu = 0.1$ and initial condition (5.5.1) using our scheme (5.1.12) with $\tau \in \{0.5, 1, 1.99\}$, $N \in \{10, 16, 24\}$, and $m = \lfloor 120N^2 \ln N \rfloor$ for random points and $m = (2N+1)^2$ for equal area points, Fekete points, and Coulomb energy points.

In our third experiment, we investigate the energy decay of our numerical scheme (5.1.12) and test the mixed quadrature-based scheme (5.4.15) discussed in Section 5.4. Since our analysis in Section 5.4 relies on quadrature exactness, we consider spherical t-designs. Recall that for $0 < \tau \leq 0.86$, our theoretical analysis demonstrates that for sufficient large N, the scheme (5.1.12) using quadrature rules of exactness degree $2N$ ensures discrete energy decay $\tilde{\mathcal{E}}(u^{n+1}) \leq \tilde{\mathcal{E}}(u^n)$ for $n \geq 0$, and it has energy decay $\mathcal{E}(u^{n+1}) \leq \mathcal{E}(u^n)$ for $n \geq 0$ if the quadrature rule (5.1.8) has exactness degree $4N$. The energy profiles of the numerical solution u^n to the Allen–Cahn equation (5.1.2) with $\nu = 0.1$ and initial condition (5.5.1) are illustrated in Figure 5.3. Despite the energy dissipation property holding for all cases, it seems that the time step significantly influences the profile of energy evolution.

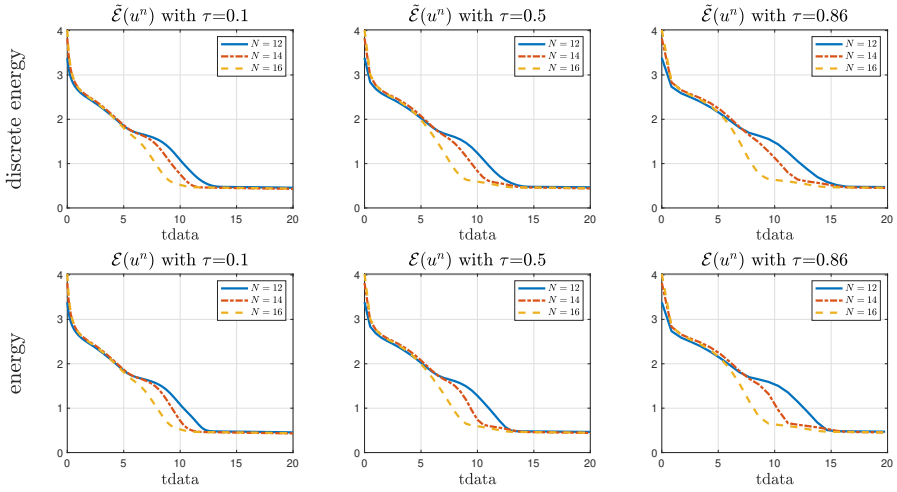


Figure 5.3: Energy profiles of the numerical solution to the Allen–Cahn equation (5.1.2) with $\nu = 0.1$ and initial condition (5.5.1) using our scheme (5.1.12) with $\tau \in \{0.1, 0.5, 0.86\}$ and $N \in \{12, 14, 16\}$. Top row: using spherical $2N$-designs; Bottom row: using spherical $4N$-designs.

It is worth noting that quadrature exactness of degree at least $2N$ is necessary for energy dissipation, as evidenced by the following counterexample. Figure 5.4 records the energy evolution of the numerical solution to the Allen–Cahn equation (5.1.2) with $\nu = 0.01$ and initial condition (5.5.1) using our scheme (5.1.12) with $\tau = 0.86$ and different values of N. If the quadrature exactness is only of degree N, as shown in the top row of Figure 5.4, the discrete energy $\tilde{E}(u^n)$ fails to dissipate, and increasing N does not resolve this issue. On the other hand, if the quadrature exactness degree is $2N$, our refined analysis in Section 5.4 guarantees that discrete energy dissipation always occurs, as demonstrated by the middle row of Figure 5.4. Furthermore, if we consider the mixed quadrature-based scheme (5.4.15) proposed in Section 5.4, where the hyperinterpolation operator with quadrature exactness N is used for projecting

u_0 to u^0 and another hyperinterpolation operator with quadrature exactness $2N$ is used in time evolution, then solutions generated by this scheme exhibit energy dissipation, as shown in the bottom row of Figure 5.4.

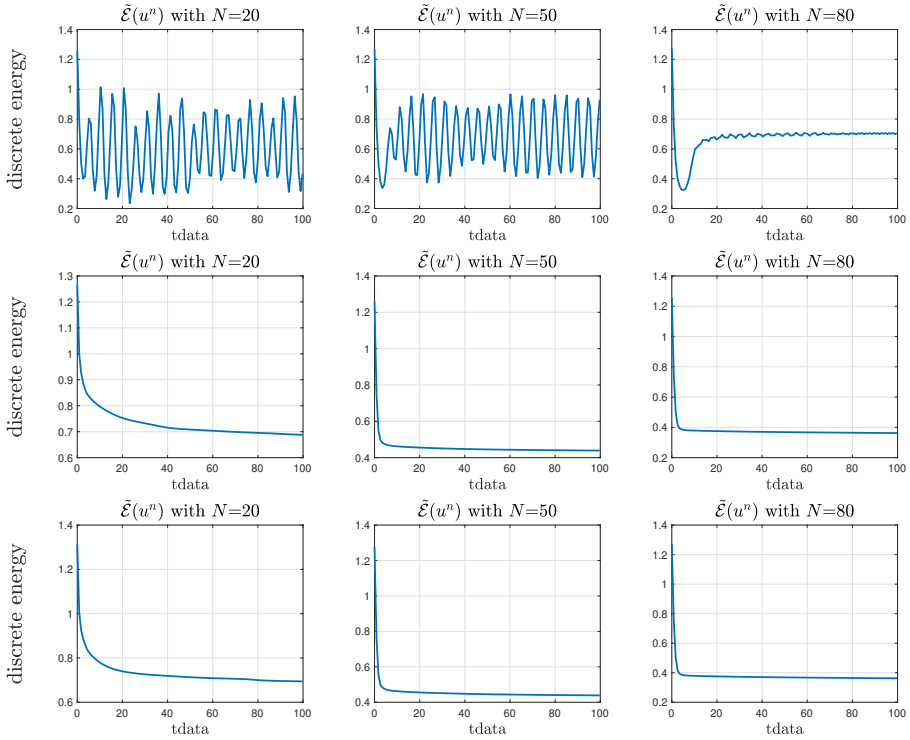


Figure 5.4: Energy profiles of the numerical solution to the Allen–Cahn equation (5.1.2) with $\nu = 0.01$ and initial condition (5.5.1) using our scheme (5.1.12) with $\tau = 0.86$ and $N \in \{20, 50, 80\}$. Top row: quadrature exactness of degree N; Middle row: quadrature exactness of degree $2N$; Bottom row: the mixed quadrature-based scheme (5.4.15).

Chapter 6

The springback model for signal reconstruction

We propose a new penalty, the springback penalty, for constructing models to recover an unknown signal from incomplete and inaccurate measurements. Mathematically, the springback penalty is a weakly convex function. It bears various theoretical and computational advantages of both the benchmark convex ℓ_1 penalty and many of its non-convex surrogates that have been well studied in the literature. We establish the exact and stable reconstruction theory for the reconstruction model using the springback penalty for both sparse and nearly sparse signals, respectively, and derive an easily implementable difference-of-convex algorithm. In particular, we show its theoretical superiority to some existing models with a sharper reconstruction bound for some scenarios where the level of measurement noise is large or the amount of measurements is limited. We also demonstrate its numerical robustness regardless of the varying coherence of the sensing matrix. The springback penalty is particularly favorable for the scenario where the incomplete and inaccurate measurements are collected by coherence-hidden or -static sensing hardware due to its theoretical guarantee of reconstruction with severe measurements, computational tractability, and numerical robustness for ill-conditioned sensing matrices.

6.1 Introduction

Signal reconstruction aims at recovering an unknown signal from its measurements, which are often incomplete and inaccurate due to technical, economical, or physical restrictions. Mathematically, a signal reconstruction problem can be expressed as estimating an unknown $\bar{x} \in \mathbb{R}^n$ from an underdetermined linear system

$$b = A\bar{x} + e, \tag{6.1.1}$$

where $A \in \mathbb{R}^{m\times n}$ is a full row-rank sensing matrix such as a projection or transformation matrix (see, e.g., [37, 43, 44]), $b \in \mathbb{R}^m \setminus \{0\}$ is a vector of measurements, $e \in \mathbb{R}^m$ is some unknown but bounded noise perturbation in

$$\mathcal{B}(\tau) := \{e \in \mathbb{R}^m : \|e\|_2 \le \tau\},$$

and the number m of measurements is considerably smaller than the size n of the signal $\bar{x}$. The set $\mathcal{B}(\tau)$ encodes both the cases of noise-free ($\tau = 0$) and noisy ($\tau > 0$) measurements.

Physically, a signal of interest, or its coefficients under certain transformation, is often sparse (see, e.g., [37]). Hence, it is natural to seek a sparse solution to the underdetermined linear system (6.1.1), though it may have infinitely many solutions. We say that $x \in \mathbb{R}^n$ is *s-sparse* if $\|x\|_0 \le s$, where $\|x\|_0$ counts the number of nonzero entries of x. To find the sparsest solution to (6.1.1), one may consider solving the following minimization problem:

$$\min_{x\in\mathbb{R}^n} \|x\|_0 \quad \text{s.t.} \quad Ax - b \in \mathcal{B}(\tau), \tag{6.1.2}$$

in which $\|x\|_0$ serves as a penalty term of the sparsity, and it is referred to as the ℓ_0 penalty for convenience. Due to the discrete and discontinuous nature of the ℓ_0 penalty, the model (6.1.2) is NP-hard [37]. This means the model (6.1.2) is computationally intractable, and this difficulty has inspired many alternatives to the ℓ_0 penalty in the literature. A fundamental proxy of the model (6.1.2) is the basis pursuit (BP) problem proposed in [58]:

$$\min_{x\in\mathbb{R}^n} \|x\|_1 \quad \text{s.t.} \quad Ax - b \in \mathcal{B}(\tau). \tag{6.1.3}$$

In this convex model,

$$\|x\|_1 := \sum_{i=1}^{n} |x_i|$$

and it is called the ℓ_1 penalty hereafter. Recall that $\|x\|_1$ is the convex envelope of $\|x\|_0$ (see, e.g., [180]), and it induces sparsity most efficiently among all convex penalties (see [37]). The BP problem (6.1.3) has been intensively studied in voluminous papers since the seminal works [42, 43, 74], in which various conditions have been comprehensively explored for the exact reconstruction via the convex model (6.1.3).

The BP problem (6.1.3) is fundamental for signal reconstruction, but its solution may be over-penalized because the ℓ_1 penalty tends to underestimate high-amplitude components of the solution, as analyzed in [83]. Hence, it is reasonable to consider

non-convex alternatives to the ℓ_1 penalty and upgrade the model (6.1.3) to achieve a more accurate reconstruction. In the literature, some non-convex penalties have been well studied, such as the smoothly clipped absolute deviation (SCAD) [83], the capped ℓ_1 penalty [244], the transformed ℓ_1 penalty [142, 243], and the ℓ_p penalty with $0 < p < 1$ [54, 55, 124]. Besides, one particular penalty is the minimax concave penalty (MCP) proposed in [240], and it has been widely shown to be effective in reducing the bias from the ℓ_1 penalty [240]. Moreover, the so-called ℓ_{1-2} penalty has been studied in the literature, e.g. [82, 238, 239], to mention a few. Some of these penalties will be summarized in Section 6.2. In a nutshell, convex penalties are more tractable in the senses of theoretical analysis and numerical computation, while they are less effective for achieving the desired sparsity (i.e., the approximation to the ℓ_0 penalty is less accurate). Non-convex penalties are generally the opposite.

Considering the pros and cons of various penalties, we are motivated to find a weakly convex penalty that can keep some favorable features from both the ℓ_1 penalty and its non-convex alternatives, and the resulting model for signal reconstruction is preferable in the senses of both theoretical analysis and numerical computation. More precisely, we propose the *springback* penalty

$$\mathcal{R}_\alpha^{\mathrm{SPB}}(x) := \|x\|_1 - \frac{\alpha}{2}\|x\|_2^2, \tag{6.1.4}$$

where $\alpha > 0$ is a model parameter, and it should be chosen meticulously. We will show later that a larger α implies a tighter stable reconstruction bound. On the other hand, a too large α may lead to negative values of $\mathcal{R}_\alpha^{\mathrm{SPB}}(x)$. Thus, a reasonable upper bound on α should be considered to ensure the well-definedness of the springback penalty (6.1.4). In the following, we will see that if the matrix A is well-conditioned (e.g., when A is drawn from a Gaussian matrix ensemble), then the requirement on α is quite loose; while if A is ill-conditioned (e.g., A is drawn from an oversampled partial DCT matrix ensemble), then generally the upper bound on α should be better discerned for the sake of designing an algorithm with theoretically provable convergence. We refer to Theorem 6.3.2, Theorem 6.4.1, Section 6.5.2, and Section 6.6.2 for more detailed discussions on the determination of α for the springback penalty (6.1.4) theoretically and numerically. With the springback penalty (6.1.4), we propose the following model for signal reconstruction:

$$\min_{x\in\mathbb{R}^n} \ \mathcal{R}_\alpha^{\mathrm{SPB}}(x) \quad \text{s.t.} \quad Ax - b \in \mathcal{B}(\tau). \tag{6.1.5}$$

Mathematically, the springback penalty (6.1.4) is a weakly convex function, and thus the *springback-penalized model* (6.1.5) can be intuitively regarded as an "average" of the convex BP model (6.1.3) and the mentioned non-convex surrogates. Recall that

a function $f : \mathbb{R}^n \to \mathbb{R}$ is α-*weakly convex* if $x \mapsto f(x) + \frac{\alpha}{2}\|x\|_2^2$ is convex. One advantage of the model (6.1.5) is that various results developed in the literature on weakly convex optimization problems (e.g., [105, 148]) can be used for both theoretical analysis and algorithmic design. Indeed, the weak convexity of the springback penalty (6.1.4) enables us to derive sharper reconstruction results with fewer measurements and to design some efficient algorithms easily.

The rest of this chapter is organized as follows. In the next section, we summarize some preliminaries for further analysis. In Sections 6.3 and 6.4, we establish the exact and stable reconstruction theory of the springback-penalized model (6.1.5) for sparse and nearly sparse signals, respectively. We also theoretically compare the springback penalty (6.1.4) with some other penalties in these two sections. In Section 6.5, we design a difference-of-convex algorithm (DCA) for the springback-penalized model (6.1.5) and study its convergence. Some numerical results are reported in Section 6.6 to verify our theoretical assertions.

6.2 Preliminaries

In this section, we summarize some preliminaries that will be used for further analysis.

6.2.1 A glance at various penalties

In the literature, there are a variety of convex and non-convex penalties. Below we list six of the most important ones, with $x \in \mathbb{R}^n$.

- ◇ The ℓ_1 penalty [37, 58]:

$$\mathcal{R}^{\ell_1}(x) := \|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

- ◇ The elastic net penalty [245]:

$$\mathcal{R}^{\mathrm{EL}}(x) := \|x\|_1 + \frac{\alpha}{2}\|x\|_2^2 = \sum_{i=1}^{n} |x_i| + \frac{\alpha}{2}\sum_{i=1}^{n} |x_i|^2.$$

- ◇ The ℓ_p penalty with parameter $0 < p < 1$ [54, 55]:

$$\mathcal{R}^{\ell_p}(x) := \|x\|_p^p = \sum_{i=1}^{n} |x_i|^p.$$

◇ The transformed ℓ_1 (TL1) with parameter $\beta > 0$ [142, 243]:

$$\mathcal{R}_\beta^{\mathrm{TL1}}(x) := \sum_{i=1}^{n} \frac{(\beta+1)|x_i|}{\beta + |x_i|}.$$

◇ The minimax concave penalty (MCP) with parameter $\mu > 0$ [240]:

$$\mathcal{R}_\mu^{\mathrm{MCP}}(x) := \sum_{i=1}^{n} \phi_\mu^{\mathrm{MCP}}(x_i), \tag{6.2.1}$$

where

$$\phi_\mu^{\mathrm{MCP}}(x_i) = \begin{cases} |x_i| - x_i^2/(2\mu), & |x_i| \le \mu, \\ \mu/2, & |x_i| \ge \mu. \end{cases}$$

◇ The ℓ_{1-2} penalty [82, 239]:

$$\mathcal{R}^{\ell_{1-2}}(x) := \|x\|_1 - \|x\|_2 = \sum_{i=1}^{n} |x_i| - \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

Note that the ℓ_1 penalty is convex, the elastic net penalty is strongly convex, and the others are non-convex.

6.2.2 Relationship among various penalties

For any nonzero vector $x \in \mathbb{R}^n$ and $\alpha > 0$, the springback penalty

$$\mathcal{R}_\alpha^{\mathrm{SPB}}(x) \to \mathcal{R}^{\ell_1}(x)$$

as $\alpha \to 0$. Besides, $\mathcal{R}_\alpha^{\mathrm{SPB}}(x)$ is reduced to the MCP in [240] within the ℓ_∞-ball $\{x \in \mathbb{R}^n : \|x\|_\infty \le \mu\}$ if $\alpha = 1/\mu$. The springback penalty appears to be a resemblance to the ℓ_{1-2} penalty, but their difference is many-sided. For instance, the gradient of $\|x\|_2$ is not defined at the origin.

Figure 6.1 displays some scalar (one-dimensional) penalties, including the ℓ_1 penalty, the $\ell_{0.5}$ penalty, the transformed ℓ_1 penalty with $\beta = 1$, the MCP with $\mu = 0.75$, and the springback penalty with $\alpha = 1/\mu$ and $\alpha = 0.15$. The ℓ_{1-2} penalty is not plotted, as it is none other than zero in the one-dimensional case. To give a better visual comparison, we scale them to attain the point $(1, 1)$. It is shown in Figure 6.1 that the springback penalty is close to the ℓ_1 penalty when $\alpha = 0.15$. The springback penalty with $\alpha = 1/\mu$ coincides with the MCP for $|x| \le \mu$ if we do not scale them. The behavior of the springback penalty for $|x| > \mu$ attracts our

interest because it turns around and heads towards the x-axis. According to Figure 6.1, this behavior is clearer in terms of the thresholding operator corresponding to the proximal mapping of the springback penalty, whose mathematical descriptions are given in Section 6.2.3.

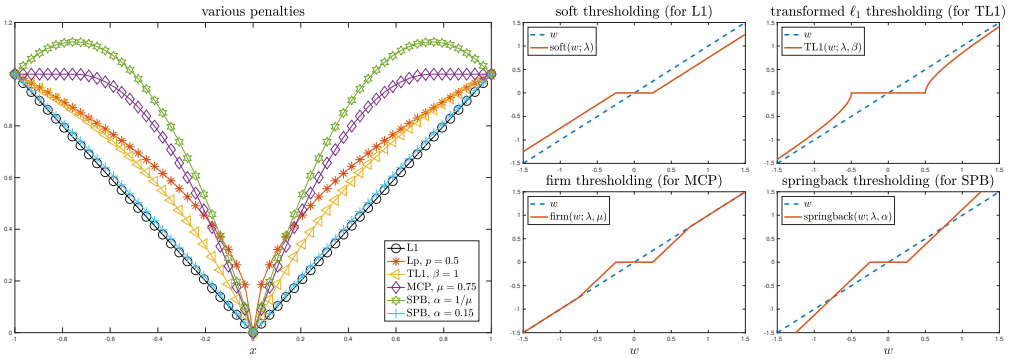


Figure 6.1: Scalar penalties and corresponding thresholding operators (for representing proximal mappings with $\lambda = 0.25$): the ℓ_1 penalty and the soft thresholding operator; the ℓ_p penalty, whose proximal mapping has no closed-form expressions (hence no thresholding operator plotted); the transformed ℓ_1 penalty with $\beta = 1$, whose proximal mapping can be expressed explicitly by a thresholding operator given in [242]; the MCP with $\mu = 0.75$ and the firm thresholding operator; and two springback penalties with $\alpha = 1/\mu$ and $\alpha = 0.15$, and the springback thresholding operator.

As mentioned, the proposed springback penalty (6.1.4) balances the approximation quality of the ℓ_0 penalty and the tractability in analysis and computation, and it is in between the convex and non-convex penalties. More specifically, it is in between the ℓ_1 penalty and the MCP. For any $x \in \mathbb{R}^n$, we can always find a parameter μ for the MCP such that $\|x\|_\infty \leq \mu$ with a resulting penalty in the form of $\|x\|_1 - \|x\|_2^2/(2\mu)$. This penalty inherits the approximation quality of the ℓ_0 penalty from the MCP and the analytical and computational advantages of the ℓ_1 penalty. Inasmuch as this penalty, we consider the more general penalty (6.1.4) in which $1/\mu$ is replaced by a more flexible parameter $\alpha > 0$.

6.2.3 Proximal mappings and thresholding operators

For a function $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}$, as defined in [150], the proximal mapping of $\mathcal{R}$ is defined as

$$\operatorname{prox}_\lambda [\mathcal{R}] (x) := \arg\min_{y \in \mathbb{R}^n} \left\{ \lambda \mathcal{R}(y) + \frac{1}{2}\|y - x\|_2^2 \right\}, \tag{6.2.2}$$

where $\lambda > 0$ is a regularization parameter. In (6.2.2), we slightly abuse the notation "$=$". This mapping takes a vector $x \in \mathbb{R}^n$ and maps it into a subset of $\mathbb{R}^n$, which might be empty, a singleton, or a set with multiple vectors; and the image of y under this mapping is a singleton if the function $\mathcal{R}$ is proper closed and convex [20]. For

a given optimization model, if the proximal mapping of its objective function has a closed-form expression, then usually it is important and necessary to consider how to take advantage of this feature for algorithmic design.

When the proximal mapping of a penalty can be represented explicitly, the closed-form representation is often called a *thresholding operator* or a shrinkage operator in the literature. For example, as analyzed in [242], with the *soft thresholding operator*

$$\mathrm{soft}(w;\lambda) = \mathrm{sgn}(w)\max\{|w|-\lambda, 0\},$$

which has been widely used in various areas such as compressed sensing and image processing, the proximal mapping (6.2.2) of the ℓ_1 penalty can be expressed explicitly by

$$\left[\mathrm{prox}_{\lambda}\left[\mathcal{R}^{\ell_1}\right](x)\right]_i = \mathrm{soft}(x_i;\lambda), \quad i = 1,\ldots,n.$$

The proximal mapping of a non-convex penalty, in general, does not have a closed-form expression; such cases include the ℓ_{1-2} penalty and the ℓ_p penalty with $0 < p < 1$. However, there are some particular non-convex penalties whose proximal mappings can still be represented explicitly. For instance, the transformed ℓ_1 penalty [242] and the MCP [240]. In particular, with the following *firm thresholding operator*

$$\mathrm{firm}(w;\lambda,\mu) = \begin{cases} 0, & |w| \le \lambda, \\ \mathrm{sgn}(w)\frac{\mu(|w|-\lambda)}{\mu-\lambda}, & \lambda \le |w| \le \mu, \\ w, & |w| \ge \mu, \end{cases}$$

which was first proposed in [95], it was further studied in [240] that the proximal mapping (6.2.2) of the MCP can be expressed explicitly by a firm thresholding operator for the case of orthonormal designs. More specifically, the proximal mapping (6.2.2) of the MCP is

$$\left[\mathrm{prox}_{\lambda}\left[\mathcal{R}_{\mu}^{\mathrm{MCP}}\right](x)\right]_i = \mathrm{firm}(x_i;\lambda,\mu), \quad i = 1,\ldots,n.$$

Below, we show that for the springback penalty (6.1.4) with a well chosen α, its proximal mapping can also be expressed explicitly.

Definition 6.2.1 *The* springback thresholding operator *is defined as*

$$\mathrm{springback}(w;\lambda,\alpha) = \begin{cases} 0, & |w| \le \lambda, \\ \mathrm{sgn}(w)\frac{|w|-\lambda}{1-\lambda\alpha}, & |w| > \lambda. \end{cases} \tag{6.2.3}$$

Proposition 6.2.2 *If $1-\lambda\alpha > 0$, then the proximal mapping of the springback penalty (6.1.4) can be represented explicitly as*

$$\left[\mathrm{prox}_{\lambda}\left[\mathcal{R}_{\alpha}^{\mathrm{SPB}}\right](x)\right]_i = \mathrm{springback}(x_i;\lambda,\alpha), \quad i=1,\ldots,n.$$

Proof. When $\mathcal{R}(x)=\mathcal{R}^{\ell_1}(x)$, it follows from (6.2.2) that, for any $z\in\mathbb{R}^n$ satisfying $0\in z-y+\lambda\partial\left(\|z\|_1\right)$, there holds $z_i=\mathrm{soft}(y_i;\lambda)$, i.e., $z=\mathrm{prox}_{\lambda}\left[\mathcal{R}^{\ell_1}\right](y)$. The assumption $1-\lambda\alpha>0$ ensures $\nabla^2\left(\frac{1}{2}\|x-y\|_2^2-\frac{\lambda\alpha}{2}\|x\|_2^2\right)=(1-\lambda\alpha)I$ to be positive definite. Thus, the optimization problem occurred in (6.2.2) is convex. When $\mathcal{R}(x)=\mathcal{R}_{\alpha}^{\mathrm{SPB}}(x)$ in (6.2.2), for any $z\in\mathbb{R}^n$ satisfying the condition $0\in z-y+\lambda\partial(\|z\|_1)-\lambda\alpha z$, which is equivalent to

$$0\in z-\frac{1}{1-\lambda\alpha}y+\frac{\lambda}{1-\lambda\alpha}\partial(\|z\|_1), \tag{6.2.4}$$

we have $z=\mathrm{prox}_{\lambda}\left[\mathcal{R}_{\alpha}^{\mathrm{SPB}}\right](y)$. It also follows from (6.2.4) that

$$z_i=\mathrm{soft}\left(\frac{y_i}{1-\lambda\alpha};\frac{\lambda}{1-\lambda\alpha}\right)=\mathrm{springback}(y_i;\lambda,\alpha).$$

Hence, the assertion is proved. □

Recall that the springback penalty (6.1.4) is a weakly convex function. Its thresholding operator defined in (6.2.3) is also in between the soft and firm thresholding operators. As $\lim_{\mu\to\infty}\mathrm{firm}(w;\lambda,\mu)=\mathrm{soft}(w;\lambda)$, a compromising μ could be large enough such that $|w|\le\mu$ and it reaches a certain compromise between the soft and firm thresholding operators. In this case, we have a particular springback thresholding operator

$$\mathrm{springback}(w;\lambda,1/\mu)=\begin{cases}0, & |w|\le\lambda,\\ \mathrm{sgn}(w)\frac{\mu(|w|-\lambda)}{\mu-\lambda}, & |w|\ge\lambda.\end{cases}$$

If $1/\mu$ is replaced by a more general $\alpha>0$, then the springback thresholding operator (6.2.3) is recovered.

6.2.4 Rationale of the name

Springback is a concept in applied mechanics (see, e.g., [215]). Figure 6.1 gives more explanations for naming (6.1.4) *springback*. With $\lambda=0.25$, Figure 6.1 displays the thresholding operators for $w\in[-1.5,1.5]$, including the soft thresholding operator,

the transformed ℓ_1 thresholding operator with $\beta = 1$, the firm thresholding operator with $\mu = 0.75$, and the springback thresholding operator with $\alpha = 1/\mu$. The transformed ℓ_1 thresholding operator enforces w with $|w| \leq \lambda(\beta+1)/\beta$ to be 0, and then its outputs approach to w as $|w|$ increases. All the other thresholding operators enforce w with $|w| \leq \lambda$ to be 0. For $w \geq \lambda$, the soft thresholding operator subtracts λ from $|w|$ and thus causes the ℓ_1 penalty to underestimate high-amplitude components; the firm thresholding operator's outputs jump from 0 to μ until $|w|$ exceeds μ, afterwards its output is w. For the springback thresholding operator, its outputs jump from 0 to μ until $|w|$ exceeds μ, and afterwards its outputs still keep going along the previous jumping trajectory.

In applied mechanics, spring is related to the process of bending some materials. When the bending process is done, the residual stresses cause the material to *spring back* towards its original shape, so the material must be *over-bent* to achieve the proper bending angle. Note that the soft thresholding operator always underestimates high-amplitude components, and the components $\|x\|_1$ and $-\frac{\alpha}{2}\|x\|_2^2$ in the springback penalty are decoupled. If we deem the soft thresholding operator as a process of over-bending, which stems for the component $\|x\|_1$, then the output of the soft thresholding operator will be sprung back toward w, which is achieved separately in consideration with the component $-\frac{\alpha}{2}\|x\|_2^2$. Such a springback process occurs for both $\lambda \leq |w| \leq \mu$ and $|w| \geq \mu$. The springback behavior is more obvious for those w with larger absolute values, and this coincides with the behavior of the springback penalty in Figure 6.1. That is, once $|x|$ exceeds μ, the penalty turns around and heads towards the x-axis. This process may also be explained as a compensation of the loss of $|w|$ with $|w| \leq \lambda$.

6.3 Springback-penalized model for sparse signal reconstruction

In this section, we focus on the reconstruction of a sparse signal using the springback-penalized model (6.1.5). After reviewing some basic knowledge of compressed sensing, we identify some conditions for exact and robust reconstruction using the springback-penalized model (6.1.5), respectively.

6.3.1 Compressed sensing basics

In the seminal compressed sensing papers [41, 74], reconstruction conditions have been established for the BP model (1.4.3). These conditions rely on the restricted isometry property (RIP) of the *sensing matrix* A, as proposed in [44].

Definition 6.3.1 *For an index set $T \subset \{1, 2, \ldots, n\}$ and an integer s with $|T| \leq s$, the* s-restricted isometry constant *(RIC) of $A \in \mathbb{R}^{m\times n}$ is the smallest $\delta_s \in (0, 1)$ such that*

$$(1 - \delta_s)\|x\|_2^2 \leq \|A_T x\|_2^2 \leq (1 + \delta_s)\|x\|_2^2$$

for all subsets T with $|T| \leq s$ and all $x \in \mathbb{R}^{|T|}$. The matrix A is said to satisfy the s-restricted isometry property *(RIP) with δ_s.*

Denoting by $x^{\rm opt}$ the minimizer of the BP problem (1.4.3), if A satisfies $\delta_{3s} < 3(1 - \delta_{4s}) - 1$, then for an s-sparse $\bar{x}$, one has

$$\|x^{\rm opt} - \bar{x}\|_2 \leq C_s \tau, \tag{6.3.1}$$

where C_s is a constant which may only depend on δ_{4s}. We refer to [42, 43] for more details. If the measurements are noise-free, i.e., $\tau = 0$, then the error bound (6.3.1) implies *exact reconstruction*. Exact reconstruction is guaranteed only in the idealized situation where $\bar{x}$ is s-sparse and the measurements are noise-free. If the measurements are perturbed by some noise, then the bound (6.3.1) is usually referred to as the *robust reconstruction* result with respect to the measurement noise. In more realistic scenarios, we can only claim that $\bar{x}$ is close to an s-sparse vector, and the measurements may also be contaminated. In such cases, we can recover $\bar{x}$ with an error controlled by its distance to s-sparse vectors, and it was proved in [42] that

$$\|x^{\rm opt} - \bar{x}\|_2 \leq C_{1,s}\tau + C_{2,s}\frac{\|\bar{x} - \bar{x}_s\|_1}{\sqrt{s}}, \tag{6.3.2}$$

where $\bar{x}_s$ is the truncated vector corresponding to the s largest values of $\bar{x}$ (in absolute value), and $C_{1,s}$ and $C_{2,s}$ are two constants which may only depend on δ_{4s}. The bound (6.3.2) is usually referred to as the *stable reconstruction* results. Reconstruction conditions for other models with different penalties are usually not as extensive as the BP model (1.4.3). Under the framework of the RIP or some generalized versions, reconstruction theory for the BP model (1.4.3) has been generalized to the ℓ_p-penalized model in [54, 92]. With the *unique representation property* of A, stable reconstruction results for the MCP-penalized model were derived in [235]. We recommend the monograph [93] for a more comprehensive and detailed exhibition on compressed sensing.

6.3.2 Reconstruction guarantee using the springback-penalized model

Still denoting by $x^{\rm opt}$ the minimizer of the springback-penalized model (6.1.5), we have the following exact and robust reconstruction results of the model (6.1.5) for an s-sparse $\bar{x}$.

Theorem 6.3.2 (reconstruction of sparse signals) *Let $\bar{x} \in \mathbb{R}^n$ be an unknown s-sparse vector to be recovered. For a given sensing matrix $A \in \mathbb{R}^{m\times n}$, let $b \in \mathbb{R}^m$ be a vector of measurements from $b = A\bar{x} + e$ with $\|e\|_2 \le \tau$, and let δ_{3s} and δ_{4s} be the $3s$- and $4s$-RIC's of A, respectively. Suppose A satisfies $\delta_{3s} < 3(1-\delta_{4s}) - 1$ and α satisfies*

$$\alpha \le \frac{\sqrt{1-\delta_{4s}}\sqrt{3s} - \sqrt{1+\delta_{3s}}\sqrt{s}}{(\sqrt{1-\delta_{4s}} + \sqrt{1+\delta_{3s}})\|x^{\rm opt}\|_2}, \tag{6.3.3}$$

then the minimizer $x^{\rm opt}$ of the problem (6.1.5) satisfies $x^{\rm opt} = \bar{x}$ when $\tau = 0$; and it satisfies

$$\|x^{\rm opt} - \bar{x}\|_2 \le \frac{\sqrt{2}}{\sqrt{D_1}}\sqrt{\tau} \tag{6.3.4}$$

when $\tau \ge 0$, where

$$D_1 = \frac{\alpha}{2}\frac{\sqrt{1-\delta_{4s}} + \sqrt{1+\delta_{3s}}}{\sqrt{3s} + \sqrt{s}}. \tag{6.3.5}$$

Proof. Let $x^{\rm opt} = \bar{x} + v$, and Λ_0 be the support of $\bar{x}$. It is clear that $v_{\Lambda_0} = x^{\rm opt}_{\Lambda_0} - \bar{x}$ and $v_{\Lambda_0^c} = x^{\rm opt}_{\Lambda_0^c}$. On the one hand, we know that

$$\|x^{\rm opt}\|_1 - \frac{\alpha}{2}\|x^{\rm opt}\|_2^2 \le \|\bar{x}\|_1 - \frac{\alpha}{2}\|\bar{x}\|_2^2.$$

On the other hand, it holds that

$$\begin{aligned}
&\|x^{\rm opt}\|_1 - \frac{\alpha}{2}\|x^{\rm opt}\|_2^2 \\
=&\|\bar{x} + v_{\Lambda_0}\|_1 + \|v_{\Lambda_0^c}\|_1 - \frac{\alpha}{2}\|\bar{x} + v\|_2^2 \\
\ge&\|\bar{x}\|_1 - \|v_{\Lambda_0}\|_1 + \|v_{\Lambda_0^c}\|_1 - \frac{\alpha}{2}\left(\|\bar{x}\|_2^2 + 2\langle\bar{x}, v\rangle + \|v\|_2^2\right).
\end{aligned}$$

Then, we have that

$$\begin{aligned}
\|v_{\Lambda_0^c}\|_1 &\le \|v_{\Lambda_0}\|_1 - \frac{\alpha}{2}\|v\|_2^2 + \alpha\|v\|_2^2 + \alpha\langle\bar{x}, v\rangle \\
&= \|v_{\Lambda_0}\|_1 - \frac{\alpha}{2}\|v\|_2^2 + \alpha\left\langle x^{\rm opt}, v\right\rangle.
\end{aligned}$$

We continue by arranging the indices in Λ_0^c in order of decreasing magnitudes (in absolute value) of $v_{\Lambda_0^c}$, and then dividing Λ_0^c into subsets of size $3s$. Set $\Lambda_0^c = \Lambda_1 \bigcup \Lambda_2 \bigcup \cdots \bigcup \Lambda_\ell$, i.e., Λ_1 contains the indices of the $3s$ largest entries (in absolute value) of $v_{\Lambda_0^c}$, Λ_2 contains the indices of the next $3s$ largest entries (in absolute value) of $v_{\Lambda_0^c}$, and so on. The cardinal number of Λ_ℓ may be less than $3s$. Denoting

$\Lambda_{01} = \Lambda_0 \bigcup \Lambda_1$ and using the RIP of A, we have

$$\begin{aligned}\|Av\|_2 &\geq \|A_{\Lambda_{01}} v_{\Lambda_{01}}\|_2 - \left\|\sum_{i=2}^{\ell} A_{\Lambda_i} v_{\Lambda_i}\right\|_2 \\ &\geq \sqrt{1-\delta_{4s}}\|v_{\Lambda_{01}}\|_2 - \sqrt{1+\delta_{3s}}\sum_{i=2}^{\ell}\|v_{\Lambda_i}\|_2.\end{aligned}$$

As the magnitude of every v_t indexed by $t \in \Lambda_{i+1}$ is less than the average of magnitudes of v_t indexed by $t \in \Lambda_i$, there holds

$$|v_t| \leq \frac{\|v_{\Lambda_i}\|_1}{3s},$$

where $t \in \Lambda_{i+1}$. Then, we have

$$\|v_{\Lambda_{i+1}}\|_2^2 \leq 3s\frac{\|v_{\Lambda_i}\|_1^2}{(3s)^2} = \frac{\|v_{\Lambda_i}\|_1^2}{3s}.$$

Together with $\|v_{\Lambda_0}\|_1 \leq \sqrt{s}\|v_{\Lambda_0}\|_2 \leq \sqrt{s}\|v_{\Lambda_{01}}\|_2$, we have

$$\sum_{i=2}^{\ell}\|v_{\Lambda_i}\|_2 \leq \sum_{i=1}^{\ell-1}\frac{\|v_{\Lambda_i}\|_1}{\sqrt{3s}} \leq \frac{1}{\sqrt{3s}}\|v_{\Lambda_0^c}\|_1 \leq \frac{1}{\sqrt{3s}}\left(\sqrt{s}\|v_{\Lambda_{01}}\|_2 - \frac{\alpha}{2}\|v\|_2^2 + \alpha\left\langle x^{\text{opt}}, v\right\rangle\right).$$

Thus, it holds that

$$\begin{aligned}\|Av\|_2 \geq &\left(\sqrt{1-\delta_{4s}} - \frac{\sqrt{1+\delta_{3s}}\sqrt{s}}{\sqrt{3s}}\right)\|v_{\Lambda_{01}}\|_2 + \frac{\alpha\sqrt{1+\delta_{3s}}}{2\sqrt{3s}}\|v\|_2^2 \\ &- \frac{\alpha\sqrt{1+\delta_{3s}}}{\sqrt{3s}}\left\langle x^{\text{opt}}, v\right\rangle.\end{aligned} \tag{6.3.6}$$

Note that

$$\begin{aligned}\|v\|_2 &\leq \|v_{\Lambda_{01}}\|_2 + \sum_{i=2}^{\ell}\|v_{\Lambda_i}\|_2 \\ &\leq \left(1 + \frac{\sqrt{s}}{\sqrt{3s}}\right)\|v_{\Lambda_{01}}\|_2 - \frac{\alpha}{2\sqrt{3s}}\|v\|_2^2 + \frac{\alpha}{\sqrt{3s}}\left\langle x^{\text{opt}}, v\right\rangle,\end{aligned}$$

and it can be written as

$$\|v_{\Lambda_{01}}\|_2 \geq \frac{\sqrt{3s}}{\sqrt{3s}+\sqrt{s}}\left(\frac{\alpha}{2\sqrt{3s}}\|v\|_2^2 + \|v\|_2 - \frac{\alpha}{\sqrt{3s}}\left\langle x^{\text{opt}}, v\right\rangle\right).$$

With the assumption $\delta_{3s} < 3(1-\delta_{4s}) - 1$ on A, the coefficient of $\|v_{\Lambda_{01}}\|_2$ in (6.3.6) is positive and thus we have

$$
\begin{aligned}
\|Av\|_2 &\geq \frac{\sqrt{1-\delta_{4s}}\sqrt{3s}-\sqrt{1+\delta_{3s}}\sqrt{s}}{\sqrt{3s}+\sqrt{s}}\left(\frac{\alpha}{2\sqrt{3s}}\|v\|_2^2+\|v\|_2-\frac{\alpha}{\sqrt{3s}}\langle x^{\mathrm{opt}},v\rangle\right)+\frac{\alpha\sqrt{1+\delta_{3s}}}{2\sqrt{3s}}\|v\|_2^2-\frac{\alpha\sqrt{1+\delta_{3s}}}{\sqrt{3s}}\langle x^{\mathrm{opt}},v\rangle \\
&= \frac{\alpha}{2}\left(\frac{\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}}}{\sqrt{3s}+\sqrt{s}}\right)\|v\|_2^2+\frac{\sqrt{1-\delta_{4s}}\sqrt{3s}-\sqrt{1+\delta_{3s}}\sqrt{s}}{\sqrt{3s}+\sqrt{s}}\|v\|_2-\alpha\left(\frac{\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}}}{\sqrt{3s}+\sqrt{s}}\right)\langle x^{\mathrm{opt}},v\rangle .
\end{aligned}
\tag{6.3.7}
$$

If $\langle x^{\mathrm{opt}}, v\rangle \leq 0$, then

$$\|Av\|_2 \geq D_1\|v\|_2^2.$$

If $\langle x^{\mathrm{opt}}, v\rangle > 0$, then the condition (6.3.3) on α guarantees

$$\begin{aligned}&\frac{\sqrt{1-\delta_{4s}}\sqrt{3s}-\sqrt{1+\delta_{3s}}\sqrt{s}}{\sqrt{3s}+\sqrt{s}}\|v\|_2-\alpha\left(\frac{\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}}}{\sqrt{3s}+\sqrt{s}}\right)\langle x^{\mathrm{opt}}, v\rangle\\ \geq&\frac{\sqrt{1-\delta_{4s}}\sqrt{3s}-\sqrt{1+\delta_{3s}}\sqrt{s}}{\sqrt{3s}+\sqrt{s}}\left(\|v\|_2-\left\langle\frac{x^{\mathrm{opt}}}{\|x^{\mathrm{opt}}\|_2}, v\right\rangle\right)\geq 0,\end{aligned}$$

where we use the Cauchy–Schwarz inequality. Hence we also have

$$\|Av\|_2 \geq D_1\|v\|_2^2.$$

When $\tau = 0$, the equality

$$Av = A(x^{\mathrm{opt}} - \bar{x}) = b - b = 0$$

renders

$$0 = \|Av\|_2 \geq D_1\|v\|_2^2,$$

which implies $\|v\|_2 = 0$. Thus $x^{\mathrm{opt}} = \bar{x}$. When $\tau > 0$, the inequality

$$\|Av\|_2 = \|Ax^{\mathrm{opt}} - A\bar{x}\|_2 \leq \|Ax^{\mathrm{opt}} - b\|_2 + \|A\bar{x} - b\|_2 \leq 2\tau$$

leads to $2\tau \geq D_1\|v\|_2^2$, which implies (6.3.4). □

In analysis of signal reconstruction models with various convex and non-convex penalties, such as the ℓ_1 penalty [43, 54] and the ℓ_{1-2} penalty [238, 239], a linear lower bound for $\|A(x^{\mathrm{opt}} - \bar{x})\|_2$ is derived somehow. The proof of Theorem 6.3.2 mainly follows the idea of [43], but we derive a quadratic lower bound for the term $\|A(x^{\mathrm{opt}} - \bar{x})\|_2$. Thus, it is worthy noting that our results cannot be reduced to the result of the BP model (6.1.3) as $\alpha \to 0$. Indeed, the quadratic bound (6.3.6) in our proof is reduced to a linear bound as $\alpha \to 0$, which then leads to the same results as the BP model (6.1.3). However, we handle our final quadratic bound by removing its linear and constant terms and hence the obtained result cannot be reduced to the result of the BP model (6.1.3) as $\alpha \to 0$.

Besides, the condition (6.3.3) on α is required for the springback-penalized model (6.1.5). It is impossible to choose an α satisfying (6.3.3) unless we have *a priori* estimation on $\|x^{\mathrm{opt}}\|_2$ before solving the problem (6.1.5). Thus, the condition (6.3.3) then can be interpreted as a *posterior verification* in the sense that it can be verified once x^{opt} is obtained by solving the problem (6.1.5).

Remark 6.3.3 (Posterior verification) *In practice, we solve the springback-penalized model* (6.1.5) *numerically and thus obtain an approximate solution, denoted by* x^**, subject to a preset accuracy* $\epsilon > 0$*. That is,*

$$\|x^{\text{opt}} - x^*\|_2 \leq \epsilon.$$

Then, the posterior verification (6.3.3) *is guaranteed if*

$$\alpha \leq \frac{\sqrt{1-\delta_{4s}}\sqrt{3s} - \sqrt{1+\delta_{3s}}\sqrt{s}}{(\sqrt{1-\delta_{4s}} + \sqrt{1+\delta_{3s}})(\|x^*\|_2 + \epsilon)}.$$

6.3.3 On the exact and robust reconstruction

In Theorem 6.3.2, we establish conditions for exact and robust reconstruction using the springback-penalized model (6.1.5). Table 6.1 lists the exact reconstruction conditions for five other popular models in the literature. In particular, the springback-penalized model (6.1.5) and the ℓ_1-penalized model, i.e., the BP model (6.1.3), have the same RIP condition. This condition is more stringent than that of the ℓ_p-penalized model ($0 < p < 1$) but weaker than those of the transformed ℓ_1- and ℓ_{1-2}-penalized models. Beside the RIP condition, there is an additional assumption $a(s) > 1$ for the ℓ_{1-2}-penalized model, where $a(s)$ was first derived in [239] and slightly improved in [238] as

$$a(s) = \left(\frac{3s-1}{\sqrt{3s} + \sqrt{4s-1}}\right)^2.$$

Note that $a(s) < 3$ was shown in [238, 239] for both the cases.

Table 6.1: Exact reconstruction conditions reconstruction models with various penalties.

Penalty	RIP condition
ℓ_1 [43]	$\delta_{3s} < 3(1-\delta_{4s}) - 1$
ℓ_p $(0 < p < 1)$ [54]	$\delta_{3s} < 3^{(2-p)/p}(1-\delta_{4s}) - 1$
transformed ℓ_1 [243]	$\delta_{3s} < \left(\frac{\beta}{\beta+1}\right)^2 3(1-\delta_{4s}) - 1$
ℓ_{1-2} [238, 239]	$\delta_{3s} < a(s)(1-\delta_{4s}) - 1$
springback	$\delta_{3s} < 3(1-\delta_{4s}) - 1$

We then discuss robust reconstruction results. If $\alpha \to 0$, then the result (6.3.4) cannot provide any information as

$$\frac{\sqrt{2}}{\sqrt{D_1}} \to \infty.$$

However, for an appropriate α, the bound (6.3.4) is informative and attractive. The robust reconstruction results of the ℓ_1-, ℓ_p-, transformed ℓ_1- and ℓ_{1-2}-penalized models were shown to be linear with respect to the level of noise τ [43, 54, 238, 239, 243], in the sense of

$$\|x^{\text{opt}} - \bar{x}\|_2 \leq C_s \tau, \tag{6.3.8}$$

where C_s is some constant. Thus, under the conditions of Theorem 6.3.2, the bound (6.3.4) for the springback-penalized model (6.1.5) is tighter than (6.3.8) in the sense of

$$\frac{\sqrt{2}}{\sqrt{D_1}}\sqrt{\tau} \leq C_s \tau \tag{6.3.9}$$

if the level of noise τ satisfies

$$\tau > \frac{2}{D_1 C_s^2}. \tag{6.3.10}$$

Assume that the reconstruction conditions listed in Table 6.1 are satisfied for each model, respectively. Then, we can summarize their corresponding ranges of τ in Table 6.2 such that the robust reconstruction bound (6.3.4) of the springback-penalized model (6.1.5) is tighter than all the others in the sense of (6.3.9).

These ranges on τ look complicated. To have a better idea, we consider a toy example with $s = 20$, $\delta_{3s} = 1/4$, $\delta_{4s} = 1/3$, $\alpha = 1$ for the spingback penalty (6.1.4), and $\beta = 1$ for the transformed ℓ_1 penalty. Then, the springback-penalized model (6.1.5) gives a tighter bound in the sense of (6.3.9) than the ℓ_1-, $\ell_{0.2}$-, $\ell_{0.5}$-, $\ell_{0.999}$-, transformed ℓ_1-, and ℓ_{1-2}-penalized models if $\tau > 0.1385$, 0.0271, 0.2333, 0.1391, 0.0807, and 2.8652×10^{-4}, respectively.

Can we further improve the robust reconstruction result (6.3.4) in Theorem 6.3.2? The following proposition suggests a potential improvement. Moreover, without any requirement on α, this proposition also means, even if the posterior verification (6.3.3) is violated sometimes, the springback-penalized model (6.1.5) may still give a good reconstruction. Note that this proposition is only of conceptual sense, because its assumption

$$\left\langle x^{\text{opt}}, x^{\text{opt}} - \bar{x} \right\rangle \leq 0$$

is not verifiable. Nevertheless, it helps us discern a possibility of achieving a better reconstruction bound than (6.3.4).

Table 6.2: Ranges of the level of noise such that the springback bound (6.3.4) is tighter than the bound (6.3.8) in the sense of (6.3.9).

Penalty	When the springback bound (6.3.4) is tighter than the bound (6.3.8)
ℓ_1 [42, 43]	$\tau > \frac{(\sqrt{3s}+\sqrt{s})(\sqrt{3}\sqrt{1-\delta_{4s}}-\sqrt{1+\delta_{3s}})^2}{4\alpha(\sqrt{1+\delta_{3s}}+\sqrt{1-\delta_{4s}})}$
ℓ_p $(0<p<1)$ [183]	$\tau > \frac{(\sqrt{3s}+\sqrt{s})\left((1-\delta_{4s})^{p/2}-(1+\delta_{3s})^{p/2}3^{p/2-1}\right)^{2/p}}{\alpha(\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}})\left(1+\frac{1}{(2/p-1)3^{2/p-1}}\right)}$
transformed ℓ_1 [243]	$\tau > \frac{4(\sqrt{3s}+\sqrt{s})(1-\delta_{3s})\left(\frac{\beta}{\beta+1}\sqrt{3}\sqrt{1-\delta_{4s}}-\sqrt{1+\delta_{3s}}\right)^2}{\alpha(\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}})\left(\frac{\beta}{\beta+1}\sqrt{3}\sqrt{1-\delta_{4s}}-\sqrt{1+\delta_{3s}}+\sqrt{3s}\sqrt{1-\delta_{3s}}\right)^2}$
ℓ_{1-2} [238]	$\tau > \frac{(\sqrt{3s}+\sqrt{s})(\sqrt{a(s)(1-\delta_{4s})}-\sqrt{1+\delta_{3s}})^2}{\alpha(\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}})(\sqrt{3s}-\sqrt{s\cdot a(s)})^2}$

Proposition 6.3.4 *Let $\bar{x} \in \mathbb{R}^n$ be an unknown s-sparse vector to be recovered. For a given sensing matrix $A \in \mathbb{R}^{m\times n}$, let $b \in \mathbb{R}^m$ be a vector of measurements from $b = A\bar{x} + e$ with $\|e\|_2 \le \tau$, and let δ_{3s} and δ_{4s} be the $3s$- and $4s$-RIC's of A, respectively. Let $x^{\rm opt}$ be the minimizer of the problem (6.1.5) and assume $\langle x^{\rm opt}, x^{\rm opt} - \bar{x}\rangle \le 0$. Suppose A satisfies $\delta_{3s} < 3(1-\delta_{4s}) - 1$, then $x^{\rm opt} = \bar{x}$ when $\tau = 0$; and $x^{\rm opt}$ satisfies*

$$\|x^{\rm opt} - \bar{x}\|_2 \le \sqrt{\frac{D_2^2}{4D_1^2} + \frac{2}{D_1}\tau} - \frac{D_2}{2D_1} \tag{6.3.11}$$

when $\tau \ge 0$, where D_1 is the constant (6.3.5) given in Theorem 6.3.2 and

$$D_2 = \frac{\sqrt{3}\sqrt{1-\delta_{4s}} - \sqrt{1+\delta_{3s}}}{\sqrt{3}+1}. \tag{6.3.12}$$

Proof. In the case of $\langle x^{\rm opt}, v\rangle \le 0$, it follows straightforwardly from (6.3.7) that

$$\begin{aligned}\|Av\|_2 &\ge \frac{\alpha}{2}\left(\frac{\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}}}{\sqrt{3s}+\sqrt{s}}\right)\|v\|_2^2 + \frac{\sqrt{1-\delta_{4s}}\sqrt{3s}-\sqrt{1+\delta_{3s}}\sqrt{s}}{\sqrt{3s}+\sqrt{s}}\|v\|_2 \\ &:= D_1\|v\|_2^2 + D_2\|v\|_2.\end{aligned}$$

The assumption $\delta_{3s} < 3(1-\delta_{4s}) - 1$ guarantees $D_2 > 0$. Hence, when $\tau = 0$, as $Av = A(x^{\rm opt} - \bar{x}) = 0$, we have

$$0 = \|Av\|_2 \ge D_1\|v\|_2^2 + D_2\|v\|_2,$$

which implies $\|v\|_2 = 0$. When $\tau > 0$, the inequality

$$\|Av\|_2 = \|Ax^{\rm opt} - A\bar{x}\|_2 \le \|Ax^{\rm opt} - b\|_2 + \|A\bar{x} - b\|_2 \le 2\tau$$

implies

$$\|v\|_2 \le \frac{\sqrt{D_2^2 + 8D_1\tau} - D_2}{2D_1}.$$

The assertion is proved. □

Remark 6.3.5 *The robust reconstruction result (6.3.11) is always better than (6.3.4) in Theorem 6.3.2 due to the subadditivity of the square root function. Under the conditions of Proposition 6.3.4, the bound (6.3.11) for the springback-penalized model (6.1.5) is tighter than (6.3.8) in the sense of*

$$\sqrt{\frac{D_2^2}{4D_1^2} + \frac{2}{D_1}\tau} - \frac{D_2}{2D_1} < C_s\tau,$$

if the level of noise τ satisfies

$$\tau > \frac{2 - D_2 C_s}{D_1 C_s^2} = \left(1 - \frac{D_2 C_s}{2}\right) \frac{2}{D_1 C_s^2}.$$

Comparing with (6.3.10), this improvement enlarges the value range of τ. For example, if C_s is the coefficient in the result (6.3.1) of the BP model (6.1.3) , then $1 - D_2 C_s/2$ is approximately 0.2679.

6.4 Springback-penalized model for nearly sparse signal reconstruction

We then study the stable reconstruction of the springback-penalized model (6.1.5) when $\bar{x}$ is nearly sparse and the measurements are noisy.

6.4.1 Reconstruction guarantee using the springback-penalized model

If the signal $\bar{x}$ to be recovered is nearly s-sparse, then we have the following stable reconstruction theorem for the springback-penalized model (6.1.5).

Theorem 6.4.1 (reconstruction of nearly sparse signals) *Let $\bar{x} \in \mathbb{R}^n$ be an unknown vector to be recovered. For a given sensing matrix $A \in \mathbb{R}^{m\times n}$, let $b \in \mathbb{R}^m$ be a vector of measurements from $b = A\bar{x} + e$ with $\|e\|_2 \le \tau$, and let δ_{3s} and δ_{4s} be the $3s$- and $4s$-RIC's of A, respectively. Let $\bar{x}_s \in \mathbb{R}^n$ be the truncated vector corresponding to the s largest values of $\bar{x}$ (in absolute value). Suppose A satisfies $\delta_{3s} < 3(1 - \delta_{4s}) - 1$ and α satisfies (6.3.3), then the minimizer $x^{\rm opt}$ of the problem (6.1.5) satisfies*

$$\|x^{\rm opt} - \bar{x}\|_2 \le \sqrt{\frac{2}{D_1}\tau + \frac{4}{\alpha}\|\bar{x} - \bar{x}_s\|_1}, \tag{6.4.1}$$

where D_1 is the constant (6.3.5) given in Theorem 6.3.2.

Proof. Let $x^{\rm opt} = \bar{x} + v$, and Λ_0 be the support of $\bar{x}_s$. It is clear that $v_{\Lambda_0} = x^{\rm opt}_{\Lambda_0} - \bar{x}_s$ and $v_{\Lambda_0^c} = x^{\rm opt}_{\Lambda_0^c} - \bar{x}_{\Lambda_0^c}$. We know that

$$\|x^{\rm opt}\|_1 - \frac{\alpha}{2}\|x^{\rm opt}\|_2^2 \le \|\bar{x}\|_1 - \frac{\alpha}{2}\|\bar{x}\|_2^2 = \|\bar{x}_s\|_1 + \|\bar{x}_{\Lambda_0^c}\|_1 - \frac{\alpha}{2}\|\bar{x}\|_2^2.$$

On the other hand, it holds that

$$\begin{aligned}
&\|x^{\rm opt}\|_1 - \frac{\alpha}{2}\|x^{\rm opt}\|_2^2 \\
=&\|\bar{x}_s + v_{\Lambda_0}\|_1 + \|\bar{x}_{\Lambda_0^c} + v_{\Lambda_0^c}\|_1 - \frac{\alpha}{2}\|\bar{x} + v\|_2^2 \\
\ge&\|\bar{x}_s\|_1 - \|v_{\Lambda_0}\|_1 + \|v_{\Lambda_0^c}\|_1 - \|\bar{x}_{\Lambda_0^c}\|_1 - \frac{\alpha}{2}\left(\|\bar{x}\|_2^2 + 2\langle \bar{x}, v\rangle + \|v\|_2^2\right).
\end{aligned}$$

Then, v satisfies the following estimation:

$$\begin{aligned}\|v_{\Lambda_0^c}\|_1 \leq & \|v_{\Lambda_0}\|_1 + 2\|\bar{x} - \bar{x}_s\|_1 - \frac{\alpha}{2}\|v\|_2^2 + \alpha\|v\|_2^2 + \alpha\langle \bar{x}, v\rangle \\ = & \|v_{\Lambda_0}\|_1 + 2\|\bar{x} - \bar{x}_s\|_1 - \frac{\alpha}{2}\|v\|_2^2 + \alpha\left\langle x^{\text{opt}}, v\right\rangle.\end{aligned}$$

We divide Λ_0^c into subsets of size $3s$, $\Lambda_0^c = \Lambda_1 \bigcup \Lambda_2 \bigcup \cdots \bigcup \Lambda_\ell$, in terms of decreasing order of magnitudes (in absolute value) of $v_{\Lambda_0^c}$. Denoting $\Lambda_{01} = \Lambda_0 \bigcup \Lambda_1$ and using the RIP of A, we have

$$\|Av\|_2 \geq \|A_{\Lambda_{01}} v_{\Lambda_{01}}\|_2 - \left\|\sum_{i=2}^{\ell} A_{\Lambda_i} v_{\Lambda_i}\right\|_2 \geq \sqrt{1-\delta_{4s}}\|v_{\Lambda_{01}}\|_2 - \sqrt{1+\delta_{3s}}\sum_{i=2}^{\ell}\|v_{\Lambda_i}\|_2.$$

As proved for Theorem 6.3.2, we have

$$\sum_{i=2}^{\ell}\|v_{\Lambda_i}\|_2 \leq \frac{\|v_{\Lambda_0^c}\|_1}{\sqrt{3}}$$

and $\|v_{\Lambda_0}\|_1 \leq \sqrt{s}\|v_{\Lambda_{01}}\|_2$. Thus, we obtain

$$\sum_{i=2}^{\ell}\|v_{\Lambda_i}\|_2 \leq \frac{1}{\sqrt{3s}}\left(\sqrt{s}\|v_{\Lambda_{01}}\|_2 + 2\|\bar{x} - \bar{x}_s\|_1 - \frac{\alpha}{2}\|v\|_2^2 + \alpha\left\langle x^{\text{opt}}, v\right\rangle\right).$$

Furthermore, it holds that

$$\begin{aligned}\|Av\|_2 \geq & \left(\sqrt{1-\delta_{4s}} - \frac{\sqrt{1+\delta_{3s}}\sqrt{s}}{\sqrt{3s}}\right)\|v_{\Lambda_{01}}\|_2 - \frac{2\sqrt{1+\delta_{3s}}}{\sqrt{3s}}\|\bar{x} - \bar{x}_s\|_1 \\ & + \frac{\alpha\sqrt{1+\delta_{3s}}}{2\sqrt{3s}}\|v\|_2^2 - \frac{\alpha\sqrt{1+\delta_{3s}}}{\sqrt{3s}}\left\langle x^{\text{opt}}, v\right\rangle.\end{aligned} \tag{6.4.2}$$

As

$$\begin{aligned}\|v\|_2 \leq & \|v_{\Lambda_{01}}\|_2 + \sum_{i=2}^{\ell}\|v_{\Lambda_i}\|_2 \\ \leq & \left(1 + \frac{\sqrt{s}}{\sqrt{3s}}\right)\|v_{\Lambda_{01}}\|_2 + \frac{2}{\sqrt{3s}}\|\bar{x} - \bar{x}_s\|_1 - \frac{\alpha}{2\sqrt{3s}}\|v\|_2^2 + \frac{\alpha}{\sqrt{3s}}\left\langle x^{\text{opt}}, v\right\rangle,\end{aligned}$$

we have

$$\|v_{\Lambda_{01}}\|_2 \geq \frac{\sqrt{3s}}{\sqrt{3s}+\sqrt{s}}\left(\frac{\alpha}{2\sqrt{3s}}\|v\|_2^2 + \|v\|_2 - \frac{\alpha}{\sqrt{3s}}\left\langle x^{\text{opt}}, v\right\rangle - \frac{2}{\sqrt{3s}}\|\bar{x} - \bar{x}_s\|_1\right).$$

Recall the assumption $\delta_{3s} < 3(1-\delta_{4s}) - 1$. The coefficient of $\|v_{\Lambda_{01}}\|_2$ in (6.4.2) is positive, and it follows that

$$
\begin{aligned}
\|Av\|_2 \geq & \frac{\sqrt{1-\delta_{4s}}\sqrt{3s}-\sqrt{1+\delta_{3s}}\sqrt{s}}{\sqrt{3s}+\sqrt{s}}\left(\frac{\alpha\|v\|_2^2}{2\sqrt{3s}}+\|v\|_2-\frac{\alpha}{\sqrt{3s}}\left\langle x^{\mathrm{opt}},v\right\rangle-\frac{2}{\sqrt{3s}}\|\bar{x}-\bar{x}_s\|_1\right)+\frac{\alpha\sqrt{1+\delta_{3s}}}{2\sqrt{3s}}\|v\|_2^2 \\
& -\frac{\alpha\sqrt{1+\delta_{3s}}}{\sqrt{3s}}\left\langle x^{\mathrm{opt}},v\right\rangle-\frac{2\sqrt{1+\delta_{3s}}}{\sqrt{3s}}\|\bar{x}-\bar{x}_s\|_1 \\
= & \frac{\alpha}{2}\left(\frac{\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}}}{\sqrt{3s}+\sqrt{s}}\right)\|v\|_2^2+\frac{\sqrt{1-\delta_{4s}}\sqrt{3s}-\sqrt{1+\delta_{3s}}\sqrt{s}}{\sqrt{3s}+\sqrt{s}}\|v\|_2-\alpha\left(\frac{\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}}}{\sqrt{3s}+\sqrt{s}}\right)\left\langle x^{\mathrm{opt}},v\right\rangle \\
& -2\left(\frac{\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}}}{\sqrt{3s}+\sqrt{s}}\right)\|\bar{x}-\bar{x}_s\|_1.
\end{aligned}
\tag{6.4.3}
$$

If $\left\langle x^{\mathrm{opt}},v\right\rangle \leq 0$, then

$$\|Av\|_2 \geq D_1\|v\|_2^2-\frac{4}{\alpha}D_1\|\bar{x}-\bar{x}_s\|_1.$$

If $\langle x^{\text{opt}}, v\rangle > 0$, then the condition (6.3.3) on α guarantees

$$\frac{\sqrt{1-\delta_{4s}}\sqrt{3s}-\sqrt{1+\delta_{3s}}\sqrt{s}}{\sqrt{3s}+\sqrt{s}}\|v\|_2-\alpha\left(\frac{\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}}}{\sqrt{3s}+\sqrt{s}}\right)\langle x^{\text{opt}}, v\rangle \geq 0,$$

which is shown in the proof of Theorem 6.3.2. Hence, we also have

$$\|Av\|_2 \geq D_1\|v\|_2^2-\frac{4}{\alpha}D_1\|\bar{x}-\bar{x}_s\|_1.$$

As

$$\|Av\|_2 = \|Ax^{\text{opt}}-A\bar{x}\|_2 \leq \|Ax^{\text{opt}}-b\|_2+\|A\bar{x}-b\|_2 \leq 2\tau,$$

we have

$$2\tau \geq D_1\|v\|_2^2-\frac{4}{\alpha}D_1\|\bar{x}-\bar{x}_s\|_1,$$

which implies (6.4.1). □

Similar to the improvement in Proposition 6.3.4, the above stable reconstruction result can be improved as follows.

Proposition 6.4.2 *Let $\bar{x}\in\mathbb{R}^n$ be an unknown vector to be recovered. For a given sensing matrix $A\in\mathbb{R}^{m\times n}$, let $b\in\mathbb{R}^m$ be a vector of measurements from $b=A\bar{x}+e$ with $\|e\|_2\leq\tau$, and let δ_{3s} and δ_{4s} be the $3s$- and $4s$-RIC's of A, respectively. Let x^{opt} be the minimizer of the problem (6.1.5) and assume $\langle x^{\text{opt}}, x^{\text{opt}}-\bar{x}\rangle \leq 0$. Let $\bar{x}_s\in\mathbb{R}^n$ be the truncated vector corresponding to the s largest values of $\bar{x}$ (in absolute value). Suppose A satisfies $\delta_{3s}<3(1-\delta_{4s})-1$, then x^{opt} satisfies*

$$\|x^{\text{opt}}-\bar{x}\|_2 \leq \sqrt{\frac{D_2^2}{4D_1^2}+\frac{2}{D_1}\tau+\frac{4}{\alpha}\|\bar{x}-\bar{x}_s\|_1}-\frac{D_2}{2D_1},$$

where D_1 and D_2 are the constants (6.3.5) and (6.3.12) given in Theorem 6.3.2 and Proposition 6.3.4, respectively.

Proof. In the case of $\langle x^{\text{opt}}, v\rangle \leq 0$, it follows straightforwardly from the estimation (6.4.3) that

$$\|Av\|_2 \geq D_1\|v\|_2^2+D_2\|v\|_2-\frac{4}{\alpha}D_1\|\bar{x}-\bar{x}_s\|_1.$$

The assumption $\delta_{3s}<3(1-\delta_{4s})-1$ guarantees $D_2>0$. Therefore, it follows from the triangle inequality that

$$\|Av\|_2 = \|Ax^{\text{opt}}-A\bar{x}\|_2 \leq \|Ax^{\text{opt}}-b\|_2+\|A\bar{x}-b\|_2 \leq 2\tau.$$

We thus have

$$D_1\|v\|_2^2 + D_2\|v\|_2 - \frac{4}{\alpha}D_1\|\bar{x} - \bar{x}_s\|_1 \leq 2\tau, \tag{6.4.4}$$

which gives the improved result by solving the system of inequalities (6.4.4) and $\|v\|_2 \geq 0$. □

6.4.2 On the stable reconstruction

If $\bar{x}$ is known to be s-sparse, then the estimation (6.4.1) in Theorem 6.4.1 is reduced to (6.3.4) in Theorem 6.3.2; and if the measurements are additionally noise-free, then both the estimations (6.3.4) and (6.4.1) imply exact reconstruction of the signal $\bar{x}$. We compare the estimation (6.4.1) with the estimation (6.3.2) for the BP model (6.1.3). The following comparison is based on theoretical error bounds. We are interested in the case where the estimation (6.4.1) is tighter than the estimation (6.3.2) in the sense of

$$\sqrt{\frac{2}{D_1}\tau + \frac{4}{\alpha}\|\bar{x} - \bar{x}_s\|_1} \leq C_{1,s}\tau + C_{2,s}\frac{\|\bar{x} - \bar{x}_s\|_1}{\sqrt{s}}, \tag{6.4.5}$$

which is equivalent to

$$\frac{s^{1/4}}{\sqrt{\alpha}}\sqrt{\frac{4(\sqrt{3}+1)}{\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}}}\tau + \frac{4\|\bar{x} - \bar{x}_s\|_1}{\sqrt{s}}} \leq C_{1,s}\tau + C_{2,s}\frac{\|\bar{x} - \bar{x}_s\|_1}{\sqrt{s}}. \tag{6.4.6}$$

Note that s takes values among $\{1, 2, \ldots, n\}$ and the right-hand side of (6.4.6) decreases as s increases. If the left-hand side of (6.4.6) is smaller than the right-hand side of (6.4.6) for $s = 1$ and the left-hand side is larger than the right-hand side for $s = n$, then there must exist a constant C such that the inequality (6.4.5) holds for $s \leq C$. Besides, if $\bar{x}$ is known to be s-sparse, then $\|\bar{x} - \bar{x}_s\|_1 = 0$ and thus (6.4.6) implies the existence of C without any assumption. Therefore, we have the following corollary.

Corollary 6.4.3 *If $\bar{x}$ is s-sparse, then there exists a constant C such that the inequality (6.4.5) holds for $s \leq C$, where*

$$C = \alpha^2 C_{1,s}^4 \tau^2 \left(\frac{\sqrt{1-\delta_{4s}}+\sqrt{1+\delta_{3s}}}{4(\sqrt{3}+1)}\right)^2. \tag{6.4.7}$$

When no information of the sparsity of $\bar{x}$ is known, if α satisfies

$$\frac{\frac{4(\sqrt{3}+1)}{\sqrt{1-\delta_4}+\sqrt{1+\delta_3}}\tau + 4\|\bar{x} - \bar{x}_1\|_1}{(C_{1,1}\tau + C_{2,1}\|\bar{x} - \bar{x}_1\|_1)^2} \leq \alpha \leq \frac{1}{C_{1,n}^2\tau}\frac{4(\sqrt{3}+1)\sqrt{n}}{\sqrt{1-\delta_{4n}}+\sqrt{1+\delta_{3n}}},$$

then there exists a constant C such that the inequality (6.4.5) holds for $s \leq C$, where C depends on α, $\bar{x}$, τ, δ_{3s}, and δ_{4s}.

In virtue of random matrix theory, we give two examples to show that the condition $\delta_{3s} < 3(1-\delta_{4s}) - 1$ on A in Theorems 6.3.2 and 6.4.1 holds.

- *Random Gaussian matrices*: the entries of A are i.i.d. Gaussian with mean zero and variance $1/m$. It was shown in [43, 44] that the condition $\delta_{3s} < 3(1-\delta_{4s})-1$ holds with overwhelming probability when $s \leq C'm/\log(n/m)$, where C' is a constant. Similar results were extended to sub-gaussian matrices in [145].
- *Fourier ensemble*: A is obtained by selecting m rows from the $n \times n$ discrete Fourier transform and renormalizing the columns so that they are unit-normed. If the rows are selected at random, the condition $\delta_{3s} < 3(1-\delta_{4s}) - 1$ holds with overwhelming probability for $s \leq C'm/(\log(n))^4$, where C' is a constant. This was initially considered in [45] and then improved in [181].

Remark 6.4.4 *Assume that α satisfies the conditions in Theorem 6.4.1 and Corollary 6.4.3. For a random Gaussian sensing matrix A, if*

$$s \leq C'm/\log(n/m),$$

then the RIP condition $\delta_{3s} < 3(1-\delta_{4s}) - 1$ on A holds with high probability; and additionally if $C'm/\log(n/m) \leq C$, i.e.,

$$m \exp\left(\frac{C'}{C}m\right) \leq n,$$

then the estimation (6.4.1) is tighter than the estimation (6.3.2) in the sense of (6.4.5). For a randomly subsampled Fourier sensing matrix A, if

$$s \leq C'm/(\log(n))^4,$$

then the RIP condition on A holds with overwhelming probability; and additionally if $C'm/(\log(n))^4 \leq C$, i.e.,

$$m \leq \frac{C}{C'}(\log(n))^4,$$

then the estimation (6.4.1) is tighter than the estimation (6.3.2) in the sense of (6.4.5). In a nutshell, for a sensing matrix satisfying the RIP condition, if the number m of observation data is limited, where "limited" can be characterized as the fact that m is less than some constant depending on n, C, and C', then the stable reconstruction using the springback-penalized model (6.1.5) is guaranteed by a tighter

bound than that of BP model (6.1.3) *in the sense of* (6.4.5). *These results can be extended to general orthogonal sensing matrices [45]. Similar comparative results with other reconstruction models may also be derived if the reconstruction error bounds of these models are linear to* τ *and* $\|\bar{x} - \bar{x}_s\|_1$, *e.g., the* ℓ_{1-2}*-penalized model [238].*

6.5 Computational aspects of the springback-penalized model

Now we focus on computational aspects for the springback-penalized model (6.1.5). We first design an algorithm for solving (6.1.5) in Section 6.5.1, and then discuss its convergence in Section 6.5.2 and elaborate on how to solve its subproblems in Section 6.5.3.

6.5.1 DCA-springback: An algorithm for the springback penalized model

Some well-developed algorithms for solving difference-of-convex (DC) optimization problems can be easily implemented to solve the springback-penalized model (6.1.5). We focus on the simplest DCA in [212, 213] without any line-search step, which has been shown to be efficient for solving signal reconstruction problems, see, e.g., [116, 239, 243].

Recall a standard DC optimization problem

$$\min_{x\in\mathbb{R}^n} \ f(x) := g(x) - h(x), \tag{6.5.1}$$

where g and h are lower semicontinuous proper convex functions on $\mathbb{R}^n$. Here, f is called a DC function, and $g - h$ is a DC decomposition of f. At each iteration, the DCA replaces the concave part $-h$ with a linear majorant and solves the resulting convex problem. That is, the DCA generates a sequence $\{x^k\}$ by solving the following subproblem iteratively:

$$x^{k+1} \in \arg\min_{x\in\mathbb{R}^n} \ \left\{ g(x) - \langle x - x^k, \xi^k \rangle \right\},$$

where $\xi^k \in \partial(h(x^k))$. Note that the springback-penalized model (6.1.5) can be written as

$$\arg\min_{x\in\mathbb{R}^n} F(x) := (\|x\|_1 + \chi_\Omega(x)) - \frac{\alpha}{2}\|x\|_2^2, \tag{6.5.2}$$

where $\Omega := \{x \in \mathbb{R}^n : \|Ax - b\|_2 \leq \tau\}$ and

$$\chi_\Omega(x) := \begin{cases} 0, & x \in \Omega, \\ +\infty, & x \notin \Omega, \end{cases}$$

is the indictor function of the set Ω. Thus, the DCA iterate scheme for solving (6.5.2) reads as

$$\begin{aligned} x^{k+1} &\in \arg\min_x \left\{ (\|x\|_1 + \chi_\Omega(x)) - \langle x - x^k, \xi^k \rangle \right\} \\ &= \arg\min_x \left\{ \|x\|_1 - \langle x - x^k, \xi^k \rangle \quad \text{s.t.} \quad x \in \Omega \right\}. \end{aligned}$$

More specifically, the resulting DCA is listed in the following algorithm, where $\epsilon_{\text{outer}} > 0$ is the preset tolerance for iterations, and "MaxIt" means the maximal number of iterations set beforehand.

Algorithm 1: DCA-springback: Solving the constrained springback model (6.1.5) via DCA

Input: Model parameters: $\alpha > 0$ satisfying the condition (6.5.6);
Stopping criterion: $\epsilon_{\text{outer}} > 0$, MaxIt > 0;
Initialization: $k = 0$, x^0 satisfying $\|Ax - b\|_2 \leq \tau$;

while $k <$ MaxIt and $\min\left\{\|x^{k+1} - x^k\|_2, \|x^{k+1} - x^k\|_2 / \|x^k\|_2\right\} > \epsilon_{\text{outer}}$ **do**
  $\xi^k = \alpha x^k$;
  $x^{k+1} \in \arg\min_x \left\{ \|x\|_1 - \langle x - x^k, \xi^k \rangle \text{ s.t. } \|Ax - b\|_2 \leq \tau \right\}$;
  $k \leftarrow k + 1$;
end

6.5.2 Convergence

Recall that the *modulus of strong convexity* of a convex function f on $\mathbb{R}^n$, denoted by $d(f)$, is defined as

$$d(f) := \sup\{\nu > 0 : f(\cdot) - \frac{\nu}{2}\|\cdot\|_2^2 \text{ is convex on } \mathbb{R}^n\}.$$

Then, according to [213, Proposition A.1], for a general DC function $f = g - h$, any sequence $\{x^k\}$ generated by the DCA satisfies

$$f(x^k) - f(x^{k+1}) \geq \frac{d(g) + d(h)}{2}\|x^{k+1} - x^k\|_2^2, \tag{6.5.3}$$

which immediately implies the decreasing property of $\{f(x^k)\}$ if at least one of g and h is strongly convex. Note that $\frac{\alpha}{2}\|x\|_2^2$ is strongly convex with modulus α. Thus,

starting with a feasible x^0, we have the decreasing property

$$F(x^k) - F(x^{k+1}) \geq \frac{\alpha}{2}\|x^{k+1} - x^k\|_2^2, \tag{6.5.4}$$

where F is defined as (6.5.2). However, the decreasing property (6.5.4) of F is not sufficient to ensure the convergence of DCA-springback. The function F could be negative if α is inappropriately large. Note that for any x^k, we have

$$\|Ax^k\|_2 - \|b\|_2 \leq \|Ax^k - b\|_2 \leq \tau.$$

Moreover, as A is assumed to be full rank, we have $\sigma_{\min}(A) > 0$. It follows from the geometric interpretation of the SVD [223, Lecture 4] that

$$\|Ax\|_2 \geq \sigma_{\min}(A)$$

for any $x \in \mathbb{R}^n$ on the unit sphere $\{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. Thus, it holds that

$$0 < \sigma_{\min}(A) \leq \min_{x\in\mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2} = \min_{\|x\|_2=1} \|Ax\|_2,$$

and we have

$$\|x^k\|_2 \leq \frac{\|b\|_2 + \tau}{\sigma_{\min}(A)}. \tag{6.5.5}$$

Note that $\|x\|_1 - \frac{\alpha}{2}\|x\|_2^2 \geq 0$ and hence F is non-negative if $\|x\|_2 \leq 2/\alpha$. Clearly, if

$$\alpha \leq \frac{2\sigma_{\min}(A)}{\|b\|_2 + \tau}, \tag{6.5.6}$$

then $F(x^k) \geq 0$ for any $k \geq 0$ because all iterates x^k satisfy (6.5.5). Together with the decreasing property (6.5.4), we can establish the convergence of DCA-springback easily by following the analytical framework in [212, 213]. Moreover, it follows the convergence of $\{F(x^k)\}$ and (6.5.4) that $\|x^{k+1} - x^k\|_2 \to 0$ as $k \to \infty$.

Remark 6.5.1 *Note that the condition* (6.3.3) *depends on the RIP condition of A, and* (6.5.6) *depends on the conditioning of A. It is easy to deduce that if*

$$\frac{\sqrt{1-\delta_{4s}}\sqrt{3s} - \sqrt{1+\delta_{3s}}\sqrt{s}}{\sqrt{1-\delta_{4s}} + \sqrt{1+\delta_{3s}}} \leq \frac{2\sigma_{\min}(A)\|x^{\text{opt}}\|_2}{\|b\|_2 + \tau}, \tag{6.5.7}$$

then the condition (6.5.6) *is implied by* (6.3.3). *Otherwise, it can be verified that the condition* (6.3.3) *is implied by* (6.5.6).

6.5.3 Solving the subproblem of DCA-springback

For the proposed DCA-springback, its subproblem at each iteration is

$$\min_x \ \|x\|_1 - \left\langle x - x^k, \xi^k \right\rangle \quad \text{s.t.} \quad \|Ax - b\|_2 \le \tau. \tag{6.5.8}$$

This problem can be easily solved by, e.g., the ADMM, which was originally proposed in [99] and had been well developed in the literature such as [56, 107]. Some details are given for completeness. Note that the subproblem (6.5.8) can be reformulated as

$$\begin{aligned} \min_{x,y,z} \ & \|y\|_1 - \langle x - x^k, \xi^k \rangle \\ \text{s.t.} \ & y = x, \\ & z = Ax - b, \\ & z \in \mathcal{B}(\tau), \end{aligned}$$

where $y, z \in \mathbb{R}^n$ are two auxiliary variables. With some trivial details skipped, the iterative scheme of the (scaled) ADMM for the subproblem (6.5.8) reads as

$$\begin{cases} x^{j+1} = (\rho A^{\mathrm{T}} A + \zeta I)^{-1} \left(\rho A^{\mathrm{T}} (b + z^j - \eta^j) + \xi^k + \zeta (y^j - u^j) \right), \\ y_i^{j+1} = \mathrm{soft}(x_i^{j+1} + u_i^j; 1/\zeta) \text{ for } i = 1, \ldots, n, \\ z^{j+1} = \mathcal{P}_{\mathcal{B}(\tau)}(Ax^{j+1} - b + \eta^j), \\ u^{j+1} = u^j + x^{j+1} - y^{j+1}, \\ \eta^{j+1} = \eta^{j+1} + Ax^{j+1} - b - z^{j+1}, \end{cases} \tag{6.5.9}$$

where $u \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^m$ are the Lagrange multipliers, $\zeta > 0$ and $\rho > 0$ are penalty parameters, and $\mathcal{P}_{\mathcal{B}(\tau)}(\cdot)$ is the projection operator onto the ball $\mathcal{B}(\tau)$. If the measurement process is noise-free, i.e., $\tau = 0$, then z^j is always set as zero and the projection of the z-subproblem in (6.5.9) is not necessary.

6.6 Numerical experiments

In this section, we implement the DCA-springback to the constrained springback-penalized model (6.1.5) with simulated data. All codes were written by MATLAB R2022a, and all numerical experiments were conducted on a laptop (16 GB RAM, Intel® CoreTM i7-9750H Processor) with macOS Monterey 12.4.

We mainly show the effectiveness of the model (6.1.5) for some specific scenarios and demonstrate the efficiency of the DCA-springback. Several state-of-the-art signal reconstruction solvers listed below are also tested for comparison.

1) The accelerated iterative hard thresholding (AIHT) algorithm in [26]: solving the constrained model

$$\min_{x\in\mathbb{R}^n} \ \|Ax-b\|_2^2 \quad \text{s.t.} \quad \|x\|_0 \le s$$

by the accelerated iterative hard thresholding, where s is set beforehand to estimate the sparsity of x. For simplicity, we only choose the fundamental AIHT in [26], and refer to, e.g., [91, 115, 119, 120, 155, 161], for various other more sophisticated algorithms.

2) ADMM-ℓ_1 [99]: solving the unconstrained ℓ_1-penalized problem by the ADMM.

3) IRLS-ℓ_p $(0<p<1)$ [125]: smoothing the unconstrained ℓ_p-penalized model as

$$\min_{x\in\mathbb{R}^n} \ \frac{1}{2}\|Ax-b\|_2^2+\lambda\|x\|_{p,\epsilon}^p \quad \text{with} \quad \|x\|_{p,\epsilon}^p := \sum_{j=1}^n (x_j^2+\epsilon^2)^{p/2},$$

where $\epsilon>0$, and implementing the iteratively reweighted least squares (IRLS) algorithm.

4) DCA-TL1 [243]: solving the unconstrained transformed ℓ_1-penalized model with parameter β by DCA and implementing the ADMM for its subproblems.

5) DCA-ℓ_{1-2} [239]: solving the unconstrained ℓ_{1-2}-penalized model by DCA and implementing the ADMM for its subproblems.

6) DCA-MCP [209]: solving the unconstrained MCP-penalized model by DCA and implementing the ADMM for its subproblems (the authors in [209] consider the ℓ_1-norm data fidelity term instead of the ℓ_2 norm, but the implementation of the MCP term is similar).

Note that the AIHT solves the ℓ_0-penalized model directly; the ADMM-ℓ_1 solves a convex surrogate model, and the others solve different non-convex approximate models.

6.6.1 Setup

We consider both incoherent and coherent sensing matrices to generate synthetic data for simulation. In the incoherent regime, we use random Gaussian matrices and random partial discrete cosine transform (DCT) matrices. For the former kind, its columns are generated by

$$A_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_m/m), \quad i=1,\dots,n,$$

where $\mathcal{N}(0, I_m/m)$ is the multivariate Gaussian distribution with location 0 and covariance I_m/m. For the latter kind, its columns are generated by

$$A_i = \frac{1}{\sqrt{m}} \cos(2i\pi\chi_i), \quad i = 1, \ldots, n,$$

where $\chi_i \in \mathbb{R}^m \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1]^m)$ is uniformly and independently sampled from $[0,1]$. Note that both kinds of matrices have small RIP constants with high probability. The coherent regime consists of more ill-conditioned sensing matrices with higher coherence, and it is represented by the randomly oversampled partial DCT matrix in our experiments. A randomly oversampled partial DCT matrix is defined as

$$A_i = \frac{1}{\sqrt{m}} \cos(2i\chi_i/\mathcal{F}), \quad i = 1, \ldots, n,$$

where $\mathcal{F} \in \mathbb{N}$ is the *refinement factor*. As $\mathcal{F}$ increases, A becomes more coherent. A matrix sampled in this way cannot satisfy an RIP, and the sparse reconstruction with such a matrix is possible only if the non-zero elements of the ground-truth $\bar{x}$ are sufficiently separated. Technically, we select the elements of $\operatorname{supp}(\bar{x})$ such that

$$\min_{j,k \in \operatorname{supp}(\bar{x})} |j - k| \geq L,$$

where L is characterized as the *minimum separation*.

We generate a ground-truth vector $\bar{x} \in \mathbb{R}^n$ with sparsity s supported on a random index set (for incoherent matrices) or an index set satisfying the required minimum separation (for coherent matrices) with non-zero entries i.i.d. drawn from the normal distribution. We then compute $b = A\bar{x}$ as the measurements, and apply each solver to produce a reconstruction vector x^* of $\bar{x}$. A reconstruction is considered successful if the relative error satisfies

$$\frac{\|x^* - \bar{x}\|_2}{\|\bar{x}\|_2} < 10^{-3}.$$

We test some cases with different sparsity s of $\bar{x}$, different levels of noise, or different numbers of measurements. We run 100 times independently for each scenario and report the success rate, which is the ratio of the number of successful trials over 100. All experiments are run in parallel with the MATLAB Parallel Computing Toolbox.

The initial guess for all tested algorithms is $x^0 = 0$. The choice of the parameter α in the springback penalty is discussed in Section 6.6.2. For outer iterates of the DCA-springback, we set $\rho = 10^5$, MaxIt $= 10$, and $\epsilon_{\text{outer}} = 10^{-5}$ (for noise-free measurements) or 10^{-3} (for noisy measurements). To implement the ADMM (6.5.9)

for subproblems, we set $\zeta = 10^{-5}$, $\tau = \|A\bar{x} - b\|_2$, and the stopping criterion as either

$$\frac{\|x^{j+1} - x^j\|_2}{\max\{\|x^{j+1}\|_2, \|x^j\|_2\}} < 10^{-5}$$

or the iteration number exceeds 500. The DCA-TL1, the DCA-ℓ_{1-2}, and the DCA-MCP are solved by DCA and their subproblems are also solved by the ADMM. We thus set the regularization parameter $\lambda = 10^{-6}$ and adopt the same parameters of the rest and stopping criterion as the DCA-springback. In particular, the parameter β in the transformed ℓ_1 penalty is set as 1 for the DCA-TL1, following [243], and the parameter μ in the MCP is set as $1/\alpha$ for the DCA-MCP. For the AIHT, we set all parameters as [26]. For the ADMM-ℓ_1, we set $\lambda = 10^{-6}$, $\zeta = 10^{-5}$, $\epsilon_{\text{outer}} = 10^{-5}$ (for noise-free measurements) and 10^{-3} (for noisy measurements), and MaxIt = 5000. For IRLS-ℓ_p, we set $p = 0.5$, $\lambda = 10^{-6}$, $\epsilon_{\text{outer}} = 10^{-8}$, and MaxIt = 1000.

6.6.2 A subroutine for choosing the model parameter

Let us focus on the parameter α of the springback penalty (6.1.4). For an 128×512 random Gaussian matrix, we test the DCA-springback with different α varying among $\{0.2, 0.4, 0.6, 0.8, 1\}$, and different levels of sparsity s among $\{25, 27, \ldots, 65\}$. The DCA-springback with $\alpha = 0.6$ or 0.8, indicated by success rates in Figure 6.2, has the best performance. For small α such as 0.2 and 0.4, the DCA-springback is not satisfactory because the springback penalty performs similarly to the ℓ_1 penalty. For $\alpha = 1$, its performance is also inferior since the convergence condition of the DCA-springback or the posterior verification (6.3.3) can be easily violated with a large α. We refer to the latter reason as the "violating behavior" of the DCA-springback. An "unsuccessful" trial is recognized due to unsatisfactory (but reasonable) reconstruction or violating behavior. Thus, success rates cannot fully reflect "violating behavior," and we also plot the relative errors in Figure 6.2. Indeed, the "violating behavior" often occurs when s becomes large. Performance of $\alpha = 0.8$ and 1 is generally inferior, and also there are few such cases when $\alpha = 0.6$. Thus, we adopt a `safeguard` for $\alpha = 0.7$, a compromise between 0.6 and 0.8. If $\alpha = 0.7$ violates the condition (6.5.6), then we replace 0.7 with the largest constant complying with this condition (6.5.6). That is, we choose

$$\alpha = \min\{0.7, 2\sigma_{\min}(A)/(\|b\|_2 + \tau)\}.$$

Success rates and relative errors with `safeguarded` $\alpha = 0.7$ are also displayed in Figure 6.2, indicating that there is no violating behavior.

Though a reasonable upper bound of α is needed, behaviors for $\alpha = 0.2$ and

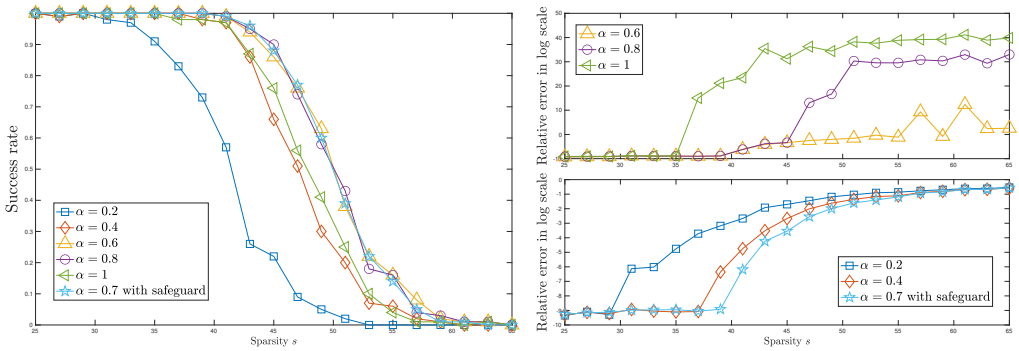


Figure 6.2: Success rates and relative errors in a natural logarithmic scale of reconstruction using DCA-springback under 128×512 random Gaussian sensing matrix, with various α.

0.4 suggest that a lower bound for α should be taken to maintain the satisfactory performance of the DCA-springback in terms of success rates. Especially if A is ill-conditioned in the sense that its singular values lie within a wide range of values, i.e., $\sigma_{\min}(A)$ could be very small, then the condition on α could be pretty stringent. To maintain the success rates of the DCA-springback, we adopt an `efficiency detection` step as follows. If the condition number

$$\mathrm{cond}(A) := \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

is greater than 5 (or other values set by the user), then we start an `efficiency detection` to enforce α to be greater than an *efficiency detection factor* ω. Thus, we suggest choosing α as the following subroutine:

$$\alpha = \begin{cases} \min\left\{0.7, \frac{2\sigma_{\min}(A)}{\|b\|_2+\tau}\right\}, & \text{if } \mathrm{cond}(A) \le 5, \\ \max\left\{\omega, \min\{0.7, \frac{2\sigma_{\min}(A)}{\|b\|_2+\tau}\}\right\}, & \text{otherwise.} \end{cases} \tag{6.6.1}$$

In short, the `safeguard` step suffices to guarantee convergence of the DCA-springback; and the `efficiency detection` step is adopted to maintain the success rates of the DCA-springback for ill-conditioned sensing matrices.

6.6.3 Exact reconstruction of sparse vectors

We first compare the DCA-springback with some state-of-the-art solvers mentioned above for noise-free measurements. We consider both the incoherent and coherent sensing matrices, respectively.

Tests on incoherent matrices. We first consider a ground-truth vector and display its reconstructions by the ADMM-ℓ_1, the DCA-TL1, the DCA-ℓ_{1-2}, the

DCA-MCP, and the DCA-springback. Let the sensing matrix $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix with $(m, n) = (64, 250)$, and the ground-truth $\bar{x} \in \mathbb{R}^{250}$ be a 22-sparse vector with nonzero entries drawn from the standard normal distribution and set the efficiency detection factor as $\omega = 0.5$. The ground-truth and its reconstructions are displayed in Figure 6.3. We see that the DCA-springback, the DCA-MCP, and the DCA-TL1 produce better reconstructions than the ADMM-ℓ_1 and the DCA-ℓ_{1-2}.

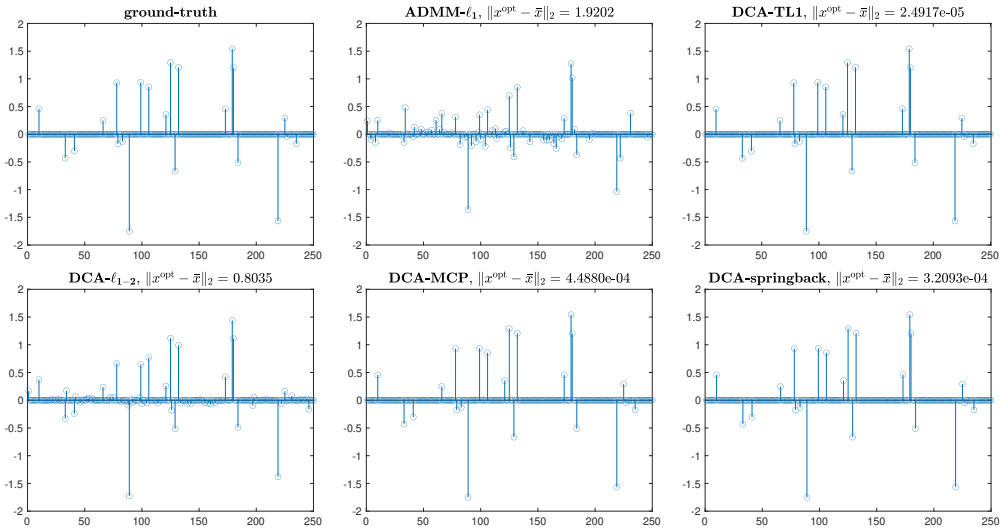


Figure 6.3: A ground-truth and its reconstructions using random Gaussian sensing matrices and noise-free measurements.

We then conduct a more comprehensive study and involve more solvers. We choose the sensing matrix $A \in \mathbb{R}^{m\times n}$ as a random Gaussian matrix and random partial DCT matrices with $(m, n) = (64, 160)$, $(64, 320)$, and $(64, 640)$, and set the efficiency detection factor as $\omega = 0.5$. Different levels of sparsity s varying among $\{6, 8, 10, \ldots, 40\}$ are tested. The success rates of each solver are plotted in Figure 6.4. For both the Gaussian and partial DCT matrices, the IRLS-ℓ_p with $p = 0.5$ has the best performance, followed by the DCA-TL1, the DCA-MCP, and the DCA-springback. In particular, the performances of the DCA-MCP and the DCA-springback are very close because we let the parameter μ in the MCP be $1/\alpha$. The DCA-ℓ_{1-2} performs moderately well, outperforming both the ADMM-ℓ_1 and the AIHT. Our numerical results are consistent with some observations in the literature (e.g., [239, 243]).

Tests on coherent matrices. Now, we choose the sensing matrix $A \in \mathbb{R}^{100\times 1500}$ as a randomly oversampled partial DCT matrix with various refinement factors $\mathcal{F} = 4, 6, 8, 10, 12, 16$ and minimum separation $L = 2\mathcal{F}$, with the sparsity

s varying among $\{5, 7, 9, \ldots, 35\}$. The efficiency detection factor is set as $\omega = 0.5$. The success rates of each solver are plotted in Figure 6.5. This figure suggests that the DCA-TL1, the DCA-MCP, and the DCA-springback are robust regardless of the varying coherence of sensing matrix A. Moreover, when the coherent of A is modest, e.g. $\mathcal{F} = 6, 8$, the DCA-MCP and the DCA-springback perform better than others. In the coherent regime, the DCA-springback is comparable with the DCA-ℓ_{1-2}, and it outperforms the DCA-TL1, the ADMM-ℓ_1, the IRLS-ℓ_p, and the AIHT. However, the best-performance solver IRLS-ℓ_p in the incoherent regime becomes inefficient as A becomes coherent.

6.6.4 Robust reconstruction in the presence of noise

We then consider noisy measurements. The noisy measurements b are obtained by `b = awgn(Ax̄,snr)`, a subroutine of the MATLAB Communication Toolbox, where `snr` corresponds to the value of signal-to-noise ratio (SNR) measured in dB. The larger the value of SNR is, the lighter the noise is added on.

We first consider a ground-truth vector with noisy measurements and display its reconstructions by the ADMM-ℓ_1, the DCA-TL1, the DCA-ℓ_{1-2}, the DCA-MCP, and the DCA-springback. Let the sensing matrix $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix with $(m, n) = (64, 250)$, and the ground-truth $\bar{x} \in \mathbb{R}^{250}$ be a 20-sparse vector with nonzero entries drawn from the standard normal distribution and set the efficiency detection factor as $\omega = 0.4$. The measurement vector $b = A\bar{x}$ is contaminated by 30 dB noise. The ground-truth and its reconstructions are displayed in Figure 6.6. In particular, we see that the DCA-springback works better on small perturbations than the other solvers.

We test both the random Gaussian matrix and the randomly oversampled partial DCT matrix with different levels of noise in dB. For Gaussian measurements, we choose $n = 64$, $m = 128$, and $s = 25$. For the oversampled partial DCT measurements, we test $n = 1500$, $m = 128$, $s = 30$, and $\mathcal{F} = 8$. We run 100 times for each scenario and record the average errors. The efficiency detection factor is set as $\omega = 0.4$.

Once we adopt the `efficiency detection` step, a single "violating behavior" could lift the mean error to a pretty large level. To overcome this computational myopia, we only reserve the accepted results, where a result of the DCA-springback is considered "accepted" if the absolute error $\|x^* - \bar{x}\|_2$ is ten times less than the absolute error of the ADMM-ℓ_1. In addition to errors displayed in Figure 6.7, we report the *acceptance rates* of the DCA-springback, which are ratios of the number of accepted trials over 100.

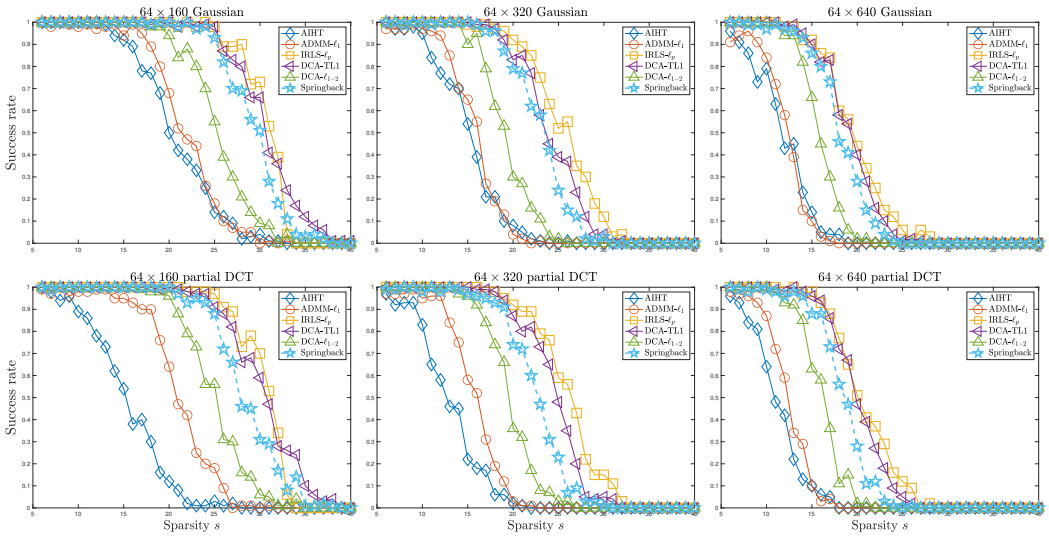


Figure 6.4: Success rates using random Gaussian and partial DCT sensing matrices.

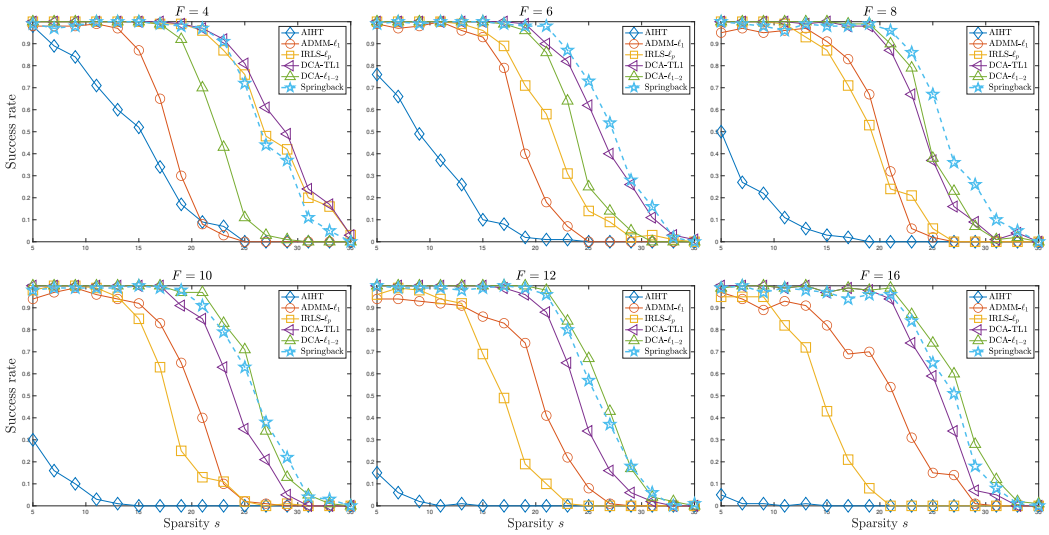


Figure 6.5: Success rates using randomly oversampled partial DCT matrices in $\mathbb{R}^{100\times 1500}$.

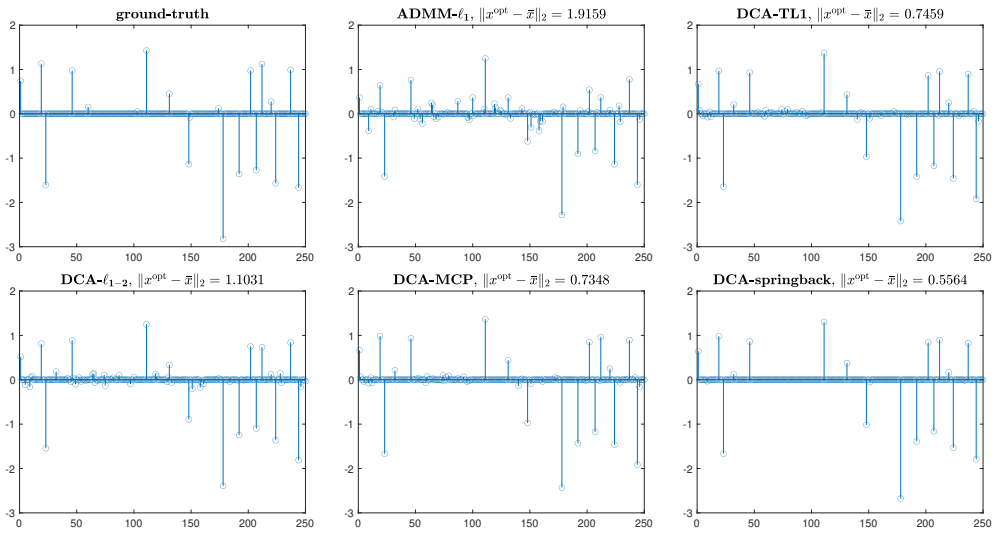


Figure 6.6: A ground-truth and its reconstructions using random Gaussian sensing matrices and noisy measurements.

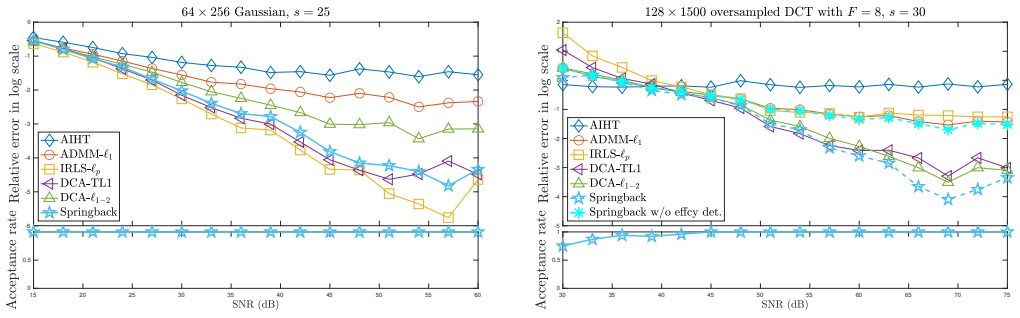


Figure 6.7: Robust reconstruction results with randomly Gaussian and oversampled partial DCT measurements.

According to our experiments, we observe no "violating behaviors" with the Gaussian measurements but a few cases with the oversampled partial DCT measurements when the noise level is relatively large. To illustrate the necessity of the `efficiency detection` step and to validate the convergence condition (6.5.6), we test the DCA-springback without the `efficiency detection` for the randomly oversampled partial DCT measurements, and we do not remove unaccepted trials. The results are labeled as "DCA-springback w/o effcy det." in Figure 6.7, as we see that the DCA-springback only performs slightly better than the ADMM-ℓ_1. Figure 6.7 also shows that the DCA-ℓ_{1-2} and the IRLS-ℓ_p are still sensitive to the coherence of A. For Gaussian measurements, the IRLS-ℓ_p with $p = 0.5$ has the best performance, followed by the DCA-TL1, the DCA-MCP, the DCA-springback, the DCA-ℓ_{1-2}, and the ADMM-ℓ_1. For oversampled DCT measurements, the DCA-springback appears to be the best solver, followed by the DCA-MCP, the DCA-ℓ_{1-2}, and the DCA-TL1, because the noise level is considered in solving the subproblems of the DCA-springback. In both cases, the DCA-springback consistently performs better than the ADMM-ℓ_1 and the DCA-ℓ_{1-2}. AIHT appears not to perform well for both matrices. Based on the plots of the DCA-springback and the DCA-springback without the `efficiency detection`, the model parameter α matters for the same solver.

We also validate some theoretical results proved in Section 6.4.2, with Gaussian measurements perturbed by 45 dB noise. We first study $m = 50$, $n = 160$, and s varying among $\{10, 11, \ldots, 40\}$, and then consider $n = 160$, $s = 20$, and m varying among $\{50, 51, \ldots, 120\}$. Errors of the ADMM-ℓ_1, the DCA-ℓ_{1-2}, and the DCA-springback are plotted in Figure 6.8, and the acceptance rates of the DCA-springback are also displayed. According to our analysis in Section 6.4.2, for an RIP sensing matrix A and an s-sparse $\bar{x}$, when $s \leq C$ (C is given in (6.4.7)) or m is limited by some constant, the estimation (6.4.1) of the springback-penalized model is tighter than the estimation (6.3.2) of the ℓ_1- and ℓ_{1-2}-penalized models in the sense of

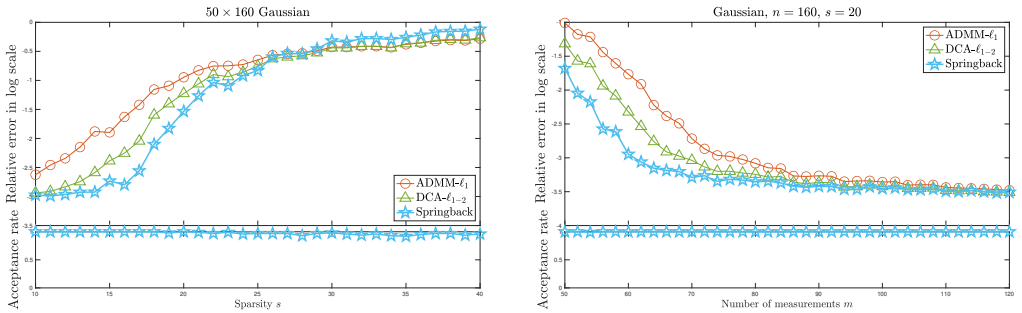


Figure 6.8: Numerical validation of theoretical results in Section 6.4.2.

(6.4.5). We see in the left plot of Figure 6.8 that the error of the DCA-springback is less than the others for small s, and it becomes larger than the others when s exceeds some constant. The right plot also indicates that the error of the DCA-springback is less than the others when m is relatively small.

6.6.5 Remarks on numerical results

As observed in the literature, reconstruction results by different models may vary for different scenarios, and no one can unanimously outperform all the others for all scenarios. For instance, the IRLS-ℓ_p prevails in the incoherent regime but quickly fails in the coherent regime, see [125, 239]. For incoherent sensing matrices, the IRLS-ℓ_p and the DCA-TL1 perform better than the DCA-ℓ_{1-2} and the ADMM-ℓ_1, while the DCA-ℓ_{1-2} performs the best for coherent sensing matrices; see [239, 243]. The DCA-TL1 is robust, and it performs well for both incoherent and coherent sensing matrices, while it is less efficient than either the IRLS-ℓ_p in the incoherent regime or the DCA-ℓ_{1-2} in the coherent regime.

Together with these known facts and our numerical observations, we have the following remarks on the numerical performance of the DCA-springback:

- For an incoherent sensing matrix: the DCA-springback performs slightly worse than the IRLS-ℓ_p and the DCA-TL1.
- For a coherent sensing matrix: the DCA-springback performs slightly worse than the DCA-ℓ_{1-2} but better than the DCA-TL1.
- For a sensing matrix with modest coherence: the DCA-springback performs comparably with the DCA-MCP, and they perform better than the others.

Similar comparison results are also observed when the measurements are contaminated by some noise. For all the three scenarios, the DCA-springback and the DCA-MCP perform comparably if the parameter μ of the MCP is set as $1/\alpha$, and their performances with well-tuned parameters are also comparable. Moreover, we see that only the DCA-springback, the DCA-MCP, and the DCA-TL1 are robust with respect to the coherence of the sensing matrix. The DCA-springback and the DCA-MCP perform better than the DCA-TL1 in the coherent regime but worse in the incoherent regime. When the coherence of the sensing matrix is unknown, for example, when the sensing hardware cannot be modified or upgraded, coherence-robust algorithms such as the DCA-springback and the DCA-MCP are preferred for signal reconstruction.

Chapter 7

The Enhanced Total Variation Model for Image Reconstruction

The total variation (TV) regularization has phenomenally boosted various variational models for image processing tasks. We propose to combine the backward diffusion process in the earlier literature on image enhancement with the TV regularization, and show that the resulting enhanced TV minimization model is particularly effective for reducing the loss of contrast. The main purpose of this chapter is to establish stable reconstruction guarantees for the enhanced TV model from noisy subsampled measurements with two sampling strategies, non-adaptive sampling for general linear measurements and variable-density sampling for Fourier measurements. In particular, under some weaker restricted isometry property conditions, the enhanced TV minimization model is shown to have tighter reconstruction error bounds than various TV-based models for the scenario where the level of noise is significant and the amount of measurements is limited. The advantages of the enhanced TV model are also numerically validated by preliminary experiments on the reconstruction of some synthetic, natural, and medical images.

7.1 Introduction

Since the work of Rudin, Osher, and Fatemi [182], various variational models based on the total variation (TV) have been intensively studied for image processing problems; see, e.g., [48, 50] for reviews. Given linear measurements $y \in \mathbb{C}^m$ observed via

$$y = \mathcal{M}\bar{X} + e \tag{7.1.1}$$

from an unknown image $\bar{X} \in \mathbb{C}^{N\times N}$, where $\mathcal{M} : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ is a linear operator defined component-wisely by

$$[\mathcal{M}(\bar{X})]_j := \langle M_j, \bar{X}\rangle = \mathrm{tr}(M_j\bar{X}^*),$$

for suitable matrices M_j with m considerably smaller than N^2, and $e \in \mathbb{C}^m$ is a noise term bounded by $\|e\|_2 \leq \tau$ with level $\tau \geq 0$, reconstruction of the unknown $\bar{X}$ can be modeled as the following TV minimization problem:

$$\min_{X \in \mathbb{C}^{N \times N}} \|X\|_{\mathrm{TV}} \quad \text{s.t.} \quad \|\mathcal{M}X - y\|_2 \leq \tau, \tag{7.1.2}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the TV semi-norm. Note that the TV semi-norm can be mainly categorized as the isotropic [46] and anisotropic [47] cases for discrete images. In this chapter, we discuss how to enhance the canonical constrained TV model (7.1.2) by the proposed springback regularization in Chapter 6 for image reconstruction and establish stable reconstruction guarantees.

As profoundly analyzed in [157], the constrained TV model (7.1.2) has the advantage of reconstructing high-quality images from a relatively small number of measurements. Theoretical analysis in [157] is mainly based on the seminal compressed sensing (CS) works [41, 74]. Note that the classic CS theory assumes the sparsity of the (vector) signal of interest or its coefficients under certain transformations. Correspondingly the signal reconstruction can be modeled as some ℓ_1-norm minimization problems. The CS theory can be extended to image reconstruction because natural images usually have (approximately) sparse gradients. Indeed, mathematically the TV semi-norm of a discrete image $X \in \mathbb{C}^{N \times N}$ is just the sum of the magnitudes of all entries $|[\nabla X]_{j,k}|$ in its gradient $\nabla X \in \mathbb{C}^{N \times N \times 2}$. That is,

$$\|X\|_{\mathrm{TV}} := \|\nabla X\|_1 = \sum_{j,k} |[\nabla X]_{j,k}|, \tag{7.1.3}$$

where the definitions of ∇X and $|[\nabla X]_{j,k}|$ are given as follows. For any image $X \in \mathbb{C}^{N \times N}$ represented by an $N \times N$ block of pixel intensities with all intensities $X_{j,k}$ in $[0, 1]$, the discrete directional derivatives of $X \in \mathbb{C}^{N \times N}$ are defined in a pixel-wise manner as

$$\begin{aligned} X_x &: \mathbb{C}^{N \times N} \to \mathbb{C}^{(N-1) \times N}, \quad (X_x)_{j,k} := X_{j+1,k} - X_{j,k}, \\ X_y &: \mathbb{C}^{N \times N} \to \mathbb{C}^{N \times (N-1)}, \quad (X_y)_{j,k} := X_{j,k+1} - X_{j,k}. \end{aligned}$$

The discrete gradient transform $\nabla : \mathbb{C}^{N \times N} \to \mathbb{C}^{N \times N \times 2}$ is defined in a matrix form as

$$[\nabla X]_{j,k} := \begin{cases} ((X_x)_{j,k}, (X_y)_{j,k}), & 1 \leq j \leq N-1,\ 1 \leq k \leq N-1, \\ (0, (X_y)_{j,k}), & j = N,\ 1 \leq k \leq N-1, \\ ((X_x)_{j,k}, 0), & 1 \leq j \leq N-1,\ k = N, \\ (0, 0), & j = k = N. \end{cases}$$

If the magnitude

$$|[\nabla X]_{j,k}| = |(X_x)_{j,k}| + |(X_y)_{j,k}|,$$

then it leads to the *anisotropic TV semi-norm* $\|\cdot\|_{\mathrm{TV}_a}$ as defined in [47, 81], that is, the sum of the magnitudes of its discrete gradient

$$\|X\|_{\mathrm{TV}_a} := \sum_{j,k} |(X_x)_{j,k}| + |(X_y)_{j,k}|. \tag{7.1.4}$$

If

$$|[\nabla X]_{j,k}| = \sqrt{(X_x)_{j,k}^2 + (X_y)_{j,k}^2},$$

then it leads to the *isotropic TV semi-norm* $\|\cdot\|_{\mathrm{TV}_i}$ as defined in [46]:

$$\|X\|_{\mathrm{TV}_i} := \sum_{j,k} \sqrt{(X_x)_{j,k}^2 + (X_y)_{j,k}^2}.$$

If we regard ∇X as an $N^2 \times 2$ matrix, then $\|X\|_{\mathrm{TV}_a}$ and $\|X\|_{\mathrm{TV}_i}$ are the $\ell_{1,1}$ and $\ell_{2,1}$ norms of ∇X, respectively. Since both TV semi-norms are equivalent subject to a factor of $\sqrt{2}$ in the sense that

$$\|X\|_{\mathrm{TV}_i} \le \|X\|_{\mathrm{TV}_a} \le \sqrt{2}\|X\|_{\mathrm{TV}_i}, \tag{7.1.5}$$

similarly to [157], we only consider the anisotropic case for succinctness and the following discussion can be extended to the isotropic case analogously. Moreover, note that

$$\|\nabla X\|_2 = \left(\sum_{j,k} (X_x)_{j,k}^2 + (X_y)_{j,k}^2\right)^{1/2}$$

in the second component of the enhanced TV regularization (7.1.6) is the $\ell_{2,2}$ norm of ∇X.

Models using the ℓ_1-norm are fundamental to various CS problems, while solutions to such models may be over-penalized because the ℓ_1 regularization tends to underestimate the high-amplitude components of the solution, as analyzed in [83]. Accordingly, many non-convex alternatives have been proposed in the literature to overcome this pitfall and thus promote sparsity more firmly; see, e.g., the ℓ_p $(0 < p < 1)$ regularization [54, 92], the ℓ_{1-2} regularization [238, 239], and the transformed ℓ_1 regularization [242, 243]. The non-convexity feature in image processing has also been emphasized in various papers; see, e.g., [159]. Besides, the springback regularization proposed in Chapter 6, and it can be generalized as the following for

discrete images:

$$\mathcal{R}_\alpha(X) := \|\nabla X\|_1 - \frac{\alpha}{2}\|\nabla X\|_2^2, \tag{7.1.6}$$

where $\alpha > 0$ is a meticulously-chosen parameter to ensure the positiveness or the well-definedness of (7.1.6), $\|\nabla X\|_1$ is the TV term (7.1.3) and we focus on the anisotropic definition (7.1.4) in this chapter, and $\|\nabla X\|_2^2$ is the sum of the squared magnitudes of ∇X. Note that the springback regularization (7.1.6) is of difference-of-convex. To some extent, it keeps both the nice recoverability of various non-convex surrogates of the TV regularization and the computability of the original TV regularization. To be consistent with the TV literature, we call (7.1.6) an *enhanced TV* regularization in this chapter.

Non-convex penalties proposed in the CS literature are mainly rooted in the field of statistics, and they are usually applied in straightforward ways in the image processing literature. Interestingly, as elaborated in Section 7.1.1, the enhanced TV regularization (7.1.6) has some intrinsic interpretations from the perspective of image processing. We are thus encouraged to consider the enhanced TV model

$$\min_{X\in\mathbb{C}^{N\times N}} \mathcal{R}_\alpha(X) \quad \text{s.t.} \quad \|\mathcal{M}X - y\|_2 \le \tau \tag{7.1.7}$$

for image reconstruction, and we aim at establishing some stable reconstruction guarantees theoretically. It is worth noting that, despite the theoretical reconstruction guarantees established in Chapter 6 for sparse signals or signals that are sparse after an orthonormal transform, the guarantees established in Chapter 6 are not applicable to the enhanced TV model (7.1.7). The reason is that the gradient transform $\nabla : X \to \nabla X$ fails to be orthonormal, as mentioned in [157]. Also, we notice that the idea of enhancing the TV regularization (the isotropic version) with a subtraction of a squared norm of the image gradient was skated over in [148], and it was empirically tested for some image denoising problems despite the lack of rigorous study for reconstruction guarantees from a few measurements.

7.1.1 An image processing view of the enhanced TV regularization

Solutions to TV-based models may lose contrast across edges. That is, the contrast of the regions on both sides of an edge may be reduced, and thus blur may occur near the edge. We refer the reader to [21, 208] for discussions on the loss of contrast caused by various image processing models using TV regularization.

Partial differential equations (PDEs) and variational approaches have been intensively investigated to enhance the contrast. On the PDE side, some well-known approaches were proposed to tackle the loss of contrast for image enhancement. For

example, the shock filter was proposed in [160] to deal with blur-like image degradations, creating strong discontinuities at image edges and flattening the image within homogeneous regions. Afterwards, the shock filter has been generalized in many ways; see, e.g., [6, 229]. Another important example is the forward-and-backward (FAB) diffusion scheme proposed in [98] to simultaneously remove the noise and enhance the contrast. Since then, a number of influential works regarding the FAB diffusion have been conducted; see, e.g., [226, 228, 230]. Despite that different PDE schemes were designed, a common feature of these works is that the *backward diffusion process* is adopted to enhance the contrast of the edges in a concerning image. Since backward diffusion is a classic example of an ill-posed problem [214], most of these PDE schemes sound numerically challenging; we refer the reader to [51, 52, 227] on discretizing and solving these PDEs efficiently. On the variational side, there are attempts to add negative terms into the variational model to maximize the contrast, see, e.g., [94, 162], though their connections with the TV regularization are not considered.

We remark that the enhanced TV model (7.1.7) has a connection to the backward diffusion approach from the PDE perspective. A detailed explanation in the context of the Euler–Lagrange (E–L) equation in a continuum setting is included in Section 7.6.1. Briefly speaking, the term $-\frac{\alpha}{2}\|\nabla X\|_2^2$ generates an additional backward diffusion term $-\alpha\Delta X$ into the E–L equation corresponding to the classic TV regularization. In Figure 7.1, we empirically illustrate that the enhanced TV regularization (7.1.6) is very effective for some fundamental denoising and deblurring problems. Figure 7.1 clearly shows that the enhanced TV regularization (7.1.6) outperforms the original TV regularization in removing noise, reducing loss of contrast, and maintaining smoothness inside homogeneous regions. These compelling performances clearly motivate us to consider theoretical reconstruction guarantees for the enhanced TV model (7.1.7). Implementation details for reproducing Figure 7.1 are enclosed in Section 7.6.2.

In Figure 7.1, we also note that the enhanced TV regularization (7.1.6) may not ideally overcome another drawback of TV: the *staircase effect*. That is, solutions to TV-based models may have stair-like edges. Many efforts are trying to avoid this effect, including the replacement of the TV regularization with an exponentiation term of it [24], the usage of the infimal convolution of functionals with first- and second-order derivatives as regularizer [49], the addition of some higher-order terms into the E–L equation corresponding to the variational TV model [53], the total generalized variation [36], the usage of some modified infimal convolutions [185, 186] regarding [49], and many others.

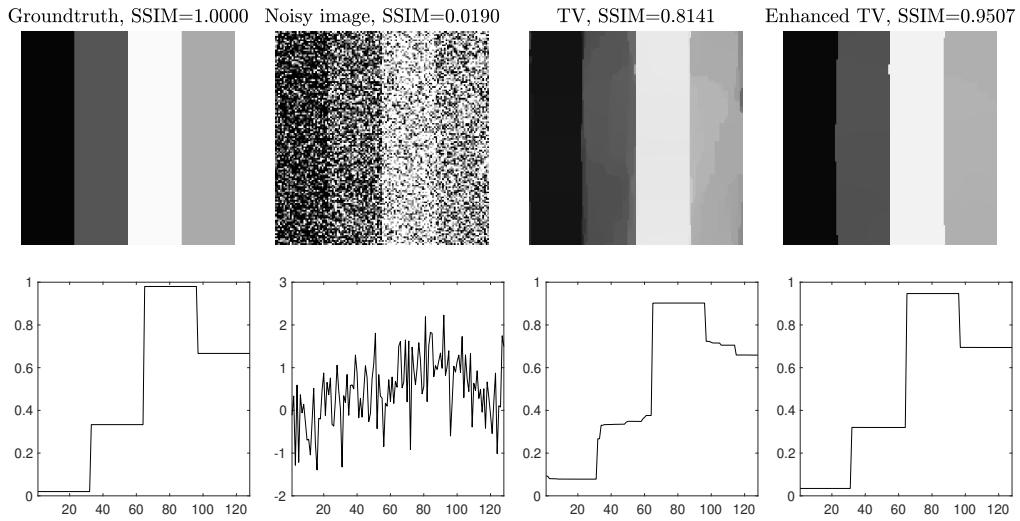


Figure 7.1: Illustration of the TV and enhanced TV regularization for image denoising. Top row: SSIM values of each image; Bottom row: intensity profiles of each image along the horizontal straight line splitting the image equally.

7.1.2 A compressed sensing view of the enhanced TV regularization

In addition to the PDE and variational perspectives, another interpretation of the enhanced TV regularization (7.1.6) can be given from the perspective of CS. As previously discussed, an image X is mostly sparse after the gradient transform $\nabla : X \to \nabla X$. Mathematically, CS amounts to minimizing the ℓ_0 norm of the image gradient, i.e., $\|\nabla X\|_0$, which counts the number of non-zero entries of ∇X. To bypass the NP-hard nature of the ℓ_0 norm, we typically seek its alternatives which lead to more tractable models. In the context of image reconstruction, we have the TV regularization [47, 81, 182]

$$\|X\|_{\mathrm{TV}} = \sum_{j,k} |[\nabla X]_{j,k}|\,,$$

which corresponds to the ℓ_1 penalty in CS. We also have the transformed TV regularization [117]

$$\|X\|_{\mathrm{TTV}} = \sum_{j,k} \frac{(\beta+1)\,|[\nabla X]_{j,k}|}{\beta + |[\nabla X]_{j,k}|}$$

with $\beta > 0$, which corresponds to the transformed ℓ_1 regularization [243] in CS. Moreover, we have the weighted difference of anisotropic and isotropic TV regularization [139]

$$\|X\|_{\mathrm{TV}_a} - \alpha\|X\|_{\mathrm{TV}_i} = \sum_{j,k} \left(|(X_x)_{j,k}| + |(X_y)_{j,k}| - \alpha\sqrt{(X_x)_{j,k}^2 + (X_y)_{j,k}^2} \right)$$

and the minimax concave penalty (MCP) [240]

$$\|X\|_{\mathrm{MCP-TV}} = \sum_{j,k} \phi_\mu(|[\nabla X]_{j,k}|),$$

where $\mu > 0$ and

$$\phi_\mu(x) = \begin{cases} |x| - x^2/(2\mu), & |x| \le \mu, \\ \mu/2, & |x| \ge \mu. \end{cases}$$

Our enhanced TV regularization (7.1.6) can also be written as

$$\mathcal{R}_\alpha(X) = \sum_{j,k} \left[|[\nabla X]_{j,k}| - \frac{\alpha}{2} \left((X_x)_{j,k}^2 + (X_y)_{j,k}^2 \right) \right].$$

In image reconstruction, it is desirable for regularization terms to generate reasonably close approximations of $\|\nabla X\|_0$. Since all the regularization terms mentioned above are separable, we can compare their behavior in terms of each component. We adopt the anisotropic definition $|[\nabla X]_{j,k}| = |(X_x)_{j,k}| + |(X_y)_{j,k}|$, except for the $\mathrm{TV}_a - \alpha\mathrm{TV}_i$ regularization. We set β, μ, and α to 1 for all regularization terms. We plot the level curves of each component with respect to $|(X_x)_{j,k}|$ and $|(X_y)_{j,k}|$ in Figure 7.2. Note that the axes of color bars are intended not to be unified for better visualization. The level lines of the ℓ_0 norm are 0 at the origin, 1 at the axes, and 2 elsewhere. Apart from the convex anisotropic TV regularization, all other regularization terms are non-convex and promote the approximation behavior to the ℓ_0 norm. We observe from Figure 7.2 that all regularization terms preserve 0 at the origin, indicating that they behave similarly within homogeneous regions of images. Additionally, our enhanced TV regularization is closer to the ℓ_0 norm than the other terms at both axes. This suggests that the enhanced TV regularization performs analogously to the ℓ_0 norm around horizontal and vertical edges. In comparison, the $\mathrm{TV}_a - \mathrm{TV}_i$ regularization yields 0 at both axes. Moreover, we note that the transformed TV regularization behaves like plain shrinkage from the anisotropic TV. Furthermore, the truncated definition of the MCP-TV regularization provides it with a closer approximation to the ℓ_0 norm within the non-axis area than other regularization terms, suggesting that this regularization may preserve the behavior of the ℓ_0 norm along oblique edges. However, the truncated definition of the MCP-TV may confound oblique edges with horizontal/vertical edges because it has the same values from the end of both axes and non-axis areas. Meanwhile, the enhanced TV regularization performs better than the MCP-TV along the horizontal or vertical edge because the enhanced TV regularization preserves the behavior of the ℓ_0 norm at the end of box axes better than the MCP-TV. These observations

suggest that the enhanced TV regularization may be a good proxy of the ℓ_0 norm in the context of image reconstruction. To compare scalar regularization terms and shrinkage operators for corresponding proximal mappings, we refer to Chapter 6.

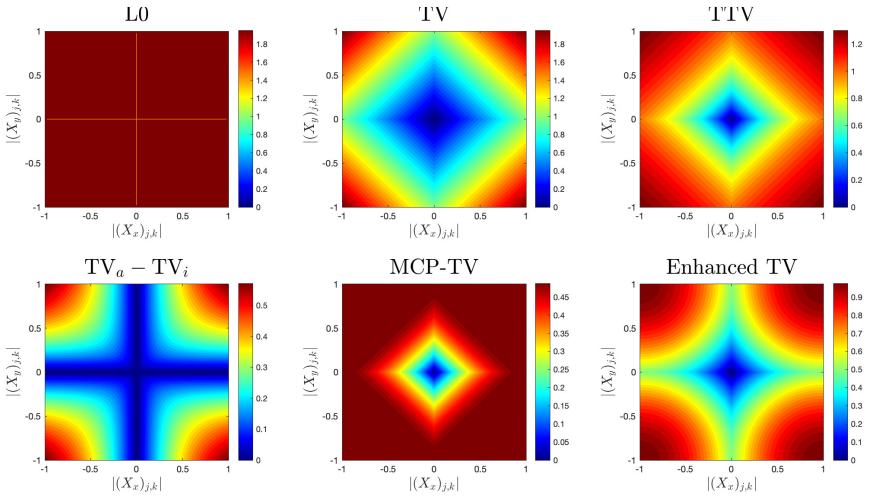


Figure 7.2: Level curves of different regularization terms with anisotropic definitions: $|[\nabla X]_{j,k}| = |(X_x)_{j,k}| + |(X_y)_{j,k}|$.

We focus on the anisotropic version of the TV regularization because it is the ℓ_1 norm of the image gradient ∇X, when viewed as a vector. This fact makes the anisotropic TV regularization better suited for image reconstruction than the isotropic version. As evidence, the level curves in Figure 7.3 demonstrate that the isotropic TV regularization provides a worse approximation to the ℓ_0 norm than the anisotropic one. Similar results will also be presented in Section 7.5.

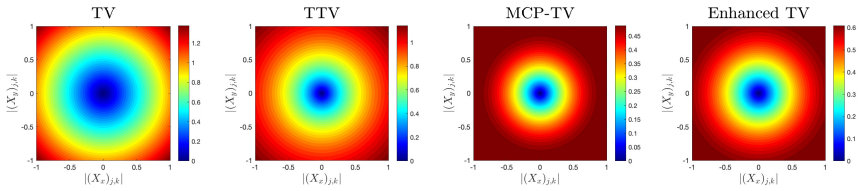


Figure 7.3: Level curves of different regularization terms with isotropic definitions: $|[\nabla X]_{j,k}| = \sqrt{(X_x)_{j,k}^2 + (X_y)_{j,k}^2}$.

7.1.3 Contributions

In the CS context, it is possible to *exactly* recover a signal if the signal is sparse and its measurements are noise-free; otherwise, we can only establish *stable* recovery guarantees. The term *stable* in this chapter is mainly concerned with both inexact sparsity and measurement noise. Our analysis is conducted under the *restricted*

isometry property (RIP) framework studied in [44]. We say that a linear operator $\mathcal{A}: \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$ has the RIP of order s and level $\delta \in (0,1)$ if

$$(1-\delta)\|X\|_2^2 \leq \|\mathcal{A}X\|_2^2 \leq (1+\delta)\|X\|_2^2 \quad \forall s-\text{sparse } X \in \mathbb{C}^{n_1 \times n_2}, \tag{7.1.8}$$

and the smallest δ for (7.1.8) is said to be the *restricted isometry constant* (RIC) associated with $\mathcal{A}$.

We first investigate *non-adaptive* subsampled linear RIP measurements of an image $\bar{X} \in \mathbb{C}^{N \times N}$ with noise level $\tau > 0$. By "non-adaptive," we mean that the sampling strategy is not designed with specific structures or under certain distributions. In Theorem 7.3.6, we show that the enhanced TV model (7.1.7) can stably reconstruct an image $\bar{X} \in \mathbb{C}^{N \times N}$ from some non-adaptive subsampled linear RIP measurements which are contaminated by noise, with the RIP order $\mathcal{O}(s)$, the RIP level $\delta < 0.6$, and the noise level $\tau > 0$. Moreover, the required RIP level $\delta < 1/3$ derived in [157] for the TV model (7.1.2) is weakened to $\delta < 0.6$ for the enhanced TV model (7.1.7) under the additional condition (7.3.8) for the parameter α. We also show in Theorem 7.3.9 that the reconstruction error bound in Theorem 7.3.6 can be further improved if more measurements are allowed.

The above reconstruction guarantees for non-adaptive measurements require the subsampled measurements and the Haar wavelet basis to be sufficiently incoherent. This requirement is satisfied by many kinds of measurements except for the Fourier frequency measurements, because low-order wavelets and Fourier measurements are highly correlated, as analyzed in [123]. Fourier measurements play essential roles in many imaging tasks. For example, as discussed in [84, 123], the measurement process of various image processing procedures such as radar, sonar, and computer tomography can be modeled (with appropriate approximation and discretization) by taking samples from weighted discrete Fourier transforms. It is also known (see, e.g., [141]) that measurements taken for magnetic resonance imaging (MRI) can be well modeled as Fourier coefficients of the desired image.

On the other hand, many empirical pieces of evidence, including the first works [140, 141] for compressed sensing MRI, have shown that better reconstruction quality is possible by subsampling Fourier frequency measurements with a preference for low frequencies over high frequencies. Thus, we follow the *density-variable* sampling strategy proposed in [123] and choose Fourier measurements randomly according to an *inverse square law density*. We show that from at least $m \gtrsim s\log^3(s)\log^5(N)$ such subsampled Fourier measurements with $s \gtrsim \log(N)$, the enhanced TV model (7.1.7) reconstructs an unknown image $\bar{X}$ stably with high probabilities. We also show that the least amount of Fourier measurements required by the enhanced TV

model (7.1.7) is only $(0.6/(1/3))^{-2} \approx 30.86\%$ of that by the TV model (7.1.2) as established in [123].

7.1.4 Related works

We briefly review some TV-related works on image reconstruction. The reconstruction of a one-dimensional image in $\mathbb{C}^N$ with an exactly s-sparse gradient from noise-free, uniformly subsampled Fourier measurements was considered in [41], without stability analysis concerning the inexact sparsity or noise. It was shown that this one-dimensional image could be recovered exactly by solving the corresponding TV model with high probabilities, provided that the number of measurements m satisfies

$$m \gtrsim s \log(N).$$

The reconstruction of a one-dimensional image using noisy measurements was then considered in [38]. The stability of the reconstruction of approximately sparse images from noisy measurements was first shown in [157] for two-dimensional images and soon extended to higher-dimensional cases in [156]. More specifically, it was asserted in [157] that, from some non-adaptive subsampled linear RIP measurements of an image $\bar{X} \in \mathbb{C}^{N\times N}$ with the RIP order $\mathcal{O}(s)$, the RIP level $\delta < 1/3$, and the noise level $\tau > 0$, the solution X^{opt} to the TV model (7.1.2) satisfies

$$\|\bar{X} - X^{\text{opt}}\|_2 \lesssim \log\left(\frac{N^2}{s}\right)\left(\frac{\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1}{\sqrt{s}} + \tau\right), \tag{7.1.9}$$

where $(\nabla\bar{X})_s$ is the best s-sparse approximation to the discrete gradient ∇X. Moreover, with more measurements, it was shown in [157] that the log factor in the bound (7.1.9) could be removed, and thus the bound (7.1.9) can be improved as

$$\|\bar{X} - X^{\text{opt}}\|_2 \lesssim \frac{\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1}{\sqrt{s}} + \tau. \tag{7.1.10}$$

In comparison with the bound (7.1.10), the reconstruction error bound for the enhanced TV model (7.1.7) in Theorem 7.3.9 is tighter if the level of noise τ is relatively large and the number of measurements m is limited. More discussions can be found in Section 7.3.3. Besides, the RIP level is assumed to satisfy $\delta < 1/3$ in [157] for the TV model (7.1.2), while we weaken it to $\delta < 0.6$ for the enhanced TV model (7.1.7). Though $\delta < 1/3$ can be improved, as remarked in [157], the reconstruction error bounds (7.1.9) and (7.1.10) for the TV model (7.1.2) tend to be infinity if $\delta \to 0.6$ (cf. the proof of Proposition 3 in [157]). On the other hand, the bounds in Theorems 7.3.6 and 7.3.9 for the enhanced TV model (7.1.7) are still

reasonably valid when $\delta \to 0.6$; meanwhile, the upper bound required for α tends to be 0 correspondingly. Thus, as $\delta \to 0.6$, the bounds (7.3.11) and (7.3.13) in Theorems 7.3.6 and 7.3.9 for the enhanced TV model (7.1.7) assert the stability of the TV model (7.1.2) in image reconstruction from a few linear RIP measurements.

As mentioned, guarantees for non-adaptive measurements require the subsampled measurements and the Haar wavelet basis to be sufficiently incoherent. Thus, the mentioned guarantees in [156, 157] cannot be directly applied to the situation of Fourier measurements. The first results on image reconstruction from Fourier measurements were derived in [123] and [164], in which *uniform* and *non-uniform*[1] reconstruction guarantees are considered, respectively. More specifically, the approach in [123] requires a larger number of measurements than [164], while its reconstruction error bound is sharper than that in [164]. In [123], uniform reconstruction guarantees were derived for two-dimensional images from noisy Fourier measurements, chosen randomly according to an inverse square law density. Specifically, from at least

$$m \gtrsim s \log^3(s) \log^5(N)$$

such subsampled Fourier measurements with $s \gtrsim \log(N)$, the reconstruction error bound for the TV model (7.1.2) was derived in the same form of (7.1.10). We refer to, e.g., [2, 3, 121], for more discussions. As we focus on the uniform reconstruction from non-adaptive measurements, we follow the approach in [123] to consider Fourier measurements.

7.1.5 Outline of the chapter

The rest of this chapter is organized as follows. In the next section, we summarize some notation and technical backgrounds. In Section 7.3, we establish stable image reconstruction guarantees for the enhanced TV model (7.1.7) from non-adaptive subsampled linear RIP measurements and variable-density subsampled Fourier measurements, respectively. Proofs of the results in Section 7.3 are presented in Section 7.4. In Section 7.5, we report some numerical results when the enhanced TV model (7.1.7) is applied to some image reconstruction problems. Different kinds of images with subsampled Fourier measurements are tested.

[1] In the context of compressed sensing, a *uniform reconstruction guarantee* indicates that a single random draw of a given measurement operator suffices to recover all sparse or approximately sparse vectors. In contrast, a *non-uniform recovery guarantee* states that a single random draw is sufficient for recovery of a fixed vector.

7.2 Preliminaries

We first summarize some notation and recall some preliminary technical backgrounds.

7.2.1 Notation

For a matrix $X \in \mathbb{R}^{m\times n}$, let $\mathrm{supp}(X) := \{(j,k) : X_{j,k} \neq 0\}$ be the support of X, and $\|X\|_0$ be the cardinality of $\mathrm{supp}(X)$. X is said to be s-sparse if $\|X\|_0 \leq s$. Let

$$\|X\|_{p,q} := \left(\sum_{j=1}^{m}\left(\sum_{k=1}^{n}|X_{j,k}|^p\right)^{q/p}\right)^{1/q}$$

be the entry-wise $\ell_{p,q}$ norm ($p,q \geq 1$) of X. If $p = q$, $\|X\|_{p,p}$ is denoted by $\|X\|_p$ for short. In particular, the $\ell_{2,2}$ norm is also known as the Frobenius norm, which is induced by the inner product $\langle X, Y\rangle := \sum_{j=1}^{m}\sum_{k=1}^{n} X_{i,j}Y_{i,j} = \mathrm{tr}(XY^*)$ for any $X, Y \in \mathbb{C}^{m\times n}$, where X^* denotes the adjoint of the matrix X. For an index set $S \subset \{1,2,\ldots,m\}\times\{1,2,\ldots,n\}$, let $X_S \in \mathbb{R}^{m\times n}$ be the matrix with the same entries as X on indices S and zero entries on indices S^c. The only exception is $\mathcal{F}_\Omega$. We denote by $\mathcal{F}_\Omega$ the restriction of the bivariate discrete Fourier transform $\mathcal{F}$ to a subset $\Omega \subset \{-N/2+1,\ldots,N/2\}^2$. Logarithm without indicating base is with respect to base 2.

7.2.2 Haar wavelet system

The Haar wavelet system provides a simple yet powerful sparse approximation of digital images. The following descriptions on this system can be found in, e.g., [157]. The *univariate* Haar wavelet system is a complete orthonormal system of square-integrable functions on the unit interval, consisting of the constant function

$$H^0(t) = \begin{cases} 1, & 0 \leq t < 1, \\ 0, & \text{otherwise}, \end{cases}$$

the mother wavelet

$$H^1(t) = \begin{cases} 1, & 0 \leq t < 1/2, \\ -1, & 1/2 \leq t < 1, \end{cases}$$

and the dyadic dilations and translates of the mother wavelet

$$H_{j,k}(t) = 2^{j/2}H^1(2^j t - k)$$

for $j \in \mathbb{N}$, $0 \le k < 2^j$. The *bivariate* Haar wavelet system is an orthonormal system for the space $L_2(Q)$ of square-integrable functions on the unit square $Q = [0,1)^2$, and it is derived from the univariate Haar system by tensor product. The bivariate Haar system consists of the constant function and all functions

$$x = (u, v), \quad H^{\ell}_{j,k}(x) = 2^j H^{\ell}(2^j x - k),$$

for $j \ge 0$, $k \in \mathbb{Z}^2 \cap 2^j Q$, and $\ell \in V := \{\{0,1\},\{1,0\},\{1,1\}\}$, where

$$H^{\ell}(u, v) = H^{\ell_1}(u) H^{\ell_2}(v)$$

and $\ell = (\ell_1, \ell_2) \in V$. Discrete images are isometric to the space $\Sigma_N \subset L_2(Q)$ of piecewise-constant functions

$$\Sigma_N = \left\{ f \in L_2(Q) : \ f(u,v) = c_{j,k}, \ \frac{j-1}{N} \le u < \frac{j}{N}, \ \frac{k-1}{N} \le v < \frac{k}{N} \right\} \tag{7.2.1}$$

with $c_{j,k} = N X_{j,k}$. If $N = 2^n$, then the bivariate Haar basis is restricted to the $2^n \times 2^n = N^2$ basis functions $\{H^{\ell}_{j,k} : \ j \le n-1\}$ and identified as some discrete images $h^{\ell}_{j,k}$ via (7.2.1) forms an orthonormal basis for $\mathbb{C}^{N \times N}$. For any given $\ell = (\ell_1, \ell_2) \in V$, we denote by $\mathcal{H}$ the bivariate Haar transform

$$X \mapsto (\langle X, h^{\ell}_{j,k} \rangle)_{j,k}.$$

By a slight abuse of notation, we also denote by $\mathcal{H}$ the unitary matrix representing this bivariate Haar transform. That is, we denote by $\mathcal{H}X$ the matrix product that generates $(\langle X, h^{\ell}_{j,k} \rangle)_{j,k}$.

Some properties of the bivariate Haar wavelet system are summarized below, and the proofs can be found in [157].

Lemma 7.2.1 *Suppose $X \in \mathbb{C}^{N \times N}$ is mean-zero, and let $c_{(k)}(X)$ be the bivariate Haar coefficient of X having the kth largest magnitude, or the entry of the bivariate Haar transform $\mathcal{H}X$ having the kth largest magnitude. Then, for all $k \ge 1$,*

$$|c_{(k)}(X)| \le \tilde{C} \frac{\|\nabla X\|_1}{k},$$

where $\tilde{C} > 0$ is some constant.

Lemma 7.2.2 *Let $N = 2^n$. For any indices (j,k) and $(j,k+1)$, there are at most $6n$ bivariate Haar wavelets which are not constant on these indices, i.e., $|h^{\ell}_{j,k}(j,k+1) - h^{\ell}_{j,k}(j,k)| > 0$.*

Lemma 7.2.3 *The bivariate Haar wavelets satisfy* $\|\nabla h_{j,k}^{\ell}\|_1 \leq 8$ *for all* j, k, ℓ.

7.2.3 Discrete Fourier system

In addition to general RIP measurements, we particularly investigate Fourier measurements. Let $N = 2^n$ be a power of 2, where $n \in \mathbb{N}$. The following facts of Fourier basis and transform in the context of imaging can be found in, e.g., [123]. The *univariate* discrete Fourier basis of $\mathbb{C}^N$ consists of vectors

$$\varphi_k(t) = \frac{1}{\sqrt{N}} e^{i2\pi tk/N}, \quad -N/2+1 \leq t \leq N/2,$$

indexed by the discrete frequencies in the range of $-N/2+1 \leq k \leq N/2$. The *bivariate* discrete Fourier basis of $\mathbb{C}^{N\times N}$ is a tensor product of univariate bases, i.e.,

$$\varphi_{j,k}(u,v) = \frac{1}{N} e^{i2\pi(ju+kv)/N}, \quad -N/2+1 \leq u, v \leq N/2,$$

indexed by the discrete frequencies in the range of $-N/2+1 \leq j, k \leq N/2$.

We denote by $\mathcal{F}$ the bivariate discrete Fourier transform

$$X \mapsto (\langle X, \varphi_{k_1,k_2}\rangle)_{k_1,k_2}.$$

Again, by a slight abuse of notation, we denote by $\mathcal{F}$ the unitary matrix representing this linear map. That is, we denote by $\mathcal{F}X$ the matrix product that generates $(\langle X, \varphi_{k_1,k_2}\rangle)_{k_1,k_2}$. Moreover, since limited measurements are considered, we denote by $\mathcal{F}_\Omega$ the restriction of $\mathcal{F}$ to a subset of frequencies $\Omega \subset \{-N/2+1, \ldots, N/2\}^2$.

7.3 Main results

We now establish reconstruction guarantees for the enhanced TV model (7.1.7) from non-adaptive linear RIP measurements and variable-density Fourier measurements, respectively. The following proposition generalizes Theorem 6.4.1 in Chapter 6 for signal recovery, and it allows us to bound the norm of an image D when it is close to the null space of an RIP operator.

Proposition 7.3.1 *Let* $\gamma \geq 1$, $k > 0$, $\delta < 0.6$, $\beta_1 > 0$, $\beta_2 > 0$, *and* $\varepsilon \geq 0$, *and let* $\mathcal{A}$ *be some linear operator* $\mathcal{A} : \mathbb{C}^{n_1\times n_2} \to \mathbb{C}^{\tilde{m}}$, *where* $n_1, n_2, \tilde{m} \in \mathbb{N}$. *Suppose that* $\mathcal{A}$ *has the RIP of order* $k + 4k\gamma^2$ *and level* δ, *and that the image* $D \in \mathbb{C}^{N\times N}$ *satisfies the tube constraint*

$$\|\mathcal{A}D\|_2 \leq \varepsilon. \tag{7.3.1}$$

Suppose further that for a subset S of cardinality $|S| \leq k$, D satisfies the cone constraint

$$\|D_{S^c}\|_1 \leq \gamma \|D_S\|_1 - \frac{\beta_1}{2}\|D\|_2^2 + \sigma + \beta_2 \langle E_1, E_2 \rangle, \tag{7.3.2}$$

where E_1, E_2 could be scalars, vectors, or matrices, and E_2 is assumed to satisfy $\|E_2\|_2 = \|D\|_2$. Here $\|\cdot\|_2$ denotes the absolute value for scalars, the usual ℓ_2 vector norm for vectors, and the $\ell_{2,2}$ norm (Frobenius norm) for matrices. If β_2 satisfies the posterior verification

$$\beta_2 \leq \frac{\gamma\sqrt{k}}{2K_2\|E_1\|_2}, \tag{7.3.3}$$

then it holds that

$$\|D\|_2 \leq \sqrt{\frac{\gamma\sqrt{k}K_1}{\beta_1 K_2}\varepsilon + \frac{2}{\beta_1}\sigma} \lesssim \sqrt{\frac{\gamma\sqrt{k}}{\beta_1}\varepsilon + \frac{1}{\beta_1}\sigma}, \tag{7.3.4}$$

where

$$K_1 := \frac{3}{2\sqrt{1-\delta} - \sqrt{1+\delta}} \quad \text{and} \quad K_2 := \frac{\sqrt{1+\delta}}{4}\left(K_1 + \frac{1}{\sqrt{1+\delta}}\right).$$

Furthermore, we have

$$\begin{aligned} \|D\|_1 &\leq \frac{(2K_2+1)\gamma\sqrt{k} + 2K_2\sqrt{k}}{2K_2}\sqrt{\frac{\gamma\sqrt{k}K_1}{\beta_1 K_2}\varepsilon + \frac{2}{\beta_1}\sigma} + \sigma \\ &\lesssim \gamma\sqrt{k}\sqrt{\frac{\gamma\sqrt{k}}{\beta_1}\varepsilon + \frac{1}{\beta_1}\sigma} + \sigma. \end{aligned} \tag{7.3.5}$$

Corollary 7.3.2 *There is a linear term of σ in (7.3.5). If*

$$\|D\|_2 \geq \sqrt{\frac{2\sigma}{\beta_1}},$$

which is compatible with (7.3.4), then this linear term can be removed. This corollary will be proved after Proposition 7.3.1.

Remark 7.3.3 *In the proof of Proposition 7.3.1, we need to ensure*

$$\sqrt{1-\delta} - \frac{\sqrt{1+\delta}}{2} > 0,$$

and this is where the requirement $\delta < 0.6$ for the RIP level stems from. Since

$$\lim_{\delta \to 0.6} \frac{K_1}{K_2} = \lim_{\delta \to 0.6} \frac{4}{\sqrt{1+\delta} + 1/K_1} = 10, \tag{7.3.6}$$

the bounds on $\|D\|_2$ and $\|D\|_1$ are still reasonable as $\delta \to 0.6$. As the whole analysis below rests upon Proposition 7.3.1, this fact (7.3.6) suggests that the following reconstruction error bounds (7.3.11), (7.3.13), and (7.3.18) are all reasonable as $\delta \to 0.6$.

Remark 7.3.4 *If $\mathcal{A}$ is assumed to have the RIP of order $5k\gamma^2 \geq k + 4k\gamma^2$, then Proposition 7.3.1 still holds. Thus, we assume the order $5k\gamma^2$ for simplicity in the following theorems.*

For any image $X \in \mathbb{C}^{N\times N}$, its derivatives X_x and X_y belong to $\mathbb{C}^{(N-1)\times N}$ and $\mathbb{C}^{N\times(N-1)}$, respectively. Thus, it is convenient to consider the matrices Π_0 and Π^0 obtained from a matrix Π by concatenating a row of zeros to the bottom and top of Π, respectively. More concretely, for a matrix $\Pi \in \mathbb{C}^{(N-1)\times N}$, we denote by $\Pi^0 \in \mathbb{C}^{N\times N}$ the augmented matrix with entries

$$(\Pi^0)_{j,k} = \begin{cases} 0, & j = 1, \\ \Pi_{j-1,k}, & 2 \leq j \leq N. \end{cases}$$

Similarly, we denote by $\Pi_0 \in \mathbb{C}^{N\times N}$ the matrix constructed from adding a row of zeros to the bottom of Π. For a linear operator $\mathcal{A} : \mathbb{C}^{(N-1)\times N} \to \mathbb{C}^m$ with $[\mathcal{A}(X)]_j = \langle A_j, X\rangle$, we denote by $\mathcal{A}^0 : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ the linear operator with $[\mathcal{A}^0(X)]_j = \langle A_j^0, X\rangle$. We denote by $\mathcal{A}_0 : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ similarly. It was shown in [157] that the entire image and its gradients could be related as follows.

Lemma 7.3.5 ([157]) *Given $X \in \mathbb{C}^{N\times N}$ and $\Pi \in \mathbb{C}^{(N-1)\times N}$,*

$$\langle \Pi, X_x\rangle = \langle \Pi^0, X\rangle - \langle \Pi_0, X\rangle \quad \text{and} \quad \langle \Pi, X_y^{\mathrm{T}}\rangle = \langle \Pi^0, X^{\mathrm{T}}\rangle - \langle \Pi_0, X^{\mathrm{T}}\rangle,$$

where X^{T} denotes the (non-conjugate) transpose of the matrix X.

7.3.1 Reconstruction from non-adaptive linear RIP measurements

We are prepared to state our first result on stable image reconstruction from non-adaptive linear RIP measurements.

Theorem 7.3.6 *Let $N = 2^n$ be a power of two, where $n \in \mathbb{N}$. Let $\mathcal{A} : \mathbb{C}^{(N-1)\times N} \to \mathbb{C}^{m_1}$ and $\mathcal{A}' : \mathbb{C}^{(N-1)\times N} \to \mathbb{C}^{m_1}$ be such that the concatenated operator $[\mathcal{A}, \mathcal{A}']$ has the RIP of order $5s$ and level $\delta < 0.6$. Let $\mathcal{H} : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ be the orthonormal bivariate Haar wavelet transform, and $\mathcal{B} : \mathbb{C}^{N\times N} \to \mathbb{C}^{m_2}$ be such that the composite operator $\mathcal{B}\mathcal{H}^* : \mathbb{C}^{N\times N} \to \mathbb{C}^{m_2}$ has the RIP of order $2s+1$ and level $\delta < 1$. Let $m = 4m_1 + m_2$, and consider the linear operator $\mathcal{M} : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ with components*

$$\mathcal{M}(X) = \left(\mathcal{A}^0(X), \mathcal{A}_0(X), \mathcal{A}'^0(X^{\mathrm{T}}), \mathcal{A}_0'(X^{\mathrm{T}}), \mathcal{B}(X)\right). \tag{7.3.7}$$

Let $\bar{X} \in \mathbb{C}^{N\times N}$ be an image and X^{opt} the solution to the enhanced TV model (7.1.7) with $\mathcal{M}$ defined as (7.3.7). If α satisfies

$$\alpha \leq \frac{\sqrt{s}}{2K_2\|\nabla X^{\text{opt}}\|_2}, \tag{7.3.8}$$

then we have the stable gradient reconstruction results

$$\|\nabla\bar{X} - \nabla X^{\text{opt}}\|_2 \lesssim \sqrt{\frac{\sqrt{s}}{\alpha}\tau + \frac{1}{\alpha}\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1} \tag{7.3.9}$$

and

$$\|\nabla\bar{X} - \nabla X^{\text{opt}}\|_1 \lesssim \sqrt{s}\sqrt{\frac{\sqrt{s}}{\alpha}\tau + \frac{1}{\alpha}\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1} + \|\nabla\bar{X} - (\nabla\bar{X})_s\|_1, \tag{7.3.10}$$

and the stable image reconstruction result

$$\begin{aligned}\|\bar{X} - X^{\text{opt}}\|_2 \lesssim{}& \log\left(\frac{N^2}{s}\right)\sqrt{\frac{\sqrt{s}}{\alpha}\tau + \frac{1}{\alpha}\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1} \\ &+ \log\left(\frac{N^2}{s}\right)\frac{\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1}{\sqrt{s}} + \tau.\end{aligned} \tag{7.3.11}$$

Corollary 7.3.7 *Enlightened by Corollary 7.3.2, if*

$$\|\nabla\bar{X} - \nabla X^{\text{opt}}\|_2 \geq \sqrt{\frac{2}{\alpha}\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1},$$

which is compatible with (7.3.9), then the linear term $\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1$ in (7.3.10) and hence the term $\log(\frac{N^2}{s})\frac{\|\nabla\bar{X}-(\nabla\bar{X})_s\|_1}{\sqrt{s}}$ in (7.3.11) can be removed. This corollary will be proved after Theorem 7.3.6.

Remark 7.3.8 *The proof of Theorem 7.3.6 is inspired by the proof in [157] for the TV model (7.1.2), in which it was conjectured that the $4m_1$ measurements derived from $\mathcal{A}$ in the construction (7.3.7) of $\mathcal{M}$ are artifacts of the proof. The components $\mathcal{A}^0(X)$, $\mathcal{A}_0(X)$, $\mathcal{A}'^0(X^{\text{T}})$, and $\mathcal{A}'_0(X^{\text{T}})$ are only used for deriving the stable gradient reconstruction bounds (7.3.9) and (7.3.10). On the other hand, component $\mathcal{B}(X)$ only helps us derive the bound (7.3.11) from (7.3.9) and (7.3.10).*

If more measurements are allowed, then the bound (7.3.11) can be further improved, the requirement (7.3.8) on α can be relaxed, and the artificial components in $\mathcal{M}$ can be removed.

Theorem 7.3.9 *Let $N = 2^n$ be a power of two, where $n \in \mathbb{N}$. Let $\mathcal{H} : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ be the orthonormal bivariate Haar wavelet transform, and $\mathcal{M} : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ be such that the composite operator $\mathcal{M}\mathcal{H}^* : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ has the RIP of order*

$Cs\log^3(N)$ and level $\delta < 0.6$. Let $\bar{X} \in \mathbb{C}^{N\times N}$ be a mean-zero image or an image containing some zero-valued pixels, and X^{opt} be the solution to the enhanced TV model (7.1.7). If α satisfies

$$\alpha \leq \frac{\sqrt{48s\log(N)}}{K_2\|\nabla X^{\text{opt}}\|_2}, \tag{7.3.12}$$

then we have

$$\|\bar{X} - X^{\text{opt}}\|_2 \lesssim \sqrt{\frac{\sqrt{s}}{\alpha}\tau + \frac{1}{\alpha}\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1}. \tag{7.3.13}$$

Remark 7.3.10 *The RIP requirements in both theorems above indicate that the linear measurements should be generated from standard RIP matrix ensembles, which are incoherent with the Haar wavelet system. Many classes of random matrices can be used to generate RIP matrix ensembles. For example, a matrix in $\mathbb{R}^{m\times N^2}$ with i.i.d. normalized Gaussian random entries has a small RIP constant $\delta_s < c$ with high probabilities if*

$$m \gtrsim c^{-2}s\log(N^2/s),$$

as shown in [44]. Similar results were extended to sub-Gaussian matrices in [145]. If

$$m \gtrsim s\log^4(N),$$

then it was proved in [45, 181] that the RIP holds with overwhelming probabilities for a partial Fourier matrix $\mathcal{F}_\Omega \in \mathbb{R}^{m\times N^2}$. The RIP also holds for randomly generated circulant matrices (see [174]) and randomly subsampled bounded orthonormal systems (see [175]). Most of these mentioned measurements are incoherent with the Haar wavelet system, but the partial Fourier matrix with uniformly subsampled rows is an exception. Thus, some specific sampling strategies for Fourier measurements should be considered. For example, it was asserted in [122] that $\mathcal{F}_\Omega \in \mathbb{R}^{m\times N^2}$ with

$$m \gtrsim s\log^4(N)$$

and randomized column signs *has the RIP; it was also shown in [123] that $\mathcal{F}_\Omega$ with rows subsampled according to some power-law densities is incoherent with the Haar wavelet system after preconditioning.*

7.3.2 Reconstruction from variable-density Fourier measurements

As shown in [123], if the measurements are sampled according to appropriate power-law densities, then they are incoherent with the Haar wavelet system. We consider a particular variable-density sampling strategy proposed in [123] and derive a partial stable image reconstruction theorem tailored for Fourier measurements. Following the idea of [123], our guarantees are based on a *weighted ℓ_2-norm* in measuring noise

such that high-frequency measurements have a higher sensitivity to noise; that is, the ℓ_2-norm in the constraint $\|\mathcal{M}X - y\|_2 \leq \tau$ of the enhanced TV model (7.1.7) is replaced by a weighted ℓ_2-norm model. For the particular scenario with Fourier measurements, the general linear operator $\mathcal{M}$ is specified as $\mathcal{F}_\Omega$, which is the restriction of the Fourier transform matrix to a set Ω of frequencies as defined in Section 7.2.3.

Theorem 7.3.11 *Let $N = 2^n$ be a power of 2, where $n \in \mathbb{N}$. Let m and s satisfy $s \gtrsim \log(N)$ and*

$$m \gtrsim s \log^3(s) \log^5(N). \tag{7.3.14}$$

Select m frequencies $\{(\omega_1^j, \omega_2^j)\}_{j=1}^m \subset \{-N/1+2, \ldots, N/2\}^2$ i.i.d. according to

$$\mathbb{P}[(\omega_1^j, \omega_2^j) = (k_1, k_2)] = C_N \min\left(C, \frac{1}{k_1^2 + k_2^2}\right) =: \eta(k_1, k_2) \tag{7.3.15}$$

for $-N/2 + 1 \leq k_1, k_2 \leq N/2$, where C is an absolute constant and C_N is chosen such that η is a probability distribution. Consider the weight vector $\rho = (\rho_j)_{j=1}^m$ with

$$\rho_j = \left[\frac{1}{\eta(\omega_1^j, \omega_2^j)}\right]^{1/2}.$$

Then we have the following assertion for all mean-zero or zero-valued pixel-containing images $\bar{X} \in \mathbb{C}^{N \times N}$ with probability exceeding $1 - N^{-C \log^3(s)}$: Given noisy partial Fourier measurements $b = \mathcal{F}_\Omega \bar{X} + e$, if

$$\alpha \leq \frac{\sqrt{48 s \log(N)}}{K_2 \|\nabla X^{\text{opt}}\|_2}, \tag{7.3.16}$$

then the solution X^{opt} to the model

$$\min_{X \in \mathbb{C}^{N \times N}} \|\nabla X\|_1 - \frac{\alpha}{2}\|\nabla X\|_2^2 \quad \text{s.t.} \quad \|\rho \circ (\mathcal{F}_\Omega X - b)\|_2 \leq \tau\sqrt{m} \tag{7.3.17}$$

satisfies

$$\|\bar{X} - X^{\text{opt}}\|_2 \lesssim \sqrt{\frac{\sqrt{s}}{\alpha}\tau + \frac{1}{\alpha}\|\nabla \bar{X} - (\nabla \bar{X})_s\|_1}. \tag{7.3.18}$$

7.3.3 Further discussion

We supplement more details about the theoretical results presented in Sections 7.3.1 and 7.3.2.

The *a posterior* verification on α. Three conditions (7.3.8), (7.3.12), and (7.3.16) on α are required in Theorems 7.3.6, 7.3.9, and 7.3.11, respectively. Determining the

value of α is possible only if we have *a priori* estimation on $\|X^{\text{opt}}\|_2$. Thus, these conditions can be interpreted as *a posterior* verification because they can be verified once X^{opt} is obtained by solving the model (7.1.7). In practice, we solve the model (7.1.7) numerically and thus obtain an approximate solution, denoted by X^*, subject to a preset accuracy $\epsilon > 0$. That is, $\|X^{\text{opt}} - X^*\|_2 \leq \epsilon$. Then, if

$$\alpha \leq \frac{\sqrt{s}}{2K_2(\|\nabla X^*\|_2 + \epsilon)},$$

then (7.3.8) is guaranteed; if

$$\alpha \leq \frac{\sqrt{48s\log(N)}}{K_2(\|\nabla X^*\|_2 + \epsilon)},$$

then (7.3.12) and (7.3.16) are satisfied.

The RIP level $\delta < 0.6$ in Theorems 7.3.6 and 7.3.9. The bound 0.6 is sharp, as we need to ensure

$$\sqrt{1-\delta} - \frac{\sqrt{1+\delta}}{2} > 0$$

(cf. proof in Section 7.4.1). For the reconstruction guarantees derived in [157] for the TV model (7.1.2), the level is assumed to satisfy $\delta < 1/3$, and it is not sharp as remarked in [157]. Though $\delta < 1/3$ can be improved, the reconstruction error bound in [157] for the TV model (7.1.2) tends to be infinity if $\delta \to 0.6$. In light of Remark 7.3.3, the bounds (7.3.11) and (7.3.13) are still valid in this case, and the upper bound required for α tends to 0 correspondingly with consideration of the behavior of K_2. That is, Theorems 7.3.6 and 7.3.9 can guarantee the stability of the TV model (7.1.2) when $\delta \to 0.6$, resulting in reconstruction error bounds in forms of (7.3.11) and (7.3.13).

The required amount m of Fourier measurements in Theorem 7.3.11. The RIP level δ does not appear explicitly in Theorem 7.3.11, while we shall assume

$$m \gtrsim s\delta^{-2}\log^3(s)\log^5(N)$$

and the constant δ is eliminated in such an inequality with $\gtrsim$; see our proof in Section 7.4.4. The least required amount m for the TV model (7.1.2) shall also satisfy this relation with s, N, and δ, as proved in [123]. Since the upper bound on the RIP level δ is enlarged from 1/3 for the TV model (7.1.2) (see [123]) to 0.6 for the enhanced TV model (7.1.7), the least amount of Fourier measurements required for the enhanced TV model (7.1.7) should be

$$(0.6/(1/3))^{-2} \approx 30.86\%$$

of the least amount of Fourier measurements required in [123] for the TV model (7.1.2).

Inconsistency when $\alpha \to 0$. The enhanced TV regularization (7.1.6) tends to be the anisotropic TV term as $\alpha \to 0$. At the same time, the reconstruction error bounds (7.3.11), (7.3.13), and (7.3.18) do not reduce to the corresponding bounds (7.1.9) and (7.1.10) for the TV model (7.1.2). Note that the bounds (7.3.13) and (7.3.18) are of the same form. To explain this inconsistency, note that Proposition 7.3.1 is a pillar of the proofs of Theorems 7.3.6, 7.3.9, and 7.3.11. In contrast, the proof for the TV model (7.1.2) in [157] relies on the following fact: If D satisfies the tube constraint (7.3.1) and the cone constraint $\|D_{S^c}\|_1 \leq \gamma\|D_S\|_1 + \sigma$, then it was shown in [157] that

$$\|D\|_2 \lesssim \frac{\sigma}{\gamma\sqrt{k}} + \varepsilon \quad \text{and} \quad \|D\|_1 \lesssim \sigma + \gamma\sqrt{k}\varepsilon. \tag{7.3.19}$$

Indeed, the left-hand side of the estimation (7.4.2) in the proof of Proposition 7.3.1 contains a quadratic term $\|D\|_2^2$ and a linear term $\|D\|_2$, and only the linear term remains if $\beta_1, \beta_2 \to 0$, which then leads to the same result as (7.3.19). However, in the proof of Proposition 7.3.1, we remove this linear term and keep the quadratic term, and hence the obtained result cannot be reduced to the result (7.3.19) as $\beta_1, \beta_2 \to 0$. Such an inconsistent situation is also encountered by the springback model in Chapter 6.

Comparison between (7.1.10) and (7.3.13). We are interested in whether or not the bound (7.3.13) (as well as the bound (7.3.18), which shares the same form as (7.3.13)) can be tighter than (7.1.10) in the sense of

$$\sqrt{\frac{\sqrt{s}}{\alpha}\tau + \frac{1}{\alpha}\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1} \lesssim \frac{\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1}{\sqrt{s}} + \tau, \tag{7.3.20}$$

with a given $\alpha > 0$. If the image $\bar{X}$ is known to have an s-sparse gradient, then the comparison (7.3.20) is reduced to

$$\sqrt{s} \lesssim \alpha\tau.$$

As s is fixed in this scenario, we can claim that the estimation (7.3.13) is tighter than the estimation (7.1.10) in the sense of (7.3.20) if

$$\tau \gtrsim \sqrt{s}/\alpha,$$

i.e., the level of noise τ is *relatively large*. If the sparsity of $\nabla\bar{X}$ is not assumed, but

the linear measurements are noise-free, i.e., $\tau = 0$, then the comparison (7.3.20) is reduced to

$$\frac{s}{\|\nabla \bar{X} - (\nabla \bar{X})_s\|_1} \lesssim \alpha, \tag{7.3.21}$$

in which the left-hand side of (7.3.21) is an increasing function of s. In order to discern the scenario where (7.3.21) holds, a key fact from Remark 7.3.10 should be noticed: for RIP measurements mentioned there, a small number m of measurements admits an RIP with a small s. The bound $\mathcal{O}(s \log(N^2/s))$ for Gaussian measurements appears not to be monotonic with respect to s. On the other hand, with the implicit constant factors derived in [181], this bound is indeed monotonically increasing with respect to s. Thus, if *the number of measurements m is limited*, which only renders an RIP with a small s, then (7.3.21) holds. This situation coincides with the intuition that, as the term $\|\nabla \bar{X} - (\nabla \bar{X})_s\|_1 \gg 1$ for many digital images, especially when the number of measurements is limited (so that s is small), taking a square root shall lead to a smaller bound than that without doing so.

Together with both scenarios, we can claim that if the level of noise τ is *relatively large* and *the number of measurements m is limited*, then the enhanced TV model (7.1.7) performs better than the TV model (7.1.2) in the sense of (7.3.20), because (7.3.20) is guaranteed to hold when

$$\sqrt{\frac{\sqrt{s}}{\alpha}\tau} + \sqrt{\frac{1}{\alpha}\|\nabla \bar{X} - (\nabla \bar{X})_s\|_1} \lesssim \frac{\|\nabla \bar{X} - (\nabla \bar{X})_s\|_1}{\sqrt{s}} + \tau,$$

and we can study

$$\sqrt{\frac{\sqrt{s}}{\alpha}\tau} \lesssim \tau$$

and

$$\sqrt{\frac{1}{\alpha}\|\nabla \bar{X} - (\nabla \bar{X})_s\|_1} \lesssim \frac{\|\nabla \bar{X} - (\nabla \bar{X})_s\|_1}{\sqrt{s}}$$

separately.

This comparison can be analogously extended to other cases for which the corresponding reconstruction error bounds are also linear with respect to terms $\|\nabla \bar{X} - (\nabla \bar{X})_s\|_1/\sqrt{s}$ and τ. Such examples include the model in [139], which has the regularization term $\|X\|_{\mathrm{TV}_a} - \|X\|_{\mathrm{TV}_i}$. For the model in [139], it seems that reconstruction guarantees leading to an error bound without the log factor $\log(N^2/s)$ are still missing. Note that this log factor also occurs in the bound (7.1.9) for the TV model (7.1.2) and the bound (7.3.11) for the enhanced TV model (7.1.7), but it is removed if the required RIP order increases from $\mathcal{O}(s)$ to $\mathcal{O}(s \log^3(N))$, and

then both bounds can be improved to (7.1.10) and (7.3.13), respectively. Reconstruction guarantees for the model in [139] have been investigated in [135]. However, the derived error bound (see Theorem 3.8 in [135]) still fails to remove the log factor $\log(N^2/s)$, despite that the subsampled measurements are required to have the RIP of order $\mathcal{O}(s^2\log(N))$ with a more complicated level δ which depends on N, s, and the constant $\tilde{C}$ in Lemma 7.2.1.

7.4 Proofs of the main results

In this section, we present the proofs for the theoretical results in Section 7.3.

7.4.1 Proofs of Proposition 7.3.1 and Corollary 7.3.2

Proof of Proposition 7.3.1. We arrange the indices in S^c in order of decreasing magnitudes (in absolute value) of D_{S^c} and divide S^c into subsets of size $4k\gamma^2$, i.e., $S^c = S_1 \bigcup S_2 \bigcup \cdots \bigcup S_r$, where

$$r = \left\lfloor \frac{N^2 - |S|}{4k\gamma^2} \right\rfloor.$$

In other words, $D_{S^c} = D_{S_1} + D_{S_2} + \cdots + D_{S_r}$, where D_{S_1} consists of the $4k\gamma^2$ largest-magnitude components of D over S^c, D_{S_2} consists of the next $4k\gamma^2$ largest-magnitude components of D over $S^c \backslash S_1$, and so forth. As the magnitude of each component of D_{S_j} is less than the average magnitude $\|D_{S_{j-1}}\|_1/(4k\gamma^2)$ of components of $D_{S_{j-1}}$,

$$\|D_{S_j}\|_2^2 \leq 4k\gamma^2 \left(\frac{\|D_{S_{j-1}}\|_1}{4k\gamma^2} \right)^2 = \frac{\|D_{S_{j-1}}\|_1^2}{4k\gamma^2}, \quad j = 2, 3, \ldots, r.$$

Thus, combining $\|D_{S_j}\|_2 \leq \frac{\|D_{S_{j-1}}\|_1}{2\gamma\sqrt{k}}$ with the cone constraint (7.3.2), we have

$$\sum_{j=2}^{r} \|D_{S_j}\|_2 \leq \frac{1}{2\gamma\sqrt{k}} \|D_{S^c}\|_1 \leq \frac{\|D_S\|_1}{2\sqrt{k}} - \frac{\beta_1}{4\gamma\sqrt{k}} \|D\|_2^2 + \frac{\sigma}{2\gamma\sqrt{k}} + \frac{\beta_2}{2\gamma\sqrt{k}} \langle E_1, E_2 \rangle .$$

The assumption $|S| \leq k$ leads to $\|D_S\|_1 \leq \sqrt{|S|}\|D_S\|_2 \leq \sqrt{k}\|D_S\|_2 \leq \sqrt{k}\|D_S + D_{S_1}\|_2$, hence we have

$$\sum_{j=2}^{r} \|D_{S_j}\|_2 \leq \frac{\|D_S + D_{S_1}\|_2}{2} - \frac{\beta_1}{4\gamma\sqrt{k}} \|D\|_2^2 + \frac{\sigma}{2\gamma\sqrt{k}} + \frac{\beta_2}{2\gamma\sqrt{k}} \langle E_1, E_2 \rangle . \tag{7.4.1}$$

Together with the bound (7.4.1), the constraint (7.3.1), and the RIP of $\mathcal{A}$, we have

$$
\begin{aligned}
\varepsilon \geq & \|\mathcal{A}D\|_2 \geq \|\mathcal{A}(D_S + D_{S_1})\|_2 - \sum_{j=2}^{r} \|\mathcal{A}D_{S_j}\|_2 \\
\geq & \sqrt{1-\delta}\|D_S + D_{S_1}\|_2 - \sqrt{1+\delta}\sum_{j=2}^{r} \|D_{S_j}\|_2 \\
\geq & \sqrt{1-\delta}\|D_S + D_{S_1}\|_2 - \sqrt{1+\delta}\left(\frac{\|D_S + D_{S_1}\|_2}{2} - \frac{\beta_1\|D\|_2^2}{4\gamma\sqrt{k}} + \frac{\sigma}{2\gamma\sqrt{k}} + \frac{\beta_2\langle E_1, E_2\rangle}{2\gamma\sqrt{k}}\right) \\
= & \left(\sqrt{1-\delta} - \frac{\sqrt{1+\delta}}{2}\right)\|D_S + D_{S_1}\|_2 + \frac{\beta_1\sqrt{1+\delta}\|D\|_2^2}{4\gamma\sqrt{k}} - \frac{\sigma\sqrt{1+\delta}}{2\gamma\sqrt{k}} - \frac{\beta_2\sqrt{1+\delta}\langle E_1, E_2\rangle}{2\gamma\sqrt{k}}.
\end{aligned}
$$

The assumption $\delta < 0.6$ ensures $\sqrt{1-\delta} - \sqrt{1+\delta}/2 > 0$. Hence, we have

$$
\begin{aligned}
&\|D_S + D_{S_1}\|_2 \leq \\
&\qquad \frac{2}{2\sqrt{1-\delta} - \sqrt{1+\delta}}\left(\varepsilon - \frac{\beta_1\sqrt{1+\delta}\|D\|_2^2}{4\gamma\sqrt{k}} + \frac{\sigma\sqrt{1+\delta}}{2\gamma\sqrt{k}} + \frac{\beta_2\sqrt{1+\delta}}{2\gamma\sqrt{k}}\langle E_1, E_2\rangle\right).
\end{aligned}
$$

As $\|D\|_2$ is bounded by the sum of $\|D_S + D_{S_1}\|_2$ and $\sum_{j=2}^{r}\|D_{S_j}\|_2$, it satisfies

$$
\begin{aligned}
\|D\|_2 \leq & \frac{3}{2}\|D_S + D_{S_1}\|_2 - \frac{\beta_1}{4\gamma\sqrt{k}}\|D\|_2^2 + \frac{\sigma}{2\gamma\sqrt{k}} + \frac{\beta_2}{2\gamma\sqrt{k}}\langle E_1, E_2\rangle \\
\leq & \frac{3\varepsilon}{2\sqrt{1-\delta}-\sqrt{1+\delta}} + \left(\frac{3}{2\sqrt{1-\delta}-\sqrt{1+\delta}} + \frac{1}{\sqrt{1+\delta}}\right)\left(-\frac{\beta_1\sqrt{1+\delta}}{4\gamma\sqrt{k}}\|D\|_2^2 + \frac{\sqrt{1+\delta}}{2\gamma\sqrt{k}}\sigma + \frac{\beta_2\sqrt{1+\delta}}{2\gamma\sqrt{k}}\langle E_1, E_2\rangle\right) \\
:= & K_1\varepsilon - \frac{\beta_1 K_2}{\gamma\sqrt{k}}\|D\|_2^2 + \frac{2K_2}{\gamma\sqrt{k}}\sigma + \frac{2\beta_2 K_2}{\gamma\sqrt{k}}\langle E_1, E_2\rangle .
\end{aligned}
$$

Thus, we have the quadratic inequality

$$
\frac{\beta_1 K_2}{\gamma\sqrt{k}}\|D\|_2^2 + \|D\|_2 - \frac{2\beta_2 K_2}{\gamma\sqrt{k}}\langle E_1, E_2\rangle - K_1\varepsilon - \frac{2K_2}{\gamma\sqrt{k}}\sigma \leq 0. \tag{7.4.2}
$$

The requirement (7.3.3) on β_2 ensures that

$$\|D\|_2 - \frac{2\beta_2 K_2}{\gamma\sqrt{k}} \langle E_1, E_2 \rangle \geq \|D\|_2 - \left\langle \frac{E_1}{\|E_1\|_2}, E_2 \right\rangle \geq 0,$$

where the last inequality is due to Cauchy–Schwarz inequality and $\|E_2\|_2 = \|D\|_2$. Then, we have

$$\frac{\beta_1 K_2}{\gamma\sqrt{k}} \|D\|_2^2 - K_1\varepsilon - \frac{2K_2}{\gamma\sqrt{k}}\sigma \leq 0,$$

which yields the estimation (7.3.4). Finally, we derive (7.3.5). As $|S| \leq k$, we have $\|D_S\|_1 \leq \sqrt{k}\|D_S\|_2$. Then, together with the requirement (7.3.3) on β_2 and the cone constraint (7.3.2), we have

$$\begin{aligned}
\|D\|_1 &\leq (\gamma+1)\|D_S\|_1 - \frac{\beta_1}{2}\|D\|_2^2 + \sigma + \beta_2 \langle E_1, E_2 \rangle \\
&\leq (\gamma+1)\|D_S\|_1 + \sigma + \frac{\gamma\sqrt{k}}{2K_2}\|D\|_2 \\
&\leq (\gamma+1)\sqrt{k}\|D_S\|_2 + \sigma + \frac{\gamma\sqrt{k}}{2K_2}\|D\|_2 \\
&\leq (\gamma+1)\sqrt{k}\|D\|_2 + \sigma + \frac{\gamma\sqrt{k}}{2K_2}\|D\|_2 \\
&= \frac{(2K_2+1)\gamma\sqrt{k} + 2K_2\sqrt{k}}{2K_2}\|D\|_2 + \sigma,
\end{aligned} \tag{7.4.3}$$

which completes the proof of Proposition 7.3.1. □

Proof of Corollary 7.3.2. In the second inequality of (7.4.3), we use the fact $-\frac{\beta_1}{2}\|D\|_2^2 \leq 0$. If $\|D\|_2$ satisfies $\|D\|_2 \geq \sqrt{2\sigma/\beta_1}$, then $-\frac{\beta_1}{2}\|D\|_2^2 + \sigma \leq 0$ and it follows from (7.4.3) that

$$\begin{aligned}
\|D\|_1 &\leq (\gamma+1)\|D_S\|_1 - \frac{\beta_1}{2}\|D\|_2^2 + \sigma + \beta_2 \langle E_1, E_2 \rangle \\
&\leq (\gamma+1)\|D_S\|_1 + \frac{\gamma\sqrt{k}}{2K_2}\|D\|_2 \\
&\leq \frac{(2K_2+1)\gamma\sqrt{k} + 2K_2\sqrt{k}}{2K_2}\|D\|_2,
\end{aligned}$$

which completes the proof of Corollary 7.3.2. □

7.4.2 Proof of Theorem 7.3.6 and Corollary 7.3.7

We first prove the stable gradient reconstruction results (7.3.9) and (7.3.10), and then obtain the stable image reconstruction result (7.3.11) with the aid of a strong

Sobolev inequality. The following Sobolev inequality was derived in [157] for images with multivariate generalization given in [156].

Lemma 7.4.1 (Strong Sobolev inequality) *Let $\mathcal{B} : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ be a linear map such that $\mathcal{B}\mathcal{H}^* : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ has the RIP of order $2s+1$ and level $\delta < 1$, where $\mathcal{H} : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ is the bivariate Haar transform. Suppose that $D \in \mathbb{C}^{N\times N}$ satisfies the tube constraint $\|\mathcal{B}D\|_2 \le \varepsilon$. Then*

$$\|D\|_2 \le C_2 \left[\left(\frac{\|\nabla D\|_1}{\sqrt{s}} \right) \log\left(\frac{N^2}{s} \right) + \varepsilon \right].$$

Proof of Theorem 7.3.6. The proof is divided into the stable gradient and image reconstructions, respectively.

Stable gradient reconstruction. We plan to apply Proposition 7.3.1 to the term $\nabla(X^{\text{opt}} - \bar{X})$. Let $V = X^{\text{opt}} - \bar{X}$ and $L = (V_x, V_y^{\mathrm{T}})$. For convenience, let P denote the mapping of indices which maps the index of a nonzero entry in ∇V to its corresponding index in L. By the definition of ∇, L has the same norm as ∇V, i.e., $\|L\|_2 = \|\nabla V\|_2$ and $\|L\|_1 = \|\nabla V\|_1$. Thus, it suffices to apply Proposition 7.3.1 to L. Let $A_1, A_2, \ldots, A_{m_1}, A_1', A_2', \ldots, A_{m_1}'$ be such that $[\mathcal{A}(Z)]_j = \langle A_j, Z\rangle$ and $[\mathcal{A}'(Z)]_j = \left\langle A_j', Z \right\rangle$.

Cone constraint. Let S denote the support of the largest s entries of $\nabla\bar{X}$. On one hand, it holds that

$$\begin{aligned}\|\nabla X^{\text{opt}}\|_1 - \frac{\alpha}{2}\|\nabla X^{\text{opt}}\|_2^2 &\le \|\nabla\bar{X}\|_1 - \frac{\alpha}{2}\|\nabla\bar{X}\|_2^2 \\ &= \|(\nabla\bar{X})_S\|_1 + \|(\nabla\bar{X})_{S^c}\|_1 - \frac{\alpha}{2}\|\nabla\bar{X}\|_2^2.\end{aligned}$$

On the other hand, we have

$$\begin{aligned}&\|\nabla X^{\text{opt}}\|_1 - \frac{\alpha}{2}\|\nabla X^{\text{opt}}\|_2^2 \\ &\quad= \|(\nabla\bar{X})_S + (\nabla V)_S\|_1 + \|(\nabla\bar{X})_{S^c} + (\nabla V)_{S^c}\|_1 - \frac{\alpha}{2}\|\nabla\bar{X} + \nabla V\|_2^2 \\ &\quad\ge \|(\nabla\bar{X})_S\|_1 - \|(\nabla V)_S\|_1 + \|(\nabla V)_{S^c}\|_1 - \|(\nabla\bar{X})_{S^c}\|_1 \\ &\qquad - \frac{\alpha}{2}\left(\|\nabla\bar{X}\|_2^2 + 2\left\langle\nabla\bar{X}, \nabla V\right\rangle + \|\nabla V\|_2^2\right).\end{aligned}$$

Thus, we obtain

$$\begin{aligned}\|(\nabla V)_{S^c}\|_1 &\le \|(\nabla V)_S\|_1 + 2\|(\nabla\bar{X})_{S^c}\|_1 + \frac{\alpha}{2}\|\nabla V\|_2^2 + \alpha\left\langle\nabla\bar{X}, \nabla V\right\rangle \\ &= \|(\nabla V)_S\|_1 + 2\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1 - \frac{\alpha}{2}\|\nabla V\|_2^2 + \alpha\left\langle\nabla X^{\text{opt}}, \nabla V\right\rangle.\end{aligned}$$

As L contains all the same nonzero entries as ∇V, it satisfies the following cone constraint:

$$\|L_{P(S)^c}\|_1 \leq \|L_{P(S)}\|_1 + 2\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1 - \frac{\alpha}{2}\|L\|_2^2 + \alpha\left\langle \nabla X^{\mathrm{opt}}, \nabla V\right\rangle.$$

Tube constraint. We note that V satisfies a tube constraint as

$$\begin{aligned}\|\mathcal{M}V\|_2^2 &= \|(\mathcal{M}X^{\mathrm{opt}} - y) - (\mathcal{M}\bar{X} - y)\|_2^2\\ &\leq 2\|\mathcal{M}X^{\mathrm{opt}} - y\|_2^2 + 2\|\mathcal{M}\bar{X} - y\|_2^2\\ &\leq 4\tau^2.\end{aligned}$$

Then, it follows from Lemma 7.3.5 that

$$\begin{aligned}|\left\langle A_j, V_x\right\rangle|^2 &= |\left\langle [A_j]^0, V\right\rangle - \left\langle [A_j]_0, V\right\rangle|^2\\ &\leq 2|\left\langle [A_j]^0, V\right\rangle|^2 + 2|\left\langle [A_j]_0, V\right\rangle|^2\end{aligned}$$

and

$$\begin{aligned}|\left\langle A_j', V_y^{\mathrm{T}}\right\rangle|^2 &= |\left\langle [A_j']^0, V^{\mathrm{T}}\right\rangle - \left\langle [A_j']_0, V^{\mathrm{T}}\right\rangle|^2\\ &\leq 2|\left\langle [A_j']^0, V^{\mathrm{T}}\right\rangle|^2 + 2|\left\langle [A_j']_0, V^{\mathrm{T}}\right\rangle|^2.\end{aligned}$$

Thus, L also satisfies a tube constraint:

$$\|[\mathcal{A}\ \mathcal{A}']L\|_2^2 = \sum_{j=1}^{m} |\left\langle A_j, V_x\right\rangle|^2 + |\left\langle A_j', V_y^{\mathrm{T}}\right\rangle|^2 \leq 2\|\mathcal{M}(V)\|_2^2 \leq 8\tau^2.$$

By virtue of Proposition 7.3.1 with $\gamma = 1$, $k = s$, $\beta_1 = \beta_2 = \alpha$, $\sigma = 2\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1$, $\varepsilon = 2\sqrt{2}\tau$, $E_1 = \nabla X^{\mathrm{opt}}$ and $E_2 = \nabla V$, the requirement (7.3.8) of α ensures that

$$\|\nabla X^{\mathrm{opt}} - \nabla\bar{X}\|_2 = \|L\|_2 \leq \sqrt{\frac{2\sqrt{2}\sqrt{s}K_1\tau}{\alpha K_2} + \frac{4}{\alpha}\|\nabla X - (\nabla X)_s\|_1}.$$

Furthermore, by (7.3.5), we have $\|\nabla X^{\mathrm{opt}} - \nabla\bar{X}\|_1 = \|L\|_1$ and

$$\|L\|_1 \leq \frac{(4K_2+1)\sqrt{s}}{2K_2}\sqrt{\frac{2\sqrt{2}\sqrt{s}K_1\tau}{\alpha K_2} + \frac{4}{\alpha}\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1} + 2\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1, \tag{7.4.4}$$

which completes the proof of the stable gradient reconstruction results (7.3.9) and (7.3.10).

Stable image reconstruction. We now apply the strong Sobolev inequality given in Lemma 7.4.1 to $X^{\text{opt}} - \bar{X}$. As

$$\|\mathcal{B}(X^{\text{opt}} - \bar{X})\|_2 \leq \|\mathcal{M}(X^{\text{opt}} - \bar{X})\|_2 \leq 2\tau,$$

we have

$$\|X^{\text{opt}} - \bar{X}\|_2 \lesssim \log\left(\frac{N^2}{s}\right)\left(\frac{\|\nabla X^{\text{opt}} - \nabla\bar{X}\|_1}{\sqrt{s}}\right) + \tau.$$

Together with the bound (7.3.10), we have the stable image reconstruction result (7.3.11). □

Proof of Corollary 7.3.7. If

$$\|\nabla\bar{X} - \nabla X^{\text{opt}}\|_2 \geq \sqrt{\frac{2}{\alpha}\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1},$$

then it follows from Corollary 7.3.2 that the linear term of $\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1$ in the estimation (7.4.4) can be removed. Thus, from (7.4.4) to (7.3.11), the term $\log(\frac{N^2}{s})\frac{\|\nabla\bar{X}-(\nabla\bar{X})_s\|_1}{\sqrt{s}}$ in (7.3.11) can be also removed. □

7.4.3 Proof of Theorem 7.3.9

We apply Proposition 7.3.1 to $c = \mathcal{H}V$ as opposed to ∇V. Some properties of the bivariate Haar wavelet system, characterized as Lemmas 7.2.1, 7.2.2, and 7.2.2, are needed in the proof. Besides, a classical Sobolev inequality weaker than the strong Sobolev inequality in Lemma 7.4.1 is needed.

Lemma 7.4.2 ([157]) *Let $X \in \mathbb{C}^{N\times N}$ be a mean-zero image or contain some zero-valued pixels. Then*

$$\|X\|_2 \leq \|\nabla X\|_1. \tag{7.4.5}$$

Proof of Theorem 7.3.9. Let $V = X^{\text{opt}} - \bar{X}$, and apply Proposition 7.3.1 to $c = \mathcal{H}V$, where $c_{(1)} := c_{(1)}(V)$ denotes the Haar coefficient corresponding to the constant wavelet, and $c_{(j)} := c_{(j)}(V)$, $j \geq 2$, denotes the $(j-1)$-st largest-magnitude Haar coefficient among the remaining. We use this ordering because Lemma 7.2.1 applies only to mean-zero images. Let $h_{(j)}$ denote the Haar wavelet associated with $c_{(j)}$. We have assumed that the composite operator $\mathcal{M}\mathcal{H}^* : \mathbb{C}^{N\times N} \to \mathbb{C}^m$ has the RIP of order $Cs\log^3(N)$ and level $\delta < 0.6$, and we now derive the constant C.

Cone constraint on $c = \mathcal{H}V$. As shown in Section 7.4.2, we have

$$\|(\nabla V)_{S^c}\|_1 \leq \|(\nabla V)_S\|_1 + 2\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1 - \frac{\alpha}{2}\|\nabla V\|_2^2 + \alpha\left\langle\nabla X^{\text{opt}}, \nabla V\right\rangle. \tag{7.4.6}$$

Recall that S is the index set of s largest-magnitude entries of ∇V. It follows from Lemma 7.2.2 that the set Ω of wavelets which are non-constant over S has the cardinality at most $6s\log(N)$, i.e., $|\Omega| \leq 6s\log(N)$. Decompose V as

$$V = \sum_j c_{(j)} h_{(j)} = \sum_{j\in\Omega} c_{(j)} h_{(j)} + \sum_{j\in\Omega^c} c_{(j)} h_{(j)} =: V_\Omega + V_{\Omega^c}.$$

Because of the linearity of ∇, we have $\nabla V = \nabla V_\Omega + \nabla V_{\Omega^c}$. By the construction of Ω, we have $(\nabla V_{\Omega^c})_S = 0$, which leads to $(\nabla V)_S = (\nabla V_\Omega)_S$. Then, it follows from Lemma 7.2.3 that

$$\|(\nabla V)_S\|_1 = \|(\nabla V_\Omega)_S\|_1 \leq \|\nabla V_\Omega\|_1 \leq \sum_{j\in\Omega} |c_{(j)}| \|\nabla h_{(j)}\|_1 \leq 8 \sum_{j\in\Omega} |c_{(j)}|.$$

Let $k = 6s\log(N)$, $\|c_\Omega\|_1$ and $\|c_{\Omega^c}\|_1$ denote $\sum_{j\in\Omega} |c_{(j)}|$ and $\sum_{j\in\Omega^c} |c_{(j)}|$, respectively. Concerning the decay of the wavelet coefficients in Lemma 7.2.1, we have $|c_{(j+1)}| \leq \tilde{C}\|\nabla V\|_1/j$. Together with the cone constraint (7.4.6) for ∇V, we have

$$\begin{aligned}
\|c_{\Omega^c}\|_1 \leq & \sum_{j=s+1}^{N^2} |c_{(j)}| \leq \tilde{C} \sum_{j=s+1}^{N^2} \frac{\|\nabla V\|_1}{j-1} \overset{(\diamond)}{\leq} C' \log\left(\frac{N^2}{s}\right) \\
\leq & C' \log\left(\frac{N^2}{s}\right) \left(2\|(\nabla V)_S\|_1 + 2\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1 - \frac{\alpha}{2}\|\nabla V\|_2^2 + \alpha\left\langle \nabla X^{\text{opt}}, \nabla V\right\rangle\right) \\
\leq & C' \log\left(\frac{N^2}{s}\right) \left(16\|c_\Omega\|_1 + 2\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1 - \frac{\alpha}{2}\|\nabla V\|_2^2 + \alpha\|\nabla X^{\text{opt}}\|_2\|\nabla\|_2\|V\|_2\right) \\
\overset{(*)}{\leq} & C' \log\left(\frac{N^2}{s}\right) \left(16\|c_\Omega\|_1 + 2\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1 - \frac{\alpha}{2}\|\nabla V\|_2^2 + \alpha\sqrt{8}\|\nabla X^{\text{opt}}\|_2\|V\|_2\right),
\end{aligned}$$

where $(\diamond)$ is due to the property of partial sum of harmonic series [64], and $(*)$ is due to the fact $\|\nabla\|_2^2 \leq 8$ [46]. As we prepare to apply Proposition 7.3.1 to $c = \mathcal{H}V$, we need to bound $\|\nabla V\|_2$ below in terms of $\|V\|_2 = \|c\|_2$, where $\|V\|_2 = \|c\|_2$ is due to Parseval's identity and the fact that $\{h_{(j)}\}$ forms an orthonormal basis for $\mathbb{C}^{N\times N}$. As $\|\nabla V\|_2 \geq \frac{1}{\sqrt{2}N}\|\nabla V\|_1$, the classical Sobolev inequality (7.4.5) implies

$$\|\nabla V\|_2 \geq \frac{1}{\sqrt{2}N}\|V\|_2. \tag{7.4.7}$$

Thus we have

$$\begin{aligned}
\|c_{\Omega^c}\|_1 \leq & C' \log\left(\frac{N^2}{s}\right) \left(16\|c_\Omega\|_1 + 2\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1 - \frac{\alpha\|c\|_2^2}{4N^2} + \right. \\
& \left. \alpha\sqrt{8}\|\nabla X^{\text{opt}}\|_2\|c\|_2\right).
\end{aligned} \tag{7.4.8}$$

Tube constraint $\|\mathcal{M}\mathcal{H}^* c\|_2 \leq 2\tau$. As $\bar{X}$ and X^{opt} are in the feasible region of the model (7.1.7), for $c = \mathcal{H}V = \mathcal{H}X^{\text{opt}} - \mathcal{H}\bar{X}$, we have

$$\|\mathcal{M}\mathcal{H}^* c\|_2 = \|\mathcal{M}X^{\text{opt}} - \mathcal{M}\bar{X}\|_2 \leq \|\mathcal{M}X^{\text{opt}} - y\|_2 + \|\mathcal{M}\bar{X} - y\|_2 \leq 2\tau.$$

Under the derived cone and tube constraints on c, along with the RIP condition on $\mathcal{M}\mathcal{H}^*$, Theorem 7.3.9 is proved by applying Proposition 7.3.1 and using $\gamma = 16C' \log(N^2/s) \leq 32C' \log(N)$, $k = 6s\log(N)$, $\sigma = 2C' \log\left(N^2/s\right) \|\nabla\bar{X} - (\nabla\bar{X})_s\|_1$, $E_1 = \sqrt{8}\|\nabla X^{\text{opt}}\|_2$, $E_2 = \|c\|_2$, $\beta_1 = \alpha C' \log\left(N^2/s\right)/(2N^2)$, and $\beta_2 = \alpha C' \log\left(N^2/s\right)$. In fact, $5k\gamma^2$ with both particular k and γ leads to the required RIP order $Cs\log^3(N)$ for $\mathcal{M}\mathcal{H}^*$. Together with all these factors and Proposition 7.3.1, we know that if

$$\alpha \leq \frac{\sqrt{8}\sqrt{6s\log(N)}}{K_2\|\nabla X^{\text{opt}}\|_2},$$

then it holds that

$$\|V\|_2 = \|c\|_2 \leq \sqrt{\frac{64N^2\sqrt{6s\log(N)}K_1}{\alpha K_2}\tau + \frac{8N^2}{\alpha}\|\nabla\bar{X} - (\nabla\bar{X})_s\|_1},$$

which leads to the estimation (7.3.13). □

7.4.4 Proof of Theorem 7.3.11

The proof of Theorem 7.3.11 follows the approach of Theorem 7.3.9, in which the *local coherence* of the sensing basis (Fourier) with respect to the sparsity basis (Haar wavelet) plays a major role.

Definition 7.4.3 (Local coherence [123]) *The* local coherence *of an orthonormal basis* $\Phi = \{\phi_j\}_{j=1}^N$ *of* $\mathbb{C}^N$ *with respect to the orthonormal basis* $\Psi = \{\psi_k\}_{k=1}^N$ *of* $\mathbb{C}^N$ *is the function* $\mu^{\text{loc}}(\Phi, \Psi) \in \mathbb{R}^N$ *defined coordinate-wise by*

$$\mu_j^{\text{loc}}(\Phi, \Psi) = \sup_{1\leq k\leq N} |\langle \phi_j, \psi_k \rangle|, \quad j = 1, 2, \ldots, N.$$

The following result indicates that, with high probabilities, signals can be stably reconstructed from subsampled measurements with the local coherence function appropriately used. It can be deemed as a finite-dimensional analog to [175, Theorem 2.1], and a proof can be found in [123].

Lemma 7.4.4 *Let* $\Phi = \{\phi_j\}_{j=1}^N$ *and* $\Psi = \{\psi_k\}_{k=1}^N$ *be two orthonormal bases of* $\mathbb{C}^N$. *Assume the local coherence of* Φ *with respect to* Ψ *is point-wise bounded by the*

function κ in the sense of

$$\sup_{1\le k\le N} |\langle \phi_j, \psi_k\rangle| \le \kappa_j.$$

Fix $\delta > 0$ and integers N, m, and s such that $s \gtrsim \log(N)$ and

$$m \gtrsim \delta^{-2}\|\kappa\|_2^2 s \log^3(s)\log(N),$$

and choose m (possibly not distinct) indices $j \in \Omega \subset \{1,2,\dots,N\}$ i.i.d. from the probability measure ν on $\{1,2,\dots,N\}$ given by

$$v(j) = \frac{\kappa_j^2}{\|\kappa\|_2^2}.$$

Consider the matrix $A \in \mathbb{C}^{m\times N}$ with entries $A_{j,k} = \langle \phi_j, \psi_k\rangle$, $j \in \Omega$, $k \in \{1,2,\dots,N\}$, and consider the diagonal matrix $G = \mathrm{diag}(g) \in \mathbb{C}^{m\times m}$ with

$$g_j = \frac{\|\kappa\|_2}{\kappa_j}, \quad j = 1,\dots,m.$$

Then with probability at least $1 - N^{-c\log^3(s)}$, the RIC δ_s of the preconditioned matrix $\frac{1}{\sqrt{m}}GA$ satisfies $\delta_s \le \delta$.

In particular, the following result describes the local coherence of the orthonormal Fourier basis with respect to the orthonormal Haar wavelet basis, which was initially occurred in [123].

Lemma 7.4.5 (Theorem 4 in [123], slightly modified) *Let $N = 2^n$ be a power of 2, where $n \in \mathbb{N}$. The local coherence μ^{loc} of the orthonormal two-dimensional Fourier basis $\{\varphi_{k_1,k_2}\}$ with respect to the orthonormal bivariate Haar wavelet basis $\{h^\ell_{j,k}\}$ in $\mathbb{C}^{N\times N}$ is bounded by*

$$\begin{aligned}\mu^{\mathrm{loc}}_{k_1,k_2} \le \kappa(k_1,k_2) &:= \min\left(1, \frac{18\pi}{\max(|k_1|,|k_2|)}\right)\kappa'(k_1,k_2)\\ &:= \min\left(1, \frac{18\pi\sqrt{2}}{(|k_1|^2+|k_2|^2)^{1/2}}\right),\end{aligned}$$

and one has $\|\kappa\|_2 \le \|\kappa'\|_2 \le \sqrt{17200 + 502\log(N)}$.

Remark 7.4.6 *For Theorem 4 in [123], $n \ge 8$ was assumed to ensure*

$$17200 + 502\log(N) \le 2700\log(N)$$

and hence $\|\kappa\|_2 \le \|\kappa'\|_2 \le 52\sqrt{\log(N)}$. We regard the assumption as a restriction

on the size $N \times N$ of images, thus we remove this assumption and adopt the bound $\sqrt{17200 + 502 \log N}$ in our following proof. Besides, it was conjectured in [123] that the factor 2700 is due to lack of smoothness for the Haar wavelets, and this factor might be removed by considering smoother wavelets.

Proof of Theorem 7.3.11. Let $P \in \mathbb{C}^{m\times m}$ be the diagonal matrix encoding the weights in the noise model. That is, $P = \operatorname{diag}(\rho)$, where, for κ' as in Lemma 7.4.5, $\rho \in \mathbb{C}^m$ is a vector converted from the matrix

$$\rho(k_1, k_2) = \frac{\|\kappa'\|_2}{\kappa'(k_1, k_2)} = C\sqrt{1 + \log(N)} \max\left(1, \frac{(|k_1|^2 + |k_2|^2)^{1/2}}{18\pi}\right)$$

with $(k_1, k_2) \in \Omega$. Note that $Pg = \rho \circ g$ for $g \in \mathbb{C}^m$. Together with the particular incoherence estimate in Lemma 7.4.5, Lemma 7.4.4 implies that with probability at least $1 - N^{-2c\log^3(s)}$ (as c is a generic constant, the factor 2 of c is removed in the statement of Theorem 7.3.11), $\mathcal{A} := \frac{1}{\sqrt{m}} P\mathcal{F}_\Omega \mathcal{H}^*$ has the RIP of order s and level $\delta < 0.6$ once $s \gtrsim \log(N^2) \gtrsim \log(N)$ and

$$m \gtrsim s\delta^{-2} \log^3(s) \log^2(N^2) \gtrsim s\delta^{-2} \log^3(s) \log^2(N).$$

By the assumption $m \gtrsim s \log^3(s) \log^5(N)$ (in fact, $m \gtrsim s\delta^{-2} \log^3(s) \log^5(N)$ should be assumed), we can assume that $\mathcal{A}$ has the RIP of order $\bar{s} = Cs \log^3(N)$ and level $\delta < 0.6$, where C is the constant derived in Theorem 7.3.9. Moreover, let $V = X^{\text{opt}} - \bar{X}$ and apply Proposition 7.3.1 again to $c = \mathcal{H}V$, where $c_{(1)} := c_{(1)}(V)$ denotes the Haar coefficient corresponding to the constant wavelet, and $c_{(j)} := c_{(j)}(V)$ $(j \geq 2)$ denotes the $(j-1)$-st largest-magnitude Haar coefficient among the remaining. To apply Proposition 7.3.1, we need to find cone and tube constraints for $c = \mathcal{H}V$.

Cone constraint on $c = \mathcal{H}V$: which is the same as (7.4.8) in the proof of Theorem 7.3.9.

Tube constraint: $\|\mathcal{A}c\|_2 = \|\mathcal{A}\mathcal{H}V\|_2 \leq \sqrt{2}\tau$, since

$$\begin{aligned} m\|\mathcal{A}\mathcal{H}V\|_2^2 = \|P\mathcal{F}_\Omega \mathcal{H}^*\mathcal{H}V\|_2^2 &= \|\rho \circ (\mathcal{F}_\Omega V)\|_2^2 \\ &\leq \|\rho \circ (\mathcal{F}_\Omega X^{\text{opt}} - b)\|_2^2 + \|\rho \circ (\mathcal{F}_\Omega \bar{X} - b)\|_2^2 \\ &\leq 2m\tau^2. \end{aligned}$$

The rest is similar to the proof of Theorem 7.3.9, and the only trivial difference is the tube constraint, where 2τ there is replaced by $\sqrt{2}\tau$ here. Hence, we omit the following steps, and the estimation for the setting in this theorem, with constants removed, is the same as (7.3.13). □

7.5 Numerical experiments

We now report some experimental results to validate the effectiveness and numerical solvability of the enhanced TV model (7.1.7). As previously mentioned, the model (7.1.7) is of difference-of-convex, and it can be solved by some well-developed algorithms in the literature. We include the details of an algorithm in Section 7.6.3. For comparison, we consider the TV model (7.1.2) and the $\mathrm{TV}_a - \mathrm{TV}_i$ model in [139]. In our experiments, the TV model (7.1.2) is solved by the split Bregman method studied in [100], and the $\mathrm{TV}_a - \mathrm{TV}_i$ model is solved by the difference-of-convex functions algorithm (DCA) with subproblems solved by the split Bregman method in [139]. Details of the tuned parameters of these algorithms are stated in Section 7.6.3.

As displayed in Figure 7.4, we test the standard *Shepp–Logan phantom*, three more synthetic piecewise-constant images (*Shape*, *Circle*, and *USC Mosaic*), two natural images (*Pepper* and *Clock*), and two medical images (*Spine* and *Brain*). Two sampling strategies are considered in our experiments. The first one is the *radial lines* sampling, and the other one is the strategy (7.3.15) proposed in Theorem 7.3.11, which is referred to as the *MRI-desired* sampling strategy below. All codes were written by MATLAB R2021b, and all numerical experiments were conducted on a laptop (16 GB RAM, Intel CoreTM i7-9750H Processor) with macOS Monterey 12.1.

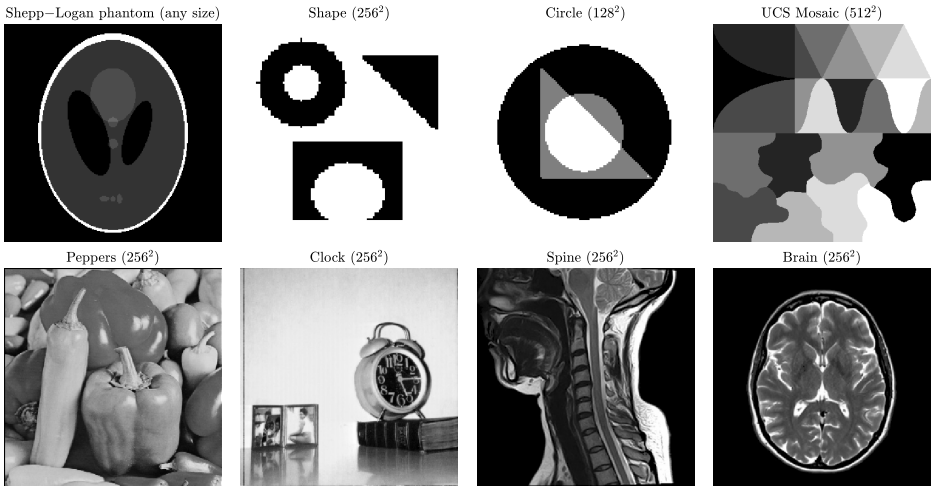


Figure 7.4: Test images.

Example #1: Shepp–Logan phantom. The Shepp–Logan phantom is a standard image in the filed of image reconstruction. Our experiments for this image are organized into three parts. The first part concentrates on the reconstruction of the Shepp–Logan phantom of size 256×256 from noise-free measurements, and α is fixed as 0.8 in the enhanced TV model (7.1.7). We sample along 15, 8, and 7 radial lines, corresponding to sampling rates 6.44%, 3.98%, and 3.03%, respectively. We also take MRI-desired measurements with rates 2.29%, 1.91%, and 1.53%. As shown in Figure 7.5, the enhanced TV model (7.1.7) produces accurate reconstruction in all six sampling settings, and reconstruction quality is much better than those in comparison when the amount of samples is limited (e.g., 7 radial lines and 1.53% MRI-desired measurements). This observation confirms the result presented in Section 7.3.3, which state that the reconstruction error bound (7.3.18) for the enhanced TV model (7.1.7) is tighter than (7.1.10) for the TV model (7.1.7) with a limited amount of noise-free measurements. As mentioned in Section 7.3.3, such a result also pertains to the comparison between the enhanced TV model (7.1.7) and the $\mathrm{TV}_a - \mathrm{TV}_i$ model in [139].

Table 7.1 presents the relative errors in the Frobenius sense and SSIM values in the format of "relative error (SSIM value)" for comparison. The advantages of the enhanced TV model (7.1.7) become apparent when the available measurements are limited (e.g., when the sampling rate is below 3.03%). However, when measurements are relatively sufficient, as in the cases of 15 lines and eight lines, the enhanced TV model (7.1.7) does not produce the least error reconstruction. Notably, though the outperformance of the enhanced TV model (7.1.7) is not sustained as measurements become sufficient, the difference of three models is too tiny to be visually observed. Furthermore, it is worth noting that the SSIM values are 1.0000 for the enhanced TV model (7.1.7) in all six sampling settings, indicating that this model's stability with respect to the number of measurements is well illustrated for the Shepp–Logan phantom image. We also report the performance of the enhanced isotropic TV model (labeled as "Enhanced TV-isotropic" in Table 7.1), using the algorithm described in Section 7.6.3. However, we observe that the enhanced isotropic TV model does not perform better than the enhanced anisotropic TV model (7.1.7), and it even fails for the case of 7 lines, reporting 8.794E+13 (0.0000) and implying that $\alpha = 0.8$ is severely large for it. If $\alpha = 0.6$ for the enhanced isotropic TV model, then 0.3970 (0.6288) is reported for the case of 7 lines. This observation can be partially explained by our discussion in Section 7.1.2 and partially explained by the fact that the value of isotropic TV is less than that of anisotropic TV. In the following experiments, we investigate only the enhanced anisotropic TV model (7.1.7).

The second part illustrates the robustness of the enhanced TV model (7.1.7)

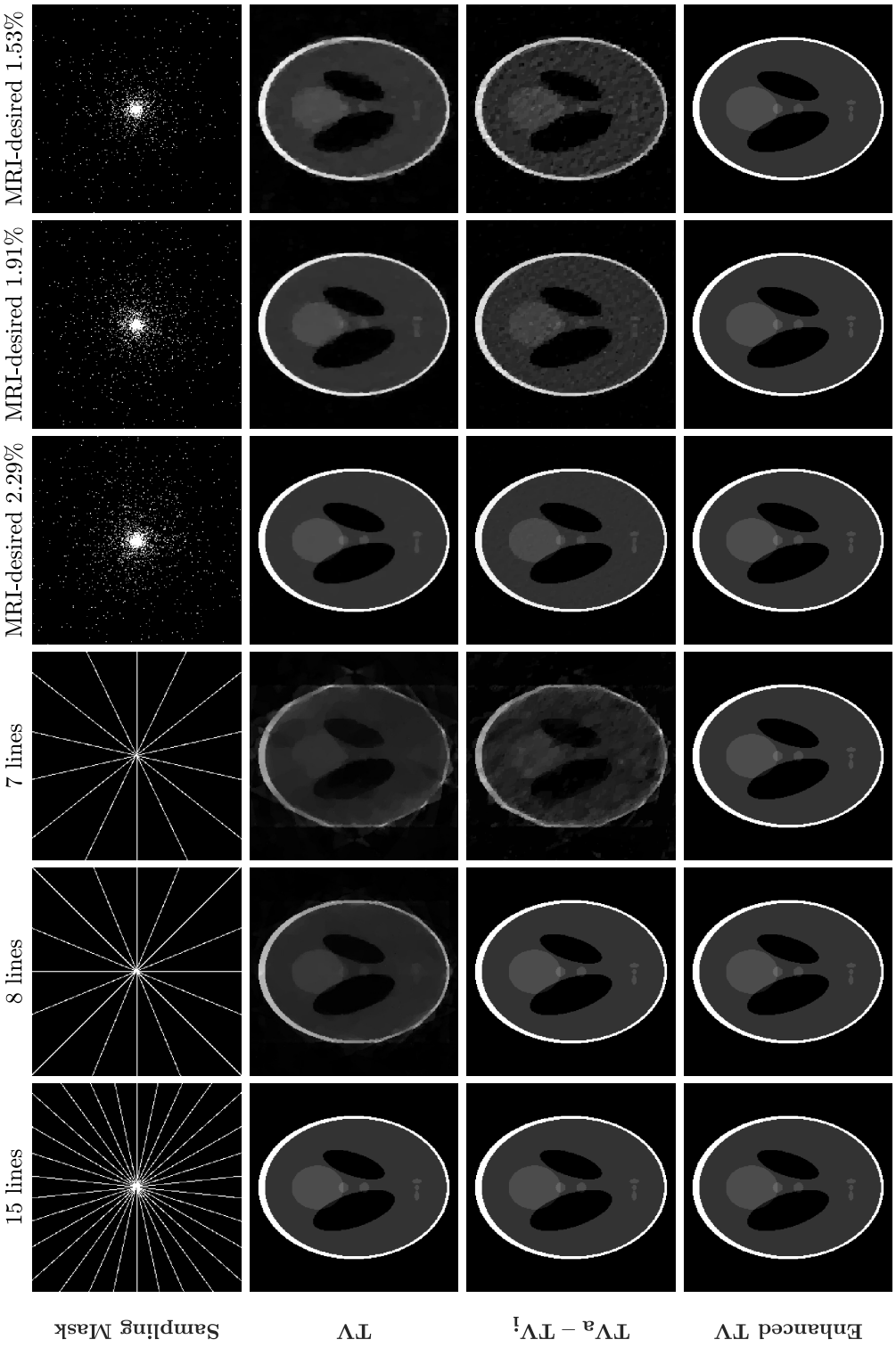


Figure 7.5: Shepp–Logan phantom: Comparison of three models with radial line-sampled and MRI-desired measurements.

Table 7.1: Relative errors and SSIM values of the reconstructed images in Figure 7.5 and images reconstructed using the enhanced isotropic TV regularization.

	TV	$\mathrm{TV}_a - \mathrm{TV}_i$	Enhanced TV	Enhanced TV-isotropic
15 lines (6.44%)	1.924E-13 (1.0000)	7.845E-14 (1.0000)	2.977E-12 (1.0000)	1.212E-07 (1.0000)
8 lines (3.98%)	0.2456 (0.6764)	3.852E-09 (1.0000)	7.841E-07 (1.0000)	1.233E-03 (1.0000)
7 lines (3.03%)	0.4819 (0.4612)	0.3968 (0.5209)	1.608E-06 (1.0000)	/
MRI-desired (2.29%)	0.0415 (0.9890)	0.0266 (0.9896)	8.069E-06 (1.0000)	8.069E-03 (0.9994)
MRI-desired (1.91%)	0.1575 (0.8937)	0.1837 (0.8404)	2.324E-05 (1.0000)	1.017E-02 (0.9990)
MRI-desired (1.53%)	0.2826 (0.7473)	0.2983 (0.7374)	8.456E-05 (1.0000)	1.584E-02 (0.9978)

Table 7.2: Relative errors and SSIM values of the reconstructed images in Figure 7.5, with three levels of noise std = 0.04, 0.06, and 0.08.

	TV	$\mathrm{TV}_a - \mathrm{TV}_i$	Enhanced TV
15 lines (6.44%), std = 0.04	0.1796 (0.5759)	0.1860 (0.4534)	0.0921 (0.9531)
15 lines (6.44%), std = 0.06	0.2506 (0.4866)	0.2748 (0.3161)	0.1038 (0.9490)
15 lines (6.44%), std = 0.08	0.3111 (0.4265)	0.3535 (0.2448)	0.1496 (0.9359)
MRI-deisred (6.50%), std = 0.04	0.1041 (0.7322)	0.1376 (0.5721)	0.0873 (0.9588)
MRI-deisred (6.50%), std = 0.06	0.1498 (0.6101)	0.2082 (0.4179)	0.1393 (0.9477)
MRI-deisred (6.50%), std = 0.08	0.1914 (0.5213)	0.2764 (0.3243)	0.1674 (0.9396)

with respect to noise. In this case, we still set α to 0.8 in the model (7.1.7), and we take measurements along 15 lines (corresponding to a 6.44% sampling rate) and use 6.5% MRI-desired samples. The Fourier measurements are perturbed by Gaussian noise with standard derivations ("`std`" for short) of 0.04, 0.06, and 0.08, respectively. The contamination process is implemented in MATLAB commands: For any image X with size $N \times N$, we first compute its Fourier measurements by the fast Fourier transform (FFT), i.e., `F=fft2(X)/N`. Then we perturb `F` by `F=F+1/sqrt(2)*(std*randn(size(F))+std*1i*randn(size(F)))`.

The relative errors and SSIM values listed in Table 7.2 show that the enhanced TV model (7.1.7) is the most robust one. In particular, in terms of SSIM values, the enhanced TV model (7.1.7) produces much better reconstruction quality, and the superiority is more apparent when the level of noise increases. These results assert the theoretical result in Section 7.3.3 that the enhanced TV model (7.1.7) has a tighter reconstruction error bound than the TV model (7.1.2) and the $\text{TV}_a - \text{TV}_i$ model in [139] when the level of noise is relatively large.

The third part focuses on the phase transition of reconstruction success rates. A reconstruction is considered *successful* if the relative error of the reconstructed image is less than 10^{-3}. We consider the Shepp–Logan phantom with size 64×64 in this part. We choose α among $\{0.7, 0.8, \ldots, 2.7\}$ for the enhanced TV model (7.1.7). We choose the number of measurements m from 3 to 12 radial lines for radial sampling and among $\{100, 140, 180, \ldots, 900\}$ for MRI-desired sampling. For each case, we test five times and report the success rate. According to Theorem 7.3.11, stable reconstruction can be achieved if samples are enough in the sense of (7.3.14) and the model parameter α is bounded in the sense of (7.3.16). The results in Figure 7.6 assert that a successful reconstruction via the enhanced TV model (7.1.7) requires relatively sufficient samples and a reasonably bounded parameter α, thus validating results in Theorem 7.3.11.

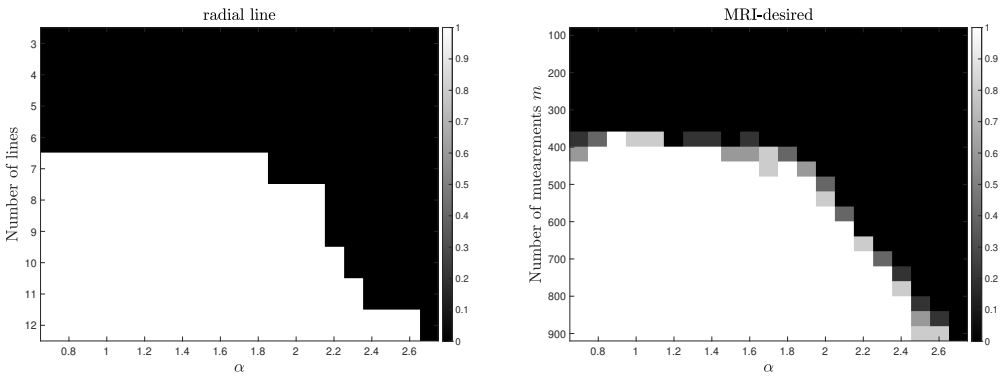


Figure 7.6: Phase transitions with respect to m and α.

Example #2: Synthetic images. Example #1 shows the superiority of the enhanced TV model (7.1.7) for Shepp–Logan phantom with limited samples. The purpose of this study is to further assert this superiority. We consider the radial line sampling and validate this superiority by testing three synthetic images: Shape, Circle, and USC Mosaic. We also fix $\alpha = 0.8$ in the enhanced TV model (7.1.7). When the number of measurements is limited enough, all three models fails to generate good reconstructions. Bearing in mind that the criteria of the limitation on the amount of measurements are different for three models, we now show some cases that the reconstruction via the enhanced TV model (7.1.7) is particularly good while those via the TV model (7.1.2) and the $\text{TV}_a - \text{TV}_i$ model in [139] may fail. The reconstruction results are displayed in Figure 7.7, and the relative errors and SSIM values are reported in Table 7.3. From both Figure 7.7 and Table 7.3, the reconstruction of the enhanced TV model (7.1.7) is significantly better than the other two models.

We also take this example to test how the inner iterations can affect the overall performance of the algorithms under comparison. The algorithm presented in Section 7.6.3 adopts DCA as the outer iteration and uses the ADMM to solve each DCA subproblem. When the maximum number of inner ADMM iterations is increased from 1,000 to 2,000, the numerical results are reported in the fifth column of Figure 7.7, labeled as "Enhanced TV-2,000". We see that even if the enhanced TV model (7.1.7) with at most 1,000 inner iterations is good enough to generate a satisfactory reconstruction, e.g., for Circle and USC Mosaic, increasing the number of inner iterations can further reduce the relative errors by up to several orders of magnitude. This observation provides a simple recipe for higher-accuracy reconstruction.

Example #3: Natural images. We further validate the superiority of the enhanced TV model (7.1.7) by testing it on two natural images: Peppers and Clock. We set α to 1 in the enhanced TV model (7.1.7) for both images and display the reconstruction results from 9.16% MRI-desired samples in Figure 7.8. We also report the relative errors in the Frobenius sense and SSIM values for each reconstruction from MRI-desired samples of rates 9.16%, 13.7%, 18.3%, and 22.9% in Table 7.4.

However, it is worth noting that the enhanced TV model (7.1.7) may not perform as effectively for natural images as it does for the images in Examples #1 and #2 due to the more complicated (non-piecewise-constant) edges in natural images. Nonetheless, this observation is not surprising as the enhanced TV model (7.1.7) is a generalization of the TV model (7.1.2), which performs better for piecewise-constant images than natural images. The enhanced TV model (7.1.7) retains the main feature of the TV regularization while also reduces the loss of contrast.

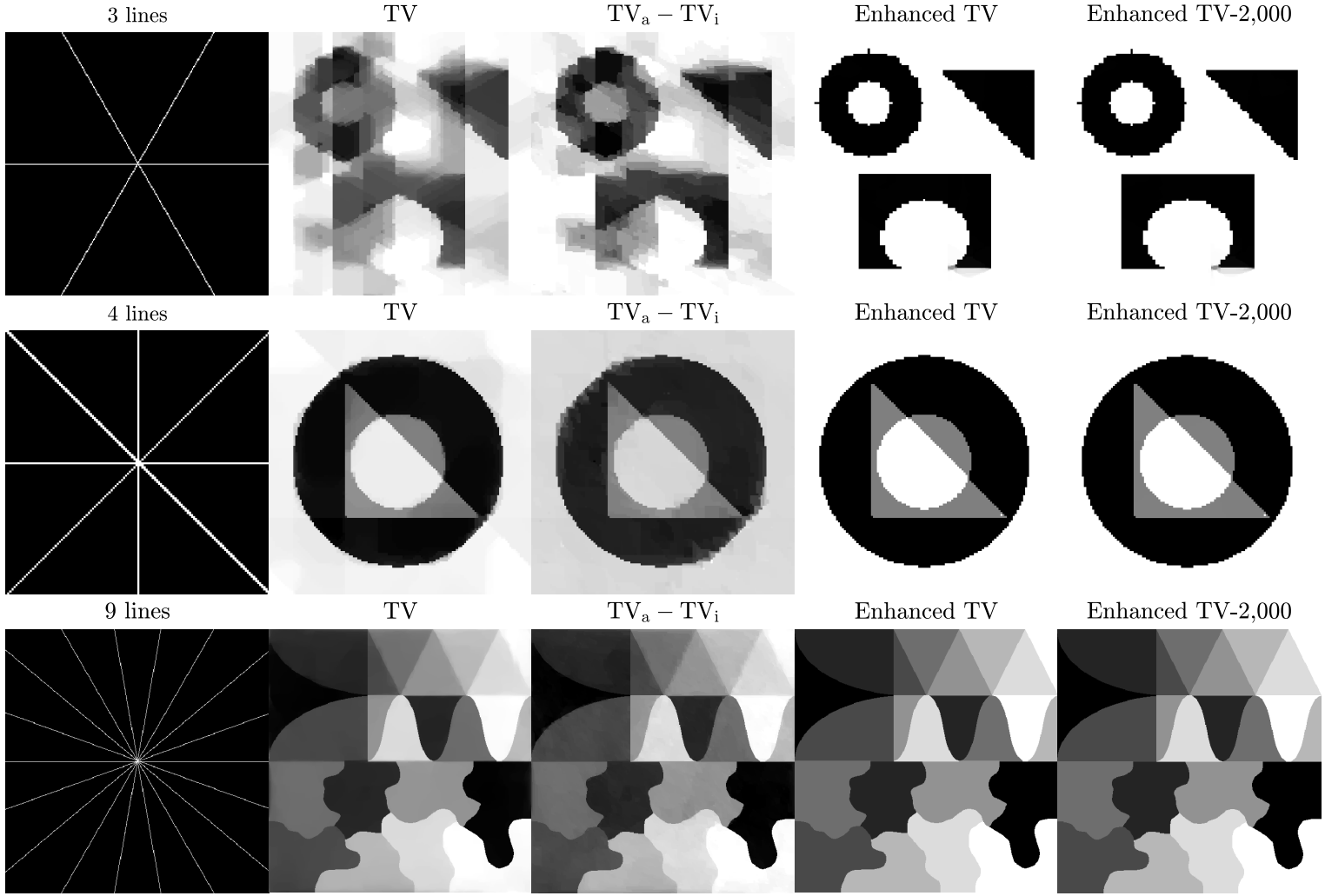


Figure 7.7: Shape, Circle, and USC Mosaic: Comparison of three models with limited measurements.

Table 7.3: Relative errors and SSIM values of the reconstructed images in Figure 7.7.

	TV	$\mathrm{TV}_a - \mathrm{TV}_i$	Enhanced TV	Enhanced TV-2,000
Shape (1.29%)	0.3094 (0.5466)	0.2503 (0.5458)	0.0266 (0.9932)	0.0261 (0.9937)
Circle (3.86%)	0.0394 (0.9705)	0.0498 (0.9430)	7.411E-08 (1.0000)	6.815E-13 (1.0000)
USC Mosaic (1.95%)	0.0405 (0.9032)	0.0439 (0.9024)	8.013E-05 (1.0000)	4.206E-07 (1.0000)

Table 7.4: Relative errors and SSIM values of reconstructions of two natural images with various sampling rates.

	TV	$\mathrm{TV}_a - \mathrm{TV}_i$	Enhanced TV
Peppers (9.16%)	0.0771 (0.8327)	0.0823 (0.7748)	0.0718 (0.8435)
Peppers (13.73%)	0.0597 (0.8793)	0.0624 (0.8409)	0.0536 (0.8908)
Peppers (18.31%)	0.0447 (0.9139)	0.0498 (0.8800)	0.0414 (0.9208)
Peppers (22.89%)	0.0388 (0.9292)	0.0424 (0.9035)	0.0351 (0.9358)
Clock (9.16%)	0.0404 (0.9010)	0.0440 (0.8297)	0.0379 (0.9124)
Clock (13.73%)	0.0288 (0.9356)	0.0319 (0.8884)	0.0272 (0.9421)
Clock (18.31%)	0.0213 (0.9563)	0.0246 (0.9218)	0.0203 (0.9592)
Clock (22.89%)	0.0182 (0.9647)	0.0205 (0.9393)	0.0169 (0.9674)

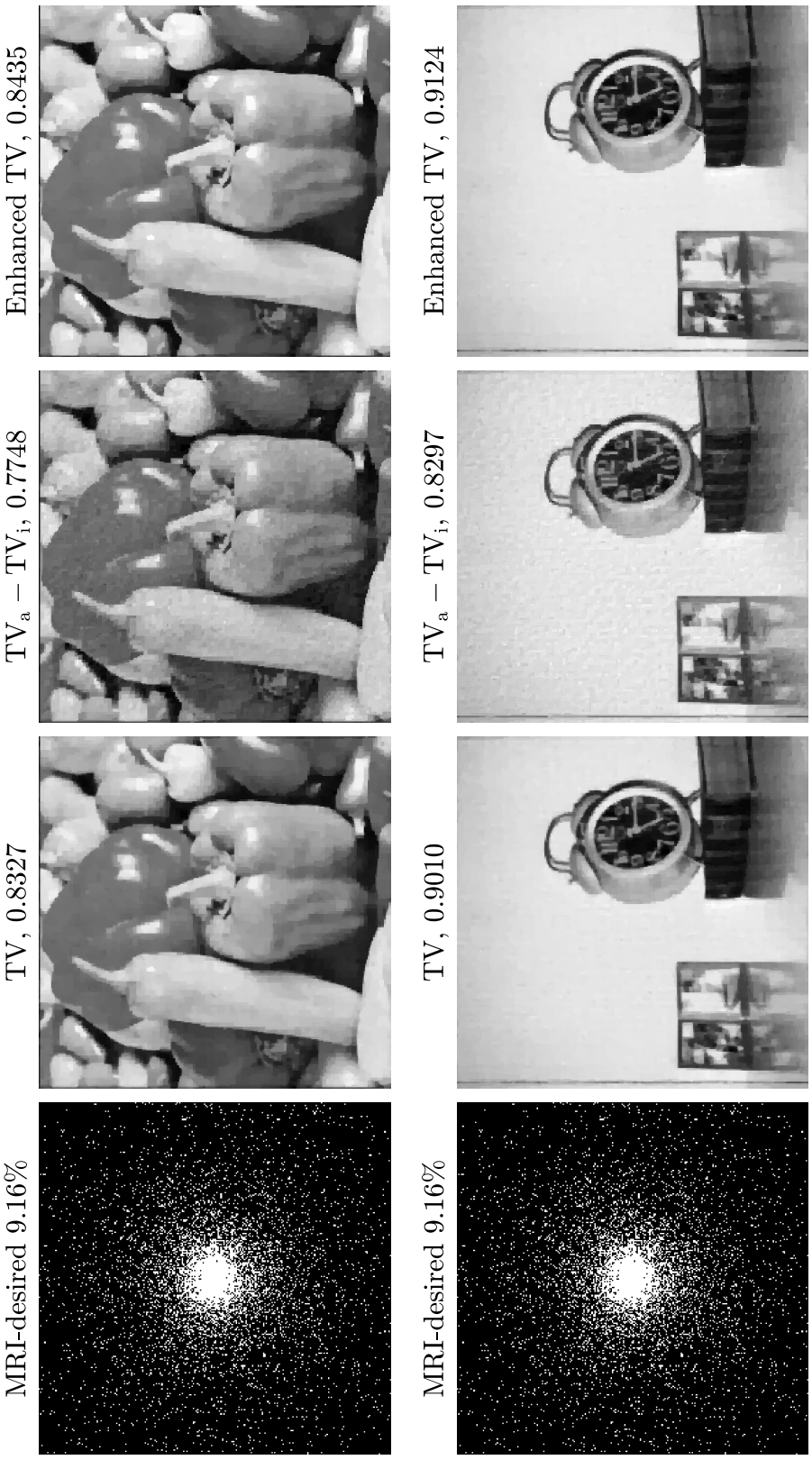


Figure 7.8: Peppers and Clock: Comparison of three models with the MRI-desired sampling with SSIM values reported.

Example #4: Medical images. Finally, we apply the enhanced TV model (7.1.7) to two medical images: Spine and Brain. We again set α to 1 and use 15.3% MRI-desired samples for the reconstruction of Spine and 9.16% for Brain. The reconstructed images are displayed in Figure 7.10, and it is evident that that the enhanced TV model (7.1.7) produces better reconstructions compared to the other models. We also test more sampling rates and report the SSIM values of reconstructions with each rate in Figure 7.9. We observe that the superiority of the enhanced TV model (7.1.7) is more apparent when the sampling rate is relatively low. Therefore, the enhanced TV model (7.1.7) is preferred when measurements are limited. Similar to Example #3, the enhanced TV model (7.1.7) performs less effectively for Example #4 than Examples #1 and #2 due to the non-piecewise-constant edges of these medical images.

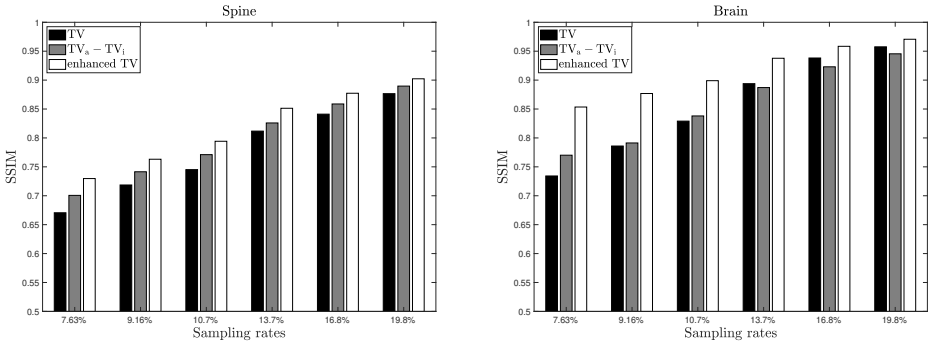


Figure 7.9: Spine and Brain: SSIM values of reconstructions with various sampling rates.

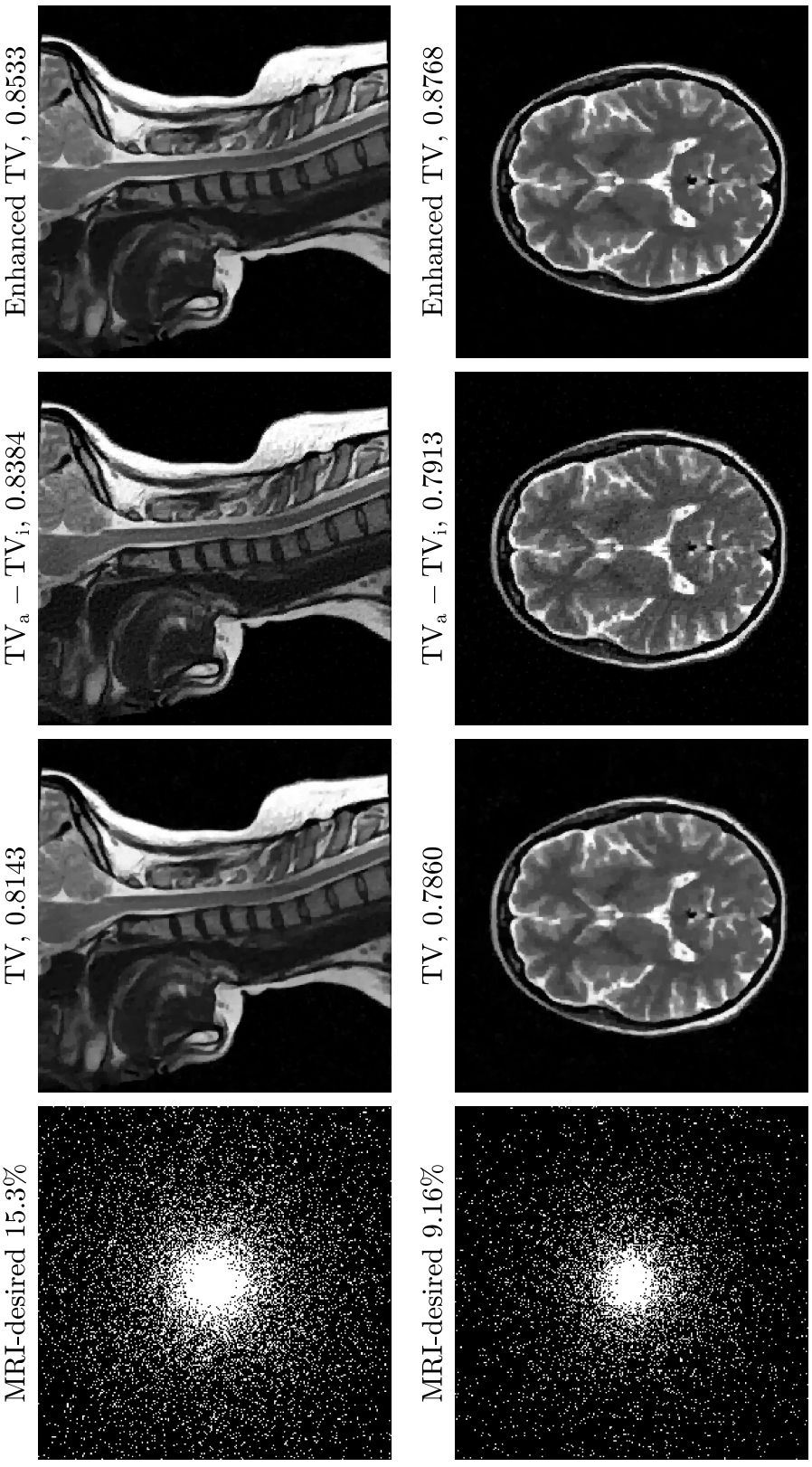


Figure 7.10: Spine and Brain: Comparison of three models on medical images with the MRI-desired sampling with SSIM values reported.

7.6 Supplementary sections

7.6.1 The enhanced TV model in a continuum setting

Let $u : \Omega \to \mathbb{R}$ be an image, where the image domain Ω is a bounded and open subset of $\mathbb{R}^2$. The TV denoising model in [182] for a noisy image $u_0 : \Omega \to \mathbb{R}$ is formulated as

$$\min_u \ \mathcal{E}_{\mathrm{TV}}(u) := \int_\Omega |\nabla u| \mathrm{d}x + \frac{\mu}{2} \int_\Omega (u(x) - u_0(x))^2 \mathrm{d}x, \tag{7.6.1}$$

where $x = (x_1, x_2) \in \Omega$, $|\nabla u| = \sqrt{(\partial_{x_1} u)^2 + (\partial_{x_2} u)^2}$, and $\mu > 0$ balances the TV term and the data fidelity term. Note that the isotropic TV proposed in [182] is used in the model (7.6.1). Though the anisotropic TV defined in [81] is used in the enhanced TV regularization (7.1.6), the main purpose of this section is to explain how the TV is enhanced in the sense of (7.1.6). Thus, we adopt the model (7.6.1) for simplicity. We refer the reader to [147] for the anisotropic TV flow. More specifically, the enhanced (isotropic) TV denoising model in a continuum setting can be written as

$$\min_u \ \mathcal{E}_{\mathrm{ETV}}(u) := \int_\Omega |\nabla u| \mathrm{d}x - \frac{\alpha}{2} \int_\Omega |\nabla u|^2 \mathrm{d}x + \frac{\mu}{2} \int_\Omega (u(x) - u_0(x))^2 \mathrm{d}x. \tag{7.6.2}$$

Then, by computing the first-order variation of the functional, the Euler–Lagrange equation associated with the energy functional $\mathcal{E}_{\mathrm{ETV}}(u)$ in the distributional sense is

$$0 = -\nabla \cdot \left[\frac{\nabla u}{|\nabla u|} \right] + \alpha \Delta u + \mu(u - u_0) \quad \text{with} \quad \left. \frac{\partial u}{\partial \mathbf{n}} \right|_{\partial\Omega} = 0, \tag{7.6.3}$$

where $\mathbf{n}$ denotes the outer normal derivative along the boundary $\partial\Omega$ of Ω.

Alternatively, as [182], we could use the *gradient descent marching with artificial time* t. That is, the solution procedure of the Euler–Lagrange equation (7.6.3) uses a parabolic equation with time t as an evolution parameter. This means, for $u : \Omega \times [0, T] \to \mathbb{R}$, we solve

$$u_t = -\frac{\partial \mathcal{E}_{\mathrm{ETV}}}{\partial u} = \nabla \cdot \left[\frac{\nabla u}{|\nabla u|} \right] - \alpha \Delta u - \mu(u - u_0) \quad \text{for } t > 0, \ x \in \Omega, \tag{7.6.4}$$

with a given initial condition $u(x, 0)$ and the boundary condition

$$\frac{\partial u}{\partial \mathbf{n}}|_{\partial\Omega} = 0.$$

Note that there is a backward diffusion term $-\alpha \Delta u$ in the evolution equation (7.6.4). Thus, as t increases, we approach a denoised and deblurred version of the image if the blur is assumed to follow such a diffusion process.

If the energy functional $\mathcal{E}_{\mathrm{ETV}}(u)$ has a minimum, then the minimizer must satisfy the Euler–Lagrange equation (7.6.4). Certainly, the existence of the minimizer of $\mathcal{E}_{\mathrm{ETV}}$ is unknown for an arbitrary α. On the other hand, with

$$\alpha < \mu \inf_{x\in\Omega} \frac{|u(x)|^2}{|\nabla u(x)|^2},$$

the Lagrangian

$$\mathcal{L}_{\mathrm{ETV}}(\nabla u, u, x) := |\nabla u| - \frac{\alpha}{2}|\nabla u|^2 + \frac{\mu}{2}(u(x) - u_0(x))^2$$

is bounded below by $|\nabla u(x)| + \frac{\mu-\alpha}{2}|u(x)|^2 - \mu u(x)u_0(x) + |u_0(x)|^2$, which is a convex function with respect to variables ∇u and u. Hence, $\mathcal{E}_{\mathrm{ETV}}$ is bounded below, and any stationary point u^* of $\mathcal{E}_{\mathrm{ETV}}$ (including global and local minimizers) must be finite and satisfy the corresponding Euler–Lagrange equation (7.6.4) involving the backward diffusion term. This requirement on α explains the rationale of the assumption on the upper bound of α in Theorems 7.3.6, 7.3.9, and 7.3.11 (e.g., $\alpha \le \frac{\sqrt{48s\log(N)}}{K_2\|\nabla X^{\mathrm{opt}}\|_2}$ in Theorems 7.3.9 and 7.3.11).

7.6.2 Implementation details for enhanced TV denoising

For denoising, let the noisy image be $y = \bar{X} + e \in \mathbb{C}^{N\times N}$. The denoising model using the enhanced TV regularization (7.1.6) is formulated as

$$\min_{X\in\mathbb{C}^{N\times N}} \|\nabla X\|_1 - \frac{\alpha}{2}\|\nabla X\|_2^2 + \frac{\mu}{2}\|y - X\|_2^2, \tag{7.6.5}$$

where $\mu > 0$ is a parameter balancing the enhanced TV regularization term and the data fidelity term. Note that the model (7.6.5) is the discretization of the model (7.6.2). The model (7.6.5) can be solved by the DCA in [212, 213], and its subproblems can be solved by the splitting Bregman iteration in [100]. We summarize the resulting algorithm as Algorithm 2 below, in which MaxDCA denotes the maximum number of the DCA iterations and MaxBreg denotes is the maximum number of the Bregman iterations.

To reproduce Figure 7.1, we test the noisy *Strip* image (displayed in Figure 7.1) with size 128×128. The parameters for Algorithm 2 are set as $\alpha = 1.2$, $\mu = 0.8$, $\beta = 1$, MaxDCA $= 10$, and MaxBreg $= 1{,}000$. We contaminate the test image by adding random values onto each pixel from a normal distribution with mean 0 and standard deviation 0.6, without normalizing all pixel intensities such that they are in the range of $[0, 1]$. To match the total number of iterations, we adopt the same

Algorithm 2: Solving the unconstrained denoising model (7.6.5)

Input: Define $X^0 = 0$, $z = 0$, $k = 0$, $d_x = d_y = 0$, MaxDCA and MaxBreg

while $k <$ MaxDCA **do**
  $b_x = b_y = 0$, $p = 0$;
  while $p <$ MaxBreg **do**
    $u = \left(\mu + \beta\nabla^{\mathrm{T}}\nabla\right)^{-1}\left(\mu y + \beta D_x^{\mathrm{T}}(d_x - b_x) + \beta D_y^{\mathrm{T}}(d_y - b_y)\right)$;
    $d_x = \text{shrink}\left(D_x u + b_x + \alpha D_x X^k/\beta, 1/\beta\right)$;
    $d_y = \text{shrink}\left(D_y u + b_y + \alpha D_y X^k/\beta, 1/\beta\right)$;
    $b_x = b_x + D_x u - d_x$;
    $b_y = b_y + D_y u - d_y$;
    $p \leftarrow p + 1$;
  end
  $X^k = u$;
  $k \leftarrow k + 1$;
end

parameters for the splitting Bregman iteration for solving the TV denoising model except that the number of the splitting Bregman iterations is set as 10,000.

7.6.3 DCA for the enhanced TV model

We apply the mentioned DCA in [212, 213] to solve the enhanced TV model (7.1.7). We denote by $D_x X$ and $D_y X$ the horizontal and vertical components of ∇X, respectively, where D_x and D_y can be deemed as two operators. The DCA replaces the second component $\frac{\alpha}{2}\|\nabla X\|_2^2$ of the enhanced TV regularization term (7.1.6) by a linear majorant $\langle X - X^k, \xi^k\rangle$, where

$$\xi^k \in \partial\left(\frac{\alpha}{2}\|\nabla X\|_2^2\right) = \{\alpha\nabla^{\mathrm{T}}\nabla X^k\},$$

and then solves the resulting convex optimization problem to generate the iterate X^{k+1}. Ignoring the constant term $\langle X^k, \xi^k\rangle$ in the objective function, the iterative scheme of the DCA reads as finding X^{k+1} that minimizes

$$\begin{aligned} \min_{X\in\mathbb{C}^{N\times N}} \quad & \|D_x X\|_1 + \|D_y X\|_1 - \alpha\langle D_x X, D_x X^k\rangle - \alpha\langle D_y X, D_y X^k\rangle \\ \text{s.t.} \quad & \|\mathcal{M}X - y\|_2 \le \tau. \end{aligned} \tag{7.6.6}$$

Convergence of the DCA (7.6.6) can be found in, e.g., [13, 212, 213]. Recall that a convex function $F : \mathbb{R}^d \to \mathbb{R}$ is said to be *ρ-strongly convex* if $F(x) - \frac{\rho}{2}\|x\|_2^2$ is convex on $\mathbb{R}^d$. A simple but critical fact ensuring the convergence is that the component $\frac{\alpha}{2}\|\nabla X\|_2^2$ is strongly convex either if X is mean-zero or if X contains zero-valued pixels (cf. the classical Sobolev inequality (7.4.5) and Equation (7.4.7)).

To solve (7.6.6), we suggest using the benchmark alternating direction method of multipliers (ADMM) in [99]. Clearly, X^{k+1} is also a solution to the reformulated problem

$$\begin{aligned} \min_{X\in\mathbb{C}^{N\times N}} \quad & \|d_x\|_1 + \|d_y\|_1 - \alpha\langle d_x, D_x X^k\rangle - \alpha\langle d_y, D_y X^k\rangle, \\ \text{s.t.} \quad & \mathcal{M}X - y - z = 0, \\ & z \in \mathcal{B}(0,\tau) := \{x \in \mathbb{R}^m : \|x\|_2 \le \tau\}, \\ & D_x X = d_x, \quad D_y X = d_y. \end{aligned}$$

Introducing three Lagrange multipliers λ, b_x, and b_y, we write the augmented Lagrangian function of (7.6.7) as

$$\begin{aligned} \mathcal{L}_{\beta,\mu}(X, d_x, d_y, z, b_x, b_y, \lambda) := {} & \|d_x\|_1 + \|d_y\|_1 - \alpha\langle d_x, D_x X^k\rangle - \alpha\langle d_y, D_y X^k\rangle \\ & + \frac{\mu}{2}\|z - (\mathcal{M}X - y) - \lambda\|_2^2 + \frac{\beta}{2}\|d_x - D_x X - b_x\|_2^2 + \frac{\beta}{2}\|d_y - D_y X - b_y\|_2^2, \end{aligned}$$

where $\mu, \beta > 0$ are penalty parameters. Implementations of the ADMM to (7.6.6) are included as Algorithm 3 below, in which MaxDCA denotes the maximum number of the DCA iterations, MaxADMM is the maximum number of the ADMM iterations for (7.6.7) with a given X^k, and "tol" is the tolerance for the DCA iterations.

Algorithm 3: DCA for the enhanced TV model (7.1.7)

Input: Define $X^0 = 0$, $z = 0$, $k = 0$, $d_x = d_y = 0$, MaxDCA, MaxADMM, and tol

while $k <$ MaxDCA *and* $\|X^k - X^{k-1}\|_2 >$ tol **do**
  $b_x = b_y = 0$, $p = 0$;
  while $p <$ MaxADMM **do**
    $u = \left(\mu\mathcal{M}^*\mathcal{M} + \beta\nabla^{\mathrm{T}}\nabla\right)^{-1}\left(\mu\mathcal{M}^*(y - z - \lambda) + \beta D_x^{\mathrm{T}}(d_x - b_x) + \beta D_y^{\mathrm{T}}(d_y - b_y)\right)$;
    $d_x = \text{shrink}\left(D_x u + b_x + \alpha D_x X^k/\beta, 1/\beta\right)$;
    $d_y = \text{shrink}\left(D_y u + b_y + \alpha D_y X^k/\beta, 1/\beta\right)$;
    $z = \mathcal{P}_{\mathcal{B}(0,\tau)}(\mathcal{M}u - y + \lambda)$;
    $b_x = b_x + D_x u - d_x$;
    $b_y = b_y + D_y u - d_y$;
    $\lambda = \lambda + (\mathcal{M}u - y) - z$;
    $p \leftarrow p + 1$;
  end
  $X^k = u$;
  $k \leftarrow k + 1$;
end

If the TV term $\|X\|_{\mathrm{TV}_a} = \|\nabla X\|_1$ is replaced by the isotropic version, then $D_x X$ and $D_x X$ in (7.6.6) or d_x and d_y in (7.6.7) do not decouple, but we can still update d_x and d_y in a closed-form manner. Like the extension from the anisotropic TV to the isotropic one in [100], to solve the enhanced isotropic TV model, we merely need to replace lines 5 and 6 in Algorithm 3 with

$$
\begin{array}{l}
s = \sqrt{|D_x u + b_x + \alpha D_x X^k/\beta|.^2 + |D_y u + b_y + \alpha D_y X^k/\beta|.^2};\\
d_x = \max(s - 1/\beta, 0). * (D_x u + b_x + \alpha D_x X^k/\beta)./s;\\
d_y = \max(s - 1/\beta, 0). * (D_y u + b_y + \alpha D_y X^k/\beta)./s;
\end{array}
$$

where the point . before operations denotes entry-wise operations.

In our numerical experiments, to implement Algorithm 3, we set $\mu = 10^3$, $\beta = 10$, MaxDCA $= 15$, tol $= 10^{-10}$ (for noise-free measurements) or 10^{-3} (for noisy measurements), and MaxADMM $= 1{,}000$. For the $\mathrm{TV}_a - \mathrm{TV}_i$ model in [139], we use the same penalty parameters and stopping criterion for running the DCA; and for the split Bregman method in solving the DCA subproblem, we set the maximum numbers of outer and inner iterations as 50 and 20, respectively. The parameters for Bregman iterations were suggested in [139], and they coincide with the maximum number of the inner ADMM iterations in Algorithm 3, as $50 \times 20 = 1,000$. For the TV model (7.1.2), we adopt the same penalty parameters and tolerance for outer iterations. We set the maximal numbers of outer and inner iterations to be 50 and 200, respectively; both numbers were suggested in [139].

Intentionally blank page.

Chapter 8

Conclusion and Outlook

In this chapter we make some conclusions and discuss possible directions of future work involving deep neural networks.

8.1 Conclusion

In **Chapter 2**, we analyze an exactness-relaxing hyperinterpolation and point out a potential development of hyperinterpolation that the exactness requirement may be dismissed with the stability and convergence resulting maintained. We claim that the required exactness degree $2n$ in constructing an original hyperinterpolation in [196] can be relaxed to $n+k$ with $0 < k \leq n$. Such relaxation is valid for k at least to $n+1$, because the projection property $\mathcal{L}_n f = f$ for all $f \in \mathbb{P}_k$ does not maintain for any non-trivial polynomial spaces.

In **Chapter 3**, we propose efficient hyperinterpolation to approximate singular and oscillatory functions in the spirit of the product-integration rule. This approximation scheme is new and easy to be implemented. We also obtain error bounds in cases of $K \in L^1(\Omega)$, $L^2(\Omega)$, and $C(\Omega)$, respectively. Our theoretical analysis and numerical experiments make it legitimate to apply the proposed scheme to solve problems involving singularities and oscillatory behaviors. On the other hand, efficient hyperinterpolation heavily relies on the accurate or stable evaluation of the modified moments. Thus, much more effort is necessary to understand our scheme's implementation to approximate the function $F = Kf$ with various singular and oscillatory terms K.

In **Chapter 4**, we investigate the approximation scheme of hyperinterpolation on the sphere. The quadrature rules used in the construction of hyperinterpolation are not required to be exact for any polynomials but only to satisfy the Marcinkiewicz–Zygmund property, and we give the corresponding error estimate.

Such an approximation scheme without the quadrature exactness assumption is referred to as the unfettered hyperinterpolation. If the quadrature rules use QMC designs, then the error estimate can be refined. To emphasize the particularity of QMC designs, we refer to the hyperinterpolation using QMC designs as quadrature points as the QMC hyperinterpolation. Note that the QMC hyperinterpolation can be regarded as a special case in the general framework of the unfettered hyperinterpolation. The general and refined estimates are split into two terms: a term representing the error estimate of the original hyperinterpolation of full quadrature exactness and another term introduced as compensation for the loss of exactness degrees. The newly introduced term may not converge to zero as the degree of hyperinterpolation tends to ∞, and we need to control it in practice. The numerical experiments show that the construction of hyperinterpolation using quadrature rules without exactness is feasible, and they verify the error estimates given in Sections 4.3 and 4.4. The general framework of the unfettered hyperinterpolation on the sphere may be extended to the scheme of hyperinterpolation on other regions, such as the disk [106], the square [39], the cube [40, 224], and the spherical triangles [206].

In **Chapter 5**, we propose and investigate a quadrature-based spectral method (5.1.12) for solving the Allen–Cahn equation (5.1.2). The purpose of investigating spectral methods is to lift the stringent condition imposed on the time stepping size τ in the literature on numerical methods for the Allen–Cahn and related phase-field equations. Instead, we impose mild conditions on the degree N of the polynomial numerical solutions. The motivation of involving quadrature rules is to provide more precise analysis for the numerical solutions that are computed on computers and to confront the practical situation where the data samples may not be obtained from quadrature points where we desire. Thanks to the recent development of numerical integration on the sphere, our scheme (5.1.12) is an intrinsic methods on the sphere, which is different from the numerical methods for equations on the Euclidean spaces. Namely, we use coordinates intrinsic to the sphere and a sphere-based mesh to discretize the PDE rather than considering the parameterization of the sphere or extending into a narrow band domain around the sphere and then using the extrinsic coordinates and an Euclidean-based mesh for discretization. As a result, our scheme (5.1.12) is consistent with the dimension of the original problem and maintains the intrinsic properties on the sphere. This consistency also suggests that our scheme (5.1.12) can also be extended to closed smooth surfaces diffeomorphic to the sphere (see, e.g., the manipulation of the change of variables in [102]). Moving forward, our approach can be applied to other PDEs in the form of (5.1.1), where the nonlinear part $\mathbf{N}(u)$ is linearized by its hyperinterpolation. The theoretical analysis of the

resulting scheme may follow the procedure in this chapter, with modifications involving of the definition of certain PDEs. The implementation process is also similar to that in this chapter. Furthermore, our idea on the sphere can also be extended to other compact manifolds and bounded, closed regions where hyperinterpolation has been or can be established. For such an extension, the Marcinkiewicz–Zygmund inequality has been investigated on compact manifolds in [89]. However, we utilize the property that our basis functions, the spherical harmonics, happen to be the eigenfunctions of the negative Laplace-Beltrami operator, and the eigenvalues have explicit expressions. Hence a potential obstacle to this extension is the underdeveloped spectral theory of the Laplace-Beltrami operator on these manifolds, which may complicate implementation in the sense that the differential operators should be discretized in an additional step. Nonetheless, our method shows promise for numerically solving a wide range of semi-linear PDEs in bounded, closed regions of $\mathbb{R}^d$ where hyperinterpolation can be defined.

In **Chapter 6**, we proposed a weakly convex penalty, named the springback penalty, for signal reconstruction from incomplete and inaccurate measurements. The springback penalty inherits major theoretical and numerical advantages from the convex ℓ_1 penalty and its various non-convex alternatives. We established exact and stable reconstruction results for the springback-penalized model (6.1.5) under the same RIP condition as the BP model (6.1.3); both the sparse and nearly sparse signals are considered. The springback-penalized model (6.1.5) is particularly suitable for signal reconstruction with a large level of noise or a limited number of measurements. We verified the effectiveness of the model and its computational tractability. The springback penalty provides a new tool to construct effective models for various sparsity-driven reconstruction problems arising in many areas such as compressed sensing, signal processing, image processing, and least-squares approximation.

In **Chapter 7**, we focused on enhancing the canonical constrained total variational (TV) minimization model for image reconstruction by the spingback regularization Chapter 6. The enhanced TV model improves the original TV model by adding a backward diffusion process to further reduce the loss of contrast. The reconstruction guarantees of the enhanced TV model (7.1.7) for non-adaptive subsampled linear RIP and variable-density subsampled Fourier measurements were theoretically established. For non-adaptive linear RIP measurements, the RIP level δ requirement was relaxed from $\delta < 1/3$ (which was derived for the TV model (7.1.2) in [157]) to $\delta < 0.6$. The reconstruction error bounds estimated in Theorems 7.3.6 and 7.3.9 also imply reasonable reconstruction error estimations for the TV model (7.1.2) when $\delta \to 0.6$. In contrast, the bounds derived in [157] for the TV model (7.1.2) tend to be infinity as $\delta \to 0.6$. For variable-density sampled Fourier measurements, the

required minimum number of measurements of the enhanced TV model (7.1.7) was shown to be around 30.86% of that established in [123] for the TV model (7.1.2). This improvement is due to the relaxation of the requirement on δ. It is worth noting that we only consider the anisotropic TV, and proofs of the main theoretical results can be easily extended to the isotropic TV case. In addition, our results can also be generalized in several other ways. For example, one can consider other sampling strategies, such as those in [2, 164], for Fourier samples as considered in Theorem 7.3.11. For the guarantees analysis with Fourier measurements, noise is measured by the weighted ℓ_2-norm (see (7.3.17)). One can consider some other norms to measure noise, such as those in [2, 164]. Our theoretical results for two-dimensional images can also be extended to higher dimensional signals, as considered in [2, 156]. Furthermore, it also seems promising to apply the enhanced TV model (7.1.7) to other problems such as image inpainting and super-resolution problems, combining the enhanced TV regularization (7.1.6) with other data fidelity terms to model some problems such as image segmentation and motion estimation, and using the enhanced TV regularization (7.1.6) in combination with other widely-used convex and/or non-convex regularizers to model various more challenging image processing problems.

8.2 Towards deep neural networks

Numerical analysis involves designing and analyzing algorithms that solve continuous mathematics problems using numerical approximation. These algorithms generate approximate but accurate solutions to challenging problems in natural sciences and engineering whose exact solutions may be impossible or prohibitively expensive to calculate.

Apart from the information-based situation considered in this thesis, another challenge arising from real-world applications for numerical analysis is the high dimensionality. High-dimensional problems always occur for various reasons, including using models with a large number of variables and measuring sampled data by hundreds of different quantities. Moreover, high-dimensional problems also imply the information-based situation, for example, data in high-dimensional cases is often unstructured. Meanwhile, numerical analysts are confronted by at least two tricky issues on the way to high-dimensional problems:

Firstly and notably, the difficulty of solving high-dimensional problems by classic numerical methods grows extraordinarily rapidly as the number of variables (or the dimension) increases, which is now widely known as the curse of dimensionality. A common experience is that the cost of an algorithm grows exponentially with

dimension, making it prohibitive in the regime of moderate or large dimensions. To a certain degree, high dimensions can be compensated by a sufficient degree of smoothness, e.g., the assumption $u_0 \in H^s(\mathbb{S}^{d-1})$ with $s > d-1$ in Chapter 5. Yet, making a smoothness assumption is impractical because the smoothness of function is inherited from real-world problems that we cannot control (e.g., images and cartoon-like functions). Secondly, modeling complex behavior in nature requires a sufficiently large number of parameters, expecting adequate data samples that match the model complexity. However, data sampling may be prohibitive or expensive, and only a limited number of samples may be available for further investigation.

Hence, an efficient method for high-dimensional problems should lift the curse of dimensionality and be overparameterized. The past decade has witnessed the astonishing success of artificial intelligence (AI) in science and engineering. Deep neural networks, considered the "workhorse" leading this massive wave of artificial intelligence, happen to satisfy the above mentioned two preferred properties. In practice, deep neural networks seem to perform incredibly well on problems where the input dimensions are very high, and this surprising performance cannot be explained within a classical approximation framework since classical results always suffer from the curse of dimensionality; see reviews in [104, 163]. Another mystery of deep neural networks is the observation that highly overparameterized deep neural networks generalize well [158]; that is, they do not fit training data too tightly (known as overfitting) and hence perform well on test data.

Deep neural network-based methods (or deep learning methods) fill the gap between high-dimensional problems and classical numerical methods. However, it is widely acknowledged in academia that a convincing mathematical explanation of the enormous success of deep learning needs to be developed. Some early-stage results on the theoretical side of deep learning are recently summarized in [22]. Thus, a possible future direction following this thesis would be confined to the intersection between numerical analysis and deep learning, and the core problems to be investigated are

Why do neural networks perform well in very high-dimensional environments, and how can we utilize this property for high-dimensional numerical analysis and natural sciences?

Some specific problems with related backgrounds, significance, and potential methods are presented in the following context.

Within the classic framework of learning theory [65, 66], the performance of a learning algorithm can be measured by a sum of the approximation error, generalization error, and optimization error, which corresponds to three research directions, namely, expressivity, generalization, and learning/optimization. Related to topics in

this thesis, the first two may be considered. Though we work within the framework of learning theory, it should be noted that the classical theory cannot satisfactorily explain the success of deep learning; see some justification on why a new theory is needed in [22].

Expressivity. The expressivity of deep learning may be the wealthiest area in terms of mathematical theories at present; see a recent survey article [71]. It helps to examine the approximation power of various neural networks and enhance our understanding of their architectures from a synergistic view of applied harmonic analysis and approximation theory. Perhaps, the most common approach to obtain a neural network approximation error bound is to show that the target function has a decomposition in terms of some fundamental building blocks with controllable decomposition coefficients and then to show that each of these blocks can be efficiently captured by deep networks. Such decomposition could be achieved by wavelets or their mathematical cousins, such as shearlets, ridgelets or curvelets, or by global representations such as (generalized) Fourier series. For target functions in classical smoothness spaces, however, the approximation by their decomposition suffers from the curse of dimensionality. That is, the neural network approximation error bounded using such ideas also suffers from the curse of dimensionality, which sounds impractical for high-dimensional problems. Regarding this issue, we may consider the following two research questions.

Can we introduce new function spaces or classes of target functions to (partially) eliminate the curse of dimensionality?

Answering this question provides a remedy for the curse of dimensionality because the search for new high-dimensional function spaces or classes is always a driving issue when dealing with high-dimensional approximation, as commented in [71, Sec. 8.10.2]. Some new spaces or classes may explain why neural networks are efficient tools in high-dimensional function approximation, and such a search shall provide more reliable and accurate guides for deep learning in real-world applications.

Can we improve the current results when the target functions belong to some classical smoothness spaces?

Answering this question, though not breaking the curse of dimensionality, links deep learning methods to traditional mathematical problems with more solid analysis because many mathematical theories are derived in some classical smoothness spaces. For example, we may provide more solid analysis for deep learning methods for solving partial differential equations (PDEs) in combination with the PDE theory, most of which is, however, developed in some classical smoothness spaces.

Generalization. In contrast to expressivity, generalization of deep learning may be the least understood part of deep learning theory. It helps to describe the out-of-sample performance of learning algorithms. This property has been intensively studied in the field of statistical learning theory: assuming the training samples are *i.i.d.* drawn from a probability distribution, one can make use of concentration inequalities to bound the generalization error. If the set of training samples is a sequence of samples drawn from a product probability space, this advanced version of the learning task is known as online learning. We propose the following two research questions regarding generalization in two different settings.

Can we establish an online deep learning theory when the set of training samples is a sequence of samples drawn from the product probability space with possibly different Borel probability measures?

Can we quantify the performance of a deep learning algorithm if the training samples are given in the sense that we do not have further information (e.g., probability distribution) on these samples or the luxury to query more?

Answering both questions advances our understanding of the generalization of neural networks towards a more general setting: one from batch learning to online learning, and the other from *i.i.d.* training samples to a much more practical situation.

Apart from the above questions within the framework of deep learning theory, the following research topic may help to explore the enhancement of classic numerical methods by deep learning.

PDEs and integral equations. As long-standing topics in numerical analysis, PDEs and integral equations are proposed to describe the physical world. Classical numerical methods for PDEs and integral equations suffer from the curse of dimensionality, which may need to be improved in high-dimensional regimes. Due to the approximation power of neural networks for high-dimensional functions, it is not surprising that a neural network ansatz could successfully solve equation models. Since 2017, deep learning-based methods for PDEs have significantly advanced, with highlights including the Deep Ritz Method [79], the Deep Galerkin method [194], and the Physics Informed Neural Networks (PINNs) [172]. The common way is to approximate the solution of a PDE by a deep neural network, which is trained by minimizing a loss function incorporating the equation itself and the initial/boundary conditions. Though these methods have been a popular topic during the past five years, there is growing concern about whether these methods are indeed practical, because only a few efforts were made to analyze these methods rigorously. We aim to address the following questions.

Can we provide theoretical numerical analysis of deep learning methods for PDEs?

Can we develop deep learning methods for integral equations?

Apart from some modern methods for integral equations, answering this question may help to solve PDEs more efficiently in light of the integral equation method for PDEs. In such a case, the considered PDE can be reformulated as an integral equation by some integral transformations. Numerically, there is a potential benefit for the integral equation method: differential operators are numerically unstable due to unbounded norms of their inverse, but integral operators occurring in integral equations are usually compact. Both questions help the development of deep learning methods for high-dimensional numerical analysis.

Deep learning tasks in various manifolds. Real-world applications may be modeled on various regions. For example, geomathematicians may always consider spherical problems, since the Earth's potato shape can be mapped to a 2-sphere by an appropriate smooth mapping. However, most of deep learning theory focuses on target functions defined on a cube $[0,1]^d$, and transferring theoretical results from Euclidean spaces to manifolds is not a direct extension. As mentioned in, e.g., [27, 188], such a transfer can be achieved by utilizing the chart–atlas definition of manifolds, namely, finding neural network approximation on each chart and then summing up estimates on each chart. The derived error bound, which is a sum of error bounds on each chart, relies on the size of the atlas, which depends on the dimension (cf. [188, Equation (34)]). Moreover, manifolds may have their intrinsic properties and the mapping to a union of Euclidean charts may not well maintain these properties. Thus the final specific question we aim to tackle in the future is the following.

Can we analyze the performance of deep learning and its applications to natural sciences with problems defined on manifolds without using the chart–atlas definition?

If we do not introduce the atlas size into the error bound, we shall make use of some developed tools on specific manifolds. For example, our spectral method in Chapter 5 is an intrinsic method on the sphere, and spherical approximation theory and neural networks with spherical inputs can be related.

Appendix A

The Rhythms of History

"I consider it immoral to discuss a topic without connecting to predecessors."

Lloyd N. Trefethen FRS, *Ten themes of how I do research*, 2006

Motivating and supporting my research presented in this thesis, a list of significant achievements in approximation theory, numerical analysis, and computational harmonic analysis is summarized in the following chronology, based on their publication dates.

1914 Gronwall examined the uniform norm of the L^2 orthogonal projection operator on $\mathbb{S}^2$ [103], 103

1929 Filon developed a method for the numerical quadrature of oscillatory integrals, which is now referred to as Filon quadrature [90], 38

1937 Marcinkiewicz and Zygmund established multiple inequalities linking the accurate integrals of trigonometrical polynomials and their quadrature formulae [143], 6, 20, 40, 66, 100

1939 Szegő published his masterpiece [210] of orthogonal polynomials, 51

1960 Clenshaw and Curtis investigated a quadrature formula in Chebyshev points and experimentally observed an unexpected accuracy comparable to Gauss quadrature [62], 27

1966 Müller published his classic monograph [151] on spherical harmonics, 69

1975 Glowinski and Marrocco proposed an algorithm, named ALG2, for the numerical solution of various problems from mechanics, physics, and differential geometry, which is now commonly known as the Alternating Direction Method of Multipliers (ADMM) [99], 150, 208

1977 Delsarte, Goethals, and Seidel introduced the concept of spherical t-designs [69], 31, 56, 72, 98

1978 Sloan and Smith proposed the product-integration method for integrand containing an absolutely integrable kernel [198], 38

1979 Allen and Cahn introduced an equation to describe the process of phase separation in iron alloys [5], 7, 94

1979 Bannai and Damerell showed tight spherical t-designs exist only for a few small values of t when t is even [17], 72

1980 Bannai and Damerell extended their 1979 results to odd t [18], 72

1984 Seymour and Zaslavsky pointed out that a spherical t-design always exists if there are sufficient points [187], 72

1988 Sloan proposed the qualocation method, aiming to achieve the theoretical benefits of the Galerkin method at a computational cost comparable to the collocation method [195], 111

1991 Du and Nicolaides investigated fully discrete schemes for the Cahn–Hilliard equation that preserve the energy law at the discrete level [78], 95

1992 Rudin, Osher, and Fatemi introduced the total variation (TV) model for image denoising [182], 161

1993 Elliott and Stuart proposed the convex splitting method for gradient flows, getting rid of the expense of solving nonconvex problems [80], 95

1995 Sloan invented hyperinterpolation [196], 5, 98

1997 Tao and An proposed the DCA for solving difference-of-convex (DC) optimization problems [212], 147, 206

2001 Mhaskar, Narcowich, and Ward investigated the Marcinkiewicz–Zygmund inequality on the sphere [146], 6, 20, 40, 66, 100

2006 The field of compressed sensing was initialized by Candès, Romberg and Tao [41] and Donoho [74] in the same year, 131

2007 Brauchart and Hesse analyzed the convergence rate of integrating $f \in H^s(\mathbb{S}^d)$ using spherical t-designs for all $s > d/2$ and $d \geq 2$ [34], 73

2008 Trefethen argued by entering the complex plane that for most functions, the Clenshaw–Curtis and Gauss quadrature rules have comparable accuracy [220], 29

2011 Filbir and Mhaskar investigated the Marcinkiewicz–Zygmund inequality on compact manifolds [89], 6, 20, 40, 66, 100

2012 He and Yuan obtained the sublinear convergence rate of the Douglas–Rachford splitting method (equivalent to the ADMM) for convex programs [107], 150

2013 Bondarenko, Radchenko and Viazovska resolved the long-standing open problem that for each $m \geq ct^d$ with some positive but unknown constant $c > 0$, there exists a spherical t-design of m points on $\mathbb{S}^d$ [28], 72

2013 Needell and Ward established the compressed imaging theory for the TV minimization model [157], 11

2014 Brauchart, Saff, Sloan, and Womersley introduced the concept of QMC designs [35], 73

2022 Trefethen commented on the exactness of quadrature formulae, aligning with the trend of interest in the numerical analysis community that the quadrature exactness should be re-accessed [222], 16

Intentionally blank page.

Bibliography

[1] R. A. Adams, *Sobolev Spaces*, Pure and Applied Mathematics, Vol. 65, Academic Press, New York, 1975. → page 72

[2] B. Adcock, N. Dexter, and Q. Xu, *Improved recovery guarantees and sampling strategies for TV minimization in compressive imaging*, SIAM J. Imaging Sci., 14 (2021), pp. 1149–1183, https://doi.org/10.1137/20M136788X. → pages 171 and 214

[3] B. Adcock, A. C. Hansen, C. Poon, and B. Roman, *Breaking the coherence barrier: a new theory for compressed sensing*, Forum Math. Sigma, 5 (2017), pp. Paper No. e4, 84, https://doi.org/10.1017/fms.2016.32. → page 171

[4] C. Ahrens and G. Beylkin, *Rotationally invariant quadratures for the sphere*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 465 (2009), pp. 3103–3125, https://doi.org/10.1098/rspa.2009.0104. → page 37

[5] S. M. Allen and J. W. Cahn, *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metallurgica, 27 (1979), pp. 1085–1095, https://doi.org/10.1016/0001-6160(79)90196-2. → pages 7, 94, and 220

[6] L. Alvarez and L. Mazorra, *Signal and image restoration using shock filters and anisotropic diffusion*, SIAM J. Numer. Anal., 31 (1994), pp. 590–605, https://doi.org/10.1137/0731032. → page 165

[7] C. An, X. Chen, I. H. Sloan, and R. S. Womersley, *Well conditioned spherical designs for integration and interpolation on the two-sphere*, SIAM J. Numer. Anal., 48 (2010), pp. 2135–2157, https://doi.org/10.1137/100795140. → pages 31, 32, 57, 85, and 98

[8] C. An, X. Chen, I. H. Sloan, and R. S. Womersley, *Regularized least squares approximations on the sphere using spherical designs*, SIAM J. Numer. Anal., 50 (2012), pp. 1513–1534, https://doi.org/10.1137/110838601. → pages 6 and 31

[9] C. An and H.-N. Wu, *Lasso hyperinterpolation over general regions*, SIAM J. Sci. Comput., 43 (2021), pp. A3967–A3991, https://doi.org/10.1137/20M137793X, https://doi.org/10.1137/20M137793X. → pages 6, 22, and 37

[10] C. An and H.-N. Wu, *Tikhonov regularization for polynomial approximation problems in Gauss quadrature points*, Inverse Problems, 37 (2021), 015008 (19 pages), https://doi.org/10.1088/1361-6420/abcd44. → page 6

[11] C. An and H.-N. Wu, *Bypassing the quadrature exactness assumption of hyperinterpolation on the sphere*, arXiv preprint arXiv:2209.11012, (2022). → page 99

[12] C. An and H.-N. Wu, *On the quadrature exactness in hyperinterpolation*, BIT, 62 (2022), pp. 1899–1919, https://doi.org/10.1007/s10543-022-00935-x. → pages 41, 50, 61, 77, 97, and 99

[13] C. An, H.-N. Wu, and X. Yuan, *The springback penalty for robust signal recovery*, Appl. Comput. Harmon. Anal., 61 (2022), pp. 319–346, https://doi.org/10.1016/j.acha.2022.07.002. → page 207

[14] K. Atkinson and W. Han, *Theoretical Numerical Analysis. A Functional Analysis Framework*, vol. 39 of Texts in Applied Mathematics, Springer, Dordrecht, third ed., 2009, https://doi.org/10.1007/978-1-4419-0458-4. → page 72

[15] K. Atkinson and W. Han, *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*, vol. 2044 of Lecture Notes in Mathematics, Springer, Heidelberg, 2012, https://doi.org/10.1007/978-3-642-25983-8. → pages 36, 60, and 96

[16] I. Babuška, *Information-based numerical practice*, J. Complexity, 3 (1987), pp. 331–346, https://doi.org/10.1016/0885-064X(87)90019-7. → page 3

[17] E. Bannai and R. M. Damerell, *Tight spherical designs. I*, J. Math. Soc. Japan, 31 (1979), pp. 199–207, https://doi.org/10.2969/jmsj/03110199. → pages 72 and 220

[18] E. Bannai and R. M. Damerell, *Tight spherical designs. II*, J. London Math. Soc. (2), 21 (1980), pp. 13–30, https://doi.org/10.1112/jlms/s2-21.1.13. → pages 72 and 220

[19] J. W. Barrett, J. F. Blowey, and H. Garcke, *Finite element approximation of the Cahn–Hilliard equation with degenerate mobility*, SIAM

J. Numer. Anal., 37 (1999), pp. 286–318, https://doi.org/10.1137/S0036142997331669. → page 95

[20] A. Beck, *First-Order Methods in Optimization*, SIAM, Philadelphia; Mathematical Optimization Society, Philadelphia, 2017, https://doi.org/10.1137/1.9781611974997. → page 128

[21] M. Benning, C. Brune, M. Burger, and J. Müller, *Higher-order TV methods–enhancement via Bregman iteration*, J. Sci. Comput., 54 (2013), pp. 269–310, https://doi.org/10.1007/s10915-012-9650-3. → page 164

[22] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen, *The modern mathematics of deep learning*, in Mathematical Aspects of Deep Learning, Cambridge University Press, 2022, pp. 1–111, https://doi.org/10.1017/9781009025096.002. → pages 215 and 216

[23] A. L. Bertozzi, N. Ju, and H.-W. Lu, *A biharmonic-modified forward time stepping method for fourth order nonlinear diffusion equations*, Discrete Contin. Dyn. Syst., 29 (2010), pp. 1367–1391, https://doi.org/10.3934/dcds.2011.29.1367. → page 95

[24] P. Blomgren, T. F. Chan, P. Mulet, and C.-K. Wong, *Total variation image restoration: Numerical methods and extensions*, in Proceedings of International Conference on Image Processing, IEEE, 1997, pp. 384–387, https://doi.org/10.1109/ICIP.1997.632128. → page 165

[25] J. F. Blowey, M. Copetti, and C. M. Elliott, *Numerical analysis of a model for phase separation of a multicomponent alloy*, IMA J. Numer. Anal., 16 (1996), pp. 111–139, https://doi.org/10.1093/imanum/16.1.111. → page 95

[26] T. Blumensath and M. E. Davies, *Iterative hard thresholding for compressed sensing*, Applied and Computational Harmonic Analysis, 27 (2009), pp. 265–274, https://doi.org/10.1016/j.acha.2009.04.002. → pages 151 and 153

[27] H. Bolcskei, P. Grohs, G. Kutyniok, and P. Petersen, *Optimal approximation with sparsely connected deep neural networks*, SIAM J. Math. Data Sci., 1 (2019), pp. 8–45, https://doi.org/10.1137/18M118709X. → page 218

[28] A. Bondarenko, D. Radchenko, and M. Viazovska, *Optimal asymptotic bounds for spherical designs*, Ann. of Math. (2), 178 (2013), pp. 443–452, https://doi.org/10.4007/annals.2013.178.2.2. → pages 72, 98, and 220

[29] E. Bonnetier, E. Bretin, and A. Chambolle, *Consistency result for a non monotone scheme for anisotropic mean curvature flow*, Interfaces Free Bound., 14 (2012), pp. 1–35, https://doi.org/10.4171/ifb/272. → page 95

[30] J. S. Brauchart, *Explicit families of functions on the sphere with exactly known Sobolev space smoothness*, in Contemporary computational mathematics—a celebration of the 80th birthday of Ian Sloan. Vol. 1, 2, Springer, Cham, 2018, pp. 153–177, https://doi.org/10.1007/978-3-319-72456-0_9. → page 86

[31] J. S. Brauchart and J. Dick, *Quasi-Monte Carlo rules for numerical integration over the unit sphere* $\mathbb{S}^2$, Numer. Math., 121 (2012), pp. 473–502, https://doi.org/10.1007/s00211-011-0444-6. → page 73

[32] J. S. Brauchart, J. Dick, E. B. Saff, I. H. Sloan, Y. G. Wang, and R. S. Womersley, *Covering of spheres by spherical caps and worst-case error for equal weight cubature in Sobolev spaces*, J. Math. Anal. Appl., 431 (2015), pp. 782–811, https://doi.org/10.1016/j.jmaa.2015.05.079. → page 73

[33] J. S. Brauchart, P. J. Grabner, I. H. Sloan, and R. S. Womersley, *Needlets liberated*, arXiv preprint arXiv:2207.12838, (2022). → page 83

[34] J. S. Brauchart and K. Hesse, *Numerical integration over spheres of arbitrary dimension*, Constr. Approx., 25 (2007), pp. 41–71, https://doi.org/10.1007/s00365-006-0629-4. → pages 73 and 220

[35] J. S. Brauchart, E. B. Saff, I. H. Sloan, and R. S. Womersley, *QMC designs: optimal order quasi Monte Carlo integration schemes on the sphere*, Math. Comp., 83 (2014), pp. 2821–2851, https://doi.org/10.1090/S0025-5718-2014-02839-1. → pages 67, 73, 89, and 221

[36] K. Bredies, K. Kunisch, and T. Pock, *Total generalized variation*, SIAM J. Imaging Sci., 3 (2010), pp. 492–526, https://doi.org/10.1137/090769521. → page 165

[37] A. M. Bruckstein, D. L. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, SIAM Review, 51 (2009), pp. 34–81, https://doi.org/10.1137/060657704. → pages 8, 9, 124, and 126

[38] J.-F. Cai and W. Xu, *Guarantees of total variation minimization for signal recovery*, Inf. Inference, 4 (2015), pp. 328–353, https://doi.org/10.1093/imaiai/iav009. → page 170

[39] M. Caliari, S. De Marchi, and M. Vianello, *Hyperinterpolation on the square*, J. Comput. Appl. Math., 210 (2007), pp. 78–83, https://doi.org/10.1016/j.cam.2006.10.058. → pages 5, 49, and 212

[40] M. Caliari, S. De Marchi, and M. Vianello, *Hyperinterpolation in the cube*, Comput. Math. Appl., 55 (2008), pp. 2490–2497, https://doi.org/10.1016/j.camwa.2007.10.003. → pages 5, 49, and 212

[41] E. J. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509, https://doi.org/10.1109/TIT.2005.862083. → pages 9, 131, 162, 170, and 220

[42] E. J. Candès, J. K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics, 59 (2006), pp. 1207–1223, https://doi.org/10.1002/cpa.20124. → pages 9, 124, 132, and 139

[43] E. J. Candès, M. Rudelson, T. Tao, and R. Vershynin, *Error correction via linear programming*, in 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05), IEEE, 2005, pp. 668–681, https://doi.org/10.1109/SFCS.2005.5464411. → pages 8, 9, 124, 132, 136, 137, 138, 139, and 146

[44] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Transactions on Information Theory, 51 (2005), pp. 4203–4215, https://doi.org/10.1109/TIT.2005.858979. → pages 8, 10, 124, 131, 146, 169, and 178

[45] E. J. Candès and T. Tao, *Near-optimal signal recovery from random projections: Universal encoding strategies?*, IEEE Transactions on Information Theory, 52 (2006), pp. 5406–5425, https://doi.org/10.1109/TIT.2006.885507. → pages 146, 147, and 178

[46] A. Chambolle, *An algorithm for total variation minimization and applications*, J. Math. Imaging Vision, 20 (2004), pp. 89–97, https://doi.org/10.1023/B:JMIV.0000011320.81911.38. Special issue on mathematics and image analysis. → pages 162, 163, and 190

[47] A. Chambolle, *Total variation minimization and a class of binary MRF models*, in International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer, Berlin, Heidelberg, 2005, pp. 136–152, https://doi.org/10.1007/11585978_10. → pages 162, 163, and 166

[48] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, *An introduction to total variation for image analysis*, in Theoretical foundations and numerical methods for sparse recovery, vol. 9 of Radon Ser. Comput. Appl. Math., Walter de Gruyter, Berlin, 2010, pp. 263–340, https://doi.org/10.1515/9783110226157.263. → page 161

[49] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188, https://doi.org/10.1007/s002110050258. → page 165

[50] A. Chambolle and T. Pock, *An introduction to continuous optimization for imaging*, Acta Numer., 25 (2016), pp. 161–319, https://doi.org/10.1017/S096249291600009X. → page 161

[51] A. Chambolle and T. Pock, *Approximating the total variation with finite differences or finite elements*, in Handbook of Numerical Analysis, vol. 22, Elsevier, 2021, pp. 383–417, https://doi.org/10.1016/bs.hna.2020.10.005. → page 165

[52] A. Chambolle and T. Pock, *Learning consistent discretizations of the total variation*, SIAM J. Imaging Sci., 14 (2021), pp. 778–813, https://doi.org/10.1137/20M1377199. → page 165

[53] T. Chan, A. Marquina, and P. Mulet, *High-order total variation-based image restoration*, SIAM J. Sci. Comput., 22 (2000), pp. 503–516, https://doi.org/10.1137/S1064827598344169. → page 165

[54] R. Chartrand, *Exact reconstruction of sparse signals via nonconvex minimization*, IEEE Signal Processing Letters, 14 (2007), pp. 707–710, https://doi.org/10.1109/LSP.2007.898300. → pages 9, 125, 126, 132, 136, 137, 138, and 163

[55] R. Chartrand and V. Staneva, *Restricted isometry properties and nonconvex compressive sensing*, Inverse Problems, 24 (2008), p. 035020, https://doi.org/10.1088/0266-5611/24/3/035020. → pages 9, 125, and 126

[56] C. Chen, B. He, Y. Ye, and X. Yuan, *The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent*, Mathematical Programming, 155 (2016), pp. 57–79, https://doi.org/10.1007/s10107-014-0826-5. → page 150

[57] L. Q. Chen and J. Shen, *Applications of semi-implicit Fourier-spectral method to phase field equations*, Comput. Phys. Commun., 108 (1998), pp. 147–158, https://doi.org/10.1016/S0010-4655(97)00115-X. → pages 7 and 95

[58] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Review, 43 (2001), pp. 129–159, https://doi.org/10.1137/S003614450037906X. → pages 9, 124, and 126

[59] X. Chen and R. S. Womersley, *Existence of solutions to systems of underdetermined equations and spherical designs*, SIAM J. Numer. Anal., 44 (2006), pp. 2326–2341, https://doi.org/10.1137/050626636. → page 98

[60] A. Chernih, I. H. Sloan, and R. S. Womersley, *Wendland functions with increasing smoothness converge to a Gaussian*, Adv. Comput. Math., 40 (2014), pp. 185–200, https://doi.org/10.1007/s10444-013-9304-5. → pages 32 and 86

[61] W.-Z. Chien, *Applications of Green Functions and Variational Methods in Electromagnetic Field and Wave Computation*, Shanghai Scientific Press, Shanghai, 1989. (In Chinese). → page 36

[62] C. W. Clenshaw and A. R. Curtis, *A method for numerical integration on an automatic computer*, Numer. Math., 2 (1960), pp. 197–205, https://doi.org/10.1007/BF01386223. → pages 27 and 219

[63] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, vol. 93 of Applied Mathematical Sciences, Springer–Verlag, Cham, 2019, https://doi.org/10.1007/978-1-4614-4942-3. Fourth edition. → page 36

[64] J. H. Conway and R. K. Guy, *The Book of Numbers*, Copernicus, New York, 1996, https://doi.org/10.1007/978-1-4612-4072-3. → page 190

[65] F. Cucker and S. Smale, *On the mathematical foundations of learning*, Bull. Amer. Math. Soc. (N.S.), 39 (2002), pp. 1–49, https://doi.org/10.1090/S0273-0979-01-00923-5. → page 215

[66] F. Cucker and D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, vol. 24 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2007, https://doi.org/10.1017/CBO9780511618796. → page 215

[67] F. Dai, *On generalized hyperinterpolation on the sphere*, Proc. Amer. Math. Soc., 134 (2006), pp. 2931–2941, https://doi.org/10.1090/S0002-9939-06-08421-8. → pages 5 and 17

[68] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, Computer Science and Applied Mathematics, Academic Press, Inc., Orlando, FL, second ed., 1984, https://doi.org/10.1016/C2013-0-10566-1. → page 5

[69] P. Delsarte, J.-M. Goethals, and J. J. Seidel, *Spherical codes and designs*, Geometriae Dedicata, 6 (1977), pp. 363–388, https://doi.org/10.1007/bf03187604. → pages 29, 31, 56, 72, 98, and 219

[70] R. DeVore, S. Foucart, G. Petrova, and P. Wojtaszczyk, *Computing a quantity of interest from observational data*, Constr. Approx., 49 (2019), pp. 461–508, https://doi.org/10.1007/s00365-018-9433-7. → page 28

[71] R. DeVore, B. Hanin, and G. Petrova, *Neural network approximation*, Acta Numerica, 30 (2021), pp. 327–444, https://doi.org/10.1017/S0962492921000052. → page 216

[72] V. Domínguez, I. G. Graham, and T. Kim, *Filon–Clenshaw–Curtis rules for highly oscillatory integrals with algebraic singularities and stationary points*, SIAM J. Numer. Anal., 51 (2013), pp. 1542–1566, https://doi.org/10.1137/120884146. → page 52

[73] V. Domínguez, I. G. Graham, and V. P. Smyshlyaev, *Stability and error estimates for Filon–Clenshaw–Curtis rules for highly oscillatory integrals*, IMA J. Numer. Anal., 31 (2011), pp. 1253–1280, https://doi.org/10.1093/imanum/drq036. → page 52

[74] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306, https://doi.org/10.1109/TIT.2006.871582. → pages 9, 124, 131, 162, and 220

[75] T. A. Driscoll, N. Hale, and L. N. Trefethen, eds., *Chebfun Guide*, Pafnuty Publications, Oxford, 2014. → page 51

[76] Q. Du and X. Feng, *The phase field method for geometric moving interfaces and their numerical approximations*, in Geometric partial differential equations. Part I, vol. 21 of Handb. Numer. Anal., Elsevier, North-Holland, Amsterdam, 2020, pp. 425–508, https://doi.org/10.1007/s. → page 95

[77] Q. Du, R. Li, and L. Zhang, *Variational phase field formulations of polarization and phase transition in ferroelectric thin films*, SIAM J. Appl. Math., 80 (2020), pp. 1590–1606, https://doi.org/10.1137/19M1291431. → page 95

[78] Q. Du and R. A. Nicolaides, *Numerical analysis of a continuum model of phase transition*, SIAM J. Numer. Anal., 28 (1991), pp. 1310–1322, https://doi.org/10.1137/0728069. → pages 95 and 220

[79] W. E and B. Yu, *The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems*, 10.1007/s40304-018-0127-z, 6 (2018), pp. 1–12, https://doi.org/10.1007/s40304-018-0127-z. → page 217

[80] C. M. Elliott and A. M. Stuart, *The global dynamics of discrete semilinear parabolic equations*, SIAM J. Numer. Anal., 30 (1993), pp. 1622–1663, https://doi.org/10.1137/0730084. → pages 95 and 220

[81] S. Esedoḡlu and S. J. Osher, *Decomposition of images by the anisotropic Rudin-Osher-Fatemi model*, Comm. Pure Appl. Math., 57 (2004), pp. 1609–1626, https://doi.org/10.1002/cpa.20045. → pages 163, 166, and 205

[82] E. Esser, Y. Lou, and J. Xin, *A method for finding structured sparse solutions to nonnegative least squares problems with applications*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 2010–2046, https://doi.org/10.1137/13090540X. → pages 9, 125, and 127

[83] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association, 96 (2001), pp. 1348–1360, https://doi.org/10.1198/016214501753382273. → pages 9, 124, 125, and 163

[84] A. C. Fannjiang, T. Strohmer, and P. Yan, *Compressed remote sensing of sparse objects*, SIAM J. Imaging Sci., 3 (2010), pp. 595–618, https://doi.org/10.1137/090757034. → page 169

[85] X. Feng, R. Glowinski, and M. Neilan, *Recent developments in numerical methods for fully nonlinear second order partial differential equations*, SIAM Rev., 55 (2013), pp. 205–267, https://doi.org/10.1137/110825960. → page 95

[86] X. Feng and Y. Li, *Analysis of symmetric interior penalty discontinuous Galerkin methods for the Allen–Cahn equation and the mean curvature flow*, IMA J. Numer. Anal., 35 (2015), pp. 1622–1651, https://doi.org/10.1093/imanum/dru058. → page 95

[87] X. Feng and A. Prohl, *Numerical analysis of the Allen–Cahn equation and approximation for mean curvature flows*, Numer. Math., 94 (2003), pp. 33–65, https://doi.org/10.1007/s00211-002-0413-1. → pages 7 and 95

[88] X. Feng and A. Prohl, *Error analysis of a mixed finite element method for the Cahn–Hilliard equation*, Numer. Math., 99 (2004), pp. 47–84, https://doi.org/10.1007/s00211-004-0546-5. → page 95

[89] F. Filbir and H. N. Mhaskar, *Marcinkiewicz–Zygmund measures on manifolds*, J. Complexity, 27 (2011), pp. 568–596, https://doi.org/10.1016/j.jco.2011.03.002. → pages 6, 20, 41, 66, 100, 213, and 220

[90] L. N. G. Filon, *On a quadrature formula for trigonometric integrals*, Proc. R. Soc. Edinburgh, 49 (1929), pp. 38–47, https://doi.org/10.1017/S0370164600026262. → pages 38 and 219

[91] S. Foucart, *Hard thresholding pursuit: an algorithm for compressive sensing*, SIAM J. Numer. Anal., 49 (2011), pp. 2543–2563, https://doi.org/10.1137/100806278. → page 151

[92] S. Foucart and M.-J. Lai, *Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$*, Applied and Computational Harmonic Analysis, 26 (2009), pp. 395–407, https://doi.org/10.1016/j.acha.2008.09.001. → pages 132 and 163

[93] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Birkhäuser/Springer, New York, 2013, https://doi.org/10.1007/978-0-8176-4948-7. → page 132

[94] A. Galdran, J. Vazquez-Corral, D. Pardo, and M. Bertalmío, *Enhanced variational image dehazing*, SIAM J. Imaging Sci., 8 (2015), pp. 1519–1546, https://doi.org/10.1137/15M1008889. → page 165

[95] H.-Y. Gao and A. G. Bruce, *WaveShrink with firm shrinkage*, Statistica Sinica, (1997), pp. 855–874, https://www.jstor.org/stable/24306159. → page 129

[96] W. Gautschi, *How and how not to check Gaussian quadrature formulae*, BIT, 23 (1983), pp. 209–216, https://doi.org/10.1007/BF02218441. → page 27

[97] W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2004, https://doi.org/10.1093/oso/9780198506720.001.0001. → page 51

[98] G. Gilboa, N. Sochen, and Y. Y. Zeevi, *Forward-and-backward diffusion processes for adaptive image enhancement and denoising*, IEEE Trans. Image Process., 11 (2002), pp. 689–703, https://doi.org/10.1109/TIP.2002.

800883. → page 165

[99] R. Glowinski and A. Marrocco, *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires*, Revue Française d'Automatique, Informatique et Recherche Opérationnelle Série Rouge. Analyse Numérique, 9 (1975), pp. 41–76, https://doi.org/10.1051/m2an/197509R200411. → pages 150, 151, 208, and 219

[100] T. Goldstein and S. Osher, *The split Bregman method for L1-regularized problems*, SIAM J. Imaging Sci., 2 (2009), pp. 323–343, https://doi.org/10.1137/080725891. → pages 194, 206, and 209

[101] I. G. Graham and I. H. Sloan, *Fully discrete spectral boundary integral methods for Helmholtz problems on smooth closed surfaces in* $\mathbb{R}^3$, Numer. Math., 92 (2002), pp. 289–323, https://doi.org/10.1007/s002110100343. → pages 36 and 37

[102] I. G. Graham and I. H. Sloan, *Fully discrete spectral boundary integral methods for Helmholtz problems on smooth closed surfaces in* $\mathbb{R}^3$, Numer. Math., 92 (2002), pp. 289–323, https://doi.org/10.1007/s002110100343. → page 212

[103] T. H. Gronwall, *On the degree of convergence of Laplace's series*, Trans. Amer. Math. Soc., 15 (1914), pp. 1–30, https://doi.org/10.2307/1988688, https://doi.org/10.2307/1988688. → pages 103 and 219

[104] I. Gühring, M. Raslan, and G. Kutyniok, *Expressivity of deep neural networks*, in Mathematical Aspects of Deep Learning, Cambridge University Press, 2022, pp. 149–199, https://doi.org/10.1017/9781009025096.004. → page 215

[105] K. Guo, D. Han, and X. Yuan, *Convergence analysis of Douglas–Rachford splitting method for "strongly+weakly" convex programming*, SIAM Journal on Numerical Analysis, 55 (2017), pp. 1549–1577, https://doi.org/10.1137/16M1078604. → page 126

[106] O. Hansen, K. Atkinson, and D. Chien, *On the norm of the hyperinterpolation operator on the unit disc and its use for the solution of the nonlinear poisson equation*, IMA J. Numer. Anal., 29 (2009), pp. 257–283, https://doi.org/10.1093/imanum/drm052. → pages 5, 49, and 212

[107] B. He and X. Yuan, *On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method*, SIAM J. Numer. Anal., 50 (2012), pp. 700–709, https://doi.org/10.1137/110836936. → pages 150 and 220

[108] K. Hesse and I. H. Sloan, *Worst-case errors in a Sobolev space setting for cubature over the sphere S^2*, Bull. Austral. Math. Soc., 71 (2005), pp. 81–105, https://doi.org/10.1017/S0004972700038041. → page 73

[109] K. Hesse and I. H. Sloan, *Cubature over the sphere S^2 in Sobolev spaces of arbitrary order*, J. Approx. Theory, 141 (2006), pp. 118–133, https://doi.org/10.1016/j.jat.2006.01.004. → page 73

[110] K. Hesse and I. H. Sloan, *Hyperinterpolation on the sphere*, in Frontiers in Interpolation and Approximation, vol. 282 of Pure Appl. Math. (Boca Raton), Chapman & Hall/CRC, Boca Raton, 2007, pp. 213–248. → pages 5, 17, 37, 70, 77, 78, 79, and 103

[111] K. Hesse, I. H. Sloan, and R. S. Womersley, *Numerical integration on the sphere*, in Handbook of Geomathematics, Springer–Verlag Berlin Heidelberg, 2010, https://doi.org/10.1007/978-3-642-01546-5_40. → pages 73, 89, and 97

[112] K. Hesse, I. H. Sloan, and R. S. Womersley, *Radial basis function approximation of noisy scattered data on the sphere*, Numer. Math., 137 (2017), pp. 579–605, https://doi.org/10.1007/s00211-017-0886-6. → page 72

[113] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 2002, https://doi.org/10.1137/1.9780898718027. → page 1

[114] T. Hou, T. Tang, and J. Yang, *Numerical analysis of fully discretized Crank–Nicolson scheme for fractional-in-space Allen–Cahn equations*, J. Sci. Comput., 72 (2017), pp. 1214–1231, https://doi.org/10.1007/s10915-017-0396-9. → page 7

[115] J. Huang, Y. Jiao, Y. Liu, and X. Lu, *A constructive approach to L_0 penalized regression*, J. Mach. Learn. Res., 19 (2018), pp. 403–439, http://jmlr.org/papers/v19/17-194.html. → page 151

[116] M. Huang, M.-J. Lai, A. Varghese, and Z. Xu, *On DC based methods for phase retrieval*, in Approximation theory XVI, Springer, Cham, 2021, pp. 87–121, https://doi.org/10.1007/978-3-030-57464-2_6. → page 147

[117] L. Huo, W. Chen, H. Ge, and M. K. Ng, *Stable image reconstruction using transformed total variation minimization*, SIAM Journal on Imaging Sciences, 15 (2022), pp. 1104–1139, https://doi.org/10.1137/21M1438566. → page 166

[118] A. Iserles and S. P. Nørsett, *On quadrature methods for highly oscillatory integrals and their implementation*, BIT, 44 (2004), pp. 755–772, https://doi.org/10.1007/s10543-004-5243-3. → page 38

[119] Y. Jiao, B. Jin, and X. Lu, *A primal dual active set with continuation algorithm for the ℓ^0-regularized optimization problem*, Applied and Computational Harmonic Analysis, 39 (2015), pp. 400–426, https://doi.org/10.1016/j.acha.2014.10.001. → page 151

[120] Y. Jiao, B. Jin, and X. Lu, *Iterative soft/hard thresholding with homotopy continuation for sparse recovery*, IEEE Signal Process. Lett., 24 (2017), pp. 784–788, https://doi.org/10.1109/LSP.2017.2693406. → page 151

[121] F. Krahmer, C. Kruschel, and M. Sandbichler, *Total variation minimization in compressed sensing*, in Compressed Sensing and its Applications, Birkhäuser/Springer, Cham, 2017, pp. 333–358, https://doi.org/10.1007/978-3-319-69802-1_11. → page 171

[122] F. Krahmer and R. Ward, *New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property*, SIAM J. Math. Anal., 43 (2011), pp. 1269–1281, https://doi.org/10.1137/100810447. → page 178

[123] F. Krahmer and R. Ward, *Stable and robust sampling strategies for compressive imaging*, IEEE Trans. Image Process., 23 (2014), pp. 612–622, https://doi.org/10.1109/TIP.2013.2288004. → pages 14, 169, 170, 171, 174, 178, 180, 181, 191, 192, 193, and 214

[124] M.-J. Lai and J. Wang, *An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems*, SIAM Journal on Optimization, 21 (2011), pp. 82–101, https://doi.org/10.1137/090775397. → pages 9 and 125

[125] M.-J. Lai, Y. Xu, and W. Yin, *Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization*, SIAM Journal on Numerical Analysis, 51 (2013), pp. 927–957, https://doi.org/10.1137/110840364. → pages 151 and 160

[126] Q. T. Le Gia and H. N. Mhaskar, *Localized linear polynomial operators and quadrature formulas on the sphere*, SIAM J. Numer. Anal., 47 (2009), pp. 440–466, https://doi.org/10.1137/060678555. → pages 41, 79, and 101

[127] Q. T. Le Gia, F. J. Narcowich, J. D. Ward, and H. Wendland, *Continuous and discrete least-squares approximation by radial basis functions on spheres*, J. Approx. Theory, 143 (2006), pp. 124–133, https://doi.org/10.1016/j.jat.2006.03.007. → page 72

[128] Q. T. Le Gia and I. H. Sloan, *The uniform norm of hyperinterpolation on the unit sphere in an arbitrary number of dimensions*, Constr. Approx., 17 (2001), pp. 249–265, https://doi.org/10.1007/s003650010025. → pages 5, 37, and 49

[129] Q. T. Le Gia, I. H. Sloan, and H. Wendland, *Multiscale analysis in Sobolev spaces on the sphere*, SIAM J. Numer. Anal., 48 (2010), pp. 2065–2090, https://doi.org/10.1137/090774550. → page 86

[130] P. Leopardi, *Diameter bounds for equal area partitions of the unit sphere*, Electron. Trans. Numer. Anal., 35 (2009), pp. 1–16. → page 85

[131] D. Li, *Effective maximum principles for spectral methods*, Ann. Appl. Math., 37 (2021), pp. 131–290, https://doi.org/10.4208/aam.oa-2021-0003. → pages 8, 95, 96, 100, 108, and 109

[132] D. Li, *Why large time-stepping methods for the Cahn–Hilliard equation is stable*, Math. Comp., 91 (2022), pp. 2501–2515, https://doi.org/10.1090/mcom/3768. → page 95

[133] D. Li, Z. Qiao, and T. Tang, *Characterizing the stabilization size for semi-implicit Fourier-spectral method to phase field equations*, SIAM J. Numer. Anal., 54 (2016), pp. 1653–1681, https://doi.org/10.1137/140993193. → page 7

[134] D. Li and T. Tang, *Stability of the semi-implicit method for the Cahn–Hilliard equation with logarithmic potentials*, Ann. Appl. Math., 37 (2021), pp. 31–60, https://doi.org/10.4208/aam.OA-2020-0003. → page 106

[135] P. Li, W. Chen, H. Ge, and M. K.-P. Ng, $\ell_1 - \alpha\ell_2$ *minimization methods for signal and image reconstruction with impulsive noise removal*, Inverse Problems, 36 (2020), p. 055009, https://doi.org/10.1088/1361-6420/ab750c. → page 183

[136] H.-L. Liao, T. Tang, and T. Zhou, *On energy stable, maximum-principle preserving, second-order BDF scheme with variable steps for the Allen–Cahn equation*, SIAM J. Numer. Anal., 58 (2020), pp. 2294–2314, https://doi.org/10.1137/19M1289157. → page 7

[137] S.-B. Lin, Y. G. Wang, and D.-X. Zhou, *Distributed filtered hyperinterpolation for noisy data on the sphere*, SIAM J. Numer. Anal., 59 (2021), pp. 634–659, https://doi.org/10.1137/19M1281095. → pages 5, 37, and 49

[138] J.-L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications. Vol. I*, Die Grundlehren der mathematischen Wissenschaften, Band 181, Springer-Verlag, New York-Heidelberg, 1972, https://doi.org/10.1007/978-3-642-65161-8. Translated from the French by P. Kenneth. → page 72

[139] Y. Lou, T. Zeng, S. Osher, and J. Xin, *A weighted difference of anisotropic and isotropic total variation model for image processing*, SIAM J. Imaging Sci., 8 (2015), pp. 1798–1823, https://doi.org/10.1137/14098435X. → pages 166, 182, 183, 194, 195, 198, 199, and 209

[140] M. Lustig, D. Donoho, and J. M. Pauly, *Sparse MRI: The application of compressed sensing for rapid MR imaging*, Magn. Reson. Med., 58 (2007), pp. 1182–1195, https://doi.org/10.1002/mrm.21391. → page 169

[141] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, *Compressed sensing MRI*, IEEE Signal Process. Mag., 25 (2008), pp. 72–82, https://doi.org/10.1109/MSP.2007.914728. → page 169

[142] J. Lv and Y. Fan, *A unified approach to model selection and sparse recovery using regularized least squares*, The Annals of Statistics, 37 (2009), pp. 3498–3528, https://doi.org/10.1214/09-AOS683. → pages 9, 125, and 127

[143] J. Marcinkiewicz and A. Zygmund, *Sur les fonctions indépendantes*, Fund. Math., 29 (1937), pp. 60–90, http://eudml.org/doc/212925. → pages 6, 20, 41, 66, 100, and 219

[144] V. G. Maz'ya and T. O. Shaposhnikova, *Theory of Multipliers in Spaces of Differentiable Functions*, vol. 23 of Monographs and Studies in Mathematics, Pitman, Boston, 1985, https://doi.org/10.1070/RM1983v038n03ABEH003484. → page 72

[145] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, *Reconstruction and subgaussian operators in asymptotic geometric analysis*, Geometric and

Functional Analysis, 17 (2007), pp. 1248–1282, https://doi.org/10.1007/s00039-007-0618-7. → pages 146 and 178

[146] H. N. Mhaskar, F. J. Narcowich, and J. D. Ward, *Spherical Marcinkiewicz–Zygmund inequalities and positive quadrature*, Math. Comp., 70 (2001), pp. 1113–1130, https://doi.org/10.1090/S0025-5718-00-01240-0. → pages 6, 20, 29, 41, 66, 100, and 220

[147] J. S. Moll, *The anisotropic total variation flow*, Mathematische Annalen, 332 (2005), pp. 177–218, https://doi.org/10.1007/s00208-004-0624-0. → page 205

[148] T. Möllenhoff, E. Strekalovskiy, M. Moeller, and D. Cremers, *The primal-dual hybrid gradient method for semiconvex splittings*, SIAM Journal on Imaging Sciences, 8 (2015), pp. 827–857, https://doi.org/10.1137/140976601. → pages 126 and 164

[149] G. Montúfar and Y. G. Wang, *Distributed learning via filtered hyperinterpolation on manifolds*, Found. Comput. Math., (2021), pp. 1–53, https://doi.org/10.1007/s10208-021-09529-5. → pages 37 and 49

[150] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bulletin de la Société Mathématique de France, 93 (1965), pp. 273–299, https://doi.org/10.24033/bsmf.1625. → page 128

[151] C. Müller, *Spherical Harmonics*, vol. 17 of Lecture Notes in Mathematics, Springer-Verlag, Berlin-New York, 1966, https://doi.org/10.1007/BFb0094775. → pages 69, 96, 101, and 219

[152] F. Narcowich, P. Petrushev, and J. Ward, *Decomposition of Besov and Triebel-Lizorkin spaces on the sphere*, J. Funct. Anal., 238 (2006), pp. 530–564, https://doi.org/10.1016/j.jfa.2006.02.011. → page 83

[153] F. J. Narcowich, P. Petrushev, and J. D. Ward, *Localized tight frames on spheres*, SIAM J. Math. Anal., 38 (2006), pp. 574–594, https://doi.org/10.1137/040614359. → page 83

[154] F. J. Narcowich and J. D. Ward, *Scattered data interpolation on spheres: error estimates and locally supported basis functions*, SIAM J. Math. Anal., 33 (2002), pp. 1393–1410, https://doi.org/10.1137/S0036141001395054. → page 86

[155] D. Needell and J. A. Tropp, *CoSaMP: iterative signal recovery from incomplete and inaccurate samples*, Appl. Comput. Harmon. Anal., 26 (2009), pp. 301–321, https://doi.org/10.1016/j.acha.2008.07.002. → page 151

[156] D. Needell and R. Ward, *Near-optimal compressed sensing guarantees for total variation minimization*, IEEE Trans. Image Process., 22 (2013), pp. 3941–3949, https://doi.org/10.1109/TIP.2013.2264681. → pages 170, 171, 187, and 214

[157] D. Needell and R. Ward, *Stable image reconstruction using total variation minimization*, SIAM J. Imaging Sci., 6 (2013), pp. 1035–1058, https://doi.org/10.1137/120868281. → pages 11, 162, 163, 164, 169, 170, 171, 172, 173, 176, 177, 180, 181, 187, 189, 213, and 220

[158] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro, *Exploring generalization in deep learning*, in Advances in Neural Information Processing Systems, vol. 30, 2017, https://proceedings.neurips.cc/paper/2017/file/10ce03a1ed01077e3e289f3e53c72813-Paper.pdf. → page 215

[159] M. Nikolova, *Energy minimization methods*, in Handbook of Mathematical Methods in Imaging, Springer, New York, 2015, pp. 157–204, https://doi.org/10.1007/978-0-387-92920-0_5. → page 163

[160] S. Osher and L. I. Rudin, *Feature-oriented image enhancement using shock filters*, SIAM J. on Numer. Anal., 27 (1990), pp. 919–940, https://doi.org/10.1137/0727053. → page 165

[161] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, *Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition*, in Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, IEEE, 1993, pp. 40–44, https://doi.org/10.1109/ACSSC.1993.342465. → page 151

[162] F. Pierre, J.-F. Aujol, A. Bugeau, G. Steidl, and V.-T. Ta, *Variational contrast enhancement of gray-scale and RGB images*, J. Math. Imaging Vision, 57 (2017), pp. 99–116, https://doi.org/10.1007/s10851-016-0670-8. → page 165

[163] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao, *Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review*, Int. J. Autom. Comput., 14 (2017), pp. 503–519, https://doi.org/10.1007/s11633-017-1054-2. → page 215

[164] C. Poon, *On the role of total variation in compressed sensing*, SIAM J. Imaging Sci., 8 (2015), pp. 682–720, https://doi.org/10.1137/140978569. → pages 171 and 214

[165] M. J. D. Powell, *Direct search algorithms for optimization calculations*, Acta Numerica, (1998), pp. 287–336, https://doi.org/10.1017/S0962492900002841. → page 3

[166] M. J. D. Powell, *UOBYQA: unconstrained optimization by quadratic approximation*, Mathematical Programming, 92 (2002), pp. 555–582, https://doi.org/10.1007/s101070100290. → page 3

[167] M. J. D. Powell, *On the use of quadratic models in unconstrained minimization without derivatives*, Optimization Methods and Software, 19 (2004), pp. 399–411, https://doi.org/10.1080/10556780410001661450. → page 3

[168] M. J. D. Powell, *The NEWUOA software for unconstrained optimization without derivatives*, in Large-scale Nonlinear Optimization, Springer, 2006, pp. 255–297, https://doi.org/10.1007/0-387-30065-1_16. → page 3

[169] M. J. D. Powell, *A view of algorithms for optimization without derivatives*, Tech. Report DAMTP2007/NA03, University of Cambridge, 2007, https://www.damtp.cam.ac.uk/user/na/NA_papers/NA2007_03. → page 3

[170] M. J. D. Powell, *Developments of NEWUOA for minimization without derivatives*, IMA Journal of Numerical Analysis, 28 (2008), pp. 649–664, https://doi.org/10.1093/imanum/drm047. → page 3

[171] D. L. Ragozin, *Constructive polynomial approximation on spheres and projective spaces*, Trans. Amer. Math. Soc., 162 (1971), pp. 157–170, https://doi.org/10.2307/1995746. → page 105

[172] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations*, J. Comput. Phys., 378 (2019), pp. 686–707, https://doi.org/10.1016/j.jcp.2018.10.045. → page 217

[173] E. A. Rakhmanov, E. B. Saff, and Y. M. Zhou, *Minimal discrete energy on the sphere*, Math. Res. Lett., 1 (1994), pp. 647–662, https://doi.org/10.4310/MRL.1994.v1.n6.a3. → page 85

[174] H. Rauhut, J. Romberg, and J. A. Tropp, *Restricted isometries for partial random circulant matrices*, Appl. Comput. Harmon. Anal., 32 (2012), pp. 242–254, https://doi.org/10.1016/j.acha.2011.05.001. → page 178

[175] H. Rauhut and R. Ward, *Sparse Legendre expansions via ℓ_1-minimization*, J. Approx. Theory, 164 (2012), pp. 517–533, https://doi.org/10.1016/j.jat.2012.01.008. → pages 178 and 191

[176] M. Reimer, *Constructive Theory of Multivariate Functions: With An Application to Tomography*, Bibliographisches Institut, Mannheim, 1990. → pages 70 and 102

[177] M. Reimer, *Hyperinterpolation on the sphere at the minimal projection order*, J. Approx. Theory, 104 (2000), pp. 272–286, https://doi.org/10.1006/jath.2000.3454. → pages 5 and 37

[178] M. Reimer, *Generalized hyperinterpolation on the sphere and the Newma–Shapiro operators*, Constr. Approx., 18 (2002), pp. 183–204, https://doi.org/10.1007/s00365-001-0008-6. → pages 5 and 17

[179] R. J. Renka, *Multivariate interpolation of large sets of scattered data*, ACM Trans. Math. Software, 14 (1988), pp. 139–148, https://doi.org/10.1145/45054.45055. → page 86

[180] R. T. Rockafellar, *Convex Analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, 1970, https://doi.org/10.1515/9781400873173. → pages 9 and 124

[181] M. Rudelson and R. Vershynin, *On sparse reconstruction from Fourier and Gaussian measurements*, Communications on Pure and Applied Mathematics, 61 (2008), pp. 1025–1045, https://doi.org/10.1002/cpa.20227. → pages 146, 178, and 182

[182] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268, https://doi.org/10.1016/0167-2789(92)90242-F. Experimental mathematics: computational issues in nonlinear science (Los Alamos, NM, 1991). → pages 161, 166, 205, and 220

[183] R. Saab, R. Chartrand, and O. Yilmaz, *Stable sparse approximations via nonconvex optimization*, in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2008, pp. 3885–3888, https://doi.org/10.1109/ICASSP.2008.4518502. → page 139

[184] C.-B. Schönlieb and A. Bertozzi, *Unconditionally stable schemes for higher order inpainting*, Commun. Math. Sci., 9 (2011), pp. 413–457, https://doi.org/10.4310/CMS.2011.v9.n2.a4. → page 95

[185] S. Setzer and G. Steidl, *Variational methods with higher-order derivatives in image processing*, in Approximation theory XII: San Antonio 2007, Mod. Methods Math., Nashboro Press, Brentwood, TN, 2008, pp. 360–385. → page 165

[186] S. Setzer, G. Steidl, and T. Teuber, *Infimal convolution regularizations with discrete ℓ_1-type functionals*, Communications in Mathematical Sciences, 9 (2011), pp. 797–827, https://doi.org/10.4310/CMS.2011.v9.n3.a7. → page 165

[187] P. D. Seymour and T. Zaslavsky, *Averaging sets: a generalization of mean values and spherical designs*, Adv. in Math., 52 (1984), pp. 213–240, https://doi.org/10.1016/0001-8708(84)90022-7. → pages 72 and 220

[188] U. Shaham, A. Cloninger, and R. R. Coifman, *Provable approximation properties for deep neural networks*, Appl. Comput. Harmon. Anal., 44 (2018), pp. 537–557, https://doi.org/10.1016/j.acha.2016.04.003. → page 218

[189] L. F. Shampine, *Vectorized adaptive quadrature in Matlab*, J. Comput. Appl. Math., 211 (2008), pp. 131–140, https://doi.org/10.1016/j.cam.2006.11.021. → page 55

[190] J. Shen, T. Tang, and L.-L. Wang, *Spectral Methods: Algorithms, Analysis and Applications*, vol. 41 of Springer Series in Computational Mathematics, Springer, Heidelberg, 2011, https://doi.org/10.1007/978-3-540-71041-7. → page 2

[191] J. Shen, T. Tang, and J. Yang, *On the maximum principle preserving schemes for the generalized Allen–Cahn equation*, Commun. Math. Sci., 14 (2016), pp. 1517–1534, https://doi.org/10.4310/CMS.2016.v14.n6.a3. → page 7

[192] J. Shen, C. Wang, X. Wang, and S. M. Wise, *Second-order convex splitting schemes for gradient flows with Ehrlich–Schwoebel type energy: application to thin film epitaxy*, SIAM J. Numer. Anal., 50 (2012), pp. 105–125, https://doi.org/10.1137/110822839. → pages 7 and 95

[193] J. Shen and X. Yang, *Numerical approximations of Allen–Cahn and Cahn–Hilliard equations*, Discrete Contin. Dyn. Syst, 28 (2010), pp. 1669–1691, https://doi.org/10.3934/dcds.2010.28.1669. → pages 7 and 95

[194] J. Sirignano and K. Spiliopoulos, *DGM: A deep learning algorithm for solving partial differential equations*, J. Comput. Phys., 375 (2018), pp. 1339–1364, https://doi.org/10.1016/j.jcp.2018.08.029. → page 217

[195] I. H. Sloan, *A quadrature-based approach to improving the collocation method*, Numer. Math., 54 (1988), pp. 41–56, https://doi.org/10.1007/BF01403890. → pages 111 and 220

[196] I. H. Sloan, *Polynomial interpolation and hyperinterpolation over general regions*, J. Approx. Theory, 83 (1995), pp. 238–254, https://doi.org/10.1006/jath.1995.1119. → pages 5, 16, 18, 20, 21, 29, 32, 38, 40, 50, 61, 63, 67, 77, 98, 99, 110, 211, and 220

[197] I. H. Sloan, *Interpolation and hyperinterpolation on the sphere*, in Multivariate Approximation, vol. 101 of Mathematical Research, Akademie Verlag, Berlin, 1997, pp. 255–268. → page 37

[198] I. H. Sloan and W. E. Smith, *Product-integration with the Clenshaw–Curtis and related points. Convergence properties*, Numer. Math., 30 (1978), pp. 415–428, https://doi.org/10.1007/BF01398509. → pages 38 and 219

[199] I. H. Sloan and W. E. Smith, *Product integration with the Clenshaw–Curtis points: implementation and error estimates*, Numer. Math., 34 (1980), pp. 387–401, https://doi.org/10.1007/BF01403676. → page 38

[200] I. H. Sloan and W. E. Smith, *Properties of interpolatory product integration rules*, SIAM J. Numer. Anal., 19 (1982), pp. 427–442, https://doi.org/10.1137/0719027. → page 38

[201] I. H. Sloan and W. L. Wendland, *A quadrature-based approach to improving the collocation method for splines of even degree*, Z. Anal. Anwendungen, 8 (1989), pp. 361–376, https://doi.org/10.4171/ZAA/359. → page 111

[202] I. H. Sloan and R. S. Womersley, *The uniform error of hyperinterpolation on the sphere*, in Advances in Multivariate Approximation, vol. 107 of Mathematical Research, Wiley-VCH, Berlin, 1999, pp. 289–306. → pages 5, 29, 37, and 49

[203] I. H. Sloan and R. S. Womersley, *Constructive polynomial approximation on the sphere*, J. Approx. Theory, 103 (2000), pp. 91–118, https://doi.org/10.1006/jath.1999.3426. → pages 37, 49, 102, and 103

[204] I. H. Sloan and R. S. Womersley, *Extremal systems of points and numerical integration on the sphere*, Adv. Comput. Math., 21 (2004), pp. 107–125, https://doi.org/10.1023/B:ACOM.0000016428.25905.da. → page 85

[205] I. H. Sloan and R. S. Womersley, *Filtered hyperinterpolation: a constructive polynomial approximation on the sphere*, GEM Int. J. Geomath., 3 (2012), pp. 95–117, https://doi.org/10.1007/s13137-011-0029-7. → pages 6, 22, 37, and 49

[206] A. Sommariva and M. Vianello, *Numerical hyperinterpolation over spherical triangles*, Math. Comput. Simulation, 190 (2021), pp. 15–22, https://doi.org/10.1016/j.matcom.2021.05.003. → pages 5 and 212

[207] H. Song and C.-W. Shu, *Unconditional energy stability analysis of a second order implicit-explicit local discontinuous Galerkin method for the Cahn–Hilliard equation*, J. Sci. Comput., 73 (2017), pp. 1178–1203, https://doi.org/10.1007/s10915-017-0497-5. → page 95

[208] D. Strong and T. Chan, *Edge-preserving and scale-dependent properties of total variation regularization*, Inverse Problems, 19 (2003), pp. S165–S187, https://doi.org/10.1088/0266-5611/19/6/059. Special section on imaging. → page 164

[209] Y. Sun, H. Chen, and J. Tao, *Sparse signal recovery via minimax-concave penalty and ℓ_1-norm loss function*, IET Signal Processing, 12 (2018), pp. 1091–1098, https://doi.org/10.1049/iet-spr.2018.5130. → page 151

[210] G. Szegő, *Orthogonal Polynomials*, American Mathematical Society Colloquium Publications, Vol. 23, American Mathematical Society, New York, 1939, https://doi.org/10.1090/coll/023. → pages 51 and 219

[211] T. Tang and J. Yang, *Implicit-explicit scheme for the Allen–Cahn equation preserves the maximum principle*, J. Comput. Math., 34 (2016), pp. 471–481, https://doi.org/10.4208/jcm.1603-m2014-0017. → page 95

[212] P. D. Tao and L. T. H. An, *Convex analysis approach to DC programming: theory, algorithms and applications*, Acta Mathematica Vietnamica, 22 (1997), pp. 289–355. → pages 147, 149, 206, 207, and 220

[213] P. D. Tao and L. T. H. An, *A DC optimization algorithm for solving the trust-region subproblem*, SIAM Journal on Optimization, 8 (1998), pp. 476–505, https://doi.org/10.1137/S1052623494274313. → pages 147, 148, 149, 206, and 207

[214] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, John Wiley & Sons; Washington, D.C, 1977. Translated from the Russian, Preface by translation editor Fritz John. → page 165

[215] R. H. Todd, D. K. Allen, and L. Alting, *Manufacturing Processes Reference Guide*, Industrial Press, Inc., New York, 1994. → page 130

[216] A. Townsend, H. Wilber, and G. B. Wright, *Computing with functions in spherical and polar geometries i. the sphere*, SIAM J. Sci. Comput., 38 (2016), pp. C403–C425, https://doi.org/10.1137/15M1045855. → page 58

[217] J. F. Traub, *Information-based complexity*, in Encyclopedia of Computer Science, 2003, pp. 850–854. → page 2

[218] L. N. Trefethen, *The definition of numerical analysis*. SIAM News, November 1992. → page 1

[219] L. N. Trefethen, *Spectral Methods in MATLAB*, vol. 10 of Software, Environments, and Tools, SIAM, Philadelphia, 2000, https://doi.org/10.1137/1.9780898719598. → page 2

[220] L. N. Trefethen, *Is Gauss quadrature better than Clenshaw–Curtis?*, SIAM Rev., 50 (2008), pp. 67–87, https://doi.org/10.1137/060659831. → pages 29 and 220

[221] L. N. Trefethen, *Approximation Theory and Approximation Practice, Extended Edition*, SIAM, Philadelphia, 2019, https://doi.org/10.1137/1.9781611975949. → pages 2 and 51

[222] L. N. Trefethen, *Exactness of quadrature formulas*, SIAM Rev., 64 (2022), pp. 132–150, https://doi.org/10.1137/20M1389522. → pages 16, 27, 67, 97, and 221

[223] L. N. Trefethen and D. Bau, III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997, https://doi.org/10.1137/1.9780898719574. → pages 1 and 149

[224] H. Wang, K. Wang, and X. Wang, *On the norm of the hyperinterpolation operator on the d-dimensional cube*, Comput. Math. Appl., 68 (2014), pp. 632–638, https://doi.org/10.1016/j.camwa.2014.07.009. → pages 5, 49, and 212

[225] Y. G. Wang, Q. T. Le Gia, I. H. Sloan, and R. S. Womersley, *Fully discrete needlet approximation on the sphere*, Appl. Comput. Harmon. Anal.,

43 (2017), pp. 292–316, https://doi.org/10.1016/j.acha.2016.01.003. → pages 83 and 86

[226] M. Welk, G. Gilboa, and J. Weickert, *Theoretical foundations for discrete forward-and-backward diffusion filtering*, in International Conference on Scale Space and Variational Methods in Computer Vision, Springer, 2009, pp. 527–538, https://doi.org/10.1007/978-3-642-02256-2_44. → page 165

[227] M. Welk, G. Steidl, and J. Weickert, *Locally analytic schemes: a link between diffusion filtering and wavelet shrinkage*, Appl. Comput. Harmon. Anal., 24 (2008), pp. 195–224, https://doi.org/10.1016/j.acha.2007.05.004. → page 165

[228] M. Welk, D. Theis, T. Brox, and J. Weickert, *PDE-based deconvolution with forward-backward diffusivities and diffusion tensors*, in International Conference on Scale-Space Theories in Computer Vision, Springer, 2005, pp. 585–597, https://doi.org/10.1007/11408031_50. → page 165

[229] M. Welk, J. Weickert, and I. Galić, *Theoretical foundations for spatially discrete 1-D shock filtering*, Image Vis. Comput., 25 (2007), pp. 455–463, https://doi.org/10.1016/j.imavis.2006.06.001. → page 165

[230] M. Welk, J. Weickert, and G. Gilboa, *A discrete theory and efficient algorithms for forward-and-backward diffusion filtering*, J. Math. Imaging Vision, 60 (2018), pp. 1399–1426, https://doi.org/10.1007/s10851-018-0847-4. → page 165

[231] H. Wendland, *Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree*, Adv. Comput. Math., 4 (1995), pp. 389–396, https://doi.org/10.1007/BF02123482. → pages 32 and 86

[232] H. Wendland, *Scattered Data Approximation*, vol. 17 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2005, https://doi.org/10.1017/CBO9780511617539. → pages 3 and 4

[233] R. S. Womersley, *Efficient spherical designs with good geometric properties*, in Contemporary computational mathematics—a celebration of the 80th birthday of Ian Sloan. Vol. 1, 2, Springer, Cham, 2018, pp. 1243–1285, https://doi.org/10.1007/978-3-319-72456-0_57. → page 98

[234] R. S. Womersley and I. H. Sloan, *How good can polynomial interpolation on the sphere be?*, Adv. Comput. Math., 14 (2001), pp. 195–226, https://doi.org/10.1023/A:1016630227163. → pages 4, 5, and 37

[235] J. Woodworth and R. Chartrand, *Compressed sensing recovery via nonconvex shrinkage penalties*, Inverse Problems, 32 (2016), p. 075004, https://doi.org/10.1088/0266-5611/32/7/075004. → page 132

[236] H. Woźniakowski, *A survey of information-based complexity*, J. Complexity, 1 (1985), pp. 11–44, https://doi.org/10.1016/0885-064X(85)90020-2, https://doi.org/10.1016/0885-064X(85)90020-2. → page 2

[237] S. Xiang, Y. J. Cho, H. Wang, and H. Brunner, *Clenshaw–Curtis-Filon–type methods for highly oscillatory Bessel transforms and applications*, IMA J. Numer. Anal., 31 (2011), pp. 1281–1314, https://doi.org/10.1093/imanum/drq035. → page 52

[238] L. Yan, Y. Shin, and D. Xiu, *Sparse approximation using $\ell_1 - \ell_2$ minimization and its application to stochastic collocation*, SIAM Journal on Scientific Computing, 39 (2017), pp. A229–A254, https://doi.org/10.1137/15M103947X. → pages 9, 125, 136, 137, 138, 139, 147, and 163

[239] P. Yin, Y. Lou, Q. He, and J. Xin, *Minimization of ℓ_{1-2} for compressed sensing*, SIAM Journal on Scientific Computing, 37 (2015), pp. A536–A563, https://doi.org/10.1137/140952363. → pages 9, 125, 127, 136, 137, 138, 147, 151, 155, 160, and 163

[240] C.-H. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, Ann. Statist., 38 (2010), pp. 894–942, https://doi.org/10.1214/09-AOS729. → pages 9, 125, 127, 129, and 167

[241] J. Zhang and Q. Du, *Numerical studies of discrete approximations to the Allen–Cahn equation in the sharp interface limit*, SIAM J. Sci. Comput., 31 (2009), pp. 3042–3063, https://doi.org/10.1137/080738398. → pages 7 and 95

[242] S. Zhang and J. Xin, *Minimization of transformed L_1 penalty: closed form representation and iterative thresholding algorithms*, Commun. Math. Sci., 15 (2017), pp. 511–537, https://doi.org/10.4310/CMS.2017.v15.n2.a9. → pages 128, 129, and 163

[243] S. Zhang and J. Xin, *Minimization of transformed L_1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing*, Math. Program., 169 (2018), pp. 307–336, https://doi.org/10.1007/

s10107-018-1236-x. → pages 9, 125, 127, 137, 138, 139, 147, 151, 153, 155, 160, 163, and 166

[244] T. Zhang, *Analysis of multi-stage convex relaxation for sparse regularization*, J. Mach. Learn. Res., 11 (2010), pp. 1081–1107, http://jmlr.org/papers/v11/zhang10a.html. → pages 9 and 125

[245] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B Stat. Methodol., 67 (2005), pp. 301–320, https://doi.org/10.1111/j.1467-9868.2005.00503.x. → page 126

Index