Technische Universität München
Fakultät für Mathematik

Max-Planck-Institut für Plasmaphysik

Generalized anisotropic Hermite functions and their applications

Anna Yurova

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende :	Prof. Dr. Barbara Wohlmuth
Prüfer der Dissertation :	1. Prof. Dr. Caroline Lasser
	2. Prof. Dr. Eric Sonnendrücker
	3. Prof. Dr. Elisabeth Larsson

Die Dissertation wurde am 15.10.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 30.01.2020 angenommen.

Abstract

In this dissertation a new basis that resembles Hermite functions but introduces an anisotropy in the Gaussian part of Hermite functions is proposed together with a comprehensive theoretical framework. This basis is then used in the derivation of two new numerical methods: Firstly, a new stabilization method for the interpolation with Gaussian Radial Basis Functions that naturally extends to the case of anisotropic Gaussians is developed. Secondly, a generalized version of the Fourier–Hermite method for the Vlasov equation is introduced and analyzed.

Zusammenfassung

In dieser Dissertation wird eine neue Basis eingeführt, die ähnlich aufgebaut ist wie Hermite-Funktionen, im Gaußchen Teil jedoch eine Anisotropie hinzufügt, und es werden die dazugehörigen mathematischen Grundlagen entwickelt. Basierend auf dieser Basis werden zwei neue numerische Methoden erarbeitet: Zum einen wird eine neue Stabilisierungsmethode für die Interpolation mit Gaußschen radialen Basisfunktionen hergeleitet, die sich auf den Fall anisotroper Gaußfunktionen übertragen lässt. Zum anderen wird eine verallgemeinerte Version der Fourier-Hermite-Methode für die Vlasov-Gleichung vorgestellt und analysiert.

Contents

1. Introduction

> *“I believe that the numbers and functions of analysis are not the arbitrary product of our minds; I believe that they exist outside of us with the same character of necessity as the objects of objective reality; and we find or discover them as do the physicists, chemists and zoologists.”*
>
> Charles Hermite (1822 – 1901)
> *Correspondance d’Hermite et de Stieltjes*, vol. 2, p. 398, Gauthier-Villars, 1905.
> Cited by Morris Kline in *Mathematical Thought From Ancient to Modern Times*, vol. 3, p. 1035, Oxford University Press, 1990.

It is hardly possible to give a more accurate description of the process of finding new types of functions than the one above by Charles Hermite. What might seem arbitrary at first sight is often a part of a bigger picture that is yet to be uncovered. That is also how the work in this thesis came to be. The standard Hermite functions are usually defined in 1D as

$$\psi_\ell(x) = \frac{\pi^{-1/4}}{\sqrt{2^\ell \ell!}} h_\ell(x) \mathrm{e}^{-x^2/2}, \quad \ell \geq 0,\ x \in \mathbb{R},$$

where h_ℓ are physicists’ Hermite polynomials. In this thesis, we discovered a new type of functions that is a generalization of Hermite functions. Instead of always having the ratio of $1/\sqrt{2}$ between the arguments of the Hermite polynomial and the Gaussian, we decouple them completely. Moreover, in multiple dimensions we allow for anisotropic Gaussians $\exp(-\mathbf{x}^T E^T E \mathbf{x})$ with an arbitrary invertible shape matrix E. This opens the door to more flexibility of adapting the basis for a specific problem. In case there is underlying anisotropy in the problem itself or in the domain on which it is defined, this can be incorporated in this spectral basis. This construction was inspired by the work of Hagedorn who introduced a similar structure in the context of quantum dynamics.

Of course, one must pay a certain price for this flexibility. One of the strengths of the standard Hermite functions is that they are defined on the whole real line and are orthonormal and dense in the Hilbert space of square-integrable functions $L^2(\mathbb{R}^d)$. This makes it attractive for spectral numerical methods in a number of physical applications, including, but not limited to, quantum dynamics and computational plasma physics. Our generalization, while providing more freedom leads to the loss of the orthogonality in $L^2(\mathbb{R}^d)$. However, once we discover the corresponding functional space, everything falls back into its place and the orthogonality is preserved at last. We go on to investigate algebraic and approximation properties of the new basis within that space and then use the new basis for two applications.

First, we make use of the generalized anisotropic Hermite functions in the context of the interpolation problem. One of the common choices of basis functions for multivariate interpolation is Gaussian radial basis functions (or Gaussian RBFs) which are a set of Gaussians of the same width centered at a set of points, or *centers of RBFs*, in the interpolation domain. The interpolation with Gaussian RBFs generalizes to higher dimensions in a simple way while yielding spectral accuracy and for this reason, is of great interest in the interpolation community. Even though Gaussian RBFs are widely used for multivariate interpolation, it is well known that when Gaussians become increasingly flat, the interpolation matrix becomes ill-conditioned. A family of methods, so-called RBF-QR methods, have been developed to tackle this issue. However, the existing methods of this type are mostly handling the interpolation with isotropic Gaussians or very simple anisotropic setups with a diagonal matrix E. However, for a number of applications, such as continental-size ice sheet simulations or statistical data fitting, where there exists underlying anisotropy, it could be advantageous to use the interpolation with anisotropic Gaussians. With the help of the new generalized anisotropic Hermite basis, we derive a new RBF-QR method that allows to include the anisotropic Gaussians in the stabilization framework.

We also deploy the generalized anisotropic Hermite functions for the spectral discretization of the Vlasov equation, which is an advective equation that is used in plasma physics for the description of the time evolution of the distribution function of the plasma. The full Vlasov model is six-dimensional (or often denoted as 3d3v in the literature) with three dimensions in space and three dimensions in velocity. In this work, we consider the simplified 1d1v model, however, we expect, that a similar approach should be possible for the 3d3v case. Standard Hermite functions have been used before for the velocity discretization of the Vlasov equation. Moreover, a modification of Hermite functions with the exponential part e^{-x^2} instead of $\mathrm{e}^{-x^2/2}$ has been also considered before and proved to bring certain advantages. With the help of generalized (anisotropic) Hermite functions, we derive a generalization of these methods, which allows to consider the general form basis functions with an arbitrary width of the Gaussian. We also develop a theoretical framework where we obtain analytic formulas for the observables and their evolution over time. In the 1d1v case, we do not make use of the anisotropy of the basis, since in 1d the shape matrix E is just a number. However, in certain physical scenarios which are modeled by the full model, there is an intrinsic anisotropy due to the strong magnetic field. Therefore, in the future, when going to higher dimensions, the anisotropy feature of the new basis could be potentially beneficial.

1.1. Main results

The main contribution of this thesis is threefold. First of all, a new generalized version of Hermite functions that introduces anisotropy in the Gaussian part is proposed together with the corresponding theoretical framework. In particular, the ladder operators are introduced for the new basis and a three-term recurrence for the efficient evaluation of the basis is derived. The new theory yields an estimate of the approximation error in Theorem 3.5.1.

Second, a novel numerical method for the stable interpolation with isotropic Gaussians has been derived based on the exponential generating function of the classic Hermite polynomials. Moreover, with the help of the anisotropic generating function, the method was extended to the case of the interpolation with anisotropic Gaussians. This is the first algorithm within the family of stabilization methods that allows for anisotropic Gaussians with an arbitrary invertible shape matrix. The theoretical description of the algorithm is generic for an arbitrary number of dimensions d which provides a convenient framework for working with problems of different dimensionality. Another notable feature of the new method is the truncation estimate derived in Theorem 10.2.1 that—contrary to the existing ones—does not only account for the diagonal contributions of the stabilization but measures the impact of the full stabilization basis.

Finally, the third major contribution is the novel *generalized Fourier–Hermite* method (14.12) for the Vlasov equation, which includes the existing Fourier–Hermite methods considered in the literature as special cases. This generalization allows to find an intermediate setup, which proved to be a good compromise between the existing methods for some cases. Moreover, the method description includes analytic formulas (15.5), (15.7), (15.9) for the evolution of the observables that are frequently used for the code verification, i.e. mass, momentum, and energy. This allows to identify the cases where the corresponding conservation properties are fulfilled or otherwise to monitor the magnitude of the error.

1.2. Outline of the thesis

This thesis is organized in three parts and contains in total 18 chapters. The first part introduces the *generalized anisotropic Hermite functions* and provides a theoretical framework for the new basis. Parts two and three are dedicated to the applications of the new basis and have their own introductory remarks which connect the new developments to the relevant results in the application-related literature. Below we provide a more detailed structure of the parts.

Part 1 contains two chapters. Chapter 2 summarizes the background theory on Hermite functions. This chapter is largely based on the introduction to Hermite polynomials by Folland [Fol09, § 6.4]. The *generalized anisotropic Hermite functions* are introduced in chapter 3, along with the corresponding functional space. The development of the theoretical framework for the new basis follows the style of the monograph by Lubich [Lub08, § III.1.1] where a similar framework for Hermite functions is presented in the context of quantum dynamics.

Part 2 of this thesis is an extended version of the paper *"Stable interpolation with isotropic and anisotropic Gaussians using Hermite generating function"* by Kormann, Lasser, & Yurova that has been accepted to the SIAM Journal on Scientific Computing on 18.09.2019 [KLY19] (preprint available at `https://arxiv.org/abs/1905.09542`). In this part of the thesis, a new numerical method for the stabilization of the interpolation with isotropic and anisotropic Gaussians is introduced. It consists of 9 chapters and starts with the introduction of the interpolation problem and provides a review of the relevant literature in chapter 4. The existing stabilization methods are briefly summarized in chapter 5. In chapter 6, an expansion of the anisotropic Gaussians via generalized anisotropic Hermite functions is derived, which is the foundation for the new stabilization method to be introduced in the next chapters. For that reason, in the framework of this application, the generalized anisotropic Hermite basis is called *HermiteGF basis* and the corresponding expansion is called *HermiteGF expansion*. The convergence properties of the HermiteGF basis in the corresponding space is discussed in chapter 7. The stabilization method itself is derived in chapter 8 and the analysis of its different parts is provided in chapter 9. A novel cut-off criterion for the HermiteGF expansion that allows to determine the number of basis functions needed is derived in chapter 10. Finally, chapter 11 provides some details on the implementation and chapter 12 illustrates the new stabilization method with numerical results.

In part 3 we use the generalized (anisotropic) Hermite functions for the discretization of the 1d1v Vlasov equation. This part consists of 5 chapters. We start by introducing the continuous Vlasov–Poisson model and its properties in chapter 13. The new spectral method for the Vlasov equation, based on our basis, is derived in chapter 14. Analytic formulas for the computation of the observables for the Vlasov equation and their evolution are derived in chapter 15. We come back to the world of standard Hermite functions in chapter 16, where we derive the error estimate of the Hermite functions solution of a simple 1d advection equation. This estimate allows us to illustrate the structural difference between the methods based on standard Hermite functions and the ones based on the generalized version. We briefly review the existing methods based on the Hermite discretization as special cases of our general method in chapter 17. We conclude by presenting the relevant numerical results in chapter 18.

Part I

Generalized anisotropic Hermite functions

2. Hermite functions framework

In this chapter, we introduce Hermite polynomials and Hermite functions and review their properties that we will need later on for constructing the generalized anisotropic Hermite basis. To introduce Hermite functions, we first reproduce the standard definition of Hermite polynomials through the Rodrigues formula and their main properties, largely recapturing [Fol09, § 6.4]. We then discuss the definition of Hermite functions based on the Hermite polynomials as well as an alternative approach via ladder operators that is common in quantum dynamics. This approach provides additional machinery for further analysis of the basis. It turns out that the same technique could be propagated to the case of the generalized anisotropic Hermite functions that are the main focus of this thesis. This framework allows to derive the approximation error for the new basis which is an essential theoretical result for evaluating the quality of the new basis.

2.1. Functional spaces

Before defining Hermite functions, let us introduce the functional spaces we will be working with. In the context of Hermite functions, the Hilbert space $L^2(\mathbb{R}^d)$ of square-integrable functions with the scalar product

$$\langle f, g\rangle = \int_{\mathbb{R}^d} f(\mathbf{x})\overline{g(\mathbf{x})}\mathrm{d}\mathbf{x} \quad \forall\ f, g \in L^2(\mathbb{R}^d)$$

is usually considered. The norm, corresponding to the above-defined scalar product is defined by $\|f\|^2 = \langle f, f\rangle$. Even though some properties of the Hermite functions can be easily derived in this space, for others it is often convenient to consider a subset of the $L^2(\mathbb{R}^d)$ space of rapidly decaying smooth functions, or the *Schwartz space*.

Definition 2.1.1: Schwartz space

The Schwartz space of rapidly decaying smooth functions is defined as

$$\mathcal{S}(\mathbb{R}^d) = \left\{ f \in C^\infty(\mathbb{R}^d) \mid \sup_{\mathbf{x}\in\mathbb{R}^d} \left|\mathbf{x}^{\boldsymbol{\alpha}} D^{\boldsymbol{\beta}} f(\mathbf{x})\right| < \infty \quad \forall \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{Z}^d_+ \right\}.$$

Let us review a few basic properties of the Schwartz space which will be useful later on.

Lemma 2.1.1: Properties of the Schwartz space

The following properties hold for the space $\mathcal{S}(\mathbb{R}^d)$.

1. $\forall f \in \mathcal{S}(\mathbb{R}^d)$

$$x_i f(\mathbf{x}) \in \mathcal{S}(\mathbb{R}^d), \quad \partial_{x_i} f(\mathbf{x}) \in \mathcal{S}(\mathbb{R}^d) \quad \forall i = 1 \ldots d. \tag{2.1}$$

2. $\forall f_1, f_2 \in \mathcal{S}(\mathbb{R}^d)$

$$f_1 + f_2 \in \mathcal{S}(\mathbb{R}^d). \tag{2.2}$$

3. $\mathcal{S}(\mathbb{R}^d)$ is dense in $L^2(\mathbb{R}^d)$.

Proof. 1. see [Rud91, Theorem 7.4, (b)].

2. see [Rud91, Theorem 7.4, (b)].

3. see [Won99, p. 14-16]. □

The Schwartz space includes the set of smooth functions with compact supports $C_0^\infty(\mathbb{R}^d)$ as well as functions of the type $P(\mathbf{x})\mathrm{e}^{-|\mathbf{x}|^2}$, where $P(\mathbf{x})$ is a polynomial.

A useful generalization of the Hilbert space $L^2(\mathbb{R}^d)$ can be obtained by replacing the element $\mathrm{d}\mathbf{x}$ of the linear measure by a weighted alternative $\omega(\mathbf{x})\mathrm{d}\mathbf{x}$, with the assumption that ω is continuous and $\omega(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathbb{R}^d$ (see, for example, [Fol09, § 3.4]).

Definition 2.1.2: Weighted L^2 space $L^2_\omega(\mathbb{R}^d)$

Consider a continuous function $\omega : \mathbb{R}^d \to \mathbb{R}_+$ for all $\mathbf{x} \in \mathbb{R}^d$. The *weighted L^2 space on $\mathbb{R}^d$, or $L^2_\omega(\mathbb{R}^d)$*, is a set of all functions on $\mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} |f(\mathbf{x})|^2 \omega(\mathbf{x}) \mathrm{d}\mathbf{x} < \infty.$$

The corresponding inner product and the norm are then defined by

$$\langle f, g \rangle_\omega = \int_{\mathbb{R}^d} f(\mathbf{x})\overline{g(\mathbf{x})}\omega(\mathbf{x})\mathrm{d}\mathbf{x}, \quad \|f\|_\omega = \sqrt{\langle f, f \rangle_\omega}.$$

Now, when the required spaces are set up, we can move on to the definition of Hermite polynomials.

2.2. Hermite polynomials (1D)

We follow the classical way to define Hermite polynomials, through the *Rodrigues formula*. We first consider one-dimensional physicists' Hermite polynomials. Multivariate Hermite polynomials can be obtained as tensor-products of one-dimensional Hermite polynomials.

Definition 2.2.1: Hermite polynomials

For $\ell \in \mathbb{N}_0$, $x \in \mathbb{R}$ the ℓ-th *Hermite polynomial* can be defined as

$$h_\ell(x) = (-1)^\ell \mathrm{e}^{x^2} \frac{\partial^\ell}{\partial x^\ell} \mathrm{e}^{-x^2}. \tag{2.3}$$

This relation is also referred to as the *Rodrigues formula*.

The most attractive property of Hermite polynomials, which makes them widely used in a variety of applications, is their orthogonality on the whole real line. However, they are not orthogonal in the standard space $L^2(\mathbb{R})$. Instead, the orthogonality relations hold on the weighted L^2 space $L^2_\omega(\mathbb{R})$ with the weight function

$$\omega(x) = \mathrm{e}^{-x^2}.$$

In this section, the symbol ω is reserved for this particular weight. However, in the following parts of the thesis, for example in chapter 3, we will be using a more general version of the weight function.

Theorem 2.2.1: Orthogonality of Hermite polynomials

Hermite polynomials $\{h_\ell\}_{\ell\in\mathbb{N}_0}$ are orthogonal in $L^2_\omega(\mathbb{R})$ with the weight $\omega(x) = \mathrm{e}^{-x^2}$. In particular, for all $\ell_1, \ell_2 \in \mathbb{N}_0$

$$\langle h_{\ell_1}, h_{\ell_2}\rangle_\omega = 0 \quad \text{if} \quad \ell_1 \neq \ell_2.$$

Otherwise

$$\langle h_\ell, h_\ell\rangle_\omega = \|h_\ell\|^2_\omega = \sqrt{\pi}2^\ell \ell! \quad \text{for all} \quad \ell \in \mathbb{N}_0. \tag{2.4}$$

Proof. We briefly recap the proof from [Fol09, Theorem 6.11]. Using the Rodrigues formula (2.3) and integrating by parts ℓ_2 times, we get

$$\begin{aligned}\langle h_{\ell_1}(x), h_{\ell_2}(x)\rangle_\omega &= \int_{\mathbb{R}} h_{\ell_1}(x)h_{\ell_2}(x)\mathrm{e}^{-x^2}\mathrm{d}x = (-1)^{\ell_2}\int_{\mathbb{R}} h_{\ell_1}(x)\frac{\partial^{\ell_2}}{\partial x^{\ell_2}}\mathrm{e}^{-x^2}\mathrm{d}x \\ &= \int_{\mathbb{R}} \frac{\partial^{\ell_2} h_{\ell_1}(x)}{\partial x^{\ell_2}}\mathrm{e}^{-x^2}\mathrm{d}x,\end{aligned}$$

where we used the fact that any polynomial multiplied by a Gaussian e^{-x^2} vanishes at $\pm\infty$.

If $\ell_1 < \ell_2$, then $\frac{\partial^{\ell_2} h_{\ell_1}(x)}{\partial x^{\ell_2}} = 0$, therefore, in this case

$$\langle h_{\ell_1}, h_{\ell_2}\rangle_\omega = 0.$$

Due to the symmetricity of the scalar product in $\langle\cdot,\cdot\rangle_\omega$, we can extend this result to all $\ell_1 \neq \ell_2$.

We are left to consider the norm of the Hermite polynomials. One can see from the Rodrigues formula (2.3) that with every derivation of e^{-x^2} we have a factor $2x$. Therefore, the leading factor of the polynomial h_ℓ is $(2x)^\ell$. Hence,

$$\langle h_\ell, h_\ell\rangle_\omega = \int_{\mathbb{R}} 2^\ell \ell! \mathrm{e}^{-x^2}\mathrm{d}x = \sqrt{\pi}2^\ell \ell! \qquad \square$$

Even though we have proven that Hermite polynomials are orthogonal, we cannot yet state that they form an orthogonal basis in $L^2_\omega(\mathbb{R})$. For that, we first have to make sure that the set $\{h_\ell\}_{\ell\in N}$ is complete in $L^2_\omega(\mathbb{R})$.

Theorem 2.2.2: Completeness of Hermite polynomials in $L^2_\omega(\mathbb{R})$

The set $\{h_\ell\}_{\ell\in\mathbb{N}_0}$ is complete in $L^2_\omega(\mathbb{R})$. Therefore, Hermite polynomials form an orthogonal basis in $L^2_\omega(\mathbb{R})$.

Proof. There are multiple proofs available in the literature for this statement. See, for example, [Fol09, § 6.4, Theorem 6.12]. □

Now, when we made sure that Hermite polynomials indeed form a basis, we shall look at its other interesting properties that we will use later on.

2.2.1. Hermite generating function and other properties

Based on the Rodrigues formula, we can immediately derive a key property of Hermite polynomials which we will also heavily use in part II of this thesis. This is a fundamental relation and is sometimes used for the definition of Hermite polynomials (see, for example, [Rai71, $\S 11$]).

Theorem 2.2.3: Hermite generating function (1D)

The following relation holds for all $a \in \mathbb{R}$, $b \in \mathbb{C}$:

$$\mathrm{e}^{2ba-a^2} = \sum_{\ell\in\mathbb{N}_0} \frac{a^\ell}{\ell!} h_\ell(b). \tag{2.5}$$

Proof. We summarize the standard proof that can be found, for example, in [Fol09, § 6.4, Theorem 6.13]. Using the Rodrigues formula and the Taylor expansion of an exponential function around 0, we get

$$\begin{aligned}\mathrm{e}^{2ba-a^2} &= \mathrm{e}^{b^2}\mathrm{e}^{(a-b)^2} = \mathrm{e}^{b^2}\sum_{\ell\in\mathbb{N}_0}\frac{\partial^\ell \mathrm{e}^{-(b-a)^2}}{\partial a^\ell}\bigg|_{a=0}\frac{a^\ell}{\ell!} = \mathrm{e}^{b^2}\sum_{\ell\in\mathbb{N}_0}(-1)^\ell\frac{\partial^\ell \mathrm{e}^{-u^2}}{\partial u^\ell}\bigg|_{u=b}\frac{a^\ell}{\ell!}\\ &= \sum_{\ell\in\mathbb{N}_0}\frac{a^\ell}{\ell!}h_\ell(b).\end{aligned}$$

□

The expression above allows us to easily derive two algebraic properties of Hermite polynomials which we will need later on.

Theorem 2.2.4: Properties of Hermite polynomials

For all $x \in \mathbb{R}$, $\ell \in \mathbb{N}_0$ the following properties hold for Hermite polynomials:

1. $\frac{\partial h_\ell(x)}{\partial x} = 2\ell h_{\ell-1}(x) \quad \forall \ell \in \mathbb{N}_0.$ (2.6)

2. $h_{\ell+1}(x) = 2xh_\ell(x) - 2\ell h_{\ell-1}(x).$ (2.7)

Proof. The relation (2.6) follows from differentiating the generating function expression (2.5) with respect to b on both sides and equating the coefficients of a^ℓ.

The second relation (2.7) follows from the Rodrigues formula, in combination with (2.6). Indeed,

$$\frac{\partial h_\ell(x)}{\partial x} = 2x(-1)^\ell \mathrm{e}^{x^2} \frac{\partial^\ell}{\partial x^\ell} \mathrm{e}^{-x^2} + (-1)^\ell \mathrm{e}^{x^2} \frac{\partial^{\ell+1}}{\partial x^{\ell+1}} \mathrm{e}^{-x^2} = 2x h_\ell(x) - h_{\ell+1}(x).$$

Recalling that $\frac{\partial h_\ell(x)}{\partial x} = 2\ell h_{\ell-1}(x)$, we get the expression (2.7). □

The recursion formula (2.7) is usually referred to as the *three-term recurrence*. It is often the method of choice for the computation of Hermite polynomials in numerical applications, since it does not involve expensive computations, such as computing derivatives, and allows to reuse already computed values of the polynomials with a lower degree.

2.2.2. Mehler's formula

Along with the exponential generating function that we considered in Theorem 2.2.3, another generating function identity, that includes the products of Hermite polynomials, is also an important tool in the framework of work with Hermite polynomials. Together with the resulting sum, it is referred to as the *Mehler's formula*.

Theorem 2.2.5: Mehler's formula. 1D.

For every $x, y \in \mathbb{R}$ and $|t| < 1$ the following relation holds:

$$\sum_{\ell \in \mathbb{N}_0} \frac{h_\ell(x) h_\ell(y)}{2^\ell \ell!} t^\ell = (1 - t^2)^{-1/2} \exp\left(\frac{2xyt - (x^2 + y^2)t^2}{1 - t^2}\right), \qquad (2.8)$$

where h_ℓ are physicists' Hermite polynomials.

Proof. See, for example, [Wat33, p. 5]. □

The expression above can be used for a variety of purposes. For example, one can use it for an alternative proof for the value of the norm (2.4) of Hermite polynomials (see [Tha93, § 1.1, Lemma 1.1.2]). With $x = y$, it can also be used for computing the sum of the squares of scaled Hermite polynomials which might be useful for numerical purposes. Once the analytic value is computed for the infinite sum, one can see numerically how many scaled Hermite polynomials bring a significant input.

2.3. Hermite functions in terms of ladder operators (1D)

As we can see from the previous section, it is advantageous to consider Hermite polynomials in the framework of a weighted L^2 space $L^2_\omega(\mathbb{R})$ with the Gaussian weight. However, it turns out that we can still enjoy the corresponding properties in the non-weighted $L^2(\mathbb{R})$ space by considering *Hermite functions* instead of Hermite polynomials.

Definition 2.3.1: Hermite functions

For $\ell \in \mathbb{N}_0$, the following functions are called the ℓ-th *Hermite functions*

$$\psi_\ell(x) = \frac{\pi^{-1/4}}{\sqrt{2^\ell \ell!}} h_\ell(x) \mathrm{e}^{-x^2/2}, \tag{2.9}$$

where $h_\ell(x)$ is ℓ-th physicists' Hermite polynomial.

It follows immediately from the properties of Hermite polynomials, that Hermite functions form an orthonormal basis in $L^2(\mathbb{R})$ where the normalization was introduced into the basis based on the value of the norm of Hermite polynomials in the weighted space and the weight was incorporated into the functions. It is clear from the definition that Hermite functions also belong to the Schwartz space $\mathcal{S}(\mathbb{R})$.

As for algebraic properties and the recurrence relation, multiplying the three-term recurrence (2.7) of Hermite polynomials by $\mathrm{e}^{-x^2/2}$ and scaling accordingly, we get the three-term recurrence for Hermite functions

$$\psi_{\ell+1}(x) = \sqrt{\frac{2}{\ell+1}} x\psi_\ell(x) - \sqrt{\frac{\ell}{\ell+1}} \psi_{\ell-1}. \tag{2.10}$$

This provides us with the tool of efficient evaluation of Hermite functions based on the values with smaller indices.

Due to the fact that Hermite functions are complete and orthonormal in $L^2(\mathbb{R})$, every function f in $L^2(\mathbb{R})$ can be represented as

$$f = \sum_{\ell \in \mathbb{N}_0} \langle f, \psi_\ell \rangle \psi_\ell.$$

in the sense of the *convergence in norm*

$$\left\| f - \sum_{\ell \geq M} \langle f, \psi_\ell \rangle \psi_\ell \right\| \to 0 \quad \text{when} \quad M \to \infty.$$

Even though we know that the difference above tends to zero when $M \to \infty$, in practice, we would prefer to have an explicit estimate of the corresponding norm. We consider a subspace

$$U_M = \operatorname{span}\{\psi_\ell | \ell \leq M-1\}.$$

and take a look at the orthogonal projector

$$P_M f = \sum_{\ell<M} \langle f, \psi_\ell \rangle \psi_\ell.$$

Our goal is to estimate the approximation error $\|f - P_M f\|$. For that, we look at the Hermite functions from a different perspective that is widely used in the quantum dynamics community (see [Lub08, § III.1.1]). In particular, one can also equivalently define Hermite functions based on Dirac's *ladder operators*. This representation provides a simple way of estimating the approximation error $\|f - P_M f\|$. Let us briefly recap the definition of the ladder operators and their relevance for the error estimation properties. We follow the line of narration of [Lub08, § III.1.1].

Definition 2.3.2: Dirac's ladder operators

Dirac's ladder operators for Hermite functions are defined by

$$Af = \frac{1}{\sqrt{2}}\left(x + \frac{\partial}{\partial x}\right) f, \tag{2.11}$$

$$A^\dagger f = \frac{1}{\sqrt{2}}\left(x - \frac{\partial}{\partial x}\right) f, \tag{2.12}$$

where A and $A^\dagger$ are, the *lowering* and *raising* operators, respectively.

Let us check that the operators A, $A^\dagger$ indeed act as lowering and raising operators on Hermite functions, i.e. they lower or raise the order of the Hermite functions respectively. Using the expression for the derivative of Hermite polynomials (2.6), we get

$$A\psi_\ell(x) = \frac{\pi^{-1/4}}{\sqrt{2^{\ell+1}\ell!}}\left(xh_\ell(x)\mathrm{e}^{-x^2/2} + 2\ell h_{\ell-1}(x)\mathrm{e}^{-x^2/2} - xh_\ell(x)\mathrm{e}^{-x^2/2}\right) = \sqrt{\ell}\psi_{\ell-1}.$$

Similarly, making use of the same formula and the three-term recurrence (2.7), we obtain

$$\begin{aligned} A^\dagger\psi_\ell(x) &= \frac{\pi^{-1/4}}{\sqrt{2^{\ell+1}\ell!}}\left(xh_\ell(x)\mathrm{e}^{-x^2/2} - 2\ell h_{\ell-1}(x)\mathrm{e}^{-x^2/2} + xh_\ell(x)\mathrm{e}^{-x^2/2}\right) \\ &= \frac{\pi^{-1/4}}{\sqrt{2^{\ell+1}\ell!}}\left(2xh_\ell(x) - 2\ell h_{\ell-1}(x)\right)\mathrm{e}^{-x^2/2} = \sqrt{\ell+1}\psi_{\ell+1}(x). \end{aligned} \tag{2.13}$$

Therefore, Hermite functions could be also defined recursively as

$$\psi_{\ell+1} = \frac{1}{\sqrt{\ell+1}} A^\dagger \psi_\ell$$

with the base of recursion

$$\psi_0(x) = \pi^{-1/4}\mathrm{e}^{-x^2/2}.$$

Before deriving the approximation error estimate, let us take a look at a few useful properties of the ladder operators.

Lemma 2.3.1: Properties of ladder operators

Hermite ladder operators have the following properties

1. Hermite ladder operators map Schwartz functions to Schwartz functions.

2. Hermite ladder operators are formally adjoint in the Schwartz space $\mathcal{S}(\mathbb{R})$. i.e. for all $f, g \in \mathcal{S}(\mathbb{R})$
$$\langle A^\dagger f, g\rangle = \langle f, Ag\rangle.$$

Proof. According to the properties (2.1), (2.2) of the Schwartz space, for every $f \in \mathcal{S}(\mathbb{R})$, its derivative and $xf(x)$, as well as their sum, also belongs to the Schwartz space. Therefore, $A^\dagger f, Af \in \mathcal{S}(\mathbb{R})$ which is the first statement.

Let us now prove the second statement. Considering the explicit form of the scalar product and using the integration by parts, we get

$$\begin{aligned}\langle A^\dagger f, g\rangle &= \frac{1}{\sqrt{2}}\int_{\mathbb{R}} xf(x)\overline{g(x)} - \frac{\partial f(x)}{\partial x}\overline{g(x)}\mathrm{d}x = \frac{1}{\sqrt{2}}\int_{\mathbb{R}} xf(x)\overline{g(x)} + \frac{\partial \overline{g(x)}}{\partial x}f(x)\mathrm{d}x\\ &= \langle f, Ag\rangle,\end{aligned}$$

where we used that for $f, g \in \mathcal{S}(\mathbb{R})$ the product $f(x)g(x)$ also rapidly decays. □

With these properties at hand, we are ready to proceed to the approximation error estimation. This is a result from the Theorem 1.2 from the book [Lub08, § III.1.1].

Theorem 2.3.1: Approximation error

The following estimate holds for every $f \in S(\mathbb{R})$, $s \leq M$,
$$\|f - P_M f\| \leq \frac{1}{\sqrt{M(M-1)\dots(M-s+1)}}\|A^s f\|. \tag{2.14}$$

Proof. We reproduce the proof from [Lub08, § III.1.1], since this approach will be relevant for the analogous estimate in the next chapter.

Using the fact that the raising operator acts on Hermite functions as (2.13) and the adjointness of the raising and lowering operators, we get

$$\begin{aligned} f - P_M f &= \sum_{\ell \geq M}\langle f, \psi_\ell\rangle\psi_\ell\\ &= \sum_{\ell \geq M}\frac{1}{\sqrt{\ell(\ell-1)\dots(\ell-s+1)}}\langle f, (A^\dagger)^s\psi_{\ell-s}\rangle\psi_\ell\\ &= \sum_{\ell \geq M}\frac{1}{\sqrt{\ell(\ell-1)\dots(\ell-s+1)}}\langle A^s f, \psi_{\ell-s}\rangle\psi_\ell\end{aligned}$$

Therefore, due to the orthonormality, we get

$$\begin{aligned}\|f - P_M f\| &\leq \frac{1}{M(M-1)\dots(M-s+1)} \sum_{j\in\mathbb{N}_0} |\langle A^s f, \psi_j\rangle|^2 \\ &= \frac{1}{M(M-1)\dots(M-s+1)} \|A^s f\|^2. \end{aligned} \qquad \square$$

We now have everything we need set up in the one-dimensional case and can move on to the multivariate case.

2.4. Multivariate Hermite polynomials and Hermite functions

The most straightforward way to extend Hermite polynomials and Hermite functions to multiple dimensions is to use a tensor product structure (see, for example [Gra49]). The d-dimensional Hermite polynomial can be defined as

$$h_{\boldsymbol{\ell}} = h_{\ell_1}(x) \cdot \ldots \cdot h_{\ell_d}(x) \quad \text{for all} \quad \boldsymbol{\ell} \in \mathbb{N}_0^d,$$

where $\boldsymbol{\ell} = (\ell_1, \dots, \ell_d)$ is a multi-index. Most of the properties can be directly translated to multiple dimensions right away. Denote as $\langle i\rangle$ the i-th d-dimensional unit vector. Then, the algebraic properties in multiple dimension read as

$$\nabla h_{\boldsymbol{\ell}}(x) = 2\boldsymbol{\ell}(h_{\boldsymbol{\ell}-\langle i\rangle})_{i=1}^d \tag{2.15}$$

$$(h_{\boldsymbol{\ell}+\langle i\rangle}(\mathbf{x}))_{i=1}^d = (2x_i h_{\boldsymbol{\ell}}(\mathbf{x}))_{i=1}^d - 2(\ell_i h_{\boldsymbol{\ell}-\langle i\rangle}(\mathbf{x}))_{i=1}^d. \tag{2.16}$$

It is also clear that multivariate Hermite polynomials form an orthogonal basis in the space $L^2_\omega(\mathbb{R}^d)$ with the weight

$$\omega(\mathbf{x}) = \mathrm{e}^{-|\mathbf{x}|/2}.$$

Indeed,

$$\langle h_{\boldsymbol{\ell}_1}, h_{\boldsymbol{\ell}_2}\rangle_\omega = \int_{\mathbb{R}^d} h_{\boldsymbol{\ell}_1}(\mathbf{x}) h_{\boldsymbol{\ell}_2}(\mathbf{x}) \mathrm{e}^{-|\mathbf{x}|^2/2} \mathrm{d}\mathbf{x}.$$

Integrating component-wise d times, we get that for $\boldsymbol{\ell}_1 \neq \boldsymbol{\ell}_2$

$$\langle h_{\boldsymbol{\ell}_1}, h_{\boldsymbol{\ell}_2}\rangle_\omega = 0$$

and for all $\boldsymbol{\ell} \in \mathbb{N}_0^d$

$$\langle h_{\boldsymbol{\ell}}, h_{\boldsymbol{\ell}}\rangle_\omega = \|h_{\boldsymbol{\ell}}\|_\omega = \pi^{-d/4}\sqrt{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!},$$

where $|\boldsymbol{\ell}| = \ell_1 + \ldots + \ell_d$ and $\boldsymbol{\ell}! = \ell_1 \cdot \ldots \cdot \ell_d$. The completeness of the set $\{h_{\boldsymbol{\ell}}\}_{\boldsymbol{\ell}\in\mathbb{N}_0^d}$ also follows from the completeness of the corresponding 1D bases.

Similarly to the Hermite polynomials, Hermite functions can be also easily extended to multiple dimensions in the tensor-product way:

$$\psi_{\boldsymbol{\ell}}(\mathbf{x}) = \psi_{\ell_1}(x_1) \cdot \ldots \cdot \psi_{\ell_d}(x_d) \quad \text{for all} \quad \boldsymbol{\ell} \in \mathbb{N}_0^d.$$

The notion of the ladder operators is also easily extensible to higher dimensions.

Indeed, we get

$$Af = \frac{1}{\sqrt{2}}(\mathbf{x} + \nabla \mathbf{x})f \quad \text{and} \quad A^\dagger f = \frac{1}{\sqrt{2}}(\mathbf{x} - \nabla \mathbf{x})f$$

with

$$A_j \psi_{\boldsymbol{\ell}} = \sqrt{\ell_j}\psi_{\boldsymbol{\ell}-\langle j\rangle} \quad \text{and} \quad A_j^\dagger \psi_{\boldsymbol{\ell}} = \sqrt{\ell_j + 1}\psi_{\boldsymbol{\ell}+\langle j\rangle}.$$

As in the one-dimensional case, the ladder operators map Schwartz functions to Schwartz functions are formally adjoint on the Schwartz space $\mathcal{S}(\mathbb{R}^d)$

$$\langle A^\dagger f, g\rangle = \langle f, Ag\rangle \quad \text{for all} \quad f, g \in \mathcal{S}(\mathbb{R}^d).$$

An alternative path to multivariate Hermite polynomials and functions is to go away from the tensor form and create a fully multivariate structure. This is the approach of Hagedorn wave packets (see [Lub08, § V]). In this case, a more general setup is considered, where an anisotropic Gaussian of a certain form, together with the corresponding ladder operators are introduced. The polynomial prefactors arising from this construction can then be viewed as a generalization of Hermite polynomials. There exists a generating function expression for these polynomials. Under certain conditions, the polynomial prefactors of Hagedorn wave packets simplify to tensor-product Hermite polynomials. This is what we will use to get the multidimensional Hermite generating function.

Lemma 2.4.1: Multidimensional Hermite generating function

For all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, the following relation holds

$$\sum_{\boldsymbol{\ell}\in\mathbb{N}^d} \frac{\mathbf{a}^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} h_{\boldsymbol{\ell}}(\mathbf{b}) = \exp(2\mathbf{b}^T\mathbf{a} - \mathbf{a}^T\mathbf{a}), \tag{2.17}$$

where $h_{\boldsymbol{\ell}}$ are tensor product of 1D physicists' Hermite polynomials,

$$h_{\boldsymbol{\ell}}(\mathbf{x}) = h_{\ell_1}(x_1) \cdot \ldots \cdot h_{\ell_d}(x_d).$$

Proof. See [DKT17, Lemma 5] or [Hag15, Theorem 3.1] with $A = \mathrm{Id}_d$. □

The bilinear generating function can also be extended to multiple dimensions. One can find a derivation in [Fol89, Chapter 1, (1.87)]. However, we provide an alternative proof here.

Theorem 2.4.1: Mehler's formula (nD)

The following relation holds for $t \in \mathbb{R}, |t| < 1$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\sum_{|k|=0}^{\infty} \frac{t^{|k|}h_k(\mathbf{x})h_k(\mathbf{y})}{2^{|k|}k!} = \frac{1}{(1-t^2)^{d/2}} \exp\left(\frac{(\mathbf{x}^T\mathbf{y} + \mathbf{y}^T\mathbf{x})t - t^2(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2)}{(1-t^2)}\right). \tag{2.18}$$

Proof. We extend the 1D proof proposed by Watson in [Wat33, p. 5] to multiple dimensions. Applying the inverse Fourier transform to the Fourier transform of a normal distribution with $\sigma = 1/\sqrt{2}\mathrm{Id}_d$, we get

$$\mathrm{e}^{-\mathbf{x}^T\mathbf{x}} = \pi^{d/2}\int_{\mathbb{R}^d} \mathrm{e}^{2\pi\mathrm{i}\mathbf{x}^T\boldsymbol{\xi}}\mathrm{e}^{-\pi^2\boldsymbol{\xi}^T\boldsymbol{\xi}}\mathrm{d}\boldsymbol{\xi} = \pi^{-d/2}\int_{\mathbb{R}^d} \mathrm{e}^{2\mathrm{i}\mathbf{x}^T\boldsymbol{\xi}}\mathrm{e}^{-\boldsymbol{\xi}^T\boldsymbol{\xi}}\mathrm{d}\boldsymbol{\xi}.$$

Hence,

$$h_{\boldsymbol{\ell}}(\mathbf{x}) = \mathrm{e}^{\mathbf{x}^T\mathbf{x}}(-\nabla)^{\boldsymbol{\ell}}\mathrm{e}^{-\mathbf{x}^T\mathbf{x}} = \frac{(-2\mathrm{i})^{|\boldsymbol{\ell}|}}{\pi^{d/2}}\mathrm{e}^{\mathbf{x}^T\mathbf{x}}\int_{\mathbb{R}^d}\boldsymbol{\xi}^{\boldsymbol{\ell}}\mathrm{e}^{2\mathrm{i}\mathbf{x}^T\boldsymbol{\xi}-\boldsymbol{\xi}^T\boldsymbol{\xi}}\mathrm{d}\boldsymbol{\xi},$$

where we used the Rodrigues formula for multivariate tensor-product Hermite polynomials (see [DKT17, Expr. 11] with $M = \mathrm{Id}$). Then,

$$\begin{aligned}
&\sum_{|\boldsymbol{\ell}|=0}^{\infty}\frac{t^{|\boldsymbol{\ell}|}h_{\boldsymbol{\ell}}(\mathbf{x})h_{\boldsymbol{\ell}}(\mathbf{y})}{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!} = \\
&= \frac{\mathrm{e}^{\mathbf{x}^T\mathbf{x}+\mathbf{y}^T\mathbf{y}}}{\pi^d}\sum_{|\boldsymbol{\ell}|=0}^{\infty}\frac{(-2t)^{|\boldsymbol{\ell}|}}{\boldsymbol{\ell}!}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\boldsymbol{\xi}_{\mathbf{x}}^{\boldsymbol{\ell}}\boldsymbol{\xi}_{\mathbf{y}}^{\boldsymbol{\ell}}\mathrm{e}^{2\mathrm{i}(\mathbf{x}^T\boldsymbol{\xi}_{\mathbf{x}}+\mathbf{y}^T\boldsymbol{\xi}_{\mathbf{y}})}\mathrm{e}^{-\boldsymbol{\xi}_{\mathbf{x}}^T\boldsymbol{\xi}_{\mathbf{x}}-\boldsymbol{\xi}_{\mathbf{y}}^T\boldsymbol{\xi}_{\mathbf{y}}}\mathrm{d}\boldsymbol{\xi}_{\mathbf{x}}\mathrm{d}\boldsymbol{\xi}_{\mathbf{y}} \\
&= \frac{\mathrm{e}^{\mathbf{x}^T\mathbf{x}+\mathbf{y}^T\mathbf{y}}}{\pi^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\mathrm{e}^{-2t\boldsymbol{\xi}_{\mathbf{x}}^T\boldsymbol{\xi}_{\mathbf{y}}}\mathrm{e}^{2\mathrm{i}(\mathbf{x}^T\boldsymbol{\xi}_{\mathbf{x}}+\mathbf{y}^T\boldsymbol{\xi}_{\mathbf{y}})}\mathrm{e}^{-\boldsymbol{\xi}_{\mathbf{x}}^T\boldsymbol{\xi}_{\mathbf{x}}-\boldsymbol{\xi}_{\mathbf{y}}^T\boldsymbol{\xi}_{\mathbf{y}}}\mathrm{d}\boldsymbol{\xi}_{\mathbf{x}}\mathrm{d}\boldsymbol{\xi}_{\mathbf{y}}, \qquad (2.19)
\end{aligned}$$

where we used the Taylor series of the exponential function. Recall the following formula for a bivariate Gaussian Fourier integral (see [Wat33, p. 5]):

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_{(x,y)}(u,v)\mathrm{d}v\mathrm{d}u = \frac{\pi\mathrm{e}^{-(x^2+y^2)/2}}{\sqrt{1-t^2}}\exp\!\left(\frac{x^2-y^2}{2} - \frac{(x-yt)^2}{1-t^2}\right). \qquad (2.20)$$

with

$$g_{(x,y)}(u,v) = \exp(-u^2 - 2tuv - v^2 + 2\mathrm{i}xu + 2\mathrm{i}yv).$$

Using (2.20) d times for the integral (2.19) together with the algebraic identity

$$-\tfrac{1}{2}(x^2+y^2) + \tfrac{1}{2}(x^2-y^2) - \frac{(x-yt)^2}{1-t^2} = \frac{2txy - x^2 - y^2}{1-t^2}$$

yields (2.18). □

A different proof can be found in [Wü15, § 6] which is given in 2D, but is easily extendable to the multidimensional case.

In the next chapter, we introduce a new basis in the Hermite family. The basis is of the type tensor-product Hermite polynomial of a scaled variable times an anisotropic Gaussian. The difference of this basis to the existing Hermite functions or Hagedorn wave packets is that the scaling of the argument of the Hermite polynomial is completely decoupled from the scaling of the exponential part. Also, the new basis functions are no longer orthogonal in $L^2(\mathbb{R}^d)$, and we will introduce a special weighted space where the orthogonality is preserved.

3. Generalized anisotropic Hermite functions

Even though Hermite functions are very useful for a variety of applications, there are some cases when a more general version of the basis is a better fit. In particular, we would like to be able to represent a truly anisotropic Gaussian by the means of the basis. As we will show in part II, it turns out to be particularly useful for the stabilization of the interpolation with anisotropic Gaussians. In this chapter, we introduce a generalization of multivariate Hermite functions to the case of the anisotropic Gaussian. The new basis is of the form of scaled tensor product Hermite polynomials times an anisotropic Gaussian, augmented with some additional parameters that allow to adapt the basis for a specific problem.

We introduce two invertible parameter matrices $E, G \in \mathbb{R}^{d\times d}$ and a technical parameter t. The matrix E is a shape matrix corresponding to the width of the multivariate Gaussian, the scaling matrix G is responsible for varying the evaluation domain of the Hermite polynomials and the technical parameter t will allow us to use the Hermite polynomial's bilinear generating function (2.18) to calculate the norm of our basis when needed. Consider

$$H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) = \frac{t^{|\boldsymbol{\ell}|/2}}{\sqrt{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}} h_{\boldsymbol{\ell}}(G^T\mathbf{x}) \exp(-\mathbf{x}^T E^T E\mathbf{x}),$$

where E, $G \in \mathbb{R}^{d\times d}$ are arbitrary invertible matrices, $t > 0$ and $h_{\boldsymbol{\ell}}(\mathbf{x})$ are tensor product of 1D physicists' Hermite polynomials,

$$h_{\boldsymbol{\ell}}(\mathbf{x}) = h_{\ell_1}(x_1) \cdot \ldots \cdot h_{\ell_d}(x_d).$$

We later refer to $\{H_{\boldsymbol{\ell}}^{G,E,t}\}_{\boldsymbol{\ell}\in\mathbb{N}_0^d}$ as *generalized anisotropic Hermite functions*. We recover the standard Hermite functions up to the constant $\pi^{-d/4}$ by setting $t = 1$, $G = \mathrm{Id}_d$ and $E = 1/\sqrt{2}\mathrm{Id}$. Note, that in order to make use of Mehler's formula directly on the basis, we would need to limit the values of the parameter t to the interval $(0, 1)$. However, the theory derived in this chapter holds for arbitrary $t > 0$. We, therefore, do not constrain the parameter t unless we need to use Mehler's formula.

The new basis functions can be expressed through multivariate tensor-product Hermite functions as follows:

$$H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) = \pi^{d/4} t^{|\boldsymbol{\ell}|/2} \psi_{\boldsymbol{\ell}}(G^T\mathbf{x}) \exp\left(\mathbf{x}^T \left(\frac{1}{2}GG^T - E^T E\right)\mathbf{x}\right). \tag{3.1}$$

where $\psi_{\boldsymbol{\ell}}$ are tensor-product Hermite functions:

$$\psi_{\boldsymbol{\ell}}(\mathbf{x}) = \psi_{\ell_1}(x_1) \cdot \ldots \cdot \psi_{\ell_d}(x_d).$$

In this chapter, we translate the main properties of Hermite functions to the new basis with the final goal to estimate the approximation error of the new basis. We will define the appropriate function spaces in section 3.1, prove that the new set forms an orthogonal basis in the corresponding space in section 3.2, derive the necessary algebraic properties in section 3.3. Based on these properties, we will introduce the ladder operators in section 3.4, until finally arriving to the approximation results in section 3.5.

3.1. Weighted spaces

First and foremost, we look for a corresponding space for the generalized anisotropic Hermite functions, where they form an orthogonal basis. Similarly to Hermite polynomials, generalized anisotropic Hermite functions are not orthogonal in the standard $L^2(\mathbb{R}^d)$ space. In order to still be able to take advantage of the orthogonality, we introduce a Hilbert space associated with the new basis where we can later prove the orthogonality and some other useful properties. We consider the space

$$L^2_\omega(\mathbb{R}^d) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \int_{\mathbb{R}^d} |f(\mathbf{x})|^2 \omega(\mathbf{x}) \mathrm{d}x < \infty \right\}$$

with the weight $\omega : \mathbb{R}^d \to \mathbb{R}_+$

$$\omega(\mathbf{x}) = \pi^{-d/2} |\det(G)| \exp(2\mathbf{x}^T E^T E \mathbf{x} - \mathbf{x}^T G G^T \mathbf{x}) \tag{3.2}$$

and the inner product

$$\langle f, g \rangle_{L^2_\omega(\mathbb{R}^d)} = \int_{\mathbb{R}^d} f(\mathbf{x}) g(\mathbf{x}) \omega(\mathbf{x}) \mathrm{d}x.$$

We later refer to $\langle \cdot, \cdot \rangle_{L^2_\omega(\mathbb{R}^d)}$ as $\langle \cdot, \cdot \rangle_\omega$. In the context of this chapter, the symbol ω always corresponds to the weight function (3.2).

We limit ourselves to real-valued functions since we only focus on those in the applications considered later in this thesis.

Remark 3.1.1. *Since we only consider real-valued functions, the dot product $\langle \cdot, \cdot \rangle_\omega$ is symmetric.*

In the following sections, we look into the structure of the $L^2_\omega(\mathbb{R}^d)$ space and prove key properties that allow us to estimate the approximation error. It turns out that some properties of the new basis can be tracked from the corresponding properties of the standard Hermite functions basis. To simplify the proofs below, let us first recall the connection of the generalized Hermite functions to the standard ones (3.1):

$$\begin{aligned} H^{G,E,t}_{\boldsymbol{\ell}}(\mathbf{x}) &= \pi^{d/4} t^{|\boldsymbol{\ell}|/2} \psi_{\boldsymbol{\ell}}(G^T \mathbf{x}) \exp\left(\mathbf{x}^T \left(\frac{1}{2} G G^T - E^T E \right) \mathbf{x} \right) \\ &= \frac{t^{|\boldsymbol{\ell}|/2} \sqrt{|\det(G)|}}{\sqrt{\omega(\mathbf{x})}} \psi_{\boldsymbol{\ell}}(G^T \mathbf{x}), \end{aligned} \tag{3.3}$$

Analogously, one can express Hermite functions through the HermiteGF basis:

$$\psi_{\boldsymbol{\ell}}(\mathbf{x}) = \frac{\sqrt{\omega(G^{-T}\mathbf{x})}}{\sqrt{|\det(G)|}t^{|\boldsymbol{\ell}|/2}} H_{\boldsymbol{\ell}}^{G,E,t}(G^{-T}\mathbf{x}). \tag{3.4}$$

As in the case of Hermite functions, it is advantageous for the derivation of certain properties to consider a subspace of $L^2_\omega(\mathbb{R}^d)$. However, the regular Schwartz space is not suitable for this purpose. For this reason, we introduce a weighted alternative of the Schwartz space that accommodates the generalized basis.

3.1.1. Weighted Schwartz space

Let us now introduce the analog of the Schwartz space $\mathcal{S}(\mathbb{R})$ for the space $L^2_\omega(\mathbb{R}^d)$.

Definition 3.1.1

Consider a weight function $\omega: \mathbb{R}^d \to \mathbb{R}$. The following space we call the *weighted Schwartz space*:

$$\mathcal{S}_\omega(\mathbb{R}^d) := \left\{f : \mathbb{R}^d \to \mathbb{R} | f\sqrt{\omega} \in \mathcal{S}(\mathbb{R}^d)\right\}.$$

Even though some statements below also hold for an arbitrary weight ω, for the sake of simplicity, we assume that we work with the generalized anisotropic Hermite functions weight (3.2).

$$\omega(\mathbf{x}) = \pi^{-d/2}|\det(G)| \exp(2\mathbf{x}^T E^T E\mathbf{x} - \mathbf{x}^T GG^T\mathbf{x}).$$

Let us prove that, analogously to the Schwartz space, the generalized Schwartz space is dense in $L^2_\omega(\mathbb{R}^d)$.

Lemma 3.1.1

$\mathcal{S}_\omega(\mathbb{R}^d) \subset L^2_\omega(\mathbb{R}^d)$ and is dense in $L^2_\omega(\mathbb{R}^d)$.

Proof. We first prove that $\mathcal{S}_\omega(\mathbb{R}^d) \subset L^2_\omega(\mathbb{R}^d)$. Indeed, if $\phi \in \mathcal{S}_\omega(\mathbb{R}^d)$, then, by definition, $\phi\sqrt{\omega} \in \mathcal{S}(\mathbb{R}^d) \subset L^2(\mathbb{R}^d)$. Therefore,

$$\int_{\mathbb{R}^d} \phi(\mathbf{x})^2 \omega(\mathbf{x}) \mathrm{d}\mathbf{x} < \infty.$$

which proves that $\phi \in L^2_\omega(\mathbb{R}^d)$.

We now proceed to prove that $\mathcal{S}_\omega(\mathbb{R}^d)$ is dense in $L^2_\omega(\mathbb{R}^d)$. Let us consider $\phi \in L^2_\omega(\mathbb{R}^d)$. Then, $\phi\sqrt{\omega} \in L^2_\omega(\mathbb{R}^d)$. Therefore, there exists a sequence $\{\phi_n\}_{n=0}^\infty \subset \mathcal{S}(\mathbb{R}^d)$ such that

$$\lim_{n\to\infty} \|\phi_n - \phi\sqrt{\omega}\|^2_{L^2(\mathbb{R}^d)} = 0.$$

Then

$$\int_{\mathbb{R}^d} \left(\phi_n(\mathbf{x}) - \phi(\mathbf{x})\sqrt{\omega(\mathbf{x})}\right)^2 \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d} \left(\frac{\phi_n(\mathbf{x})}{\sqrt{\omega(\mathbf{x})}} - \phi(\mathbf{x})\right)^2 \omega(\mathbf{x})\mathrm{d}\mathbf{x} = \left\|\frac{\phi_n}{\sqrt{\omega}} - \phi\right\|_\omega^2.$$

Denote

$$\phi_n^\omega(\mathbf{x}) = \phi_n(\mathbf{x})\omega(\mathbf{x})^{-1/2}.$$

Then, by definition of $\mathcal{S}_\omega(\mathbb{R}^d)$, $\phi_n^\omega \in \mathcal{S}_\omega(\mathbb{R}^d)$. Therefore, we have found a sequence $\{\phi_n^\omega\}_{n=0}^\infty$ such that

$$\lim_{n\to\infty} \|\phi_n^\omega - \phi\|_\omega = 0. \qquad \square$$

Note that since the generalized Hermite basis can be described in terms of Hermite functions as

$$H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) = \frac{t^{|\boldsymbol{\ell}|/2}\sqrt{|\det(G)|}}{\sqrt{\omega(\mathbf{x})}}\psi_{\boldsymbol{\ell}}(G^T\mathbf{x}),$$

and due to the fact that Hermite functions are included in the Schwartz space $\mathcal{S}(\mathbb{R}^d)$, the generalized Hermite functions $H_{\boldsymbol{\ell}}^{G,E,t} \in \mathcal{S}_\omega(\mathbb{R}^d)$ for all $\boldsymbol{\ell} \in \mathbb{N}_0^d$.

3.2. Orthogonality and completeness

We now proceed to investigate the properties of the generalized anisotropic Hermite functions in the $L^2_\omega(\mathbb{R}^d)$ space. First of all, we have to prove that they indeed form an orthogonal basis in $L^2_\omega(\mathbb{R}^d)$. For that, we need to prove that generalized anisotropic Hermite functions are orthogonal in $L^2_\omega(\mathbb{R}^d)$ and that the set $\{H_{\boldsymbol{\ell}}^{G,E,t}\}_{\boldsymbol{\ell}\in\mathbb{N}_0^d}$ is complete in $L^2_\omega(\mathbb{R}^d)$.

Lemma 3.2.1: Generalized anisotropic Hermite basis

Generalized anisotropic Hermite functions are orthogonal in $L^2_\omega(\mathbb{R}^d)$:

$$\langle H_{\boldsymbol{\ell}_1}^{G,E,t}, H_{\boldsymbol{\ell}_2}^{G,E,t}\rangle_\omega = t^{(|\boldsymbol{\ell}_1|+|\boldsymbol{\ell}_2|)/2}\delta_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2}.$$

and the set of functions $\{H_{\boldsymbol{\ell}}^{G,E,t}\}_{\boldsymbol{\ell}\in\mathbb{N}_0^d}$ is complete in $L^2_\omega(\mathbb{R}^d)$:

$$f = \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d} \frac{\langle f, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega}{\langle H_{\boldsymbol{\ell}}^{G,E,t}, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega} H_{\boldsymbol{\ell}}^{G,E,t} \text{ for all } f \in L^2_\omega(\mathbb{R}^d) \qquad (3.5)$$

with $\langle H_{\boldsymbol{\ell}}^{G,E,t}, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega = t^{|\boldsymbol{\ell}|}$.

Proof. We start by proving orthogonality. We first recall that the generalized anisotropic Hermite functions can be expressed through the standard ones as

$$H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) \overset{(3.3)}{=} \frac{t^{|\boldsymbol{\ell}|/2}\sqrt{|\det(G)|}}{\sqrt{\omega(\mathbf{x})}}\psi_{\boldsymbol{\ell}}(G^T\mathbf{x}).$$

Using this relation, we get

$$\begin{aligned}\langle H_{\boldsymbol{\ell}_1}^{G,E,t}, H_{\boldsymbol{\ell}_2}^{G,E,t}\rangle_\omega &= \int_{\mathbb{R}^d} H_{\boldsymbol{\ell}_1}^{G,E,t}(\mathbf{x}) H_{\boldsymbol{\ell}_2}^{G,E,t}(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x}\\ &\overset{(3.3)}{=} t^{(|\boldsymbol{\ell}_1|+|\boldsymbol{\ell}_2|)/2}\int_{\mathbb{R}^d}\psi_{\boldsymbol{\ell}_1}(G^T\mathbf{x})\psi_{\boldsymbol{\ell}_2}(G^T\mathbf{x})\frac{|\det(G)|}{\omega(\mathbf{x})}\omega(\mathbf{x})\mathrm{d}\mathbf{x}\\ &\overset{\bar{\mathbf{x}}=G^T\mathbf{x}}{=} t^{(|\boldsymbol{\ell}_1|+|\boldsymbol{\ell}_2|)/2}\int_{\mathbb{R}^d}\psi_{\boldsymbol{\ell}_1}(\bar{\mathbf{x}})\psi_{\boldsymbol{\ell}_2}(\bar{\mathbf{x}})\mathrm{d}\bar{\mathbf{x}} = t^{(|\boldsymbol{\ell}_1|+|\boldsymbol{\ell}_2|)/2}\delta_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2}.\end{aligned}$$

We now proceed to the proof of completeness. Consider $f \in L^2_\omega(\mathbb{R}^d)$. Then

$$\int_{\mathbb{R}} f(\mathbf{x})^2\omega(\mathbf{x})\mathrm{d}\mathbf{x} < \infty.$$

This implies

$$\int_{\mathbb{R}^d} f(\mathbf{x})^2\omega(\mathbf{x})\mathrm{d}\mathbf{x} \overset{\bar{\mathbf{x}}=G^T\mathbf{x}}{=} \int_{\mathbb{R}^d} f(G^{-T}\bar{\mathbf{x}})^2\frac{\omega(G^{-T}\bar{\mathbf{x}})}{|\det(G)|}\mathrm{d}\mathbf{x} < \infty.$$

Therefore,

$$f(G^{-T}\bar{\mathbf{x}})\sqrt{\frac{\omega(G^{-T}\bar{\mathbf{x}})}{|\det(G)|}} \in L^2(\mathbb{R}^d).$$

Using the fact that the Hermite functions form a complete set in $L^2(\mathbb{R}^d)$ [Yse10, §3.4, Theorem 3.5], we get

$$f(G^{-T}\bar{\mathbf{x}})\sqrt{\frac{\omega(G^{-T}\bar{\mathbf{x}})}{|\det(G)|}} = \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d}\alpha_{\boldsymbol{\ell}}\psi_{\boldsymbol{\ell}}(\bar{\mathbf{x}}),$$

where $\psi_{\boldsymbol{\ell}}$ are tensor-product Hermite functions. The coefficients $\alpha_{\boldsymbol{\ell}}$ can be computed as

$$\begin{aligned}\alpha_{\boldsymbol{\ell}} &= \int_{\mathbb{R}^d} f(G^{-T}\bar{\mathbf{x}})\sqrt{\frac{\omega(G^{-T}\bar{\mathbf{x}})}{|\det(G)|}}\psi_{\boldsymbol{\ell}}(\bar{\mathbf{x}})\mathrm{d}\bar{\mathbf{x}}\\ &= \int_{\mathbb{R}^d} f(G^{-T}\bar{\mathbf{x}})\frac{\omega(G^{-T}\bar{\mathbf{x}})}{|\det(G)|}t^{-|\boldsymbol{\ell}|/2}H_{\boldsymbol{\ell}}^{G,E,t}(G^{-T}\bar{\mathbf{x}})\mathrm{d}\bar{\mathbf{x}}\\ &\overset{\mathbf{x}=G^{-T}\bar{\mathbf{x}}}{=} t^{-|\boldsymbol{\ell}|/2}\int_{\mathbb{R}^d} f(\mathbf{x})H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x} = t^{-|\boldsymbol{\ell}|/2}\langle f, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega.\end{aligned}$$

Using the fact that the Hermite functions basis is complete in $L^2(\mathbb{R}^d)$ and the expression (3.4), we get

$$\begin{aligned}0 &= \left\| f(G^{-T}\bar{\mathbf{x}})\sqrt{\frac{\omega(G^{-T}\bar{\mathbf{x}})}{|\det(G)|}} - \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d}\alpha_{\boldsymbol{\ell}}\psi_{\boldsymbol{\ell}}(\bar{\mathbf{x}})\right\|^2_{L^2(\mathbb{R}^d)}\\ &= \int_{\mathbb{R}^d}\left(f(G^{-T}\bar{\mathbf{x}}) - \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d}\alpha_{\boldsymbol{\ell}}t^{-|\boldsymbol{\ell}|/2}H_{\boldsymbol{\ell}}^{G,E,t}(G^{-T}\bar{\mathbf{x}})\right)^2\frac{\omega(G^{-T}\bar{\mathbf{x}})}{|\det(G)|}\mathrm{d}\bar{\mathbf{x}}\\ &= \left\| f - \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d}\langle f, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega t^{-|\boldsymbol{\ell}|}H_{\boldsymbol{\ell}}^{G,E,t}\right\|^2_\omega = \left\| f - \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d}\frac{\langle f, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega}{\langle H_{\boldsymbol{\ell}}^{G,E,t}, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega}H_{\boldsymbol{\ell}}^{G,E,t}\right\|^2_\omega.\end{aligned}$$

Therefore, generalized anisotropic Hermite functions $\{H_{\boldsymbol{\ell}}^{G,E,t}\}_{\boldsymbol{\ell}\in\mathbb{N}_0^d}$ form an orthogonal basis in $L^2_\omega(\mathbb{R}^d)$. □

Now, when we are sure that the new functions indeed form a basis, we can proceed to develop the rest of the generalized framework.

3.3. Algebraic properties and three-term recurrence

As for the case of the Hermite functions (see section 2.3), we would not want to evaluate our basis based on the definition. The more efficient way to do that is to use the *three-term recurrence*.

First, we prove several properties of the generalized anisotropic Hermite basis that are not only useful for the derivation of the three-term recurrence, but also for the ladder operator theory that we will develop in the next section. For the simplicity of the notation, we denote two auxiliary vectors

$$H_{\boldsymbol{\ell}-}^{G,E,t}(\mathbf{x}) = \begin{pmatrix} \sqrt{\ell_1}H_{\boldsymbol{\ell}-\langle 1\rangle}^{G,E,t}(\mathbf{x}) \\ \cdots \\ \sqrt{\ell_d}H_{\boldsymbol{\ell}-\langle d\rangle}^{G,E,t}(\mathbf{x}) \end{pmatrix} \quad \text{and} \quad H_{\boldsymbol{\ell}+}^{G,E,t}(\mathbf{x}) = \begin{pmatrix} \sqrt{\ell_1+1}H_{\boldsymbol{\ell}+\langle 1\rangle}^{G,E,t}(\mathbf{x}) \\ \cdots \\ \sqrt{\ell_d+1}H_{\boldsymbol{\ell}+\langle d\rangle}^{G,E,t}(\mathbf{x}) \end{pmatrix},$$

where $\langle i\rangle = (0,\dots,1,\dots,0)$ denotes the i-th d-dimensional unit vector. We investigate the relation between $H_{\boldsymbol{\ell}-}^{G,E,t}$ and $H_{\boldsymbol{\ell}+}^{G,E,t}$ in order to derive the three-term recurrence explicitly. First, let us recall the multivariate version of the three-term recurrence (2.16) of tensor-product Hermite polynomials.

$$(h_{\boldsymbol{\ell}+\langle i\rangle}(\mathbf{x}))_{i=1}^d = (2x_i h_{\boldsymbol{\ell}}(\mathbf{x}))_{i=1}^d - 2(\ell_i h_{\boldsymbol{\ell}-\langle i\rangle}(\mathbf{x}))_{i=1}^d.$$

We now proceed to the proof of the algebraic properties of the basis.

Theorem 3.3.1: Properties of generalized anisotropic Hermite functions

For all $\mathbf{x}\in\mathbb{R}^d$, invertible $E, G\in\mathbb{R}^{d\times d}$, $\boldsymbol{\ell}\in\mathbb{N}_0^d$, the following relations hold:

1. $$\nabla H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) = \sqrt{2t}(G - E^TEG^{-T})H_{\boldsymbol{\ell}-}^{G,E,t}(\mathbf{x}) - \sqrt{\tfrac{2}{t}}E^TEG^{-T}H_{\boldsymbol{\ell}+}^{G,E,t}(\mathbf{x}). \quad (3.6)$$

2. $$\mathbf{x}H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) = G^{-T}\left(\sqrt{\tfrac{t}{2}}H_{\boldsymbol{\ell}-}^{G,E,t}(\mathbf{x}) + \tfrac{1}{\sqrt{2t}}H_{\boldsymbol{\ell}+}^{G,E,t}(\mathbf{x})\right). \quad (3.7)$$

Proof. Denote $\bar{E} = E^TE$. Using the property (2.6) of Hermite polynomials and the fact that $\bar{E}$ is symmetric, we get

$$\begin{aligned}\nabla H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) &= \frac{t^{|\boldsymbol{\ell}|/2}}{\sqrt{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}\left(\nabla h_{\boldsymbol{\ell}}(G^T\mathbf{x})\mathrm{e}^{-\mathbf{x}^TE^TE\mathbf{x}} + h_{\boldsymbol{\ell}}(G^T\mathbf{x})\nabla\mathrm{e}^{-\mathbf{x}^TE^TE\mathbf{x}}\right) \\ &= \frac{t^{|\boldsymbol{\ell}|/2}}{\sqrt{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}\left(2G(\ell_i h_{\boldsymbol{\ell}-\langle i\rangle}(G^T\mathbf{x}))_{i=1}^d - 2\bar{E}\mathbf{x}h_{\boldsymbol{\ell}}(G^T\mathbf{x})\right)\mathrm{e}^{-\mathbf{x}^TE^TE\mathbf{x}}\end{aligned}$$

Now, with the help of (2.7), we obtain

$$\begin{aligned}\nabla H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) &= \frac{t^{|\boldsymbol{\ell}|/2}}{\sqrt{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}\big(2G(\ell_i h_{\boldsymbol{\ell}-\langle i\rangle}(G^T\mathbf{x}))_{i=1}^d - 2\bar{E}G^{-T}(\ell_i h_{\boldsymbol{\ell}-\langle i\rangle}(G^T\mathbf{x}))_{i=1}^d \\ &\qquad + 2\bar{E}G^{-T}(\ell_i h_{\boldsymbol{\ell}-\langle i\rangle}(G^T\mathbf{x})_{i=1}^d - 2\bar{E}\mathbf{x}h_{\boldsymbol{\ell}}(G^T\mathbf{x})\big)\mathrm{e}^{-\mathbf{x}^T E^T E\mathbf{x}} \\ &= \frac{t^{|\boldsymbol{\ell}|/2}}{\sqrt{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}\big(2(G-\bar{E}G^{-T})(\ell_i h_{\boldsymbol{\ell}-\langle i\rangle})_{i=1}^d - \bar{E}G^{-T}(h_{\boldsymbol{\ell}+\langle i\rangle}(G^T\mathbf{x})_{i=1}^d\big)\mathrm{e}^{-\mathbf{x}^T E^T E\mathbf{x}} \\ &= \sqrt{2t}(G - E^T E G^{-T})H_{\boldsymbol{\ell}-}^{G,E,t}(\mathbf{x}) - \sqrt{\frac{2}{t}}E^T E G^{-T} H_{\boldsymbol{\ell}+}^{G,E,t}(\mathbf{x}).\end{aligned}$$

Similarly, using the three-term recurrence (2.7) for Hermite polynomials, we get

$$\begin{aligned}\mathbf{x}H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) &= \frac{t^{|\boldsymbol{\ell}|/2}}{\sqrt{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}(\mathbf{x}h_{\boldsymbol{\ell}}(G^T\mathbf{x}))\mathrm{e}^{\mathbf{x}^T E^T E\mathbf{x}} \\ &= \frac{t^{|\boldsymbol{\ell}|/2}}{\sqrt{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}G^{-T}\left(\frac{1}{2}(h_{\boldsymbol{\ell}+\langle i\rangle}(G^T\mathbf{x}))_{i=1}^d + (\ell_i h_{\boldsymbol{\ell}-\langle i\rangle}(G^T\mathbf{x}))_{i=1}^d\right) \\ &= G^{-T}\left(\frac{1}{\sqrt{2t}}H_{\boldsymbol{\ell}+}^{G,E,t}(\mathbf{x}) + \sqrt{\frac{t}{2}}H_{\boldsymbol{\ell}-}^{G,E,t}(\mathbf{x})\right). \qquad \square\end{aligned}$$

One can reformulate the property (3.7) to obtain the analog of the Hermite functions three-term recurrence for the new generalized anisotropic Hermite basis.

Three-term recurrence

The following relation holds for all $\mathbf{x}\in\mathbb{R}^d$, $\boldsymbol{\ell}\in\mathbb{N}_0^d$:

$$\left(\sqrt{\ell_i+1}H_{\boldsymbol{\ell}+\langle i\rangle}^{G,E,t}(\mathbf{x})\right)_{i=1}^d = \sqrt{2t}G^T\mathbf{x}H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) - t\left(\sqrt{\ell_i}H_{\boldsymbol{\ell}-\langle i\rangle}^{G,E,t}(\mathbf{x})\right)_{i=1}^d. \tag{3.8}$$

For the case $E^T E = \frac{1}{2}\mathrm{Id}_d$, $G = \mathrm{Id}_d$, and $t = 1$, the three-term recurrence (3.8) matches the one for the tensor-product Hermite functions with the difference of the factor $\pi^{-d/4}$ for the 0-th basis function.

With everything set up, we can move on to the next step of the framework construction. In particular, to the definition of the analog of *Dirac ladder operators* for the $\{H_{\boldsymbol{\ell}}^{G,E,t}\}_{\boldsymbol{\ell}\in\mathbb{N}_0^d}$ in $L^2_\omega(\mathbb{R}^d)$.

3.4. Ladder operators

Analogously to Hermite functions, it is possible to define the appropriate ladder operators for the new basis in terms of the space $L^2_\omega(\mathbb{R}^d)$. Consider the differential operator

$$A_\omega f(\mathbf{x}) := \frac{G^{-1}}{\sqrt{2t}}(\nabla f(\mathbf{x}) + 2E^T E\mathbf{x}f(\mathbf{x})), \quad \forall\, \mathbf{x}\in\mathbb{R}^d. \tag{3.9}$$

The operator A_ω acts as a lowering operator for the generalized anisotropic Hermite functions. Indeed, using the properties (3.6), (3.7), we get

$$\begin{aligned}
A_\omega H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) &= \frac{G^{-1}}{\sqrt{2t}}(\nabla H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) + 2E^T E\mathbf{x}H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}))\\
&= \frac{G^{-1}}{\sqrt{2t}}\Big(\sqrt{2t}(G - E^T E G^{-T})H_{\boldsymbol{\ell}-}^{G,E,t}(\mathbf{x}) - \sqrt{\frac{2}{t}}E^T E G^{-T} H_{\boldsymbol{\ell}+}^{G,E,t}(\mathbf{x})\\
&+ \sqrt{2t}E^T E G^{-T} H_{\boldsymbol{\ell}-}^{G,E,t}(\mathbf{x}) + \sqrt{\frac{2}{t}}E^T E G^{-T} H_{\boldsymbol{\ell}+}^{G,E,t}(\mathbf{x})\Big)\\
&= H_{\boldsymbol{\ell}-}^{G,E,t}(\mathbf{x}).
\end{aligned}$$

Analogously one can define a raising operator for $\{H_{\boldsymbol{\ell}}^{G,E,t}\}_{\boldsymbol{\ell}\in\mathbb{N}_0^d}$:

$$A_\omega^\dagger f(\mathbf{x}) = \sqrt{\frac{t}{2}}G^{-1}\left(-\nabla f(\mathbf{x}) + 2(GG^T - E^T E)\mathbf{x}f(\mathbf{x})\right). \tag{3.10}$$

Explicit calculation yields

$$\begin{aligned}
A_\omega^\dagger H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) &= \sqrt{\frac{t}{2}}G^{-1}\Big(-\nabla H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) + 2(GG^T - E^T E)\mathbf{x}H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x})\Big)\\
&= \sqrt{\frac{t}{2}}G^{-1}\Bigg(- \Bigg(\sqrt{2t}(G - E^T E G^{-T})H_{\boldsymbol{\ell}-}^{G,E,t}(\mathbf{x}) - \sqrt{\frac{2}{t}}E^T E G^{-T} H_{\boldsymbol{\ell}+}^{G,E,t}(\mathbf{x})\Bigg)\\
&+ 2(GG^T - E^T E)G^{-T}\left(\sqrt{\frac{t}{2}}H_{\boldsymbol{\ell}-}^{G,E,t}(\mathbf{x}) + \frac{1}{\sqrt{2t}}H_{\boldsymbol{\ell}+}^{G,E,t}(\mathbf{x})\right)\Bigg)\\
&= H_{\boldsymbol{\ell}+}^{G,E,t}(\mathbf{x}).
\end{aligned}$$

Therefore, we have defined the raising and the lowering operator for the generalized anisotropic Hermite functions.

Ladder operators

The operators $A_\omega^\dagger$ and A_ω act as the raising and lowering operator for generalized anisotropic Hermite functions:

$$A_\omega^\dagger H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) = H_{\boldsymbol{\ell}+}^{G,E,t}(\mathbf{x}) \tag{3.11}$$

$$A_\omega H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) = H_{\boldsymbol{\ell}-}^{G,E,t}(\mathbf{x}) \tag{3.12}$$

We now proceed to the investigation of the properties of the ladder operators.

3.4.1. Properties of ladder operators in weighted Schwartz space

As the next step, we would like to prove that $A_\omega^\dagger$ is adjoint to A_ω in the weighted Schwartz space. However, it turns out that the relation does not hold exactly. We will prove a similar relation, with the difference of an additional factor t. We will refer to that relation as t-*adjointness*.

Definition 3.4.1: t-adjointness

Consider $t > 0$ and two differential operators $A^\dagger_\omega$, $A_\omega : \mathcal{S}_\omega(\mathbb{R}^d) \to \mathcal{S}_\omega(\mathbb{R}^d)$. Then, these operators are called t*-adjoint* if for all $f, g \in \mathcal{S}_\omega(\mathbb{R}^d)$

$$\langle (A^\dagger_\omega)_i f, g\rangle_\omega = t\langle f, (A_\omega)_i g\rangle_\omega \quad \text{for all} \quad i = 1 \dots d.$$

where $(A_\omega)_i$ and $(A^\dagger_\omega)_i$ denote the i-th components of the differential operators.

In order to prove that $A^\dagger_\omega$ is t-adjoint to A_ω in $\mathcal{S}_\omega(\mathbb{R}^d)$, we first prove that for both operators the image of $\mathcal{S}_\omega(\mathbb{R}^d)$ is also in $\mathcal{S}_\omega(\mathbb{R}^d)$.

Lemma 3.4.1

Ladder operators (3.9), (3.10) map weighted Schwartz functions to weighted Schwartz functions. For all $f \in \mathcal{S}_\omega(\mathbb{R}^d)$, $i = 1 \dots d$,

$$(A_\omega)_i f \in \mathcal{S}_\omega(\mathbb{R}^d), \quad (A^\dagger_\omega)_i f \in \mathcal{S}_\omega(\mathbb{R}^d).$$

Proof. Consider $f \in \mathcal{S}_\omega(\mathbb{R}^d)$. Denote $\bar{E} = E^T E$, $\bar{G} = GG^T$. Let us first compute

$$\begin{aligned}\nabla(f(\mathbf{x})\sqrt{\omega(\mathbf{x})}) &= \nabla f(\mathbf{x})\sqrt{\omega(\mathbf{x})} + f(\mathbf{x})\nabla\sqrt{\omega(\mathbf{x})} \\ &= \nabla f(\mathbf{x})\sqrt{\omega(\mathbf{x})} + (2\bar{E} - \bar{G})\mathbf{x} f(\mathbf{x})\sqrt{\omega(\mathbf{x})}. \qquad (3.13)\end{aligned}$$

Then

$$\begin{aligned}(A_\omega f)\sqrt{\omega} &= \frac{G^{-1}}{\sqrt{2t}}\left(\left(\nabla + 2\bar{E}\mathbf{x}\right) f(\mathbf{x})\right)\sqrt{\omega(\mathbf{x})} \\ &= \frac{G^{-1}}{\sqrt{2t}}\left(\nabla(f(\mathbf{x})\sqrt{\omega(\mathbf{x})}) + \bar{G}\mathbf{x} f(\mathbf{x})\sqrt{\omega(\mathbf{x})}\right).\end{aligned}$$

Using the fact that $f\sqrt{\omega} \in \mathcal{S}(\mathbb{R}^d)$ and the properties of the Schwartz space, we see that $(A_\omega)_i f\sqrt{\omega} \in \mathcal{S}(\mathbb{R}^d)$. Therefore, $(A_\omega)_i f \in \mathcal{S}_\omega(\mathbb{R}^d)$.

We now prove the same for the raising operator $A^\dagger_\omega$. Indeed, using the relation (3.13), we get

$$\begin{aligned}((A^\dagger_\omega)_i f(\mathbf{x}))\sqrt{\omega(\mathbf{x})} &= \sqrt{\frac{t}{2}} G^{-1}\left[\left(-\nabla + 2(\bar{G} - \bar{E})\mathbf{x}\right) f(\mathbf{x})\right]\sqrt{\omega(\mathbf{x})} \\ &= \sqrt{\frac{t}{2}} G^{-1}\left(-\nabla(f(\mathbf{x})\sqrt{\omega(\mathbf{x})}) + \bar{G}\mathbf{x} f(\mathbf{x})\sqrt{\omega(\mathbf{x})}\right).\end{aligned}$$

Using the same reasoning as before, we observe that $((A^\dagger_\omega)_i f)\sqrt{\omega} \in \mathcal{S}(\mathbb{R}^d)$ and therefore $((A^\dagger_\omega)_i f) \in \mathcal{S}_\omega(\mathbb{R}^d)$. □

Finally, we can proceed to the proof of the t-adjointness.

Lemma 3.4.2

For all f, $g \in \mathcal{S}_\omega(\mathbb{R}^d)$, $i = 1 \dots d$,

1. $\langle (A_\omega^\dagger)_i f, g\rangle_\omega = t\langle f, (A_\omega)_i g\rangle_\omega.$ (3.14)

2. $\langle f, (A_\omega^\dagger)_i g\rangle_\omega = t\langle (A_\omega)_i f, g\rangle_\omega.$ (3.15)

Proof. Denote $\bar{E} = E^T E$, $\bar{G} = GG^T$. We first observe that

$$\frac{\partial \omega(\mathbf{x})}{\partial x_i} = \pi^{d/2}|\det(G)|\frac{\partial\left(\exp(2\mathbf{x}^T\bar{E}\mathbf{x} - \mathbf{x}^T\bar{G}\mathbf{x})\right)}{\partial x_i}$$
$$= \left(\sum_{j=1}^{d}(4\bar{E}_{ij} - 2\bar{G}_{ij})x_j\right)\omega(\mathbf{x}). \tag{3.16}$$

Let us now compute the following integral

$$\int_{\mathbb{R}^d}\frac{\partial f(\mathbf{x})}{\partial x_i}g(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x} = -\int_{\mathbb{R}^d} f(\mathbf{x})\frac{\partial(g(\mathbf{x})\omega(\mathbf{x}))}{\partial x_i}\mathrm{d}\mathbf{x},$$

where we used integration by parts and the fact that for $f, g \in \mathcal{S}_\omega(\mathbb{R}^d)$ the product $f(\mathbf{x})g(\mathbf{x})\omega(\mathbf{x})$ vanishes in the limit $x_i \to \pm\infty$. With the help of the expressions above, we proceed to the proof of the first equality.

$$\langle (GA_\omega^\dagger)_i f, g\rangle_\omega = \sqrt{\frac{t}{2}}\int_{\mathbb{R}^d}\left(-\frac{\partial f(\mathbf{x})}{\partial x_i} + \left(\sum_{j=1}^{d}(2\bar{G}_{ij} - 2\bar{E}_{ij})x_j\right) f(\mathbf{x})\right) g(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x}$$
$$= \sqrt{\frac{t}{2}}\left(\int_{\mathbb{R}^d} f(\mathbf{x})\frac{\partial(g(\mathbf{x})\omega(\mathbf{x}))}{\partial x_i}\mathrm{d}\mathbf{x}\right.$$
$$\left. + \int_{\mathbb{R}^d}\left(\sum_{j=1}^{d}(2\bar{G}_{ij} - 2\bar{E}_{ij})x_j\right) f(\mathbf{x})g(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x}\right)$$
$$= \sqrt{\frac{t}{2}}\int_{\mathbb{R}^d}\left(\partial_{x_i} + 2\sum_{j=1}^{d}\bar{E}_{ij}x_j\right) g(\mathbf{x})f(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x} = t\langle f, (GA_\omega)_i g\rangle_\omega,$$

where we used the relation (3.16) for the partial derivative of $\omega(\mathbf{x})$. Due to the bilinearity of the scalar product, the expression holds also for the operators A_ω, $A_\omega^\dagger$ themselves. Indeed, denote $A_{G,\omega}^\dagger = GA_\omega^\dagger$, $A_{G,\omega} = GA_\omega$ and $G^{\mathrm{inv}} = G^{-1}$. Then,

$$(A_\omega^\dagger)_i = \sum_{j=1}^{d} G_{ij}^{\mathrm{inv}}(A_{G,\omega}^\dagger)_j \quad \text{and} \quad (A_\omega)_i = \sum_{j=1}^{d} G_{ij}^{\mathrm{inv}}(A_{G,\omega})_j$$

Therefore, the first equality (3.14) follows from

$$\langle (A_\omega^\dagger)_i f, g\rangle_\omega = \sum_{j=1}^{d} G_{ij}^{\mathrm{inv}}\langle (A_{G,\omega}^\dagger)_j f, g\rangle_\omega = \sum_{j=1}^{d} G_{ij}^{\mathrm{inv}} t\langle f, (A_{G,\omega})_j g\rangle_\omega = t\langle f, (A_\omega)_i g\rangle_\omega.$$

Since the product $\langle\cdot,\cdot\rangle_\omega$ is commutative, the second equality can be obtained as

$$\langle f, (A_\omega^\dagger)_i g\rangle_\omega = \langle (A_\omega^\dagger)_i g, f\rangle_\omega = t\langle g, (A_\omega)_i f\rangle_\omega = t\langle (A_\omega)_i f, g\rangle_\omega \qquad \square$$

With this tool, we can now proceed to the estimation of the approximation error.

3.5. Approximation properties

In this section, we derive the estimations for the approximation error in $L^2_\omega(\mathbb{R}^d)$ when using generalized anisotropic Hermite functions. We first consider the estimate that is analogous to Theorem 2.3.1 for Hermite functions. We then extend it to a more general case.

Let us first consider the 1D case. Since in this case G and E are just numbers, we introduce a simplified notation $\{H^{\gamma,\varepsilon,t}_\ell\}_{\ell\in\mathbb{N}_0}$ for this case. We denote

$$U_M = \operatorname{span}\{H^{\gamma,\varepsilon,t}_\ell | \ell \le M-1\}$$

and consider the orthogonal projector P_M onto U_M given by

$$P_M f = \sum_{\ell<M} \frac{\langle f, H^{\gamma,\varepsilon,t}_\ell\rangle_\omega}{\langle H^{\gamma,\varepsilon,t}_\ell, H^{\gamma,\varepsilon,t}_\ell\rangle_\omega} H^{\gamma,\varepsilon,t}_\ell.$$

Our goal is to estimate the error $\|f - P_M f\|_\omega$. We now prove the estimate that is analogous to the result [Lub08, § III.1.1, Theorem 1.2].

Theorem 3.5.1

For every integer $s \le M$ and every function f in the weighted Schwartz space $\mathcal{S}_\omega(\mathbb{R})$,

$$\|f - P_M f\|_\omega \le \frac{t^{s/2}}{\sqrt{M(M-1)\dots(M-s+1)}}\|A^s_\omega f\|_\omega = \frac{t^{s/2}\sqrt{(M-s)!}}{\sqrt{M!}}\|A^s_\omega f\|_\omega.$$

Proof. We follow the flow of the proof [Lub08, § III.1.1, Theorem 1.2]. Using the result of Lemma 3.4.2 together with the properties (3.11), (3.15) of the raising operator, we get

$$\begin{aligned} f - P_M f &= \sum_{\ell\ge M} \frac{\langle f, H^{\gamma,\varepsilon,t}_\ell\rangle_\omega}{\langle H^{\gamma,\varepsilon,t}_\ell, H^{\gamma,\varepsilon,t}_\ell\rangle_\omega} H^{\gamma,\varepsilon,t}_\ell \\ &\overset{(3.11)}{=} \sum_{\ell\ge M} \frac{1}{\sqrt{\ell(\ell-1)\dots(\ell-s+1)}} \frac{\langle f, (A^\dagger_\omega)^s H^{\gamma,\varepsilon,t}_{\ell-s}\rangle_\omega}{\langle H^{\gamma,\varepsilon,t}_\ell, H^{\gamma,\varepsilon,t}_\ell\rangle_\omega} H^{\gamma,\varepsilon,t}_\ell \\ &\overset{(3.15)}{=} \sum_{\ell\ge M} \frac{t^s}{\sqrt{\ell(\ell-1)\dots(\ell-s+1)}} \frac{\langle A^s_\omega f, H^{\gamma,\varepsilon,t}_{\ell-s}\rangle_\omega}{\langle H^{\gamma,\varepsilon,t}_\ell, H^{\gamma,\varepsilon,t}_\ell\rangle_\omega} H^{\gamma,\varepsilon,t}_\ell. \end{aligned}$$

Recall that

$$\langle H^{\gamma,\varepsilon,t}_\ell, H^{\gamma,\varepsilon,t}_\ell\rangle_\omega = t^\ell \quad \forall\, \ell \in \mathbb{N}_0.$$

Using the orthogonality, we get

$$\|f - P_M f\|^2_\omega = \sum_{\ell\ge M} \frac{t^{2s}}{\ell(\ell-1)\dots(\ell-s+1)} \frac{|\langle A^s_\omega f, H^{\gamma,\varepsilon,t}_{\ell-s}\rangle_\omega|^2}{t^\ell}.$$

Therefore, we can now estimate

$$\begin{aligned}\|f - P_M f\|_\omega^2 &\le \frac{t^{2s}}{M(M-1)\dots(M-s+1)} \sum_{j\ge M-s} \frac{|\langle A_\omega^s f, H_j^{\gamma,\varepsilon,t}\rangle_\omega|^2}{t^{(j+s)}} \\ &\le \frac{t^{2s}}{M(M-1)\dots(M-s+1)} \sum_{j\ge 0} \frac{|\langle A_\omega^s f, H_j^{\gamma,\varepsilon,t}\rangle_\omega|^2}{t^{(j+s)}} \\ &= \frac{t^{s}}{M(M-1)\dots(M-s+1)} \|A_\omega^s f\|_\omega^2,\end{aligned}$$

where we used Parseval's identity for the case of an orthogonal basis. □

Now, when we have the 1D estimate, we can easily extend it to higher dimensions using the tensor-product approach. Let us consider the set of indexes

$$\mathcal{K}_{\mathbf{M}} = \{\boldsymbol{\ell} \in \mathbb{N}_0^d | \ell_i < M_i \ \forall i = 1\dots d\}$$

and the complementary set

$$\bar{\mathcal{K}}_{\mathbf{M}} = \{\boldsymbol{\ell} \in \mathbb{N}_0^d | \ell_i \ge M_i \ \forall i = 1\dots d\},$$

where the integer vector $\mathbf{M} = \begin{pmatrix} M_1 & \dots & M_d \end{pmatrix}$. The corresponding orthogonal projector $P_{\mathbf{M}}$ is given by

$$P_{\mathbf{M}} f = \sum_{\boldsymbol{\ell}\in\mathcal{K}_{\mathbf{M}}} \frac{\langle f, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega}{\langle H_{\boldsymbol{\ell}}^{G,E,t}, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega} H_{\boldsymbol{\ell}}^{G,E,t}.$$

We are now ready to estimate the approximation error $\|f - P_{\mathbf{M}} f\|_\omega$ for a multivariate function $f \in \mathcal{S}_\omega(\mathbb{R}^d)$.

Theorem 3.5.2

For every integer vector $\boldsymbol{s}$ such that $s_i \le M_i$ for all i and every function f in the weighted Schwartz space $\mathcal{S}_\omega(\mathbb{R}^d)$,

$$\|f - P_{\mathbf{M}} f\|_\omega \le \frac{t^{|\boldsymbol{s}|/2}\sqrt{(\mathbf{M}-\boldsymbol{s})!}}{\sqrt{\mathbf{M}!}} \|A_\omega^{\boldsymbol{s}} f\|_\omega.$$

Proof. We follow the flow of the proof [Lub08, § III.1.1, Theorem 1.2]. Using the result of Lemma 3.4.1 with the properties (3.11), (3.15) of the raising operator, we get

$$\begin{aligned} f - P_{\mathbf{M}} f &= \sum_{\boldsymbol{\ell}\in\bar{\mathcal{K}}_{\mathbf{M}}} \frac{\langle f, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega}{\langle H_{\boldsymbol{\ell}}^{G,E,t}, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega} H_{\boldsymbol{\ell}}^{G,E,t} \\ &\overset{(3.11)}{=} \sum_{\boldsymbol{\ell}\in\bar{\mathcal{K}}_{\mathbf{M}}} \frac{1}{\sqrt{\ell_1\dots(\ell_1-s_1+1)\dots\ell_d\dots(\ell_d-s_d+1)}} \frac{\langle f, (A_\omega^\dagger)^{\boldsymbol{s}} H_{\boldsymbol{\ell}-\boldsymbol{s}}^{\gamma,\varepsilon,t}\rangle_\omega}{\langle H_{\boldsymbol{\ell}}^{G,E,t}, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega} H_{\boldsymbol{\ell}}^{G,E,t} \\ &\overset{(3.15)}{=} \sum_{\boldsymbol{\ell}\in\bar{\mathcal{K}}_{\mathbf{M}}} \frac{t^{|\boldsymbol{s}|}}{\sqrt{\ell_1\dots(\ell_1-s_1+1)\dots\ell_d\dots(\ell_d-s_d+1)}} \frac{\langle A_\omega^{\boldsymbol{s}} f, H_{\boldsymbol{\ell}-\boldsymbol{s}}^{\gamma,\varepsilon,t}\rangle_\omega}{\langle H_{\boldsymbol{\ell}}^{G,E,t}, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_\omega} H_{\boldsymbol{\ell}}^{G,E,t}. \end{aligned}$$

Using the orthogonality, we get

$$\begin{aligned}\|f - P_{\mathbf{M}} f\|_\omega^2 &= \sum_{\boldsymbol{\ell}\in\bar{\mathcal{K}}_{\mathbf{M}}} \frac{t^{2|\boldsymbol{s}|}}{\ell_1 \dots (\ell_d - s_d + 1)} \frac{|\langle A_\omega^{\boldsymbol{s}} f, H_{\boldsymbol{\ell}-\boldsymbol{s}}^{G,E,t}\rangle_\omega|^2}{t^{|\boldsymbol{\ell}|}} \\ &\leq \frac{t^{2|\boldsymbol{s}|}}{M_1(M_1-1)\dots(M_1-s_1+1)\dots M_d\dots(M_d-s_d+1)} \sum_{|\boldsymbol{j}|\geq 0} \frac{|\langle A_\omega^{\boldsymbol{s}} f, H_{\boldsymbol{j}}^{G,E,t}\rangle_\omega|^2}{t^{(|\boldsymbol{j}|+|\boldsymbol{s}|)}} \\ &= \frac{t^{|\boldsymbol{s}|}}{M_1(M_1-1)\dots(M_1-s_1+1)\dots M_d\dots(M_d-s_d+1)} \|A_\omega^{\boldsymbol{s}} f\|_\omega^2,\end{aligned}$$

where we used Parseval's identity for the case of an orthogonal basis. □

We now move a bit further and generalize the obtained result to a more general space that is defined based on the raising operator $A_\omega^\dagger$. In order to be able to define the corresponding norm, we first have to prove that the operator has a trivial kernel.

Lemma 3.5.1

The raising operator $A_\omega^\dagger$ has a trivial kernel in $L_\omega^2(\mathbb{R}^d)$.

Proof. Consider $f \in L_\omega^2(\mathbb{R}^d)$. Indeed,

$$\begin{aligned}A_\omega^\dagger f(\mathbf{x}) = 0 &\Leftrightarrow \sqrt{\frac{t}{2}} G^{-1}\left(-\nabla + 2(GG^T - E^T E)\mathbf{x}\right) f(\mathbf{x}) = 0 \\ &\Leftrightarrow \nabla f(\mathbf{x}) = 2(GG^T - E^T E)\mathbf{x} f(\mathbf{x}).\end{aligned}$$

Hence, the functions $\tilde{f}$ for which $A_\omega^\dagger \tilde{f} = 0$ have the form

$$\tilde{f}(\mathbf{x}) = f(\mathbf{0})\mathrm{e}^{\mathbf{x}^T(GG^T - E^T E)\mathbf{x}}.$$

Let us now check if $\tilde{f}$ is in $L_\omega^2(\mathbb{R}^d)$:

$$\begin{aligned}\int_{\mathbb{R}^d} |\tilde{f}(\mathbf{x})|^2 \omega(\mathbf{x}) \mathrm{d}\mathbf{x} &= \int_{\mathbb{R}^d} \pi^{d/2} f(\mathbf{0}) \mathrm{e}^{2\mathbf{x}^T(GG^T - E^T E)\mathbf{x}} \mathrm{e}^{2\mathbf{x}^T E^T E\mathbf{x} - \mathbf{x}^T GG^T \mathbf{x}} |\det(G)| \mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^d} \pi^{d/2} f(\mathbf{0}) \mathrm{e}^{\mathbf{x}^T GG^T \mathbf{x}} |\det(G)| \mathrm{d}\mathbf{x} = \infty.\end{aligned}$$

Hence, $\tilde{f} \notin L_\omega^2(\mathbb{R}^d)$. □

We now prove a more general convergence result.

3.5.1. Generalized estimate

Let us first consider the 1D case. We denote

$$W_\omega^m(\mathbb{R}) = \left\{ f \,\middle|\, (A_\omega^\dagger)^k f \in L_\omega^2(\mathbb{R}), 0 \leq k \leq m \right\}.$$

The norm of $W_\omega^m(\mathbb{R})$ is given by:

$$\|f\|_{W_\omega^m(\mathbb{R})} = \left(\sum_{k=0}^{m} \|(A_\omega^\dagger)^k f\|_\omega^2 \right)^{\frac{1}{2}}.$$

Since the operator $A_\omega^\dagger$ is linear and we have proved in Lemma 3.5.1 that the operator $A_\omega^\dagger$ has a trivial kernel, it is a valid norm. We refer to $\|f\|_{W_\omega^m(\mathbb{R})}$ as $\|f\|_{m,\omega}$. We now estimate the approximation error in this norm.

Theorem 3.5.3

$\forall f \in \mathcal{S}_\omega(\mathbb{R})$ with an integer $0 \le m \le r$ the following estimate holds

$$\|f - P_M f\|_{m,\omega} \le C(m,r) t^{r/2} \max\{t^{m/2}, 1\} M^{(m-r)/2} \|A_\omega^r f\|_\omega \qquad \forall M \ge 2r-1$$

with

$$C(m,r) = \sqrt{m+1} \cdot 2^{r/2}.$$

Proof. Consider an integer $0 \le k \le m$ and $f \in \mathcal{S}_\omega(\mathbb{R})$. Since all basis functions $H_\ell^{\gamma,\varepsilon,t} \in \mathcal{S}_\omega(\mathbb{R}^d)$, the tail $f - P_M f \in \mathcal{S}_\omega(\mathbb{R})$. Then, according to Lemma 3.4.1,

$$(A_\omega^\dagger)^k (f - P_M f) \in \mathcal{S}_\omega(\mathbb{R}).$$

Therefore, using the Parseval's identity, Lemma 3.4.1, and with the properties (3.12), (3.14) of the ladder operators, we get

$$\begin{aligned}
\|(A_\omega^\dagger)^k (f - P_M f)\|_\omega^2 &= \sum_{\ell \ge 0} \frac{|\langle (A_\omega^\dagger)^k (f - P_M f), H_\ell^{\gamma,\varepsilon,t} \rangle_\omega|^2}{t^\ell} \\
&\overset{(3.14)}{=} \sum_{\ell \ge 0} \frac{t^{2k} |\langle f - P_M f, A_\omega^k H_\ell^{\gamma,\varepsilon,t} \rangle_\omega|^2}{t^\ell} \\
&\overset{(3.12)}{=} \sum_{\ell \ge k} \ell(\ell-1)\dots(\ell-k+1) \frac{t^{2k} |\langle f - P_M f, H_{\ell-k}^{\gamma,\varepsilon,t} \rangle_\omega|^2}{t^\ell}
\end{aligned}$$

Due to orthogonality,

$$\begin{aligned}
\|(A_\omega^\dagger)^k (f - P_M f)\|_\omega^2 &= \sum_{\ell \ge M+k} \ell(\ell-1)\dots(\ell-k+1) \frac{t^{2k} |\langle f, H_{\ell-k}^{\gamma,\varepsilon,t} \rangle_\omega|^2}{t^\ell} \\
&= \sum_{j \ge M} (j+1)\dots(j+k) \frac{t^{2k} |\langle f, H_j^{\gamma,\varepsilon,t} \rangle_\omega|^2}{t^{(j+k)}} \\
&\overset{(3.11)}{=} \sum_{j \ge M} \frac{(j+1)\dots(j+k)}{j(j-1)\dots(j-r+1)} \frac{t^k |\langle f, (A_\omega^\dagger)^r H_{j-r}^{\gamma,\varepsilon,t} \rangle_\omega|^2}{t^j} \\
&\overset{(3.15)}{=} \sum_{j \ge M} \frac{(j+1)\dots(j+k)}{j(j-1)\dots(j-r+1)} \frac{t^{k+2r} |\langle A_\omega^r f, H_{j-r}^{\gamma,\varepsilon,t} \rangle_\omega|^2}{t^j}
\end{aligned}$$

Recall that we required $M \ge 2r-1$. In this case, for all $j \ge M$

$$j - r + 1 > j - r + \frac{1}{2} = \frac{2j - 2r + 1}{2} \ge \frac{j + 2r - 1 - 2r + 1}{2} = \frac{j}{2}.$$

Moreover, since $k \le m \le r$

$$j - k + 1 > j - r + 1 > r > k \qquad \forall 0 \le k \le m, \forall n \ge N.$$

Let us combine the terms from the numerator with the first k terms of the denominator:

$$\frac{(j+1)\dots(j+k)}{j(j-1)\dots(j-r+1)} = \frac{(j+1)\dots(j+k)}{j(j-1)\cdot\dots\cdot(j-k+1)} \cdot \frac{1}{(j-k)\cdot\dots\cdot(j-r+1)}$$
$$= \frac{j+k}{j}\cdot\frac{j+k-1}{j-1}\cdot\dots\cdot\frac{j+1}{j-k+1}\cdot\frac{1}{(j-k)\cdot\dots\cdot(j-r+1)}$$
$$= \left(1+\frac{k}{j}\right)\cdot\left(1+\frac{k}{j-1}\right)\cdot\dots\cdot\left(1+\frac{k}{j-k+1}\right)\cdot\frac{1}{(j-k)\cdot\dots\cdot(j-r+1)}$$
$$\le \left(1+\frac{k}{j-k+1}\right)^k\cdot\frac{1}{(j-r+1)^{r-k}} \le \frac{2^r}{j^{r-k}}.$$

Therefore for all $0 \le k \le m$,

$$\begin{aligned}\|(A_\omega^\dagger)^k(f-P_Mf)\|_\omega^2 &\le \frac{2^r}{M^{r-k}}\sum_{j\ge M}\frac{t^{k+2r}|\langle A_\omega^r f, H_{j-r}^{\gamma,\varepsilon,t}\rangle_\omega|^2}{t^j}\\ &\le \frac{2^r t^{k+r}}{M^{r-k}}\|A_\omega^r f\|_\omega^2,\end{aligned}$$

Hence,

$$\begin{aligned}\|(f-P_Mf)\|_{m,\omega}^2 = \sum_{k=0}^m\|(A_\omega^\dagger)^k(f-P_Nf)\|_\omega^2 &\le 2^r t^r\max\{t^m,1\}\sum_{k=0}^m\frac{1}{M^{r-k}}\|A_\omega^r f\|_\omega^2\\ &\le (m+1)\frac{2^r t^r\max\{t^m,1\}}{M^{r-m}}\|A_\omega^r f\|_\omega^2. \qquad \square\end{aligned}$$

Note, that for the case $m = 0$, this estimate matches the one derived in Theorem 3.5.1. Indeed, for $M \ge r$ we have the following estimate from Theorem 3.5.1

$$\|f-P_Mf\|_\omega \le \frac{t^{r/2}\sqrt{(M-r)!}}{\sqrt{M!}}\|A_\omega^r f\|_\omega.$$

If we now restrict the values of r to the ones satisfying $M > 2r-1$, then

$$\frac{(M-r)!}{M!} = \frac{1}{M(M-1)\dots(M-r+1)} \le \frac{1}{(M-r+1)^r} \le \frac{2^r}{M^r}.$$

Therefore, in this case the estimate takes the form

$$\|f-P_Mf\|_\omega \le \frac{t^{r/2}2^{r/2}}{M^{r/2}}\|A_\omega^r f\|_\omega,$$

which matches the estimate from Theorem 3.5.3 for $m = 0$.

In a similar fashion as before, namely, using the tensor-product approach, we can extend the estimate for the multidimensional case. We first extend the notions of the space and the corresponding norm. Denote

$$W_\omega^{\mathbf{m}}(\mathbb{R}^d) = \left\{f \,\middle|\, (A_\omega^\dagger)^{\mathbf{k}}f \in L_\omega^2(\mathbb{R}^d), 0 \le \mathbf{k} \le \mathbf{m}\right\}.$$

The norm of $W_\omega^{\mathbf{m}}(\mathbb{R})$ is given by:

$$\|f\|_{W_\omega^{\mathbf{m}}(\mathbb{R}^d)} = \left(\sum_{k_1=0}^{m_1}\dots\sum_{k_d=0}^{m_d}\|(A_\omega^\dagger)^{\mathbf{k}}f\|_\omega^2\right)^{\frac{1}{2}}.$$

We later refer to $\|f\|_{W^{\mathbf{m}}_{\omega}(\mathbb{R})}$ as $\|f\|_{\mathbf{m},\omega}$. Let us now prove the final approximation error estimate in this space.

Theorem 3.5.4: Approximation error

Consider $f \in \mathcal{S}_{\omega}(\mathbb{R}^d)$ and integer vectors $0 \leq \mathbf{m} \leq \mathbf{r}$ such that $2r_i - 1 \leq M_i$ for all $i = 1 \dots d$. Then, the following estimate holds

$$\|f - P_{\mathbf{M}} f\|_{\mathbf{m},\omega} \leq C(\mathbf{m}, \mathbf{r}) t^{|\mathbf{r}|/2} \max\{t^{|\mathbf{m}|/2}, 1\} \mathbf{M}^{(\mathbf{m}-\mathbf{r})/2} \|A^{\mathbf{r}}_{\omega} f\|_{\omega}.$$

with

$$C(\mathbf{m}, \mathbf{r}) = \sqrt{(m_1 + 1) \dots (m_d + 1)} \cdot 2^{|\mathbf{r}|/2}.$$

Proof. As in the 1D case, due to the fact that all basis functions $H^{G,E,t}_{\boldsymbol{\ell}} \in \mathcal{S}_{\omega}(\mathbb{R}^d)$, the tail $f - P_{\mathbf{M}} f \in \mathcal{S}_{\omega}(\mathbb{R}^d)$. Then, according to the Lemma 3.4.1,

$$(A^{\dagger}_{\omega})^{\mathbf{k}} (f - P_{\mathbf{M}} f) \in \mathcal{S}_{\omega}(\mathbb{R}^d).$$

Therefore, using the Parseval's identity, Lemma 3.4.1 and with the properties (3.12), (3.14) of the ladder operators, we get

$$\begin{aligned}
\|(A^{\dagger}_{\omega})^{\mathbf{k}} (f - P_{\mathbf{M}} f)\|^2_{\omega} &= \sum_{\boldsymbol{\ell} \in \mathbb{N}^d_0} \frac{|\langle (A^{\dagger}_{\omega})^{\mathbf{k}} (f - P_{\mathbf{M}} f), H^{G,E,t}_{\boldsymbol{\ell}} \rangle_{\omega}|^2}{\langle H^{G,E,t}_{\boldsymbol{\ell}}, H^{G,E,t}_{\boldsymbol{\ell}} \rangle_{\omega}} \\
&\overset{(3.14)}{=} \sum_{\boldsymbol{\ell} \in \mathbb{N}^d_0} \frac{t^{2|\mathbf{k}|} |\langle f - P_{\mathbf{M}} f, A^{\mathbf{k}}_{\omega} H^{G,E,t}_{\boldsymbol{\ell}} \rangle_{\omega}|^2}{t^{|\boldsymbol{\ell}|}} \\
&\overset{(3.12)}{=} \sum_{\boldsymbol{\ell} - \mathbf{k} \in \mathbb{N}^d_0} \ell_1 \dots (\ell_1 - k_1 + 1) \dots \ell_d \dots (\ell_d - k_d + 1) \frac{t^{2|\mathbf{k}|} |\langle f - P_{\mathbf{M}} f, H^{G,E,t}_{\boldsymbol{\ell} - \mathbf{k}} \rangle_{\omega}|^2}{t^{|\boldsymbol{\ell}|}} \\
&= \sum_{\boldsymbol{\ell} - \mathbf{k} \in \mathbb{N}^d_0} \frac{\boldsymbol{\ell}!}{(\boldsymbol{\ell} - \mathbf{k})!} \frac{t^{2|\mathbf{k}|} |\langle f - P_{\mathbf{M}} f, H^{G,E,t}_{\boldsymbol{\ell} - \mathbf{k}} \rangle_{\omega}|^2}{t^{|\boldsymbol{\ell}|}}
\end{aligned}$$

Due to orthogonality,

$$\begin{aligned}
\|(A^{\dagger}_{\omega})^{\mathbf{k}} (f - P_{\mathbf{M}} f)\|^2_{\omega} &= \sum_{\boldsymbol{\ell} - \mathbf{k} \in \bar{\mathcal{K}}_{\mathbf{M}}} \frac{\boldsymbol{\ell}!}{(\boldsymbol{\ell} - \mathbf{k})!} \frac{t^{2|\mathbf{k}|} |\langle f, H^{G,E,t}_{\boldsymbol{\ell} - \mathbf{k}} \rangle_{\omega}|^2}{t^{|\boldsymbol{\ell}|}} \\
&= \sum_{\mathbf{j} \in \bar{\mathcal{K}}_{\mathbf{M}}} \frac{(\mathbf{j} + \mathbf{k})!}{\mathbf{j}!} \frac{t^{2|\mathbf{k}|} |\langle f, H^{G,E,t}_{\mathbf{j}} \rangle_{\omega}|^2}{t^{(|\mathbf{j}| + |\mathbf{k}|)}} \\
&\overset{(3.11)}{=} \sum_{\mathbf{j} \in \bar{\mathcal{K}}_{\mathbf{M}}} \frac{(\mathbf{j} + \mathbf{k})!}{\mathbf{j}!} \frac{(\mathbf{j} - \mathbf{r})!}{\mathbf{j}!} \frac{t^{|\mathbf{k}|} |\langle f, (A^{\dagger}_{\omega})^{\mathbf{r}} H^{G,E,t}_{\mathbf{j} - \mathbf{r}} \rangle_{\omega}|^2}{t^{|\mathbf{j}|}} \\
&\overset{(3.15)}{=} \sum_{\mathbf{j} \in \bar{\mathcal{K}}_{\mathbf{M}}} \frac{(\mathbf{j} + \mathbf{k})!}{\mathbf{j}!} \frac{(\mathbf{j} - \mathbf{r})!}{\mathbf{j}!} \frac{t^{|\mathbf{k}| + 2|\mathbf{r}|} |\langle A^{\mathbf{r}}_{\omega} f, H^{G,E,t}_{\mathbf{j} - \mathbf{r}} \rangle_{\omega}|^2}{t^{|\mathbf{j}|}}
\end{aligned}$$

Recall that in 1D

$$\frac{(j+1) \dots (j+k)}{j(j-1) \dots (j-r+1)} = \frac{(j+k)!}{j!} \frac{(j-r)!}{j!} \leq \frac{2^r}{j^{r-k}}$$

under the conditions of the theorem. Therefore for all $0 \le \mathbf{k} \le \mathbf{m}$,

$$\begin{aligned}\|(A_\omega^\dagger)^{\mathbf{k}}(f - P_{\mathbf{M}}f)\|_\omega^2 &\le \frac{2^{|\mathbf{r}|}}{\mathbf{M}^{\mathbf{r}-\mathbf{k}}} \sum_{\mathbf{j}\in\bar{\mathcal{K}}_{\mathbf{M}}} \frac{t^{|\mathbf{k}|+2|\mathbf{r}|}|\langle A_\omega^{\mathbf{r}} f, H_{\mathbf{j}-\mathbf{r}}^{G,E,t}\rangle_\omega|^2}{t^{|\mathbf{j}|}} \\ &\le \frac{2^{|\mathbf{r}|}t^{|\mathbf{k}|+|\mathbf{r}|}}{\mathbf{M}^{\mathbf{r}-\mathbf{k}}}\|A_\omega^{\mathbf{r}} f\|_\omega^2,\end{aligned}$$

Hence,

$$\begin{aligned}\|(f - P_{\mathbf{M}}f)\|_{\mathbf{m},\omega}^2 &= \sum_{k_1=0}^{m_1} \dots \sum_{k_d=0}^{m_d} \|(A_\omega^\dagger)^{\mathbf{k}}(f - P_{\mathbf{M}}f)\|_\omega^2 \\ &\le 2^{|\mathbf{r}|}t^{|\mathbf{r}|}\max\{t^{|\mathbf{m}|}, 1\} \sum_{k_1=0}^{m_1} \dots \sum_{k_d=0}^{m_d} \frac{1}{\mathbf{M}^{\mathbf{r}-\mathbf{k}}}\|A_\omega^{\mathbf{r}} f\|_\omega^2 \\ &\le ((m_1+1)\dots(m_d+1))\frac{2^{|\mathbf{r}|}t^{|\mathbf{r}|}\max\{t^{|\mathbf{m}|}, 1\}}{\mathbf{M}^{\mathbf{r}-\mathbf{m}}}\|A_\omega^{\mathbf{r}} f\|_\omega^2. \qquad \square\end{aligned}$$

Now, with the appropriate function spaces at hand and a good understanding of the interpolation properties of the generalized anisotropic Hermite basis, we can move on to deriving numerical methods based on it.

Part II

Stable interpolation with isotropic and anisotropic Gaussians

4. Interpolation problem

In this part of the thesis, we use the basis developed in part I for the interpolation problem. The content of this part of the thesis is an extended version of the paper *"Stable interpolation with isotropic and anisotropic Gaussians using Hermite generating function"* by Kormann, Lasser, & Yurova that has been accepted to the SIAM Journal on Scientific Computing on 18.09.2019 [KLY19] (preprint available at `https://arxiv.org/abs/1905.09542`). In particular, the present chapter, chapters 6, 8, 10 and 12 reproduce the results of the paper with some additional details, whereas the other chapters provide additional information on the method.

Multivariate interpolation is a topic that is relevant for a vast number of applications. Gaussian radial basis functions (Gaussian RBFs) are a class of functions for which interpolation generalizes to higher dimensions in a simple way while yielding spectral accuracy [FHW12]. However, it is known that rather small values of the *shape parameter* $\varepsilon > 0$ (the width of the Gaussians) are often required. In this case, the Gaussians become increasingly flat, and the interpolation matrix becomes ill-conditioned. This problem has been extensively studied in the literature (see the review [FF15] by Fornberg & Flyer and [Tar85] for Tarwater's description of this phenomenon in 1985). It has been quantified by Fornberg & Zuev [FZ07], that the eigenvalues of the interpolation matrix are proportional to powers of the shape parameter, causing the notorious ill-conditioning in the flat limit regime $\varepsilon \to 0$.

A direct collocation solution of the interpolation problem, which is referred to as RBF-Direct in the literature, computes the expansion coefficients of the Gaussian interpolant by inverting the collocation matrix and then evaluating the expansion. Several algorithms have been proposed to stabilize this procedure in the flat limit regime, see [FW04, FP07, FLF11, FM12, LLHF13, DMS13, FM15, FLP13, RFK16, WF17]. A common idea of many of the stabilization algorithms—including the one proposed in this part of the thesis—is to evaluate the interpolant in a sequence of well-conditioned steps by a transformation to a different basis so that the ill-conditioning is isolated in a diagonal matrix that can be inverted analytically.

4.1. The new HermiteGF stabilization approach

In this part of the thesis, we propose a stabilizing expansion of isotropic Gaussian functions, later referred to as HermiteGF expansion, built on the exponential generating function of the classic Hermite polynomials. For certain classes of functions, anisotropic Gaussians yield improved accuracy as shown in [BDL10].

To include these cases in our description, we use an anisotropic generating function recently obtained by Dietert, Keller, & Troppmann [DKT17] as well as by Hagedorn [Hag15] and generalize our HermiteGF expansion to the anisotropic case. We also propose and analyze a novel cut-off criterion, that contrary to the existing ones does not only account for the diagonal contributions of the stabilization but measures the impact of the full stabilization basis.

4.2. Previous stabilization approaches

The first stabilization method was the Contour–Padé approximation proposed by Fornberg & Wright for multiquadrics [FW04]. The authors later enhanced it to the RBF-RA stabilization algorithm [WF17], which has a wide applicability but is limited to a small number of nodes. Later Fornberg & Piret [FP07] proposed the so-called RBF-QR method for stable interpolation with Gaussians on the sphere by expanding the Gaussians in spherical harmonics. The method has been extended to more general domains in one to three dimensions by Fornberg, Larsson, & Flyer [FLF11]. This expansion is based on a combination of Chebyshev polynomials and spherical harmonics. This method will be referred to as Chebyshev-QR in this thesis. The technique has also been used for the stable computation of RBF-generated finite differences by Larsson, Lehto, Heryudono, & Fornberg [LLHF13]. Fornberg, Lehto, & Powell [FLP13] developed an alternative stabilization technique for the same problem. To treat complex domains, the Chebyshev-QR method has been combined with a partition-of-unity approach by Larsson, Shcherbakov, & Heryudono [LSH17]. Fasshauer & McCourt [FM12] have developed another RBF-QR method, called Gauss-QR, that relies on a Mercer expansion of the Gaussian kernel. The basis transformation involves scaled Hermite polynomials. De Marchi and Santin [DMS13] considered a different construction of a stable basis based on a factorization of the kernel matrix for general radial kernels. Our new basis is similar to the one in [FM12] with the difference that it can be extended to the interpolation with anisotropic Gaussians. Moreover, the generating function framework enables us to derive a new cut-off criterion that accounts for the full Hermite basis effect. We note that a multiscale analysis as provided by Griebel, Rieger, & Zwicknagl [GRZ15] might allow to estimate the tail of the Mercer expansion of a Gaussian kernel and subsequently to derive an alternative Mercer series based truncation criterion (see [FM15, Remark 13.12]).

4.3. The interpolation problem

Before we dive into the discussion of the family of RBF-QR methods, let us formalize the interpolation problem. Given a set $\{\phi_k(\cdot)\}_{k=1}^N$ of basis functions and the values $\{f_i\}$ of the function f at points $\{\mathbf{x}_i^{\mathrm{col}}\}_{i=1}^N$ we seek to find an interpolant of the following form,

$$s(\mathbf{x}) = \sum_{k=1}^{N} \alpha_k \phi_k(\mathbf{x}), \tag{4.1}$$

such that it satisfies the N collocation conditions,

$$s(\mathbf{x}_i^{\mathrm{col}}) = f_i \quad \text{for} \quad i = 1, \ldots, N.$$

The straightforward approach is to find the coefficients $\{\alpha_i\}$ as a solution of the linear system,

$$\Phi^{\mathrm{col}} \boldsymbol{\alpha} = \boldsymbol{f}, \quad \text{with} \quad \Phi_{ij}^{\mathrm{col}} = \phi_j(\mathbf{x}_i^{\mathrm{col}}). \tag{4.2}$$

The matrix Φ^{col} is called *collocation matrix*. Then, the interpolant (4.1) can be evaluated at any point of the domain.

In this thesis, we consider Gaussian radial basis functions (isotropic Gaussians)

$$\phi_k(\mathbf{x}) = \exp(-\varepsilon^2 \|\mathbf{x} - \mathbf{x}_k\|^2)$$

with the shape parameter $\varepsilon > 0$ and anisotropic Gaussians

$$\phi_k(\mathbf{x}) = \exp(-(\mathbf{x} - \mathbf{x}_k)^T E^T E(\mathbf{x} - \mathbf{x}_k))$$

with the invertible shape matrix E. As for the other types of RBFs the collocation matrix gets severely ill-conditioned for small values of the shape parameter ε or the elements of the shape matrix E. We will first revise the existing stabilization methods that allow to tackle the ill-conditioning in the isotropic case. We then move on to derive a new algorithm that can not only be used for isotropic Gaussians, but also is easily extended to the case of the anisotropic Gaussians.

5. RBF-QR methods

Before deriving the new HermiteGF stabilization method, we introduce the general framework of RBF-QR methods and review the existing versions. The main idea of the RBF-QR methods is to expand each RBF in a more stable basis, that is different for every specific RBF-QR method. After the expansion, the terms yielding ill-conditioning should be confined in the expansion coefficients. After that, a preconditioner of a certain form is used in order to tackle those terms analytically. In this chapter, we first summarize the general workflow of RBF-QR methods in section 5.1. We then take a look at two specific expansions of RBFs that yield the two existing RBF-QR methods that we will use as a reference: the Chebyshev-QR method is described in section 5.2 and the Gauss-QR method in section 5.3.

5.1. General algorithm

In this section, we consider the general flow of RBF-QR methods. We focus on the one-dimensional case, since all steps are identical for any number of dimensions. The main idea of the methods from the RBF-QR family is to do a basis transformation to a more stable basis that still spans the same space. After this representation in the stable basis is found, the RBF interpolant can be evaluated in a sequence of well-conditioned steps. RBF-QR algorithms start off with an expansion of RBFs $\{\phi_k\}_{k=1}^N$ centered at points $\{x_k\}_{k=1}^N$ in a certain basis $\{\zeta_j\}_{j\in\mathbb{N}_0}$ that has the following form for all $x \in \mathbb{R}$

$$\phi_k(x) = \sum_{\ell\in\mathbb{N}_0} d_\ell c_\ell(x_k)\zeta_\ell(x) \tag{5.1}$$

where $d_\ell > 0$, $d_\ell \xrightarrow[\ell\to\infty]{} 0$, $\{x_k\}_{k=1}^N$ are the centers of the RBFs, $c_\ell(x_k) \in \mathbb{R}$. We will further refer to (5.1) as *RBF-expansion* and to $\{\zeta_\ell\}$ as *expansion functions*. At this point, the small ill-conditioned terms should be isolated in the coefficients $\{d_\ell\}$. Then, the expansion (5.1) for all RBFs can be written as follows:

$$\begin{pmatrix} \phi_1(x) \\ \phi_2(x) \\ \cdots \\ \phi_N(x) \end{pmatrix} = \begin{pmatrix} c_1(x_1) & c_2(x_1) & \cdots & c_M(x_1) & \cdots \\ c_1(x_2) & c_2(x_2) & \cdots & c_M(x_2) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_1(x_N) & c_2(x_N) & \cdots & c_M(x_N) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix} \begin{pmatrix} d_1 & 0 & \cdots & \cdots \\ 0 & \ddots & \cdots & \cdots \\ 0 & \cdots & d_M & \cdots \\ 0 & \cdots & \cdots & \ddots \end{pmatrix} \begin{pmatrix} \zeta_1(x) \\ \zeta_2(x) \\ \cdots \\ \zeta_M(x) \\ \cdots \end{pmatrix} \tag{5.2}$$

We denote the row vector of Gaussians with center points X^{cen} evaluated at $x \in \mathbb{R}$

$$\Phi(x, X^{\text{cen}}) = \big(\phi_1(x), \ldots, \phi_N(x)\big)$$

to keep it consistent with the definition of the collocation matrix Φ^{col} from the interpolation problem (4.2). And, in the same fashion, we denote as

$$Z(x) = \big(\zeta_1(x), \zeta_2(x), \ldots\big)$$

the vector containing the values of the expansion functions at a point $x \in \mathbb{R}$. Then the expression (5.2) takes the form

$$\Phi(x)^T = CDZ(x)^T,$$

with an $N \times \infty$ matrix C and an infinite-dimensional diagonal matrix D.

Recall that the goal is to find a basis $\{\psi_j\}$ spanning the same space as $\{\phi_k\}$ but yielding a better conditioned collocation matrix. In particular, we need an invertible matrix X such that $X^{-1}\Phi(x)^T$ is better conditioned. The idea is to perform a QR-decomposition on $C = QR$, hence the name of the method, and split away the upper left $N \times N$ blocks of resulting matrices

$$\Phi(x)^T = CDZ(x)^T = Q\begin{pmatrix} R_1 & R_2 \end{pmatrix}\begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} Z(x)^T,$$

where R_1 and D_1 are the $N \times N$ upper left blocks of matrices R and D and R_2 and D_2 contain the remaining entries. Consider $X = QR_1D_1$. The new basis can be formed as:

$$\begin{aligned}\Psi(x)^T &= X^{-1}\Phi(x) = D_1^{-1}R_1^{-1}Q^{\mathrm{H}}\Phi(x)^T = D_1^{-1}R_1^{-1}Q^{\mathrm{H}}Q\begin{pmatrix} R_1D_1 & R_2D_2 \end{pmatrix} Z(x)^T \\ &= \begin{pmatrix} \mathrm{Id} & D_1^{-1}R_1^{-1}R_2D_2 \end{pmatrix} Z(x)^T = Z_1(x)^T + (D_1^{-1}R_1^{-1}R_2D_2)Z_2(x)^T, \qquad (5.3)\end{aligned}$$

where $Z_1(x)$ contains the values of the first N expansion functions at the point $x \in \mathbb{R}$ and $Z_2(x)$ contains the rest. The action of D_1^{-1} and D_2 can be computed as the Hadamard product with

$$\tilde{D}_{ij} = \frac{D_{j+N,j+N}}{D_{ii}} \quad \text{with} \quad i \in \{1, \ldots, N\},\ j \geq 1.$$

Hence, the effect of the diagonal matrices D_1^{-1} and D_2 on $R_1^{-1}R_2$ is computed *analytically* and therefore under/overflow can be avoided. That is why, despite the harmful effects that are contained in D, the term $D_1^{-1}R_1^{-1}R_2D_2$ does not yield ill-conditioning.

Now, when we have the stable basis, the interpolation problem (4.2) takes the form:

$$\Psi^{\text{col}}\lambda = \begin{pmatrix} \Psi(x_1^{\text{col}}) \\ \Psi(x_2^{\text{col}}) \\ \cdots \\ \Psi(x_N^{\text{col}}) \end{pmatrix} \lambda = f$$

Using the relation (5.3) we get:

$$\begin{pmatrix} \Psi(x_1^{\text{col}}) \\ \cdots \\ \Psi(x_N^{\text{col}}) \end{pmatrix} = \begin{pmatrix} Z(x_1^{\text{col}}) \\ \cdots \\ Z(x_N^{\text{col}}) \end{pmatrix} \begin{pmatrix} \mathrm{Id} \\ (D_1^{-1}R_1^{-1}R_2D_2)^T \end{pmatrix}$$

Denote $R_D = D_1^{-1} R_1^{-1} R_2 D_2$, the left $N \times N$ block of the matrix $\begin{pmatrix} Z(x_1^{\mathrm{col}}) \\ \dots \\ Z(x_N^{\mathrm{col}}) \end{pmatrix}$ as Z_1 and the $N \times \infty$ matrix containing the remaining entries as Z_2. Then, we can rewrite the system above as

$$(Z_1 + Z_2 R_D^T)\lambda = f.$$

Having computed the coefficients λ, the interpolant s at a point x can be evaluated as follows:

$$s(x) = \Psi(x)\lambda.$$

Up to this point, this algorithm is still purely theoretical, since the matrices D_2 and R_2 remain infinite-dimensional, which makes the numerical computation impossible. Therefore, an algorithm for truncating the expansion (5.1) is required. One of the common ways of truncation is based on the values of the matrix $\tilde{D}$ (see, for example, [FLF11, § 5]). The idea is that once elements in $\tilde{D}$ become less than machine precision,

$$\max_{\substack{i=1\dots N, \\ j>M}} \tilde{D}_{ij} < \varepsilon_{\text{mach}}$$

we can stop. Since we know explicitly the elements of D, it is usually possible to identify analytically when the elements $\tilde{D}$ drop below machine precision and based on that deduce the number of basis functions M. Even though this approach provides a viable strategy to determine M, it only focuses on the matrix $\tilde{D}$ and ignores other parts of the expansion. For this reason, for the new RBF-QR method that we will develop in following chapters we will derive an alternative cut-off criterion. The flow of the stable interpolation procedure provided in Algorithm 1.

Let us take a look at the two main RBF-expansions used for the RBF-QR algorithms.

5.2. Chebyshev-QR

We look at the RBF-QR algorithm that is based on the expansion of Gaussian RBFs in a basis involving Chebyshev polynomials. In this case, the RBF-expansion originated from the Taylor expansion translated to polar coordinates in 2D and spherical coordinates in 3D. Further improvement of the conditioning is based on expanding the powers of the radius r in Chebyshev polynomials.

We provide the formulas of the resulting expansion functions and the coefficients in 1D. The expansion functions read as

$$\tilde{T}_\ell(x) = T_\ell(x) \exp(-\varepsilon^2 x^2),$$

with $\{T_\ell\}$ being Chebyshev polynomials. The expansion coefficients C and D

Algorithm 1 General form of the RBF-QR methods

Given the values f of the function $f(x)$ at a set of points $\{x_k^{\mathrm{col}}\}$ a stable way to compute the RBF-interpolant $s(x)$ includes the following steps:

```
1: Choose an *RBF-expansion* of the form (5.1):
       $$\phi_k(x) = \sum_{\ell=0}^{\infty} d_\ell c_\ell(x_k)\zeta_\ell(x).$$
2: Fix the truncation value $M \geq N$.
3: Formulate (5.1) in a matrix-vector form for a vector of all RBFs:
       $$\Phi(x)^T = CD\Psi(x)^T.$$
4: Perform $C = QR$.
5: Compute $R_D = D_1^{-1}R_1^{-1}R_2D_2$ without explicitly inverting $D_1$.
6: Evaluate $Z(x)$ at collocation points $\{x_k^{\mathrm{col}}\}_{k=1}^N$.
7: Assemble $\Psi^{\mathrm{col}} = Z_1 + Z_2R_D^T$.
8: Solve $\Psi^{\mathrm{col}}\lambda = f$.
9: Compute $s(x) = \Psi(x)\lambda$
```

from (5.1) are given as:

$$d_\ell = \frac{2\varepsilon^{2\ell}}{\ell!}, \quad c_\ell(x_k) = t_\ell \exp(-\varepsilon^2 x_k^2)x_k^\ell \, {}_0F_1(;\ell+1, \varepsilon^4 x_k^2), \tag{5.4}$$

where t_ℓ are fixed scalars and ${}_0F_1$ is the hypergeometric function. For the truncation, the algorithm based on D is used:

$$\frac{\max_{i>M} D_{ii}}{\min_{1\leq i\leq N} D_{ii}} < \varepsilon_{\text{mach}}. \tag{5.5}$$

The method can be extended to multiple dimensions, however the code is available in 1-3D. For higher dimensional cases, explicit formulas have to be derived. However, it is also possible to use a simple tensor-product approach based on the 1D version of the method. In this work, we use the available code as one of the references.

5.3. Gauss-QR

The Gauss-QR algorithm was developed by Fasshauer & McCourt [FM12] who noticed a connection between the RBF-QR algorithm and Mercer or Hilbert-Schmidt expansions for positive definite kernels. It turns out that the Mercer expansion [Ras03, $\S4.3$,Theorem 4.2] for Gaussian kernels can be used as an RBF-expansion and, consecutively, the Gauss-QR algorithm can be developed.

In particular, the expansion functions are given by:

$$\phi_\ell(x) = \gamma_\ell \exp(-\delta^2 x^2) H_{\ell-1}(\alpha\beta x), \quad \ell \in \mathbb{N}$$

where α and ε are parameters specified by the user, $H_{\ell-1}$ are Hermite polynomials and the parameters β, δ can be calculated from ε, α as follows:

$$\beta = \left(1 + \frac{4\varepsilon^2}{\alpha^2}\right)^{1/4}, \quad \gamma_\ell = \sqrt{\frac{\beta}{2^{\ell-1}\Gamma(\ell-1)}}, \quad \delta^2 = \frac{\alpha^2}{2}(\beta^2 - 1).$$

The Mercer expansion, which in the context of RBF-QR is used as the RBF-expansion, reads as:

$$\exp\left(-\varepsilon^2 (x-z)^2\right) = \sum_{\ell=1}^{\infty} \lambda_\ell \phi_\ell(x) \phi_\ell(z), \tag{5.6}$$

where $\lambda_\ell = \sqrt{\frac{\alpha^2}{\alpha^2+\varepsilon^2+\delta^2}} \left(\frac{\varepsilon^2}{\alpha^2+\varepsilon^2+\delta^2}\right)^{\ell-1}$. Interpreting (5.6) as an RBF-expansion (5.1) we arrive to the following formulas for the coefficients C and D from (5.1):

$$c_\ell(x_k) = \phi_\ell(x_k), \quad d_\ell = \lambda_\ell.$$

The truncation value M can be derived from the magnitude of the values of λ. In particular, in the available version of the code (available at `http://math.iit.edu/~mccomic/gaussqr/`), it is determined as the smallest M satisfying $\lambda_M < \varepsilon_{\text{mach}} \lambda_N$ which corresponds to the condition

$$\max_{\substack{i=1\ldots N,\\ j>M}} \tilde{D}_{ij} < \varepsilon_{\text{mach}}.$$

The method has been implemented for arbitrary dimension and we will use it as one of the reference methods for all relevant test cases.

6. HermiteGF expansion of Gaussians

In this chapter, we propose a new RBF expansion that works directly on general anisotropic Gaussians. In particular, we use the generalized anisotropic Hermite functions that we have introduced in chapter 3 as expansion functions. We then build a new RBF-QR method based on that. For the sake of simplicity, we first derive the expansion for Gaussians in 1D. After that, we extend our expansion to the case of multivariate anisotropic Gaussians.

6.1. HermiteGF expansion in 1D

Recall that generalized anisotropic Hermite functions in 1D are defined as

$$H_\ell^{\gamma,\varepsilon,t}(x) = \frac{t^{\ell/2}}{\sqrt{2^\ell \ell!}} h_\ell(\gamma x) \mathrm{e}^{-\varepsilon^2 x^2}, \quad \varepsilon > 0,\ \gamma > 0, t > 0,$$

where $\{h_\ell\}$ are physicists' Hermite polynomials. In the context of the interpolation problem, the parameters can have a clear interpretation. The parameter ε corresponds to the shape parameter of the original Gaussian RBFs, the parameter γ controls the evaluation domain of the Hermite polynomials, which will allow to improve conditioning. The technical parameter t will help us to control the truncation error of the stabilization expansion (see chapter 10). The cut-off algorithm will be based on the Mehler's formula (2.18), that is why in this case we limit ourselves to the values of t within the interval

$$t \in (0, 1).$$

Based on the generating function theory, we derive an infinite expansion of the one-dimensional Gaussian RBFs in the basis $\{H_\ell^{\gamma,\varepsilon,t}\}$.

Theorem 6.1.1: HermiteGF expansion (1D)

For all parameters $\varepsilon > 0$, $\gamma > 0$, $x_0 \in \mathbb{R}$, $q \in \mathbb{R}$, we have a pointwise expansion

$$\phi_q(x) = \mathrm{e}^{-\varepsilon^2 (x-q)^2} = \exp\left(\varepsilon^2 \Delta_q^2 \left(\frac{\varepsilon^2}{\gamma^2} - 1\right)\right) \sum_{\ell \geq 0} \frac{\varepsilon^{2\ell}\sqrt{2^\ell}}{\gamma^\ell \sqrt{t^\ell \ell!}} \Delta_q^\ell H_\ell^{\gamma,\varepsilon,t}(x - x_0), \tag{6.1}$$

where $\Delta_q = q - x_0$. The RBF interpolant $s(x)$ can then be pointwise computed as,

$$s(x) = \sum_{k=1}^N \alpha_k \exp\left(\varepsilon^2 \Delta_k^2 \left(\frac{\varepsilon^2}{\gamma^2} - 1\right)\right) \sum_{\ell \geq 0} \frac{\varepsilon^{2\ell}\sqrt{2^\ell}}{\gamma^\ell \sqrt{t^\ell \ell!}} \Delta_k^\ell H_\ell^{\gamma,\varepsilon,t}(x - x_0), \tag{6.2}$$

where $\Delta_k = x_k - x_0$ and $\{x_k\}_{k=1}^N$ are the centers of the RBFs.

Proof. The Hermite polynomial's generating function is given by (2.17),

$$\mathrm{e}^{2ba-a^2} = \sum_{\ell\geq 0} \frac{a^\ell}{\ell!} h_\ell(b)$$

Choosing $a = \frac{\varepsilon^2\Delta_q}{\gamma}$ and $b = \gamma(x - x_0)$, we obtain

$$\sum_{\ell\geq 0} \frac{\varepsilon^{2\ell}}{\gamma^\ell \ell!} \Delta_q^\ell h_\ell(\gamma(x - x_0)) = \exp\left(2\varepsilon^2\Delta_q(x - x_0) - \frac{\varepsilon^4\Delta_q^2}{\gamma^2}\right).$$

Hence, we get

$$\begin{aligned}
&\exp\left(\varepsilon^2\Delta_q^2\left(\frac{\varepsilon^2}{\gamma^2} - 1\right)\right) \sum_{\ell\geq 0} \frac{\varepsilon^{2\ell}\sqrt{2^\ell}}{\gamma^\ell\sqrt{t^\ell \ell!}} \Delta_q^\ell H_\ell^{\gamma,\varepsilon,t}(x - x_0)\\
&= \exp\left(\varepsilon^2\Delta_q^2\left(\frac{\varepsilon^2}{\gamma^2} - 1\right)\right) \sum_{\ell\geq 0} \frac{\varepsilon^{2\ell}}{\gamma^\ell \ell!} \Delta_q^\ell h_\ell(\gamma(x - x_0)) \mathrm{e}^{-\varepsilon^2(x-x_0)^2}\\
&= \exp\left(\varepsilon^2\Delta_q^2\left(\frac{\varepsilon^2}{\gamma^2} - 1\right) + 2\varepsilon^2\Delta_q(x - x_0) - \frac{\varepsilon^4\Delta_q^2}{\gamma^2} - \varepsilon^2(x - x_0)^2\right)\\
&= \exp\left(-\varepsilon^2\left(\Delta_q - (x - x_0)\right)^2\right) = \mathrm{e}^{-\varepsilon^2(x-q)^2},
\end{aligned}$$

which proves expansion (6.1). Using expansion (6.1) in the interpolant (4.1), we get the representation (6.2). □

In the context of the interpolation problem, we will further refer to the generalized anisotropic Hermite functions also as *HermiteGF basis functions*, since in this framework the expansion basis is bound to the Hermite generating functions expansion.

Remark 6.1.1. *Basis centering.*

Hermite polynomials have a symmetry with respect to the axis $x = 0$. Due to its growth behavior, it is advantageous to have the basis centered around this point of symmetry, that is, to use the translation $x_0 = \frac{B-A}{2}$, where $[A, B]$ is the interval of interest for evaluating the function f.

Remark 6.1.2. *The parameter γ.*

The parameter $\gamma > 0$ in the basis $\{H_\ell^{\gamma,\varepsilon,t}\}_{\ell\geq 0}$ allows control over the evaluation domain of the Hermite polynomials. When choosing it, one has to consider two counteracting effects: For small values of γ, ill-conditioning appears since the values of the basis functions at the collocation points are too similar. On the other hand, Hermite polynomials take very large values on large domains which can lead to an overflow. An optimal balance depends on the particular function and the number of basis functions. The parameter plays a similar role to that of the "global scale parameter" α in [FM12].

6.2. Multivariate HermiteGF expansion of anisotropic Gaussians

The HermiteGF expansion can be easily extended to higher dimensions for the case of isotropic Gaussians, using tensor products of 1D physicists' Hermite polynomials

$$h_{\boldsymbol\ell}(\mathbf{x}) = h_{\ell_1}(x_1) \cdot \dots \cdot h_{\ell_d}(x_d), \quad \boldsymbol\ell \in \mathbb{N}^d, \quad \mathbf{x} \in \mathbb{R}^d.$$

However, finding a stable interpolant for anisotropic Gaussian functions of the type

$$\phi_{\mathbf{q}}(\mathbf{x}) = \exp(-(\mathbf{x}-\mathbf{q})^T E^T E(\mathbf{x}-\mathbf{q})), \qquad \mathbf{x}, \mathbf{q} \in \mathbb{R}^d,$$

is a more challenging task. A similar question was raised in [FM12, § 8.5], however, without further investigation. McCourt & Fasshauer [MF17] considered anisotropic Gaussians with diagonal shape matrix E using Mercer expansion theory, but this result has not been extended to the case of arbitrary E. Analogously to the 1D case, we define the multivariate version of our *HermiteGF functions* by

$$H_{\boldsymbol\ell}^{G,E,t}(\mathbf{x}) = \frac{t^{|\boldsymbol\ell|/2}}{\sqrt{2^{|\boldsymbol\ell|}\boldsymbol\ell!}} h_{\boldsymbol\ell}(G^T\mathbf{x}) \exp(-\mathbf{x}^T E^T E\mathbf{x}),$$

where $G, E \in \mathbb{R}^{d\times d}$ are arbitrary invertible matrices.

Proposition 6.2.1: HermiteGF expansion of anisotropic Gaussians

Let $\mathbf{q} \in \mathbb{R}^d$ and $E, G \in \mathbb{R}^{d\times d}$ be invertible matrices. Consider the anisotropic Gaussian $\phi_{\mathbf{q}}(\cdot)$. Then, for any shift $\mathbf{x}_0 \in \mathbb{R}^d$ the following relation holds pointwise in $\mathbf{x} \in \mathbb{R}^d$:

$$\phi_{\mathbf{q}}(\mathbf{x}) = \exp(\boldsymbol\Delta_{\mathbf{q}}^T(\tilde{G}-E^T E)\boldsymbol\Delta_{\mathbf{q}}) \cdot \sum_{\boldsymbol\ell \in \mathbb{N}^d} \frac{(G^{-1}E^T E\boldsymbol\Delta_{\mathbf{q}})^{\boldsymbol\ell}\sqrt{2^{|\boldsymbol\ell|}}}{\sqrt{t^{|\boldsymbol\ell|}\boldsymbol\ell!}} H^{G,E,t}(\mathbf{x}-\mathbf{x}_0), \quad (6.3)$$

where $\boldsymbol\Delta_{\mathbf{q}} = \mathbf{q} - \mathbf{x}_0$, $\tilde{G} = E^T E G^{-T} G^{-1} E^T E$.

Proof. Denote $\boldsymbol\Delta_{\mathbf{q}} = \mathbf{q} - \mathbf{x}_0$, $\mathbf{b} = G^T(\mathbf{x}-\mathbf{x}_0)$, $\mathbf{a} = G^{-1}E^T E\boldsymbol\Delta_{\mathbf{q}}$. Then, using the multivariate Hermite generating function (2.17), we get

$$\sum_{\boldsymbol\ell \in \mathbb{N}_0^d} \frac{(G^{-1}E^T E\boldsymbol\Delta_{\mathbf{q}})^{\boldsymbol\ell}}{\boldsymbol\ell!} h_{\boldsymbol\ell}(G^T(\mathbf{x}-\mathbf{x}_0)) = \exp(2(\mathbf{x}-\mathbf{x}_0)^T E^T E\boldsymbol\Delta_{\mathbf{q}} - \boldsymbol\Delta_{\mathbf{q}}^T\tilde{G}\boldsymbol\Delta_{\mathbf{q}}).$$

We observe that

- $2\mathbf{x}^T E^T E\mathbf{q} = 2(\mathbf{x}-\mathbf{x}_0)E^T E\boldsymbol\Delta_{\mathbf{q}} + 2\mathbf{x}^T E^T E\mathbf{x}_0 + 2\mathbf{x}_0 E^T E\boldsymbol\Delta_{\mathbf{q}}$.
- $-\mathbf{x}^T E^T E\mathbf{x} = -(\mathbf{x}-\mathbf{x}_0)^T E^T E(\mathbf{x}-\mathbf{x}_0) - 2\mathbf{x}^T E^T E\mathbf{x}_0 + \mathbf{x}_0^T E^T E\mathbf{x}_0$.

Then

$$\begin{aligned}
&\exp(-(\mathbf{x}-\mathbf{q})^T E^T E(\mathbf{x}-\mathbf{q})) = \exp(-\mathbf{x}^T E^T E\mathbf{x} + 2\mathbf{x}^T E^T E\mathbf{q} - \mathbf{q}^T E^T E\mathbf{q})\\
&= \exp(-(\mathbf{x}-\mathbf{x}_0)^T E^T E(\mathbf{x}-\mathbf{x}_0)) \cdot \exp(-2\mathbf{x}^T E^T E\mathbf{x}_0 + \mathbf{x}_0^T E^T E\mathbf{x}_0)\\
&\quad \cdot \exp(2(\mathbf{x}-\mathbf{x}_0)E^T E\boldsymbol{\Delta}_{\mathbf{q}}) \cdot \exp(2\mathbf{x}^T E^T E\mathbf{x}_0 + 2\mathbf{x}_0 E^T E\boldsymbol{\Delta}_{\mathbf{q}}) \cdot \exp(-\mathbf{q}E^T E\mathbf{q})\\
&= \exp(-(\mathbf{x}-\mathbf{x}_0)^T E^T E(\mathbf{x}-\mathbf{x}_0)) \cdot \exp(2(\mathbf{x}-\mathbf{x}_0)E^T E\boldsymbol{\Delta}_{\mathbf{q}})\\
&\quad \cdot \exp(-\mathbf{q}E^T E\mathbf{q} + 2\mathbf{x}_0 E^T E\mathbf{q} - \mathbf{x}_0 E^T E\mathbf{x}_0)\\
&= \exp(-\boldsymbol{\Delta}_{\mathbf{q}}^T E^T E\boldsymbol{\Delta}_{\mathbf{q}} + \boldsymbol{\Delta}_{\mathbf{q}}^T \tilde{G}\boldsymbol{\Delta}_{\mathbf{q}})\\
&\quad \cdot \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d} \frac{(G^{-1}E^T E\boldsymbol{\Delta}_{\mathbf{q}})^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} h_{\boldsymbol{\ell}}(G^T(\mathbf{x}-\mathbf{x}_0)) \exp(-(\mathbf{x}-\mathbf{x}_0)E^T E(\mathbf{x}-\mathbf{x}_0))\\
&= \exp(-\boldsymbol{\Delta}_{\mathbf{q}}^T E^T E\boldsymbol{\Delta}_{\mathbf{q}} + \boldsymbol{\Delta}_{\mathbf{q}}^T \tilde{G}\boldsymbol{\Delta}_{\mathbf{q}}) \cdot \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d} \frac{(G^{-1}E^T E\boldsymbol{\Delta}_{\mathbf{q}})^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}} H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0). \quad \square
\end{aligned}$$

This expansion provides a new powerful tool of dealing with anisotropic approximation. Note that the standard multidimensional isotropic Gaussian interpolation corresponds to the following matrix E:

$$E_{\text{isotropic}} = \varepsilon \mathrm{Id}_d.$$

6.3. The absence of the scaling invariance

One can see from the derivations above that the free parameters γ/G and t allow for multiple HermiteGF expansions for one Gaussian. We illustrate it further by introducing a scaling of the parameters of the original Gaussian that does not change the final values of the Gaussian, but affects the HermiteGF expansion.

Proposition 6.3.1: Nonscalability of the HermiteGF expansion (1D)

Consider a scaling parameter $\alpha \in \mathbb{R}, \alpha > 0$. Denote

$$\bar{x} = \frac{x}{\alpha}, \quad \bar{x}_0 = \frac{x_0}{\alpha}, \quad \bar{q} = \frac{q}{a}, \quad \Delta_{\bar{q}} = \bar{q} - \bar{x}_0, \quad \bar{\varepsilon} = \alpha\varepsilon$$

with $x, x_0, q \in \mathbb{R}, \varepsilon > 0$. Then the HermiteGF expansion of $\phi_{\bar{q}}^{\bar{\varepsilon}}(\bar{x})$ with $\gamma \in \mathbb{R}, t \in (0,1)$ corresponds to the HermiteGF expansion of $\phi_q^{\varepsilon}(x)$ with $\bar{\gamma} = \gamma/\alpha$ and the same t.

Proof. Recall that for a fixed $\gamma \in \mathbb{R}, t \in (0,1)$ HermiteGF-expansion (6.1) in 1D writes as

$$\phi_q^{\varepsilon}(x) = \mathrm{e}^{-\varepsilon^2(x-q)^2} = \exp\left(\varepsilon^2\Delta_q^2\left(\frac{\varepsilon^2}{\gamma^2} - 1\right)\right)\sum_{\ell\geq 0}\frac{\varepsilon^{2\ell}\sqrt{2^\ell}}{\gamma^\ell\sqrt{t^\ell \ell!}}\Delta_q^\ell H_\ell^{\gamma,\varepsilon,t}(x - x_0).$$

At the same time,

$$\phi_{\bar{q}}^{\bar{\varepsilon}}(\bar{x}) = \mathrm{e}^{\bar{\varepsilon}^2(\bar{x}-\bar{q})^2} = \mathrm{e}^{\alpha^2\varepsilon^2\left(\frac{x}{\alpha}-\frac{q}{\alpha}\right)^2} = \phi_q^{\varepsilon}(x).$$

However, the HermiteGF expansions are different for $\phi_{\bar{q}}^{\bar{\varepsilon}}(\bar{x})$ and $\phi_q^{\varepsilon}(x)$. Indeed,

$$\begin{aligned}
\mathrm{e}^{\bar{\varepsilon}^2(\bar{x}-\bar{q})^2} &= \exp\left(\bar{\varepsilon}^2\Delta_{\bar{q}}^2\left(\frac{\bar{\varepsilon}^2}{\gamma^2}-1\right)\right)\sum_{\ell\geq 0}\frac{\bar{\varepsilon}^{2\ell}\sqrt{2^\ell}}{\gamma^\ell\sqrt{t^\ell\ell!}}\Delta_{\bar{q}}^\ell H_\ell^{\gamma,\bar{\varepsilon},t}\left(\bar{x}-\bar{x}_0\right) \\
&= \exp\left(\varepsilon^2\Delta_q^2\left(\frac{\alpha^2\varepsilon^2}{\gamma^2}-1\right)\right)\sum_{\ell\geq 0}\frac{\alpha^\ell\varepsilon^{2\ell}\sqrt{2^\ell}}{\gamma^\ell\sqrt{t^\ell\ell!}}\Delta_q^\ell H_\ell^{\gamma,\bar{\varepsilon},t}\left(\bar{x}-\bar{x}_0\right).
\end{aligned}$$

We note that

$$H_\ell^{\gamma,\bar{\varepsilon},t}\left(\bar{x}\right) = \frac{t^{\ell/2}}{\sqrt{2^\ell\ell!}}h_\ell\left(\frac{\gamma}{\alpha}(x-x_0)\right)\mathrm{e}^{-\varepsilon^2x^2} \overset{\bar{\gamma}:=\gamma/\alpha}{=} H_\ell^{\bar{\gamma},\varepsilon,t}\left(x-x_0\right).$$

Therefore,

$$\mathrm{e}^{\bar{\varepsilon}^2(\bar{x}-\bar{q})^2} = \exp\left(\varepsilon^2\Delta_q^2\left(\frac{\varepsilon^2}{\bar{\gamma}^2}-1\right)\right)\sum_{\ell\geq 0}\frac{\varepsilon^{2\ell}\sqrt{2^\ell}}{\bar{\gamma}^\ell\sqrt{t^\ell\ell!}}\Delta_q^\ell H_\ell^{\bar{\gamma},\varepsilon,t}\left(x-x_0\right),$$

which corresponds to the expansion of $\phi_q^{\varepsilon}(x)$ with $\gamma \to \gamma/\alpha$. □

In practice it means, that scaling the interpolation domain to a certain interval and compensating it with the scaling parameter, alters the expansion. Even though it is advantageous in some cases to scale the centers of RBFs, one should keep in mind that the resulting expansion differs from the original one. In order to have a one-to-one correspondence, one has to scale γ as well.

Analogously to the 1D case, the multidimensional HermiteGF expansion is also non invariant with respect to scaling.

Proposition 6.3.2: Nonscalability of the HermiteGF expansion (nD)

Consider an invertible matrix $A \in \mathbb{R}^{d\times d}$. Denote

$$\bar{\mathbf{x}} = A^{-1}\mathbf{x}, \quad \bar{\mathbf{x}}_0 = A^{-1}\mathbf{x}_0, \quad \bar{\mathbf{q}} = A^{-1}\mathbf{q}, \quad \boldsymbol{\Delta}_{\bar{\mathbf{q}}} = A^{-1}(\mathbf{q}-\mathbf{x}_0), \quad \bar{E} = EA$$

with $\mathbf{x}, \mathbf{q}, \mathbf{x}_0 \in \mathbb{R}^d$, $E, G \in \mathbb{R}^{d\times d}$ are invertible. Then the HermiteGF expansion of $\phi_{\bar{\mathbf{q}}}^{\bar{E}}(\bar{\mathbf{x}})$ with $G \in \mathbb{R}^{d\times d}$, $t \in (0,1)$ corresponds to the HermiteGF expansion of $\phi_{\mathbf{q}}^{E}(\mathbf{x})$ with $\bar{G} = A^{-T}G$ and the same t.

Proof. Recall the HermiteGF expansion in multiple dimensions (6.3)

$$\begin{aligned}
\phi_{\mathbf{q}}^{E}(\mathbf{x}) &= \exp(-(\mathbf{x}-\mathbf{q})^TE^TE(\mathbf{x}-\mathbf{q})) \\
&= \exp(-\boldsymbol{\Delta}_{\mathbf{q}}^TE^TE\boldsymbol{\Delta}_{\mathbf{q}} + \boldsymbol{\Delta}_{\mathbf{q}}\tilde{G}\boldsymbol{\Delta}_{\mathbf{q}}) \cdot \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d}\frac{(G^{-1}E^TE\boldsymbol{\Delta}_{\mathbf{q}})^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0)
\end{aligned}$$

with $\tilde{G} = E^TEG^{-T}G^{-1}E^TE$. Then

$$\phi_{\bar{\mathbf{q}}}^{\bar{E}}(\bar{\mathbf{x}}) = \exp(-(A^{-1}(\mathbf{x}-\mathbf{q}))^TA^TE^TEAA^{-1}(\mathbf{x}-\mathbf{q})) = \phi_{\mathbf{q}}^{E}(\mathbf{x}).$$

However, the HermiteGF expansion of $\phi_{\bar{\mathbf{q}}}(\bar{\mathbf{x}})$ reads as

$$\begin{aligned}\phi^{\bar{E}}_{\bar{\mathbf{q}}}(\bar{\mathbf{x}}) &= \exp\left(-\boldsymbol{\Delta}_{\bar{\mathbf{q}}}^T A^T E^T E A \boldsymbol{\Delta}_{\bar{\mathbf{q}}} + \boldsymbol{\Delta}_{\bar{\mathbf{q}}}^T A^T E^T E A G^{-T} G^{-1} A^T E^T E A \boldsymbol{\Delta}_{\bar{\mathbf{q}}}\right)\\ &\quad\cdot \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d} \frac{(G^{-1}A^T E^T E A \boldsymbol{\Delta}_{\bar{\mathbf{q}}})^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}} H^{G,\bar{E},t}_{\boldsymbol{\ell}}(\bar{\mathbf{x}}-\bar{\mathbf{x}}_0)\\ &= \exp\left(-\boldsymbol{\Delta}_{\mathbf{q}}^T E^T E \boldsymbol{\Delta}_{\mathbf{q}} + \boldsymbol{\Delta}_{\mathbf{q}}^T E^T E A G^{-T} G^{-1} A^T E^T E \boldsymbol{\Delta}_{\mathbf{q}}\right)\\ &\quad\cdot \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d} \frac{(G^{-1}A^T E^T E \boldsymbol{\Delta}_{\mathbf{q}})^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}} H^{G,\bar{E},t}_{\boldsymbol{\ell}}(A^{-1}(\mathbf{x}-\mathbf{x}_0)).\end{aligned}$$

We note that the following relation holds for the basis functions

$$H^{G,\bar{E},t}_{\boldsymbol{\ell}}(\bar{\mathbf{x}}) = \frac{t^{|\boldsymbol{\ell}|/2}}{\sqrt{2^{\boldsymbol{\ell}}\boldsymbol{\ell}!}} h_{\boldsymbol{\ell}}(G^T A^{-1}(\mathbf{x}-\mathbf{x}_0))\exp(-\mathbf{x}^T E^T E \mathbf{x}) \overset{\bar{G}=A^{-T}G}{=} H^{\bar{G},E,t}_{\boldsymbol{\ell}}(\mathbf{x}-\mathbf{x}_0).$$

Therefore,

$$\begin{aligned}\phi^{\bar{E}}_{\bar{\mathbf{q}}}(\mathbf{x}) &= \exp\left(-\boldsymbol{\Delta}_{\mathbf{q}} E^T E \boldsymbol{\Delta}_{\mathbf{q}} + \boldsymbol{\Delta}_{\mathbf{q}}^T E^T E \bar{G}^{-T} \bar{G}^{-1} E^T E \boldsymbol{\Delta}_{\mathbf{q}}\right)\\ &\quad\cdot \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d} \frac{\left(\bar{G}^{-1} E^T E \boldsymbol{\Delta}_{\mathbf{q}}\right)^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}} H^{\bar{G},E,t}_{\boldsymbol{\ell}}(\mathbf{x}-\mathbf{x}_0),\end{aligned}$$

which corresponds to the expansion of $\phi^{E}_{\mathbf{q}}(\mathbf{x})$ with $G \to A^{-T}G$. □

6.4. Bilinear generating function and the norm of the HermiteGF basis

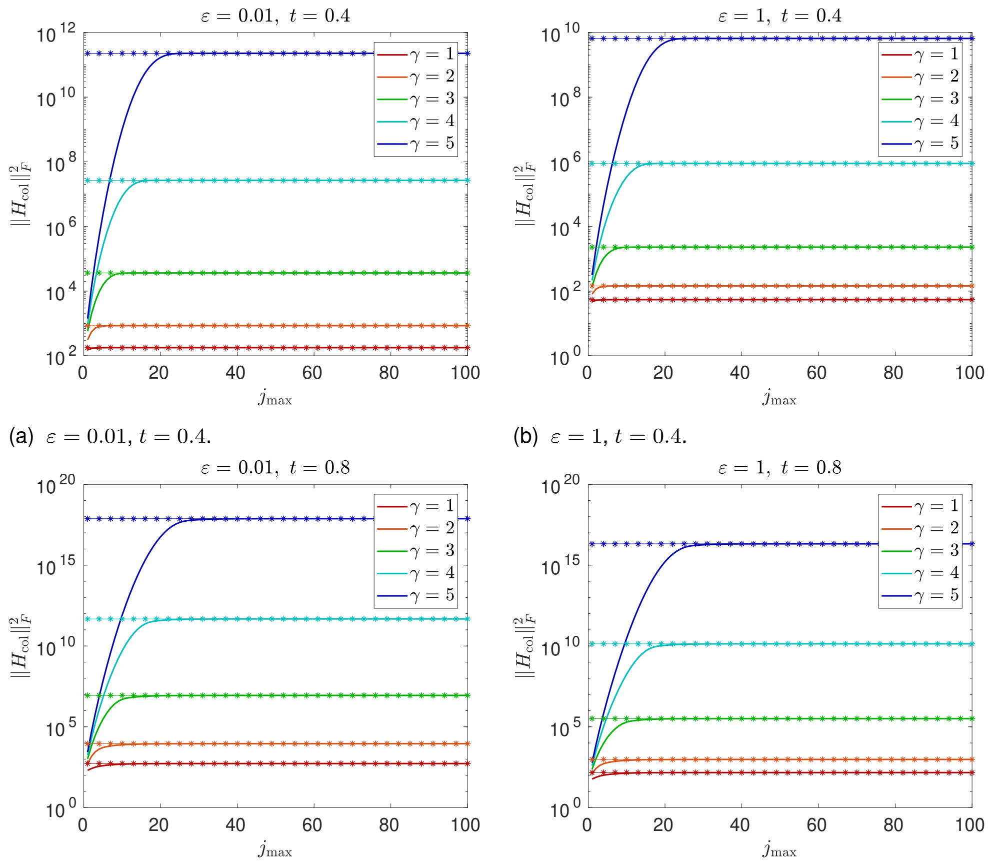
Let us now investigate the behavior of the HermiteGF basis functions with a large ℓ approaching infinity. It will become crucial later on, in chapter 10, when we have to cut our basis in order to use it for numerical purposes. We would like to get a first idea of how many basis functions bring a significant input. For that, we take a look at the magnitude of the values of the HermiteGF basis. We estimate the 2-norm of the infinite dimensional vector $H^{G,E,t}(\mathbf{x})^T$, containing the values of all basis functions at a certain point $\mathbf{x} \in \mathbb{R}^d$. It turns out that we can compute this norm analytically, using a multivariate extension of Mehler's formula for Hermite polynomials (2.18)

$$\sum_{|\boldsymbol{\ell}|=0}^{\infty} \frac{t^{|\boldsymbol{\ell}|} h_{\boldsymbol{\ell}}(\mathbf{x}) h_{\boldsymbol{\ell}}(\mathbf{y})}{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!} = \frac{1}{(1-t^2)^{d/2}} \exp\left(\frac{(\mathbf{x}^T\mathbf{y}+\mathbf{y}^T\mathbf{x})t - t^2(\|\mathbf{x}\|_2^2+\|\mathbf{y}\|_2^2)}{(1-t^2)}\right).$$

With the help of the multidimensional Mehler's formula we can now compute the square of the Euclidean norm

$$H^{G,E,t}_{\lim}(\mathbf{x}) := \sum_{|\boldsymbol{\ell}|=0}^{\infty} H^{G,E,t}_{\boldsymbol{\ell}}(\mathbf{x})^2 = \sum_{|\boldsymbol{\ell}|=0}^{\infty} \frac{t^{|\boldsymbol{\ell}|}}{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!} h^2_{\boldsymbol{\ell}}(G^T\mathbf{x}) \exp(-2\mathbf{x}^T E^T E \mathbf{x})$$

of the infinite vector containing the values of all basis functions $H^{G,E,t}_{\boldsymbol{\ell}}(\mathbf{x})$ with



(a) $\varepsilon = 0.01, t = 0.4.$

(b) $\varepsilon = 1, t = 0.4.$

(c) $\varepsilon = 0.01, t = 0.8.$

(d) $\varepsilon = 1, t = 0.8.$

Figure 1 Frobenius norm of the matrix H_{col} of values the basis functions at 100 Halton nodes $\mathbf{x}_k$ on a square domain $[-1,1] \times [-1,1]$. The solid line is the value of corresponding to only the basis functions with a degree up to $j_{\max}$. The $\star$ corresponds to the analytically computed value.

$\boldsymbol{\ell} \in \mathbb{N}^d$ at some point $\mathbf{x} \in \mathbb{R}^d$. Indeed,

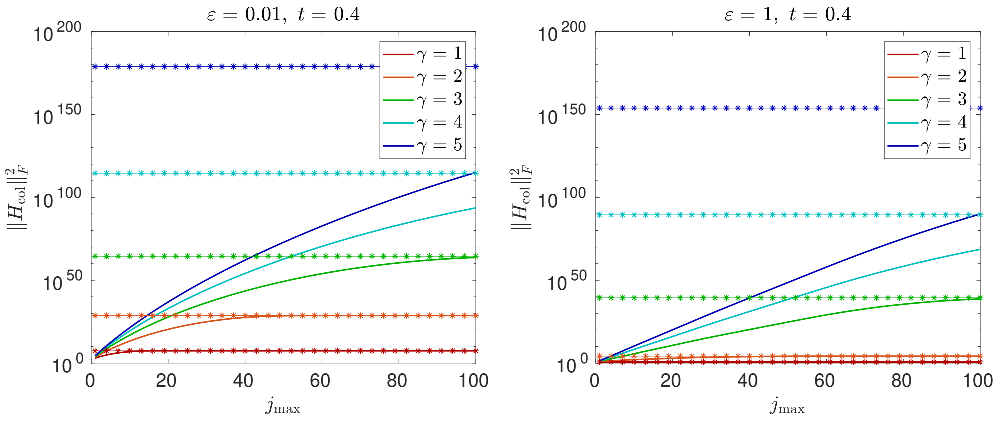
$$H_{\mathrm{lim}}^{G,E,t}(\mathbf{x}) = \frac{\exp\left(-2\mathbf{x}^T E^T E\mathbf{x} + \frac{2t\|G^T\mathbf{x}\|_2^2}{1+t}\right)}{(1-t^2)^{d/2}}. \tag{6.4}$$

One can also use the Mehler's formula for computing the Frobenius norm of the matrix $H_{\mathrm{col}}^{G,E,t}$ containing the values of all basis functions $\{H_{\boldsymbol{\ell}}^{G,E,t}\}_{\boldsymbol{\ell}\in\mathbb{N}_0^d}$ at all collocation points $\{\mathbf{x}_k\}_{k=1\ldots N}$:

$$\|H_{\mathrm{col}}^{G,E,t}\|_F^2 = \sum_{k=1}^{N} H_{\mathrm{lim}}^{G,E,t}(\mathbf{x}_k).$$

6.4.1. Analysis of $\|H_{\mathrm{col}}^{G,E,t}\|_F$

We first check if the analytically predicted value of $\|H_{\mathrm{col}}^{G,E,t}\|_F^2$ corresponds to numerical results. We look at the $\|H_{\mathrm{col}}^{G,E,t}\|_F^2$ in 2D on a unit square $[-1,1]\times[-1,1]$ with 100 Halton points in the isotropic case ($E = \varepsilon\mathrm{Id}$, $G = \gamma\mathrm{Id}$). One can see in the Figure 1 the limit value is matched for different values of ε, γ and t. For smaller values of γ the value of the norm is smaller. On the other hand, in case



(a) $\varepsilon = 0.01, t = 0.4$.

(b) $\varepsilon = 1, t = 0.4$.

Figure 2 Frobenius norm of the matrix H of values the basis functions at 100 Halton nodes $\mathbf{x}_k$ on a square domain $[-1, 1] \times [-1, 1]$. The limit value is very large for larger γ.

the basis is used for interpolation, small γ makes the evaluation points of Hermite polynomials closer to each other, which might worsen the conditioning. One can also see that the values of the norms are smaller for smaller t. However, for the interpolation problem, too small t might also make the values of the basis close to each other, since we are taking the power $|\boldsymbol{\ell}|/2$ of t in the basis. For further experiments with interpolation with different values of γ, t see chapter 12.

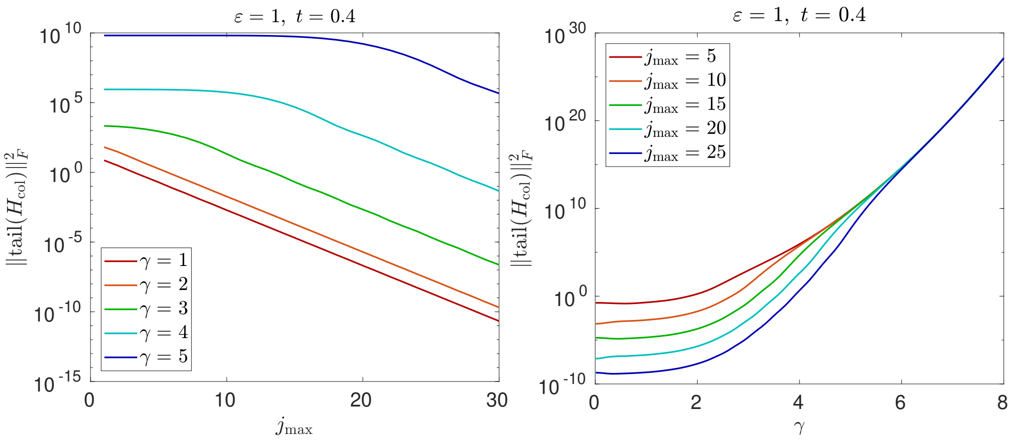
For the unit square case, the values of $\|H_{\mathrm{col}}^{G,E,t}\|_F$ were harmless for all values of γ, t. However, for a larger domain, in particular, a square $[-4, 4] \times [-4, 4]$, one can see in the Figure 2 that the norm takes extremely large values which we would like to avoid in our computations. With the help of the Mehler's formula, we can use the analytic estimation of the norm in order to prevent overflow.

Tail of the expansion

From the figures 1 and 2 one can see that the speed of convergence of the norm is different for different setups. From the computational cost perspective, we would be interested in a fast convergence, since it could allow for a smaller number of basis functions. Let us take a closer look at the tail of this expansion. In particular, we look at the values of the following expression:

$$\text{tail}(\|H_{\mathrm{col}}^{G,E,t}\|_F^2) = \sum_{k=1}^{N} \sum_{|\boldsymbol{\ell}|=j_{\max}}^{\infty} \frac{t^{|\boldsymbol{\ell}|}}{2^{\boldsymbol{\ell}}\boldsymbol{\ell}!} h_{\boldsymbol{\ell}}(G^T\mathbf{x}_k) \exp(-2\mathbf{x}_k^T E^T E\mathbf{x}_k)$$

One can see in the Figure 3a that for smaller γ the speed of convergence is way faster than for larger values. On the other hand, in the Figure 3b, one can see that after a certain value of γ the first 25 basis functions do not bring us closer to the limit value. Moreover, we can see that for all values of $j_{\max}$ the value of the tail starts to grow exponentially with γ after a certain point. Therefore, one has to



(a) Dependence of the tail from $j_{\max}$.

(b) Dependence of the tail from γ.

Figure 3 One can see that smaller values of γ provide smaller values of H_{col} and faster decay. Due to the fact that ε only enters with the exponential factor, which is constant with respect to $j_{\max}$ for every $\mathbf{x}_k$, the pictures for other values of ε are the same, up to a scalar factor.

be very careful while increasing γ. In certain setups, even a small increase might lead to a substantial quality loss.

7. Convergence of the HermiteGF basis in the space $L^2_\omega(\mathbb{R}^d)$

In chapter 3 we have proved general convergence results for the HermiteGF, or *generalized anisotropic Hermite basis* within the corresponding weighted space

$$L^2_\omega(\mathbb{R}^d) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \int_{\mathbb{R}^d} |f(\mathbf{x})|^2 \omega(\mathbf{x}) \mathrm{d}x < \infty \right\}$$

with the weight $\omega : \mathbb{R}^d \to \mathbb{R}_+$

$$\omega(\mathbf{x}) = \pi^{-d/2} |\det(G)| \exp(2\mathbf{x}^T E^T E \mathbf{x} - \mathbf{x}^T G G^T \mathbf{x})$$

and the inner product

$$\langle f, g \rangle_{L^2_\omega(\mathbb{R}^d)} = \int_{\mathbb{R}^d} f(\mathbf{x}) g(\mathbf{x}) \omega(\mathbf{x}) \mathrm{d}x.$$

However, now we want to look at it in more detail for the specific case of the HermiteGF approximation. It turns out that the HermiteGF approximation is the *best approximation* in that space. In particular, the HermiteGF approximation matches exactly the projection of the anisotropic Gaussian interpolant onto the space $L^2_\omega(\mathbb{R}^d)$. We then derive two approximation error estimates. The first one estimates the truncation error for the case when we cut the Hermite expansion based on the polynomial degree $j_{\max}$. This estimate is based on the multivariate Taylor expansion and on the multinomial theorem. The second estimate is applicable to the case when we cut our expansion based on the dimension-wise bounds on the polynomial degree. In this case, we employ the ladder operator theory developed in section 3.5.

7.1. Anisotropic Gaussian interpolant and $L^2_\omega(\mathbb{R}^d)$

Consider an anisotropic Gaussian interpolant

$$s(\mathbf{x}) = \sum_{k=1}^{N} \phi_k(\mathbf{x}) = \sum_{k=1}^{N} \alpha_k \exp(-(\mathbf{x} - \mathbf{x}_k)^T E^T E (\mathbf{x} - \mathbf{x}_k)), \tag{7.1}$$

where E is the shape matrix. So far, we have considered the infinite HermiteGF expansion of the Gaussians. However, in a numerical algorithm, we will need to truncate the anisotropic HermiteGF expansion at a certain degree $j_{\max}$. Note that the number of basis functions is then equal to

$$M = \binom{j_{\max} + d}{j_{\max}}.$$

Let us first define the approximation of the anisotropic Gaussian interpolant via the truncated HermiteGF expansion. Plugging in the corresponding expansion (6.3), that was cut at a degree $j_{\max}$, instead of each Gaussian in the interpolant

(7.1), we get

$$s_{j_{\max}}^{G,E,t}(\mathbf{x}) = \sum_{k=1}^{N} \alpha_k \exp(\boldsymbol{\Delta}_k^T(-E^TE+\tilde{G})\boldsymbol{\Delta}_k) \sum_{|\boldsymbol{\ell}|=0}^{j_{\max}} \frac{(G^{-1}E^TE\boldsymbol{\Delta}_k)^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}} H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0)$$

$$= \sum_{|\boldsymbol{\ell}|=0}^{j_{\max}} \left(\sum_{k=1}^{N} \alpha_k \exp(\boldsymbol{\Delta}_k^T(-E^TE+\tilde{G})\boldsymbol{\Delta}_k) \frac{(G^{-1}E^TE\boldsymbol{\Delta}_k)^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}} \right) H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0),$$

where each Gaussian we approximated with the truncated HermiteGF expansion. We later refer to $s_{j_{\max}}^{G,E,t}(\mathbf{x})$ as *HermiteGF interpolant*. We want to interpret $s_{j_{\max}}^{G,E,t}(\mathbf{x})$ as a partial sum of the expansion of s in the HermiteGF basis in the weighted L^2 space $L^2_\omega(\mathbb{R}^d)$. Recall that in chapter 3 we considered the weight

$$\omega(\mathbf{x}) = \pi^{-d/2}|\det(G)|\exp(2\mathbf{x}^TE^TE\mathbf{x} - \mathbf{x}^TGG^T\mathbf{x}).$$

In the L^2 space with this weight, the non-shifted HermiteGF functions are orthogonal and there is a ladder operator framework available. However, to accommodate the shift in the basis, we now have to also shift our weight ω to preserve orthogonality. We consider the weight

$$\omega_0(\mathbf{x}) = \omega(\mathbf{x}-\mathbf{x}_0)$$

and the corresponding weighted L^2 space $L^2_{\omega_0}(\mathbb{R}^d)$ instead. We now prove that the HermiteGF interpolant is the best approximation in this space. For the convenience of notation, we denote the shifted basis functions as

$$H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t}(\mathbf{x}) = H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0) \quad \text{for all} \quad \boldsymbol{\ell} \in \mathbb{N}_0^d.$$

We now proceed to proving the best approximation result.

Proposition 7.1.1: Best approximation

Denote

$$U_{j_{\max}} = \operatorname{span}\{H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0) \,\big|\, |\boldsymbol{\ell}| \le j_{\max}\} \subset L^2_{\omega_0}(\mathbb{R}^d).$$

Then $s_{j_{\max}}^{G,E,t}$ is the *best approximation* of s in $U_{j_{\max}}$:

$$\|s - s_{j_{\max}}^{G,E,t}\|_{\omega_0} = \min_{u \in U_{j_{\max}}} \|s-u\|_{\omega_0}.$$

Proof. In order to show that $s_{j_{\max}}^{G,E,t}$ is the *best approximation* in $U_{j_{\max}}$, it is enough to prove that $s_{j_{\max}}^{G,E,t}$ is the orthogonal projection of $s \in L^2_{\omega_0}(\mathbb{R}^d)$ onto $U_{j_{\max}}$. In particular, we need to prove that

$$s_{j_{\max}}^{G,E,t}(\mathbf{x}) = \sum_{|\boldsymbol{\ell}|\le j_{\max}} \frac{\langle s, H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t}\rangle_{\omega_0}}{\langle H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t}, H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t}\rangle_{\omega_0}} H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t}(\mathbf{x}). \tag{7.2}$$

Before proceeding to the evaluation of the coefficient $\langle s, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_{\omega_0}$, we compute the individual coefficients $\langle \phi_k, H_{\boldsymbol{\ell}}^{G,E,t}\rangle_{\omega_0}$:

$$\begin{aligned}\langle \phi_k, H^{G,E,t}_{\boldsymbol{\ell},\mathbf{x}_0}\rangle_{\omega_0} &= \int_{\mathbb{R}^d} \phi_k(\mathbf{x}) H^{G,E,t}_{\boldsymbol{\ell}}(\mathbf{x}-\mathbf{x}_0)\omega(\mathbf{x}-\mathbf{x}_0)\mathrm{d}\mathbf{x}\\ &= \frac{t^{|\boldsymbol{\ell}|/2}}{\pi^{d/2}\sqrt{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}} \int_{\mathbb{R}^d} h_{\boldsymbol{\ell}}(G^T\mathbf{x})\phi_k(\mathbf{x}+\mathbf{x}_0)\exp(\mathbf{x}^TE^TE\mathbf{x}-\mathbf{x}^TGG^T\mathbf{x})|\det(G)|\mathrm{d}\mathbf{x}\\ &= \frac{t^{|\boldsymbol{\ell}|/2}\mathrm{e}^{-\boldsymbol{\Delta}_kE^TE\boldsymbol{\Delta}_k}}{\pi^{d/2}\sqrt{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}} \int_{\mathbb{R}^d} h_{\boldsymbol{\ell}}(G^T\mathbf{x})\exp(2\boldsymbol{\Delta}_k^TE^TE\mathbf{x}-\mathbf{x}^TGG^T\mathbf{x})|\det(G)|\mathrm{d}\mathbf{x},\end{aligned}$$

where we used that

$$\begin{aligned}&\phi_k(\mathbf{x}+\mathbf{x}_0) = \exp((\mathbf{x}-\mathbf{x}_k+\mathbf{x}_0)^TE^TE(\mathbf{x}-\mathbf{x}_k+\mathbf{x}_0))\\ &= \exp((\mathbf{x}-\boldsymbol{\Delta}_k)^TE^TE(\mathbf{x}-\boldsymbol{\Delta}_k)) = \exp(-\boldsymbol{\Delta}_kE^TE\boldsymbol{\Delta}_k + 2\boldsymbol{\Delta}_k^TE^TE\mathbf{x} - \mathbf{x}^TE^TE\mathbf{x}).\end{aligned}$$

Changing the variable $\bar{\mathbf{x}} = G^T\mathbf{x}$, we get

$$\langle \phi_k, H^{G,E,t}_{\boldsymbol{\ell},\mathbf{x}_0}\rangle_{\omega_0} = \frac{t^{|\boldsymbol{\ell}|/2}\mathrm{e}^{-\boldsymbol{\Delta}_kE^TE\boldsymbol{\Delta}_k}}{\pi^{d/2}\sqrt{2^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}} \int_{\mathbb{R}^d} h_{\boldsymbol{\ell}}(\bar{\mathbf{x}})\exp(2\boldsymbol{\Delta}_k^TE^TEG^{-T}\bar{\mathbf{x}}-\bar{\mathbf{x}}^T\bar{\mathbf{x}})\mathrm{d}\bar{\mathbf{x}}.$$

To compute the remaining integral we recall a useful property of Hermite polynomials [GR14, § 7.374, Expr. 6]:

$$\int_{\mathbb{R}} h_n(x)\mathrm{e}^{-(x-a)^2}\mathrm{d}x = 2^n\sqrt{\pi}a^n \quad \forall x\in\mathbb{R}, n\in\mathbb{N}_0. \tag{7.3}$$

We observe that

$$2\boldsymbol{\Delta}_k^TE^TEG^{-T}\bar{\mathbf{x}}-\bar{\mathbf{x}}^T\bar{\mathbf{x}} = -(\bar{\mathbf{x}}-G^{-1}E^TE\boldsymbol{\Delta}_k)^T(\bar{\mathbf{x}}-G^{-1}E^TE\boldsymbol{\Delta}_k)+\boldsymbol{\Delta}_k^T\tilde{G}\boldsymbol{\Delta}_k,$$

with $\tilde{G} = E^TEG^{-T}G^{-1}E^TE$. Using (7.3) dimension-wise, we get

$$\langle \phi_k, H^{G,E,t}_{\boldsymbol{\ell},\mathbf{x}_0}\rangle_{\omega_0} = \exp(-\boldsymbol{\Delta}_kE^TE\boldsymbol{\Delta}_k+\boldsymbol{\Delta}_k\tilde{G}\boldsymbol{\Delta}_k)\frac{t^{|\boldsymbol{\ell}|/2}(G^{-1}E^TE\boldsymbol{\Delta}_k)^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{\boldsymbol{\ell}!}}.$$

Adding up N elements of the original interpolant, we get

$$\begin{aligned}\langle s(x), H^{G,E,t}_{\boldsymbol{\ell},\mathbf{x}_0}\rangle_{\omega_0} &= \sum_{k=1}^N \alpha_k\langle \phi_k, H^{G,E,t}_{\boldsymbol{\ell},\mathbf{x}_0}\rangle_{\omega_0}\\ &= \sum_{k=1}^N \alpha_k\exp(-\boldsymbol{\Delta}_k^TE^TE\boldsymbol{\Delta}_k+\boldsymbol{\Delta}_k^T\tilde{G}\boldsymbol{\Delta}_k)\frac{t^{|\boldsymbol{\ell}|/2}(G^{-1}E^TE\boldsymbol{\Delta}_k)^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{\boldsymbol{\ell}!}}.\end{aligned}$$

Using the fact that $\langle H^{G,E,t}_{\boldsymbol{\ell},\mathbf{x}_0}, H^{G,E,t}_{\boldsymbol{\ell},\mathbf{x}_0}\rangle_{\omega_0} = \langle H^{G,E,t}_{\boldsymbol{\ell}}, H^{G,E,t}_{\boldsymbol{\ell}}\rangle_{\omega} = t^{|\boldsymbol{\ell}|}$ and summing over $\boldsymbol{\ell} \le j_{\max}$, we arrive to (7.2) which proves the statement of the lemma. □

7.2. Convergence of the truncated anisotropic HermiteGF expansion

Even though we have proved that the HermiteGF interpolant is the *best approximation* in the corresponding subspace of $L^2_{\omega_0}(\mathbb{R}^d)$, to be able to quantify the accuracy of the HermiteGF interpolant, we seek to derive an explicit estimate of the truncation error. In order to do that, we need the following lemma.

Lemma 7.2.1: Exponential tail

Consider $\mathbf{y} \in \mathbb{R}^d$ with $y_i \geq 0$ for all $i = 1, \ldots, d$ and $j_{\max} \in \mathbb{N}$. Then,

$$\sum_{|\boldsymbol{\ell}| \geq j_{\max}} \frac{\mathbf{y}^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} \leq \sum_{|\boldsymbol{\ell}| = j_{\max}} \exp(\|\mathbf{y}\|_1) \frac{\mathbf{y}^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!}, \tag{7.4}$$

where $\|\mathbf{y}\|_1 = |y_1| + \ldots + |y_d|$.

Proof. Consider the function

$$f(\mathbf{y}) = \exp(y_1)\exp(y_2)\cdots\exp(y_d) = \exp(\|\mathbf{y}\|_1), \quad \mathbf{y} \geq 0.$$

Then the Taylor series for the function f expanded at a point $\mathbf{a} \in \mathbb{R}^d$ reads as

$$f_{\mathbf{a}}(\mathbf{y}) = \sum_{\boldsymbol{\ell} \in \mathrm{N}_0^d} \frac{\partial_{\boldsymbol{\ell}} f(\mathbf{a})}{\boldsymbol{\ell}!}(\mathbf{y} - \mathbf{a})^{\boldsymbol{\ell}} = \sum_{|\boldsymbol{\ell}| \leq j_{\max}} \frac{\partial_{\boldsymbol{\ell}} f(\mathbf{a})}{\boldsymbol{\ell}!}(\mathbf{y} - \mathbf{a})^{\boldsymbol{\ell}} + R_{j_{\max}}(\mathbf{y}),$$

where $R_j(\mathbf{y})$ is the remainder of the Taylor series.

We note that for $\mathbf{a} = \mathbf{0}$ the remainder of the Taylor series coincides with the estimated sum. Then, according to the multivariate Taylor's theorem, there exists $\boldsymbol{\xi} \in [0, \mathbf{y}]$ such that

$$\sum_{|\boldsymbol{\ell}| \geq j+1} \frac{\mathbf{y}^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} = R_{j_{\max}}(\mathbf{y}) = \sum_{|\boldsymbol{\ell}| = j+1} \exp(\|\boldsymbol{\xi}\|_1) \frac{\mathbf{y}^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!}.$$

Noting that $\exp(\|\boldsymbol{\xi}\|_1) \leq \exp(\|\mathbf{y}\|_1)$ for non-negative $\mathbf{y}$ we arrive to the estimate (7.4). □

We can now proceed to the estimation of the truncation error. We continue working in $L^2_{\omega_0}(\mathbb{R}^d)$ since in this case we can represent the norm of the difference $s - s^{G,E,t}_{j_{\max}}$ through the known coefficients $\langle s, H^{G,E,t}_{\boldsymbol{\ell}} \rangle_{\omega_0}$.

Theorem 7.2.1: Truncation error estimate

Let $s \in L^2_\omega(\mathbb{R}^d)$, $E, G \in \mathbb{R}^{d \times d}$ invertible and define $\tilde{G} = E^T E G^{-T} G^{-1} E^T E$. If $E^T E - \tilde{G}$ is positive definite, then

$$\|s - s^G_{j_{\max}}\|^2_{\omega_0} \leq \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2} \frac{(2/t)^{(j_{\max}+1)}}{(j_{\max}+1)!} \sum_{k=1}^{N} \exp((2/t)\|\mathbf{y}_k\|^2_{\boldsymbol{\ell}_2}) \|\mathbf{y}_k\|^{2(j_{\max}+1)}_{\boldsymbol{\ell}_2}, \tag{7.5}$$

where $\mathbf{y}_k = G^{-1} E^T E \boldsymbol{\Delta}_k$ and the vector $\boldsymbol{\alpha}$ contains the coefficients of the interpolant in the original basis.

Proof. Using the Parseval's identity for the orthogonal basis, we get

$$\|s - s^G_{j_{\max}}\|^2_{\omega_0} = \sum_{|\boldsymbol{\ell}| \geq j_{\max}+1} \frac{\langle s, H^{G,E,t}_{\boldsymbol{\ell},\mathbf{x}_0} \rangle^2_{\omega_0}}{\langle H^{G,E,t}_{\boldsymbol{\ell},\mathbf{x}_0}, H^{G,E,t}_{\boldsymbol{\ell},\mathbf{x}_0} \rangle_{\omega_0}}$$

Now, with the help of Proposition 7.1.1 and the Cauchy-Schwartz inequality, we obtain

$$
\begin{aligned}
\|s - s^G_{j_{\max}}\|^2_{\omega_0} &= \sum_{|\boldsymbol{\ell}|\geq j_{\max}+1} \left(\sum_{k=1}^{N} \alpha_k \underbrace{\exp(\boldsymbol{\Delta}_k^T(-E^TE + \tilde{G})\boldsymbol{\Delta}_k)}_{\leq 1} \frac{(G^{-1}E^TE\boldsymbol{\Delta}_k)^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{\boldsymbol{\ell}!}\sqrt{t^{|\boldsymbol{\ell}|}}} \right)^2 \\
&\leq \sum_{|\boldsymbol{\ell}|\geq j_{\max}+1} \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2} \sum_{k=1}^{N} \frac{(G^{-1}E^TE\boldsymbol{\Delta}_k)^{2\boldsymbol{\ell}}2^{|\boldsymbol{\ell}|}}{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}.
\end{aligned}
$$

Denote $\mathbf{y}_k = G^{-1}E^TE\boldsymbol{\Delta}_k$ and $\tilde{\mathbf{y}}_k = \left(\frac{2}{t}(y_k)^2_1 \quad \dots \quad \frac{2}{t}(y_k)^2_d\right)$. Then, using (7.4) we get

$$
\begin{aligned}
\|s - s^{G,E,t}_{j_{\max}}\|^2_{\omega} &\leq \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2} \sum_{|\boldsymbol{\ell}|\geq j_{\max}+1} \sum_{k=1}^{N} \frac{\tilde{\mathbf{y}}^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} = \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2} \sum_{k=1}^{N} \sum_{|\boldsymbol{\ell}|\geq j_{\max}+1} \frac{\tilde{\mathbf{y}}_k^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} \\
&\leq \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2} \sum_{k=1}^{N} \sum_{|\boldsymbol{\ell}|= j_{\max}+1} \exp(\|\tilde{\mathbf{y}}_k\|_1) \frac{\tilde{\mathbf{y}}_k^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} \\
&= \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2} \left(\frac{2}{t}\right)^{(j_{\max}+1)} \sum_{k=1}^{N} \exp\left(\frac{2}{t}\|\mathbf{y}_k\|^2_{\boldsymbol{\ell}_2}\right) \sum_{|\boldsymbol{\ell}|= j_{\max}+1} \frac{\mathbf{y}_k^{2\boldsymbol{\ell}}}{\boldsymbol{\ell}!}
\end{aligned}
$$

Applying the multinomial theorem to the second sum we arrive to the estimate (7.5). □

In case the nodes $\{\mathbf{y}_k\}_{k=1\dots N}$ are lying within a sphere, one can simplify the estimate above.

Corollary 7.2.1. *If* $\|G^{-1}E^TE\boldsymbol{\Delta}_k\|_{\boldsymbol{\ell}_2} < \rho$ *for all* $k = 1, \dots, N$ *then*

$$
\|s - s^G_{j_{\max}}\|_{\omega_0} \leq \|\boldsymbol{\alpha}\|_{\boldsymbol{\ell}_2}\sqrt{N}\exp\left(\frac{\rho^2}{t}\right)\frac{(2\rho^2/t)^{(j_{\max}+1)/2}}{\sqrt{(j_{\max}+1)!}}. \tag{7.6}
$$

Proof. From the estimate (7.5) we get

$$
\begin{aligned}
\|s - s^G_{j_{\max}}\|^2_{\omega_0} &\leq \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2} \left(\frac{2}{t}\right)^{(j_{\max}+1)} \sum_{k=1}^{N} \exp((2/t)\|\mathbf{y}_k\|^2_{\boldsymbol{\ell}_2}) \frac{\|\mathbf{y}_k\|^{2(j_{\max}+1)}_{\boldsymbol{\ell}_2}}{(j_{\max}+1)!} \\
&\leq \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2} N \exp\left(\frac{2\rho^2}{t}\right) \frac{(2\rho^2/t)^{(j_{\max}+1)}}{(j_{\max}+1)!}.
\end{aligned}
$$

□

One can see that in order for the method to be accurate, the value of the components of the vectors $G^{-1}E^TE\boldsymbol{\Delta}_k$ should be rather small. In particular, it is advantageous for the convergence speed to have

$$
\rho\sqrt{\frac{2}{t}} < 1.
$$

Since we are mostly focusing on the case of flat Gaussians, a lot of times it is automatically achieved by the factor E^TE with G of order one. In other cases, a large magnitude of G improves the decay of the expansion. On the other hand, for large magnitude of G, the argument of the Hermite polynomial $G^T\mathbf{x}$ gets large, which can, in turn, lead to ill-conditioning. We investigate this phenomena for isotropic Gaussians numerically in the Section 12.1.2.

As for the combinatorial constant, one can see that in 1D the factor $(j_{\max}+1)!$ easily balances N. However, with the increase of the dimensionality the rate of decay of the factor $\frac{N}{(j_{\max}+1)!}$ deteriorates dramatically. This is related to the curse of the dimensionality. Therefore in higher dimensions, one has to be particularly careful to choose a small enough ρ and a large enough $j_{\max}$ to get a good approximation quality.

For the isotropic case, one can simplify the error estimate even further.

Corollary 7.2.2. *If* $E = \varepsilon \mathrm{Id}$*,* $G = \gamma \mathrm{Id}$ *and the nodes* $\|\mathbf{x}_k\|_2 < L$ *for all* $k = 1, \ldots, N$ *then*

$$\|s - s_{j_{\max}}^{G,E,t}\|_{\omega_0} \leq \|\boldsymbol{\alpha}\|_{\boldsymbol{\ell}_2} \sqrt{N} \exp\left(\frac{\gamma^{-2}\varepsilon^4 L^2}{t}\right) \frac{(2t^{-1}\gamma^{-2}\varepsilon^4 L^2)^{(j_{\max}+1)/2}}{(j_{\max}+1)!} \tag{7.7}$$

Moreover, in the isotropic case we have more information about the coefficients α_k which allows for a more precise analysis.

7.2.1. The influence of the coefficients α_k in the isotropic case

In the isotropic case, the RBF coefficients α_k tend to infinity as $\varepsilon \to 0$. In particular, according to [LF05, § 4, Corollary 4.1] $\alpha_k \sim \varepsilon^{-2P}$ for $k = 1, \ldots, N$, where $P \in \mathbb{N}_0$ defined as follows:

$$P = \min\left\{ K \in \mathbb{N}_0 \,\middle|\, N \leq \binom{K+d}{d} \right\}.$$

Therefore, in case $j_{\max} \geq P$, the behavior of the coefficients α_k is neutralized. However, one has to make sure by choosing appropriate γ and $j_{\max}$ that the remaining decay is good enough for reaching the desired approximation quality. In particular, it could be advantageous to choose γ such that

$$\sqrt{t}\gamma > \sqrt{2}\varepsilon^2 L.$$

However, as before, one has to keep in mind that a very large magnitude of γ might lead to ill-conditioning in practical applications. For practical experiments see Section 12.1.2.

7.2.2. Refined convergence proof in 1D

In this section, we try to improve our estimate from the previous section by handling the term $\exp(-\boldsymbol{\Delta}_k^T E^T E \boldsymbol{\Delta}_k + \boldsymbol{\Delta}_k^T \tilde{G} \boldsymbol{\Delta}_k)$ more carefully. For the sake of simplicity, we consider the 1D case with the zero shift $x_0 = 0$ in order to see if it provides any significant improvement. Note that in 1D the number of the expansion functions is equal to the maximum polynomial degree $j_{\max}$.

Theorem 7.2.2: Refined estimate in 1D

If $s \in L^2_\omega(\mathbb{R})$, $\varepsilon < \gamma$, $N < j_{\max}$ then

$$\|s - s^{\gamma,\varepsilon,t}_{j_{\max}}\|_\omega \le \|\boldsymbol{\alpha}\|_{\boldsymbol{\ell}_2}\sqrt{N}\max\{\mathrm{e}^{-(j_{\max}+1)/2}, \mathrm{e}^{-\beta L^2}\}\mathrm{e}^{\frac{\rho^2}{t}}\frac{(2\rho^2/t)^{(j_{\max}+1)/2}}{\sqrt{(j_{\max}+1)!}}, \qquad (7.8)$$

where $L = \max_k\{|x_k|\}$, $\beta = \varepsilon^2\left(1 - \frac{\varepsilon^2}{\gamma^2}\right)$, $\rho = \frac{\varepsilon^2 L}{\gamma}$.

Proof. Using Proposition 7.1.1, Parseval's identity and Cauchy-Schwartz inequality, we get

$$\begin{aligned}
\|s - s^{\gamma,\varepsilon,t}_{j_{\max}}\|^2_\omega &= \sum_{\ell \le j_{\max}+1} \frac{\langle s, H^{\gamma,\varepsilon,t}_\ell\rangle^2_\omega}{\langle H^{\gamma,\varepsilon,t}_\ell, H^{\gamma,\varepsilon,t}_\ell\rangle_\omega} \\
&= \sum_{\ell \ge j_{\max}+1}\left(\sum_{k=1}^{N}\alpha_k \exp\left(\varepsilon^2 x_k^2\left(\frac{\varepsilon^2}{\gamma^2} - 1\right)\right)\frac{\sqrt{2^\ell}\varepsilon^{2\ell}x_k^\ell}{\gamma^\ell\sqrt{t^\ell \ell!}}\right)^2 \\
&\le \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2}\sum_{\ell \ge j_{\max}+1}\frac{2^\ell\varepsilon^{4\ell}}{t^\ell\gamma^{2\ell}\ell!}\left(\sum_{k=1}^{N}\exp\left(2\varepsilon^2\left(\frac{\varepsilon^2}{\gamma^2} - 1\right)x_k^2\right)x_k^{2\ell}\right).
\end{aligned}$$

Consider the function

$$g(x) = \mathrm{e}^{-2\beta x^2}x^{2\ell}, \quad \beta > 0,\ \ell \ge j_{\max} + 1.$$

Note that for $\beta = \varepsilon^2\left(1 - \frac{\varepsilon^2}{\gamma^2}\right)$ and $x = x_k$, the value $g(x_k)$ coincides with the corresponding element of the sum above. If we are able to determine the maximum of $g(x)$ in our interpolation domain, it can serve as an upper bound for all elements of the sum over k.

The derivative of the function $g(x)$ reads as

$$g'(x) = -4\beta x\mathrm{e}^{-2\beta x^2}x^{2\ell} + 2\ell x^{2\ell-1}\mathrm{e}^{-2\beta x^2} = 2\mathrm{e}^{-2\beta x^2}x^{2\ell-1}(-2\beta x^2 + \ell).$$

Since the function g is symmetric, the global maximum is reached at

$$x_* = \sqrt{\frac{\ell}{2\beta}}.$$

Since all nodes x_k are within the interval $[-L, L]$, we are only interested at the maximum at that interval. Therefore, on the interval $[-L, L]$ we have the following estimate

$$\begin{cases} |g(x)| \le g(L) & \text{if} \quad x_* > L, \\ |g(x)| \le g(x_*) & \text{if} \quad x_* < L. \end{cases}$$

The condition $x_* < L$ implies

$$\sqrt{\frac{\ell}{2\beta}} < L \ \Leftrightarrow\ \ell < 2\beta L^2.$$

Denote $\ell_0 = \lfloor 2\beta L^2 \rfloor$ and $\rho = \frac{\varepsilon^2 L}{\gamma}$. Then

$$\begin{aligned}
\|s - s^{\gamma}_{j_{\max}}\|^2_{\omega} &\le \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2} \sum_{\ell \ge j_{\max}+1} \frac{2^{\ell}\varepsilon^{4\ell}}{t^{\ell}\gamma^{2\ell}\ell!} \left(\sum_{k=1}^{N} g(x_k) \right) \\
&\le \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2} N \left(\sum_{\ell=j_{\max}+1}^{\ell_0} \frac{\mathrm{e}^{-\ell}\ell^{\ell}}{(2\beta)^{\ell}} \frac{2^{\ell}\varepsilon^{4\ell}}{t^{\ell}\gamma^{2\ell}\ell!} + \sum_{\ell \ge \ell_0+1} \mathrm{e}^{-2\beta L^2} L^{2\ell} \frac{2^{\ell}\varepsilon^{4\ell}}{t^{\ell}\gamma^{2\ell}\ell!} \right) \\
&\le \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\ell}_2} N \max\{\mathrm{e}^{-(j_{\max}+1)}, \mathrm{e}^{-2\beta L^2}\} \sum_{\ell \ge j_{\max}+1} \frac{(2\rho^2/t)^{\ell}}{\ell!},
\end{aligned}$$

where we used that $\ell/2\beta < L^2$ for $\ell < \ell_0$. Applying the estimate of the tail of the Taylor expansion of an exponential function from the Lemma 7.2.1, we arrive to the estimate (7.8). □

One can see that the estimate above is identical to the multivariate one with the only difference being the factor $\max\{\mathrm{e}^{-(j_{\max}+1)}, \mathrm{e}^{-2\beta L^2}\}$. However, since we are interested mostly in small values of ε, the value of $\beta = \varepsilon^2(\varepsilon^2/\gamma^2 - 1)$ in most common scenarios requiring stabilization is rather small. That, in turn, implies that the value of constant will be very close to 1 which was our estimate in the multivariate version. Therefore, we conclude that treating the exponential factor this way doesn't provide any significant improvement for the estimate for real life cases.

7.3. Convergence for the dimension-wise truncation

Even though we are mostly focusing on the truncation by the cumulative polynomial degree $j_{\max}$, or the *simplex* index set, it is also possible to truncate the expansion dimension-wise, based on the maximum number of elements M_i in every dimension $i = 1 \dots d$. This truncation strategy is especially useful in case the centers of RBFs are also on a tensor grid (see, for example, [RFK16, YK17]), since in this case a Kronecker product representation of the matrices can be employed which brings a significant performance gain. However, it limits one of the main features of the RBF methods, namely, the ability to work with scattered points.

In this thesis, we focus on the strategy considered in the previous sections. However, in this section we derive a truncation error estimate for the dimension-wise truncation strategy since it also brings some insights into the structure of the basis.

We consider a vector

$$\mathbf{M} = \big(M_1, \dots, M_d\big)$$

of the maximum number of basis functions in each dimension. Then, the HermiteGF interpolant takes the form

$$s_{\mathbf{M}}^{G,E,t}(\mathbf{x}) = \sum_{\ell_1=0}^{M_1} \dots \sum_{\ell_d=0}^{M_d} \sum_{k=1}^{N} \alpha_k \frac{\langle \phi_k, H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t} \rangle_{\omega_0}}{\langle H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t}, H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t} \rangle_{\omega_0}} H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0)$$

with the result of the Proposition 7.1.1 that

$$\langle \phi_k, H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t} \rangle_{\omega_0} = \exp(-\boldsymbol{\Delta}_k^T E^T E \boldsymbol{\Delta}_k + \boldsymbol{\Delta}_k^T \tilde{G} \boldsymbol{\Delta}_k) \frac{(G^{-1} E^T E \boldsymbol{\Delta}_k)^{\boldsymbol{\ell}} \sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|} \boldsymbol{\ell}!}}.$$

Analogously to the interpolant $s_{j_{\max}}^{G,E,t}$, $s_{\mathbf{M}}^{G,E,t}$ also is an orthogonal projection to the subspace of $L_{\omega_0}^2(\mathbb{R}^d)$ corresponding to the set of indexes

$$\mathcal{K}_{\mathbf{M}} = \{\boldsymbol{\ell} \in \mathbb{N}_0^d | \ell_i < M_i \ \forall i = 1 \dots d\}.$$

Then the truncation error takes the form

$$\|s - s_{\mathbf{M}}^{G,E,t}\|_{\omega_0} = \int_{\mathbb{R}^d} (s(\mathbf{x}) - s_{\mathbf{M}}^{G,E,t}(\mathbf{x}))^2 \omega_0(\mathbf{x}) \mathrm{d}\mathbf{x}$$

$$= \int_{\mathbb{R}^d} \left((s(\mathbf{x}) - \sum_{\ell_1=0}^{M_1} \dots \sum_{\ell_d=0}^{M_d} \frac{\langle s, H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t} \rangle_{\omega_0}}{\langle H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t}, H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t} \rangle_{\omega_0}} H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0) \right)^2 \omega(\mathbf{x}-\mathbf{x}_0) \mathrm{d}\mathbf{x}$$

We note that

$$\langle s, H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t} \rangle_{\omega_0} = \int_{\mathbb{R}^d} s(\mathbf{x}) H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0) \omega(\mathbf{x}-\mathbf{x}_0) \mathrm{d}\mathbf{x}$$

$$= \int_{\mathbb{R}^d} s(\mathbf{x}+\mathbf{x}_0) H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) \omega(\mathbf{x}) \mathrm{d}\mathbf{x} = \langle s_{\mathbf{x}_0}, H_{\boldsymbol{\ell}}^{G,E,t} \rangle_{\omega},$$

where $s_{\mathbf{x}_0}(\mathbf{x}) = s(\mathbf{x}+\mathbf{x}_0)$ is the shifted original interpolant. Therefore, we can interpret the sum above as

$$\|s - s_{\mathbf{M}}^{G,E,t}\|_{\omega_0} = \int_{\mathbb{R}^d} \left((s_{\mathbf{x}_0}(\mathbf{x}) - \sum_{\ell_1=0}^{M_1} \dots \sum_{\ell_d=0}^{M_d} \frac{\langle s_{\mathbf{x}_0}, H_{\boldsymbol{\ell}}^{G,E,t} \rangle_{\omega}}{\langle H_{\boldsymbol{\ell}}^{G,E,t}, H_{\boldsymbol{\ell},\mathbf{x}_0}^{G,E,t} \rangle_{\omega}} H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) \right)^2 \omega(\mathbf{x}) \mathrm{d}\mathbf{x}$$

$$= \|s_{\mathbf{x}_0} - P_{\mathbf{M}} s_{\mathbf{x}_0}\|_{\omega},$$

where $P_{\mathbf{M}}$ is the orthogonal projector in the non-shifted space $L_{\omega}^2(\mathbb{R}^d)$

$$P_{\mathbf{M}} f = \sum_{\boldsymbol{\ell} \in \mathcal{K}_{\mathbf{M}}} \frac{\langle f, H_{\boldsymbol{\ell}}^{G,E,t} \rangle_{\omega}}{\langle H_{\boldsymbol{\ell}}^{G,E,t}, H_{\boldsymbol{\ell}}^{G,E,t} \rangle_{\omega}} H_{\boldsymbol{\ell}}^{G,E,t}.$$

For this norm we have already proved in section 3.5 an estimate for all f from the weighted Schwartz space $\mathcal{S}_\omega$ that was defined as

$$\mathcal{S}_\omega(\mathbb{R}^d) := \left\{ f : \mathbb{R}^d \to \mathbb{R} | f\sqrt{\omega} \in \mathcal{S}(\mathbb{R}^d) \right\}$$

In particular, according to Theorem 3.5.2, the error estimate writes as follows

$$\|f - P_M f\|_{\omega} \leq \frac{t^{|\boldsymbol{r}|/2} \sqrt{(\mathbf{M} - \boldsymbol{r})!}}{\sqrt{\mathbf{M}!}} \|A_\omega^{\boldsymbol{r}} f\|_{\omega},$$

where $\mathbf{r}$ is an integer vector with $r_i \leq M_i$ for all $i = 1 \dots d$ and A_ω is the lowering operator

$$A_\omega f(\mathbf{x}) = \frac{G^{-1}}{\sqrt{2t}}(\nabla f(\mathbf{x}) + 2E^T E\mathbf{x} f(\mathbf{x})), \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

In order to use the estimate above, we should make sure that $s \in \mathcal{S}_\omega(\mathbb{R}^d)$. In particular, it is enough to show that Gaussians $\phi_k(\mathbf{x}+\mathbf{x}_0) \in \mathcal{S}_\omega(\mathbb{R}^d)$, then, their linear combination also belongs to $\mathcal{S}_\omega(\mathbb{R}^d)$. Recall that $f \in \mathcal{S}_\omega(\mathbb{R}^d)$ if $f\sqrt{\omega} \in \mathcal{S}(\mathbb{R}^d)$. In case of the anisotropic Gaussians $\{\phi_k\}$, we get

$$\begin{aligned}\phi_k(\mathbf{x}+\mathbf{x}_0)\sqrt{\omega} &= \pi^{-d/4}\sqrt{|\det(G)|}\exp(-(\mathbf{x}-\boldsymbol{\Delta}_k)^T E^T E(\mathbf{x}-\boldsymbol{\Delta}_k)) \\ &\quad \cdot \exp\left(\mathbf{x}^T E^T E\mathbf{x} - \frac{1}{2}\mathbf{x}^T G G^T \mathbf{x}\right) \\ &= \pi^{-d/4}\sqrt{|\det(G)|}\exp(-\boldsymbol{\Delta}_k^T E^T E\boldsymbol{\Delta}_k)\exp\left(2\mathbf{x}^T E^T E\boldsymbol{\Delta}_k - \frac{1}{2}\mathbf{x}^T G G^T \mathbf{x}\right) \\ &= \pi^{-d/4}\sqrt{|\det(G)|}\exp(\boldsymbol{\Delta}_k^T(-E^T E + \tilde{G})\boldsymbol{\Delta}_k) \\ &\quad \cdot \exp\left(-\frac{1}{2}\left(\mathbf{x} - 2G^{-T}G^{-1}E^T E\boldsymbol{\Delta}_k\right)^T GG^T\left(\mathbf{x} - 2G^{-T}G^{-1}E^T E\boldsymbol{\Delta}_k\right)\right) \in \mathcal{S}(\mathbb{R}^d),\end{aligned}$$

where, as before, G is an invertible matrix and $\tilde{G} = E^T E G^{-T} G^{-1} E^T E$. Therefore, the shifted anisotropic Gaussians $\phi_k(\mathbf{x}+\mathbf{x}_0) \in \mathcal{S}_\omega(\mathbb{R}^d)$.

Since we have made sure that $s_{\mathbf{x}_0} \in \mathcal{S}_\omega(\mathbb{R}^d)$, we can now derive an explicit form of the estimate of $\|s - s_{\mathbf{M}}^{G,E,t}\|_{\omega_0}$ based on Theorem 3.5.2.

Theorem 7.3.1

Let s be the anisotropic Gaussian interpolant (7.1), invertible $E, G \in \mathbb{R}^{d\times d}$. Consider a vector $\mathbf{M} = (M_1, \ldots, M_d)$ of maximum number of functions in each dimension and the corresponding HermiteGF interpolant $s_{\mathbf{M}}^{G,E,t}$. Then, for all integer vectors $\mathbf{r}$ such that $r_i \le M_i$ for all $i = 1\ldots d$, we have

$$\|s - s_{\mathbf{M}}^{G,E,t}\|_{\omega_0} \le \frac{\sqrt{(\mathbf{M}-\mathbf{r})!}}{\sqrt{2^{|\mathbf{r}|}\mathbf{M}!}}\sum_{k=1}^{N}|\alpha_k||(2G^{-1}E^T E\boldsymbol{\Delta}_k)^{\mathbf{r}}|\exp(\boldsymbol{\Delta}_k^T(-E^T E + 2\tilde{G})\boldsymbol{\Delta}_k). \tag{7.9}$$

Proof. As we have shown above, all functions $\phi_k(\mathbf{x}+\mathbf{x}_0) \in \mathcal{S}_\omega(\mathbb{R}^d)$, therefore, we can directly use the result of Theorem 3.5.2 on $s_{\mathbf{x}_0}(\mathbf{x}) = s(\mathbf{x}+\mathbf{x}_0)$.

$$\|s - s_{\mathbf{M}}^{G,E,t}\|_{\omega_0} = \|s_{\mathbf{x}_0} - P_{\mathbf{M}} s_{\mathbf{x}_0}\|_\omega \le \frac{t^{|\mathbf{r}|/2}\sqrt{(\mathbf{M}-\mathbf{r})!}}{\sqrt{\mathbf{M}!}}\|A_\omega^{\mathbf{r}} s_{\mathbf{x}_0}\|_\omega$$

for every $\mathbf{r} \le \mathbf{M}$ element-wise. We are now focusing on computing $\|A_\omega^{r} f\|_\omega$. First of all, we use the Minkowski inequality

$$\|A_\omega^{\mathbf{r}} s_{\mathbf{x}_0}\|_\omega = \left\|\sum_{k=1}^{N}\alpha_k A_\omega^{\mathbf{r}}\phi_{k,\mathbf{x}_0}\right\|_\omega \le \sum_{k=1}^{N}|\alpha_k|\|A_\omega^{\mathbf{r}}\phi_{k,\mathbf{x}_0}\|_\omega,$$

where $\phi_{k,\mathbf{x}_0} = \phi_k(\mathbf{x}+\mathbf{x}_0)$. We are now left with computing $\|A_\omega^{\mathbf{r}}\phi_{k,\mathbf{x}_0}\|_\omega$. Let us first apply the lowering operator A_ω to $\phi_{k,\mathbf{x}_0}$. Using the explicit formula (3.9) for A_ω, we get

$$\begin{aligned}
A_\omega\phi_{k,\mathbf{x}_0} &= \frac{G^{-1}}{\sqrt{2t}}\left(\nabla\phi_k(\mathbf{x}+\mathbf{x}_0) + 2E^TE\mathbf{x}\phi_k(\mathbf{x}+\mathbf{x}_0)\right)\\
&= \frac{G^{-1}}{\sqrt{2t}}\left(-2E^TE(\mathbf{x}-\boldsymbol{\Delta}_k) + 2E^TE\mathbf{x}\right)\phi_k(\mathbf{x}+\mathbf{x}_0)\\
&= \frac{2G^{-1}E^TE\boldsymbol{\Delta}_k}{\sqrt{2t}}\phi_k(\mathbf{x}+\mathbf{x}_0),
\end{aligned}$$

where we used that

$$\nabla\phi_k(\mathbf{x}+\mathbf{x}_0) = \nabla\exp(-(\mathbf{x}-\boldsymbol{\Delta}_k)^TE^TE(\mathbf{x}-\boldsymbol{\Delta}_k)) = -2E^TE(\mathbf{x}-\boldsymbol{\Delta}_k)\phi_k(\mathbf{x}+\mathbf{x}_0).$$

Therefore, the square of the norm $\|A_\omega^{\mathbf{r}}\phi_{k,\mathbf{x}_0}\|_\omega$ can be computed as follows

$$\begin{aligned}
&\|A_\omega^{\mathbf{r}}\phi_{k,\mathbf{x}_0}\|_\omega^2 = \int_{\mathbb{R}^d}\frac{(2G^{-1}E^TE\boldsymbol{\Delta}_k)^{2\mathbf{r}}}{(2t)^{|\mathbf{r}|}}\phi_k(\mathbf{x}+\mathbf{x}_0)^2\omega(\mathbf{x})\mathrm{d}\mathbf{x}\\
&= \pi^{-d/2}\frac{(2G^{-1}E^TE\boldsymbol{\Delta}_k)^{2\mathbf{r}}}{(2t)^{|\mathbf{r}|}}\int_{\mathbb{R}^d}\exp(-2\mathbf{x}^TE^TE\mathbf{x} + 4\mathbf{x}^TE^TE\boldsymbol{\Delta}_k - 2\boldsymbol{\Delta}_kE^TE\boldsymbol{\Delta}_k)\\
&\qquad\cdot\exp(2\mathbf{x}^TE^TE\mathbf{x} - \mathbf{x}^TGG^T\mathbf{x})|\det(G)|\mathrm{d}\mathbf{x}.
\end{aligned}$$

Recalling the Gaussian integral, that reads as

$$\int_{\mathbb{R}}\mathrm{e}^{-(x-a)^2}\mathrm{d}x = \sqrt{\pi}\quad\forall x\in\mathbb{R},$$

we get

$$\begin{aligned}
\|A_\omega^{\mathbf{r}}\phi_{k,\mathbf{x}_0}\|_\omega^2 &= \pi^{-d/2}\frac{(2G^{-1}E^TE\boldsymbol{\Delta}_k)^{2\mathbf{r}}}{(2t)^{|\mathbf{r}|}}\exp(-2\boldsymbol{\Delta}_k^TE^TE\boldsymbol{\Delta}_k)\\
&\quad\cdot\int_{\mathbb{R}^d}\exp(4\mathbf{x}^TE^TE\boldsymbol{\Delta}_k - \mathbf{x}^TGG^T\mathbf{x})|\det(G)|\mathrm{d}\mathbf{x}\\
&\overset{\bar{\mathbf{x}}=G^T\mathbf{x}}{=} \pi^{-d/2}\frac{(2G^{-1}E^TE\boldsymbol{\Delta}_k)^{2\mathbf{r}}}{(2t)^{|\mathbf{r}|}}\exp(-2\boldsymbol{\Delta}_k^TE^TE\boldsymbol{\Delta}_k)\exp(4\boldsymbol{\Delta}_k^T\tilde{G}\boldsymbol{\Delta}_k)\\
&\int_{\mathbb{R}^d}\exp\left(-(\bar{\mathbf{x}}-2G^{-1}E^TE\boldsymbol{\Delta}_k)^T(\bar{\mathbf{x}}-2G^{-1}E^TE\boldsymbol{\Delta}_k)\right)\mathrm{d}\bar{\mathbf{x}}\\
&= \frac{(2G^{-1}E^TE\boldsymbol{\Delta}_k)^{2\mathbf{r}}}{(2t)^{|\mathbf{r}|}}\exp(2\boldsymbol{\Delta}_k^T(-E^TE+2\tilde{G})\boldsymbol{\Delta}_k)
\end{aligned}$$

Using the expression above for all $k = 1\ldots N$, we arrive to the expression (7.9).

□

It was noted in [Lub08, § III.1.1] that the convergence of Hermite functions centered at zero for a shifted Gaussian with a large shift is slow. We can now see that it is also propagated to our case. Indeed, analogously to [Lub08, § III.1.1], we look for all $k = 1\ldots N$ at the term

$$\frac{\sqrt{(\mathbf{M}-\mathbf{r})!}}{\sqrt{2^{|\mathbf{r}|}\mathbf{M}!}}|(2G^{-1}E^TE\boldsymbol{\Delta}_k)^{\mathbf{r}}|$$

of the estimate (7.9). Denote the vector of absolute values of $2G^{-1}E^TE\boldsymbol{\Delta}_k$ as $\boldsymbol{\rho}_k$

with

$$(\rho_k)_i = |2(G^{-1}E^T E\boldsymbol{\Delta}_k)_i| \quad \text{for all} \quad i = 1 \ldots d.$$

One can see that the approximation would have a slow convergence for large shifts $\boldsymbol{\rho}_k$. That is why it is advisable to use the center of the domain as the shift x_0. Moreover, considering $\mathbf{r} = \mathbf{M}$ and using the Stirling's approximation of the factorial, we get

$$\frac{\sqrt{(\mathbf{M}-\mathbf{r})!}}{\sqrt{2^{|\mathbf{r}|}\mathbf{M}!}}|(2G^{-1}E^T E\boldsymbol{\Delta}_k)^{\mathbf{r}}| \sim \frac{1}{(2\pi M_1 \cdot \ldots \cdot M_d)^{1/4}} \left(\frac{\sqrt{\mathrm{e}}\boldsymbol{\rho}_k}{\sqrt{2\mathbf{M}}}\right)^{\mathbf{M}}.$$

Therefore, we would prefer to have

$$\mathbf{M} \geq \frac{\mathrm{e}}{2} \max_{k=1 \ldots N} \boldsymbol{\rho}_k.$$

This condition is not sufficient to have a small error, since the coefficients $\{\alpha_k\}_{k=1}^N$ get large in magnitude. However, it could be helpful while choosing parameters G, E and the shift x_0.

8. Stabilization of the RBF interpolation

Now, when we have studied in detail the behavior of the HermiteGF expansion, we derive an RBF-QR algorithm based on it. Recall that the main idea is to perform a basis transformation to a more stable basis (see section 5.1). We use the HermiteGF functions, or generalized anisotropic Hermite functions, $\{H_{\boldsymbol{\ell}}^{G,E,t}\}$ as the expansion functions. For appropriately chosen parameter G and t we expect the basis $\{H_{\boldsymbol{\ell}}^{G,E,t}\}$ to be better conditioned. We now first derive the HermiteGF-QR algorithm in 1D and then generalize it to a multidimensional form.

8.1. HermiteGF-QR. 1D

In 1D we denote by

$$\Phi(x, X^{\text{cen}}) = \big(\phi_1(x), \quad \dots, \quad \phi_N(x)\big)$$

the vector of Gaussians with center points X^{cen} evaluated at $x \in \mathbb{R}$ and write the stabilization expansion (6.1) as an infinite matrix-vector product

$$\Phi(x, X^{\text{cen}}) = H^{G,E,t}(x - x_0)\, B(\varepsilon, \gamma, t, X^{\text{cen}}) \tag{8.1}$$

where the vector

$$H^{G,E,t}(x - x_0) = \big(H_0^{G,E,t}(x - x_0),\, H_1^{G,E,t}(x - x_0),\, \dots\big)$$

contains all the elements of the polynomial basis $\{H_\ell^{G,E,t}\}_{\ell \geq 0}$ evaluated in the translated point $x - x_0$ and

$$B(\varepsilon, \gamma, t, X^{\text{cen}})_{\ell k} = \exp\left(\varepsilon^2 \Delta_k^2 \left(\frac{\varepsilon^2}{\gamma^2} - 1\right)\right) \frac{\varepsilon^{2\ell}\sqrt{2^\ell}}{\gamma^\ell \sqrt{t^\ell \ell!}} \Delta_k^\ell$$

is an $\infty \times N$ matrix. The major part of the ill-conditioning is now confined in the matrix B. Since B is independent of the point x where the basis function is evaluated, both the evaluation and interpolation matrix can be expressed in the form (8.1) with the same matrix B.

We follow the RBF-QR approach and further split

$$B^T = CD$$

into a well-conditioned full $N \times \infty$ matrix C and an infinite diagonal matrix D, where all harmful effects are confined in D. In the case of expansion (6.1), the following setup follows naturally from the Chebyshev-QR theory [FLF11, § 4.1.3],

$$C_{k\ell} = \exp\left(\varepsilon^2 \Delta_k^2 \left(\frac{\varepsilon^2}{\gamma^2} - 1\right)\right) \frac{\Delta_k^\ell}{L^\ell}, \quad D_{\ell\ell} = \frac{\varepsilon^{2\ell}\sqrt{2^\ell}}{\gamma^\ell \sqrt{t^\ell \ell!}}\, L^\ell.$$

Here we also divide each coefficient by the radius of the domain L containing the center points in order to avoid ill-conditioning in C coming from taking high powers of x_k. That might be dangerous when the domain is too large, however, it still extends the range of domain diameters possible.

The goal is now to find a basis $\{\psi_j\}$ spanning the same space as $\{\phi_k\}$ but yielding a better conditioned collocation matrix. In particular, we need an invertible matrix X such that $X^{-1}\Phi(x)^T$ is better conditioned. Let us perform a QR-decomposition on $C = QR$. Then we get,

$$\Phi(x)^T = CDH^{G,E,t}(x-x_0)^T = Q\begin{pmatrix}R_1 & R_2\end{pmatrix}\begin{pmatrix}D_1 & 0\\ 0 & D_2\end{pmatrix}H^{G,E,t}(x-x_0)^T$$

where R_1 and D_1 are $N \times N$ matrices containing the upper (left) block of the infinite matrices R and D, respectively, while R_2 and D_2 assemble the remaining entries. Consider $X = QR_1D_1$.[1] The new basis $\Psi(\mathbf{x})^T := X^{-1}\Phi(x)^T$ can be formed as,

$$\begin{aligned}\Psi(x)^T &= D_1^{-1}R_1^{-1}Q^{\mathrm{H}}\Phi(x)^T\\ &= D_1^{-1}R_1^{-1}Q^{\mathrm{H}}Q\begin{pmatrix}R_1D_1 & R_2D_2\end{pmatrix}H^{G,E,t}(x-x_0)^T\\ &= \begin{pmatrix}\mathrm{Id} & D_1^{-1}R_1^{-1}R_2D_2\end{pmatrix}H^{G,E,t}(x-x_0)^T.\end{aligned}$$

To avoid under/overflow in the computation of $D_1^{-1}R_1^{-1}R_2D_2$, we form the two matrices $\tilde{R} = R_1^{-1}R_2$ and $\tilde{D}$ with elements

$$\tilde{D}_{i,j} = \left(\frac{\varepsilon^2 L}{\gamma}\sqrt{\frac{2}{t}}\right)^{N+j-i}\sqrt{\frac{i!}{(N+j)!}}, \qquad 1 \le i \le N, \quad j \ge 1,$$

and compute their Hadamard product, $\tilde{R}.*\tilde{D}$. Despite the harmful effects contained in D, the resulting term $D_1^{-1}R_1^{-1}R_2D_2 = \tilde{R}.*\tilde{D}$ is then harmless.

8.2. Multivariate anisotropic HermiteGF-QR

In this section, we derive an analog of the HermiteGF-QR algorithm for the multivariate case. Since we are now dealing with matrices, in the general case, it is impossible to separate E from $\mathbf{x}_k$ in $(E\mathbf{x}_k)^{\boldsymbol{\ell}}$ as we did before. Therefore, the flow of the HermiteGF-QR does not apply directly. However, it is still possible to tackle some part of the ill-conditioning analytically. Consider the following splitting of the matrix $G^{-1}E^TE \in \mathbb{R}^{d\times d}$:

$$G^{-1}E^TE = \mathrm{Diag} + \mathrm{Rem},$$

where Diag is the $d \times d$ diagonal matrix containing the diagonal elements of $G^{-1}E^TE$ and Rem contains the remaining off-diagonal terms. Denote

$$\mathbf{v}_k = (\mathrm{Id} + \mathrm{Diag}^{-1}\mathrm{Rem})\boldsymbol{\Delta}_k \quad \text{and} \quad \boldsymbol{\Delta}_k = \mathbf{x}_k - \mathbf{x}_0.$$

[1] We assume that the matrix X is invertible. If this is not the case, then column pivoting in the QR decomposition has proved to be effective (see [FLF11, § 5]).

Then, it holds that

$$(G^{-1}E^TE\mathbf{\Delta}_k)^{\boldsymbol{\ell}} = ((\mathrm{Diag}+\mathrm{Rem})\mathbf{\Delta}_k)^{\boldsymbol{\ell}} = \left(\mathrm{Diag}\left((\mathrm{Id}+\mathrm{Diag}^{-1}\mathrm{Rem})(\mathbf{x}_k-\mathbf{x}_0)\right)\right)^{\boldsymbol{\ell}}$$
$$= \prod_{i=1}^{d}(\mathrm{Diag}_{ii}(\mathbf{v}_k)_i)^{\ell_i} = \prod_{i=1}^{d}\mathrm{Diag}_{ii}^{\ell_i}(\mathbf{v}_k)_i^{\ell_i} = \left(\prod_{i=1}^{d}\mathrm{Diag}_{ii}^{\ell_i}\right)\mathbf{v}_k^{\boldsymbol{\ell}},$$

where $\mathrm{Diag}^{-1}\mathrm{Rem}$ can be computed analytically. Denote $\mathbf{d}_{\mathrm{vec}} = \mathrm{diag}(\mathrm{Diag})$. Then, the generating function expansion (6.3) can be written as:

$$\phi_k(\mathbf{x}) = \exp(-\mathbf{\Delta}_k^TE^TE\mathbf{\Delta}_k + \mathbf{\Delta}_k^T\tilde{G}\mathbf{\Delta}_k)\sum_{\boldsymbol{\ell}\in\mathbb{N}^d}\frac{(G^{-1}E^TE\mathbf{\Delta}_k)^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}H_{\ell}^{G,E,t}(\mathbf{x}-\mathbf{x}_0)$$
$$= \exp(-\mathbf{\Delta}_k^TE^TE\mathbf{\Delta}_k + \mathbf{\Delta}_k^T\tilde{G}\mathbf{\Delta}_k)\sum_{\boldsymbol{\ell}\in\mathbb{N}^d}\frac{\mathbf{d}_{\mathrm{vec}}^{\boldsymbol{\ell}}\mathbf{v}_k^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}H_{\ell}^{G,E,t}(\mathbf{x}-\mathbf{x}_0).$$

As before, we can write the expansion above as the infinite matrix-vector product

$$\Phi(\mathbf{x}) = H^{G,E,t}(\mathbf{x}-\mathbf{x}_0)\,B(E,G,t,X^{\mathrm{cen}})$$

with

$$B(E,G,t,X^{\mathrm{cen}})_{\boldsymbol{\ell}k} = \exp(-\mathbf{\Delta}_k^TE^TE\mathbf{\Delta}_k + \mathbf{\Delta}_k^T\tilde{G}\mathbf{\Delta}_k)\frac{\mathbf{d}_{\mathrm{vec}}^{\boldsymbol{\ell}}\mathbf{v}_k^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}.$$

As before, we write the transpose of the infinite matrix B as a product CD with

$$C_{k\boldsymbol{\ell}} = \exp(-\mathbf{\Delta}_k^TE^TE\mathbf{\Delta}_k + \mathbf{\Delta}_k^T\tilde{G}\mathbf{\Delta}_k)\mathbf{v}_k^{\boldsymbol{\ell}}, \quad D_{\boldsymbol{\ell}\boldsymbol{\ell}} = \frac{\mathbf{d}_{\mathrm{vec}}^{\boldsymbol{\ell}}\sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}. \tag{8.2}$$

The $d\times d$ matrix product $\mathrm{Diag}^{-1}\mathrm{Rem}$ contained in the vectors $\mathbf{v}_k$ can be computed analytically. The diagonal part of the matrix $G^{-1}E^TE$, that is now in the matrix D can be handled in the exact same fashion as it has been done in 1D. In particular, we perform the QR-decomposition of the matrix $C = QR$, block decompose

$$R = \begin{pmatrix} R_1 & R_2 \end{pmatrix}, \qquad D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix},$$

such that the entries related to the first N stabilizing basis functions are contained in the $N \times N$ matrices R_1 and D_1, and consider the preconditioner $X = QR_1D_1$. Hence, analogously to the HermiteGF-QR case, the new basis can be formed as

$$\Psi(\mathbf{x})^T := X^{-1}\Phi(\mathbf{x})^T = \begin{pmatrix}\mathrm{Id} & D_1^{-1}R_1^{-1}R_2D_2\end{pmatrix}H^{G,E,t}(\mathbf{x}-\mathbf{x}_0)^T.$$

The action of D_1^{-1} and D_2 can again be computed as the Hadamard product with

$$\tilde{D}_{ij} = \frac{D_{j+N,j+N}}{D_{ii}} \quad \text{with} \quad i \in \{1,\dots,N\},\ j \geq 1. \tag{8.3}$$

Remark 8.2.1. *When the magnitude of $\mathbf{v}_k$ gets too large, the matrix C can also become ill-conditioned. To avoid that, one can increase the magnitude of the* *elements of G. Alternatively, one can add a scaling in C which should then be compensated in the matrix D, similarly to the scaling with the domain size L in the 1D version.*

In the new basis, we can write the equivalent formulation of the interpolant (4.1) as

$$s(\mathbf{x}) = \Psi(\mathbf{x})(\Psi^{\text{col}})^{-1}\boldsymbol{f} \quad \text{with} \quad \Psi^{\text{col}}_{ij} = \psi_j(\mathbf{x}^{\text{col}}_i). \tag{8.4}$$

Now, when we have derived the general version of the HermiteGF-QR method, let us take a look on which form it takes in the case of isotropic Gaussians.

8.2.1. Isotropic HermiteGF-QR

In the isotropic case, when $E = \varepsilon \mathrm{Id}_d$ and $G = \gamma \mathrm{Id}_d$, the expressions for matrices C and D can be written in a simpler way. In particular, in this case we have

$$\mathrm{Diag} = \gamma^{-1}\varepsilon^2 \mathrm{Id}_d, \quad \mathrm{Rem} = \mathbf{0}_d.$$

and, therefore,

$$\mathbf{v}_k = \mathbf{x}_k - \mathbf{x}_0 = \boldsymbol{\Delta}_k.$$

Hence, the diagonal vector simplifies to $\mathbf{d}_{\text{vec}} = \gamma^{-1}\varepsilon^2(1, \dots, 1)$ and the elements of matrices C and D take the form

$$C_{k\ell} = \exp\left(\varepsilon^2\left(\frac{\varepsilon^2}{\gamma^2} - 1\right)\|\boldsymbol{\Delta}_k\|^2_{\ell_2}\right)\boldsymbol{\Delta}_k^{\boldsymbol{\ell}}, \quad D_{\ell\ell} = \frac{(\sqrt{2}\gamma^{-1}\varepsilon^2)^{|\boldsymbol{\ell}|}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}.$$

8.3. Alternative splitting based on the Vandermonde matrix

Instead of using a QR-decomposition of the matrix $C \in \mathbb{R}^{N\times\infty}$ one could also split it as $C = \bar{E}W$, where $\bar{E} \in \mathbb{R}^{N\times N}$ is a diagonal matrix for the exponential part and $W \in \mathbb{R}^{N\times\infty}$ accounts for the polynomial contributions,

$$\bar{E}_{kk} = \exp(-\boldsymbol{\Delta}_k^T E^T E \boldsymbol{\Delta}_k + \boldsymbol{\Delta}_k^T \tilde{G} \boldsymbol{\Delta}_k) \quad \text{and} \quad W_{k\ell} = \mathbf{v}_k^{\boldsymbol{\ell}}. \tag{8.5}$$

We now decompose the original basis using this splitting,

$$\begin{aligned}\Phi(\mathbf{x})^T &= C\begin{pmatrix} D_1 & 0 \\ 0 & D_2\end{pmatrix} H^{G,E,t}(\mathbf{x} - \mathbf{x}_0)^T \\ &= \bar{E}\begin{pmatrix} W_1 & W_2\end{pmatrix}\begin{pmatrix} D_1 & 0 \\ 0 & D_2\end{pmatrix} H^{G,E,t}(\mathbf{x} - \mathbf{x}_0)^T \\ &= \bar{E}\begin{pmatrix} W_1 D_1 & W_2 D_2\end{pmatrix} H^{G,E,t}(\mathbf{x} - \mathbf{x}_0)^T,\end{aligned}$$

where $W_1 \in \mathbb{R}^{N\times N}$ and $W_2 \in \mathbb{R}^{N\times\infty}$. With the preconditioner $X_V = \bar{E}W_1 D_1$, the new basis reads as

$$\Psi_V(\mathbf{x})^T = \begin{pmatrix}\mathrm{Id} & D_1^{-1}W_1^{-1}W_2 D_2\end{pmatrix} H^{G,E,t}(\mathbf{x} - \mathbf{x}_0)^T.$$

One can see that

$$X_V = \bar{E}W_1D_1 = C_1D_1 = QR_1D_1 = X,$$

where C_1 is the first $N \times N$ block of C. Therefore, in exact arithmetic Ψ and Ψ_V are the same, however, in floating-point arithmetic the values of the bases might differ. However, we will use this alternative splitting for the derivation and analysis of a suitable cut-off criterion for the HermiteGF basis in chapter 10.

Remark 8.3.1. *Note that in the isotropic case the part of C containing the values of the shape parameter ε explicitly cancels out. Therefore, we expect that ε does not have a significant impact on the term $R_1^{-1}R_2$ as well. In the anisotropic case, the elements of E are included in the vectors $\mathbf{v}_k$, and therefore, the situation might differ.*

9. Analysis of the HermiteGF-QR ingredients

In this chapter, we take a look at the different matrices that are involved in the assembly of the stable basis Ψ. In particular, we aim to gain a better understanding of the influence of different parts on the final result. At the same time, we build up the knowledge base that can assist us in the derivation of the cut-off criterion in chapter 10. For the sake of simplicity, we focus on the case with the zero shift $\mathbf{x}_0 = 0$.

9.1. The matrix B

From section 8.2 we know that the original anisotropic Gaussian basis can be written though the HermiteGF basis in the matrix-vector form as follows:

$$\Phi(\mathbf{x}) = H^{G,E,t}(\mathbf{x} - \mathbf{x}_0)B(E, G, X^{\text{cen}}, t),$$

where $B(E, G, X^{\text{cen}}, t) \in \mathbb{R}^{\infty \times N}$ with the elements

$$B(E, G, t, X^{\text{cen}})_{\boldsymbol{\ell}k} = \exp(-\boldsymbol{\Delta}_k^T E^T E \boldsymbol{\Delta}_k + \boldsymbol{\Delta}_k^T \tilde{G} \boldsymbol{\Delta}_k) \frac{\mathbf{d}_{\text{vec}}^{\boldsymbol{\ell}} \mathbf{v}_k^{\boldsymbol{\ell}} \sqrt{2^{|\boldsymbol{\ell}|}}}{\sqrt{t^{|\boldsymbol{\ell}|} \boldsymbol{\ell}!}}.$$

We further refer to the matrix $B(E, G, t, X^{\text{cen}})$ as B when the parameters are fixed within the context.

In section 6.4 we proved that the norm of $H^{G,E,t}(\mathbf{x})$ converges with $\boldsymbol{\ell}$ approaching infinity and found the limit. We now consider the matrix B, containing the coefficients of the anisotropic Gaussian interpolant in the HermiteGF basis, before it has been decomposed into the product of C and D. We investigate whether there is also decay in the coefficients matrix B by taking a look at the tail of its Frobenius norm. In particular, we split the $\infty \times N$ matrix B into two parts

$$B = \begin{pmatrix} B_{j_{\max}} \\ B_\infty \end{pmatrix},$$

Here $B_{j_{\max}}$ being the upper $M_{j_{\max}} \times N$ part of the matrix B, where $M_{j_{\max}}$ corresponds to the number of basis functions with a cumulative polynomial degree $\leq j_{\max}$. In 1D, it also corresponds to the number of basis functions $M_{j_{\max}}$. In multiple dimensions, the number of basis functions $M_{j_{\max}}$ can be obtained as $M_{j_{\max}} = \begin{pmatrix} j_{\max} + d \\ d \end{pmatrix}$. The $\infty \times N$ matrix B_∞ contains the remaining entries of B. It turns out that we can compute the norm $\|B_\infty\|_F$ exactly. The value of this norm gives us an idea of the cumulative impact of all elements of B that will be cut off if we choose to truncate at the degree $j_{\max}$. We now proceed to the derivation of the value of $\|B_\infty\|_F$.

Lemma 9.1.1: The tail of $\|B\|_F$

Let $E, G \in \mathbb{R}^{d\times d}$ be invertible, $t \in (0,1)$. Denote $\theta_k = -\mathbf{\Delta}_k^T E^T E \mathbf{\Delta}_k + \mathbf{\Delta}_k^T \tilde{G} \mathbf{\Delta}_k$, $\mathbf{y}_k = G^{-1}E^T E\mathbf{x}_k$ and $\tilde{\mathbf{y}}_k = \begin{pmatrix} \frac{2}{t}(y_k)_1^2 & \dots & \frac{2}{t}(y_k)_d^2 \end{pmatrix}$. Then, the tail of the Frobenius norm of the corresponding matrix $B(E, G, t, X^{\text{cen}})$ can be evaluated as

$$\|B_\infty\|_F^2 = \sum_{k=1}^N \exp(2\theta_k) \frac{\exp(\tilde{\mathbf{y}}_k)}{j_{\max}!} \gamma(j_{\max}+1, \|\tilde{\mathbf{y}}_k\|_1), \tag{9.1}$$

where $\gamma(\cdot,\cdot)$ is the incomplete gamma function.

Proof. We follow the same flow as we did in the convergence proof in $L^2_\omega(\mathbb{R}^d)$ in the section 7.2. However, this time we compute the tail of the Taylor series exactly.

Using the integral form of the tail of the Taylor expansion [Hör90, $\S 1.1$, Expr. (1.1.7)'], we get

$$\begin{aligned}
\|B_\infty\|_F^2 &= \sum_{k=1}^N \exp(2\theta_k) \sum_{|\boldsymbol{\ell}|\geq j_{\max}+1} \frac{\tilde{\mathbf{y}}_k^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} \\
&= \sum_{k=1}^N \exp(2\theta_k)(j_{\max}+1) \sum_{|\boldsymbol{\ell}|=j_{\max}+1} \frac{\tilde{\mathbf{y}}_k^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} \int_0^1 (1-s)^{|\boldsymbol{\ell}|-1} \mathrm{D}^{\boldsymbol{\ell}} \exp(s\|\tilde{\mathbf{y}}_k\|_1) \mathrm{d}s \\
&= \sum_{k=1}^N \exp(2\theta_k)(j_{\max}+1) \int_0^1 (1-s)^{j_{\max}} \exp(s\|\tilde{\mathbf{y}}_k\|_1) \mathrm{d}s \sum_{|\boldsymbol{\ell}|=j_{\max}+1} \frac{\tilde{\mathbf{y}}_k^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!}.
\end{aligned}$$

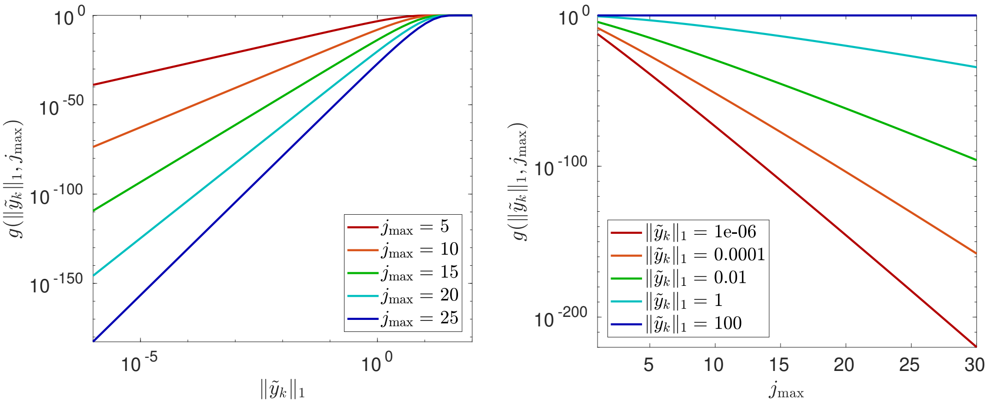
We now compute the integral over s with the help of [GR14, Expr. 3.351]

$$\begin{aligned}
\int_0^1 (1-s)^{j_{\max}} \exp(s\|\tilde{\mathbf{y}}_k\|_1) \mathrm{d}s &= \exp(\|\tilde{\mathbf{y}}_k\|_1) \int_0^1 \bar{s}^{j_{\max}} \exp(-\|\tilde{\mathbf{y}}_k\|_1 \bar{s}) \mathrm{d}\bar{s} \\
&= \exp(\|\tilde{\mathbf{y}}_k\|_1) \|\tilde{\mathbf{y}}_k\|_1^{-j_{\max}-1} \gamma(j_{\max}+1, \|\tilde{\mathbf{y}}_k\|_1),
\end{aligned}$$

where γ is the incomplete gamma function. Therefore, using the multinomial theorem, we obtain

$$\begin{aligned}
\|B_\infty\|_F^2 &= \sum_{k=1}^N \exp(2\theta_k)(j_{\max}+1) \frac{\exp(\|\tilde{\mathbf{y}}_k\|_1)}{\|\tilde{\mathbf{y}}_k\|_1^{j_{\max}+1}} \gamma(j_{\max}+1, \|\tilde{\mathbf{y}}_k)\|_1) \sum_{|\boldsymbol{\ell}|=j_{\max}+1} \frac{\tilde{\mathbf{y}}_k^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} \\
&= \sum_{k=1}^N \exp(2\theta_k)(j_{\max}+1) \frac{\exp(\|\tilde{\mathbf{y}}_k\|_1)}{\|\tilde{\mathbf{y}}_k\|_1^{j_{\max}+1}} \gamma(j_{\max}+1, \|\tilde{\mathbf{y}}\|_1) \frac{\|\tilde{\mathbf{y}}_k\|_1^{j_{\max}+1}}{(j_{\max}+1)!} \\
&= \sum_{k=1}^N \exp(2\theta_k) \frac{\exp(\|\tilde{\mathbf{y}}_k\|_1)}{j_{\max}!} \gamma(j_{\max}+1, \|\tilde{\mathbf{y}}_k\|_1). \qquad \square
\end{aligned}$$

Once in possession of the explicit formula, we can now take a look at the numerical values of the tail for different parameter values.



(a) Dependence from the value of $\|\tilde{\mathbf{y}}_k\|_1$ for different values of $j_{\max}$.

(b) Dependence from the value of $j_{\max}$ for different values of $\|\tilde{\mathbf{y}}_k\|_1$.

Figure 4 Dependence of the values of the function g from the values of $j_{\max}$ and $\|\tilde{\mathbf{y}}\|_1$. For all $j_{\max}$ the growth of the tail stagnates at 1. This indicates that at the point of stagnation there is no decrease of the tail with $j_{\max}$, which could indicate suboptimal behavior of the method for larger ε, since $\|\tilde{\mathbf{y}}_k\|_1$ grows with ε. For $\|\tilde{\mathbf{y}}_k\|_1 > 10^1$ the numerical convergence is almost non-existent.

9.1.1. The behavior of the tail of the coefficients matrix B

Let us first investigate how the speed of convergence of the tail behaves with the increasing $j_{\max}$. The faster the decay, the fewer basis functions we would need for getting an accurate representation of the anisotropic Gaussian interpolant in the HermiteGF basis. We study separately the part of the expression (9.1), that depends on $j_{\max}$, as a function of $j_{\max}$ and $\|\tilde{\mathbf{y}}_k\|_1$. In particular, we introduce

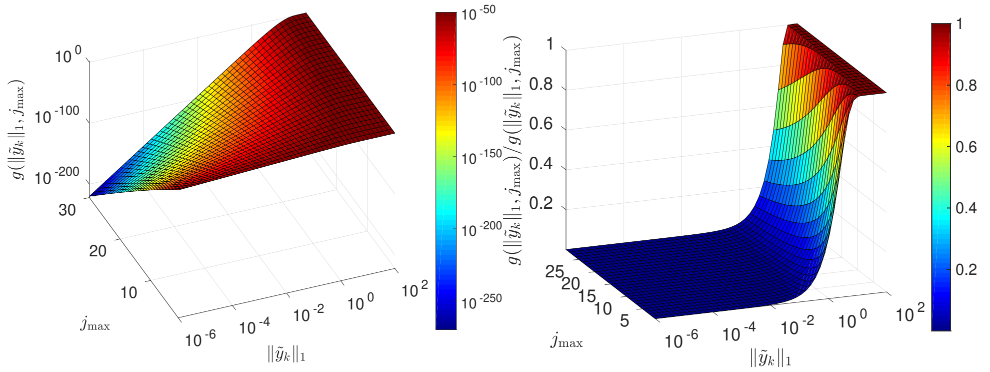
$$g(j_{\max}, \|\tilde{\mathbf{y}}_k\|_1) = \frac{\gamma(j_{\max}+1, \|\tilde{\mathbf{y}}_k\|_1)}{j_{\max}!}.$$

and examine the behavior of this function numerically.

One can see in the Figure 4a that the tail takes extremely small values for small magnitude of $\|\tilde{\mathbf{y}}_k\|_1$ and the values are exponentially growing with the increase of the magnitude of $\|\tilde{\mathbf{y}}_k\|_1$. The growth rate increases with the increase of $j_{\max}$. However, after a certain point, for all $j_{\max}$ the growth stagnates around 1. This indicates that for large $\|\tilde{\mathbf{y}}_k\|_1$ the tail does not decrease anymore with the growth of $j_{\max}$. At the same time, one can see in the Figure 4b that for larger $\|\tilde{\mathbf{y}}_k\|_1$ there is almost no decay of the g with the growth of $j_{\max}$. Interpreting it in terms of the HermiteGF-QR method and recalling that

$$\tilde{\mathbf{y}}_k = \left(\tfrac{2}{t}(G^{-1}E^T E x_k)_1^2 \quad \dots \quad \tfrac{2}{t}(G^{-1}E^T E x_k)_d^2\right),$$

we observe that for the fixed values of parameters G and t the value of $\|\tilde{\mathbf{y}}_k\|_1$ is larger for larger values of elements of E. This implies that we would need more basis functions for larger E, especially, after $\|\tilde{\mathbf{y}}_k\|_1$ gets larger than 10^1. Increasing the magnitude of G would help to reduce $\|\tilde{\mathbf{y}}_k\|_1$, however, we know from the section 6.4 that in this case, the convergence within the HermiteGF



(a) Absolute values.

(b) Relative values.

Figure 5 Dependence of the values of g from the values of $j_{\max}$ and $\|\tilde{\mathbf{y}}_k\|_1$.

basis is worse. On the other hand, we can deduce that taking extremely small values of t should be avoided. The smallest t that we consider in our experiments in chapter 12 is of order 10^{-1}. Indeed, if we take the values of t of order 10^{-2}, it is very easy to reach the state when $\|\tilde{\mathbf{y}}_k\|_1$ approaches 10^2, where we have an almost non-existent convergence.

Observation 1. *When performing the interpolation using the HermiteGF-QR method, one should avoid the setups when $\|\tilde{\mathbf{y}}_k\|_1 > 10^1$. In particular, one should not use very small values of t. We recommend to limit the range of considered values of t to $(0.1, 1)$.*

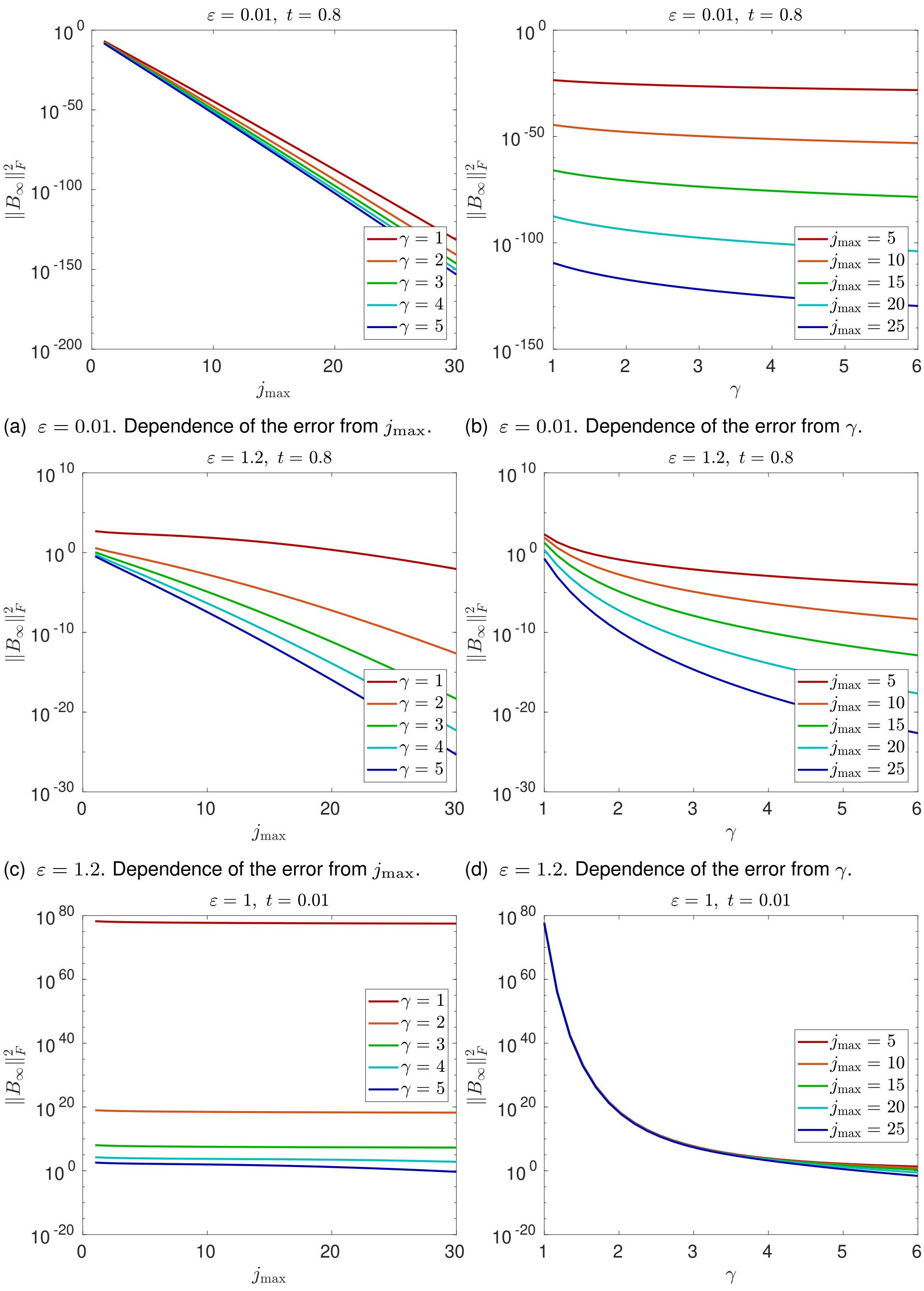
Observation 2. *When all parameters are fixed, for larger values of the shape parameter ε (or the shape matrix E), we would need more basis functions. In certain cases, when the stagnation point is reached for the function g, the method might be suboptimal.*

This behavior is consistent for all combinations of $j_{\max}$ and $\|\tilde{\mathbf{y}}_k\|_1$ as one can see on the Figure 5a. If we now take a look at the relative behavior of the tail, namely, the values of $g(\|\tilde{\mathbf{y}}_k\|_1, j_{\max}+1)/g(\|\tilde{\mathbf{y}}_k\|_1, j_{\max})$, we see that the value stagnates at some point $\|\tilde{\mathbf{y}}_k\|_1$ (see Figure 5b). This is another indicator, that at this point we do not have convergence anymore. However, this point is different for different $j_{\max}$.

Up to now, we only considered the part of the $\|B_\infty\|_F^2$ that depends on $j_{\max}$. Let us now take a look at the values of the full $\|B_\infty\|_F^2$ for a standard HermiteGF-QR test case. In particular, we consider the isotropic HermiteGF-QR with

$$G = \gamma \mathrm{Id} \quad \text{and} \quad E = \varepsilon \mathrm{Id}$$

and $N = 100$ Halton points on a unit square $[-1, 1] \times [-1, 1]$ as the nodes $\{\mathbf{x}_k\}_{k=1\ldots N}$. For different values of the shape parameter, smaller value $\varepsilon = 0.01$ and larger value $\varepsilon = 1.2$, we examine the behavior of $\|B_\infty\|_2$ with the change of



(a) $\varepsilon = 0.01$. Dependence of the error from $j_{\max}$.

(b) $\varepsilon = 0.01$. Dependence of the error from γ.

(c) $\varepsilon = 1.2$. Dependence of the error from $j_{\max}$.

(d) $\varepsilon = 1.2$. Dependence of the error from γ.

(e) $t = 0.01, \varepsilon = 1$. Dependence of the error from $j_{\max}$.

(f) $t = 0.01, \varepsilon = 1$. Dependence of the error from γ.

Figure 6 Isotropic interpolation on a unit square with 100 Halton points. For smaller ε the decay with $j_{\max}$ is the comparable for all γ whereas for larger ε larger γ provides significantly larger decay. Using the small value of $t = 0.01$ for the case of a relatively large $\varepsilon = 1$ yields divergence of the $\|B_\infty\|_F^2$.

$j_{\max}$ and γ. Since we observed above that small values of t might hinder the convergence, for this test, we choose a large enough $t = 0.8$. We can see in the Figure 6 that the observations for the behavior of the function g can be transferred to the case of the whole $\|B_\infty\|_F^2$. Note that in this case, larger magnitude of $\|\tilde{\mathbf{y}}_k\|_1$

corresponds to smaller γ or larger ε. One can see that the larger the $\|\tilde{\mathbf{y}}_k\|_1$, the worse the convergence with $j_{\max}$. Moreover, the convergence is more sensitive to the value of ε than to the value of γ. It is clear from the Figure 6a that for a small value of ε the norm of B_∞ decays very fast with $j_{\max}$ for all γ, whereas the Figure 6c illustrates that for larger ε it is crucial to choose larger γ. However, we have seen in Section 6.4 that the opposite is true for the part containing the values of $H^{G,E,t}(\mathbf{x})$. Therefore, it is important to keep both parts in mind while choosing the γ.

As for the value of the parameter t, as predicted before, if we take a small value of $t = 0.01$ for the case when the ε is of O(1), the norm of B does not converge (see figures 6e and 6f).

Note that even though the convergence of the Frobenius norm of the coefficients matrix B gets worse for larger magnitudes of $\tilde{\mathbf{y}}_k$, the full HermiteGF expansion might have a better convergence due to the properties of the decay in the HermiteGF basis itself (see Section 6.4).

9.2. The matrices D and $\tilde{R}$.

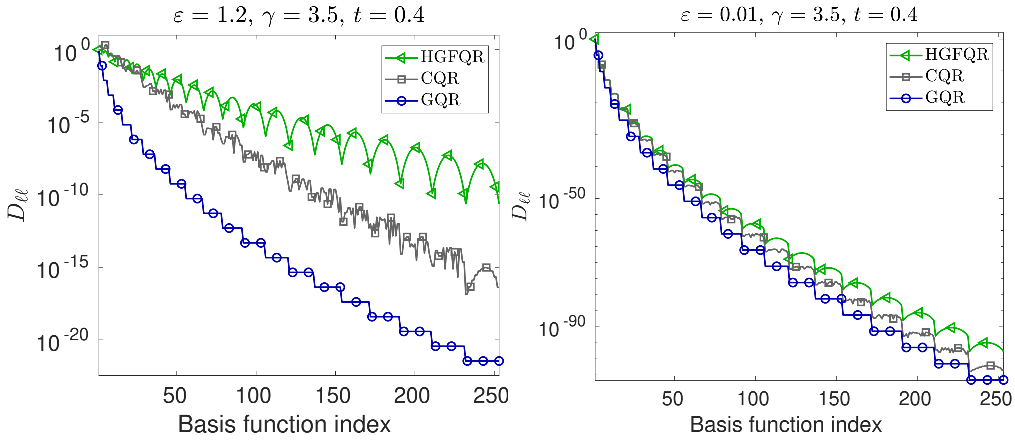
In the previous section, we thoroughly analyzed the matrix B before it was decomposed into C and D. In this section, we take a step further. First, we take a look at the elements of the matrix D. In order to be able to accurately compare the values to the corresponding ones of the reference methods, we consider the isotropic case. Recall that in this case, the matrix D has the following elements

$$D_{\boldsymbol{\ell}\boldsymbol{\ell}} = \frac{(\sqrt{2}\gamma^{-1}\varepsilon^2)^{|\boldsymbol{\ell}|}}{\sqrt{t^{|\boldsymbol{\ell}|}\boldsymbol{\ell}!}}.$$

For the reference, we use Gauss-QR and Chebyshev-QR methods. Here, for the Gauss-QR method we used $\alpha_{\text{Gauss-QR}} = \gamma$. One can see in the Figure 7 that for small values of ε the behavior is very similar for all three methods. However, for large ε the decay in D is particularly bad in our formulation. In case of HermiteGF-QR, we also see that the magnitude of oscillations in D is larger for larger ε (see Figure 7a). This is consistent with the observations from the previous sections, that when other parameters are fixed, the convergence is worse for larger ε.

Observation 3. *For the small values of ε the behavior of the elements of the matrix D is similar for all three methods. For a larger value of ε, the decay of the elements of the matrix D for the HermiteGF-QR method is worse than for the reference methods.*

We now move on to the matrix C, to investigate its impact on the HermiteGF-QR method. Since we do not use the full matrix C in the method, but perform a



(a) Larger $\varepsilon = 1$.

(b) Smaller $\varepsilon = 0.01$.

Figure 7 For $\varepsilon = 0.01$ the elements of D behave very similarly for all cases. For a larger value of ε, the elements of D in HermiteGF method show larger magnitude of oscillations.

factorization, we look at the values of $\tilde{R} = R_1^{-1}R_2$ in case of the QR splitting, or of $\tilde{R} = W_1^{-1}W_2$ in case of the Vandermonde splitting. This is the matrix that is used later on in the formation of Ψ. Recall that in the HermiteGF-QR case, Ψ can be formed as follows

$$\Psi(\mathbf{x})^T = \begin{pmatrix}\mathrm{Id} & R_D\end{pmatrix} H^{G,E,t}(\mathbf{x} - \mathbf{x}_0)^T,$$

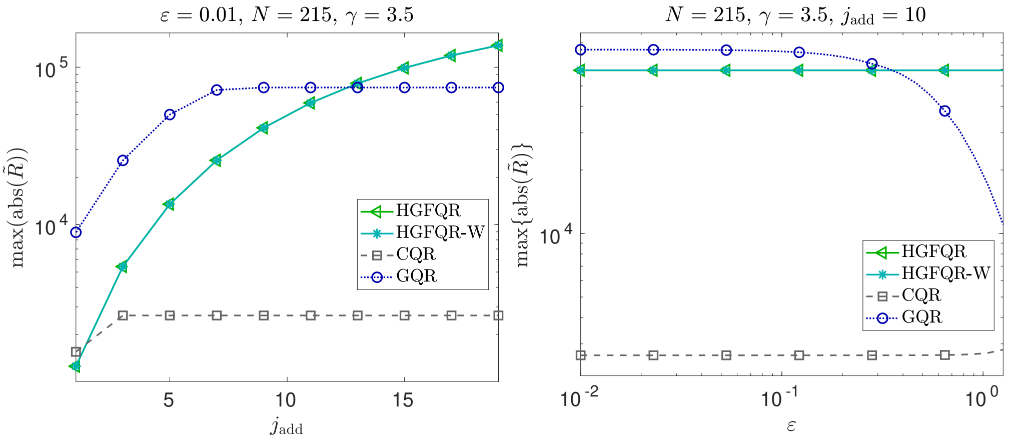
where $R_D = D_1^{-1}R_1^{-1}R_2D_2$ is computed as the Hadamard product of $\tilde{D}$ and $\tilde{R}$. We want to investigate the impact of the term R_D on Ψ. We first look at the maximum value of $\tilde{R}$. Note that the values of $\tilde{R}$ do not depend on t.

For all our tests we take a unit square domain with Halton points. We fix $\gamma = 3.5$ and $N = 215$. Since the required condition for the RBF-QR method is that the number of basis functions $M \geq N$, for this test we take

$$P = \min\left\{ K \in \mathbb{N}_0 \,\middle|\, N \leq \binom{K+d}{d} \right\}$$

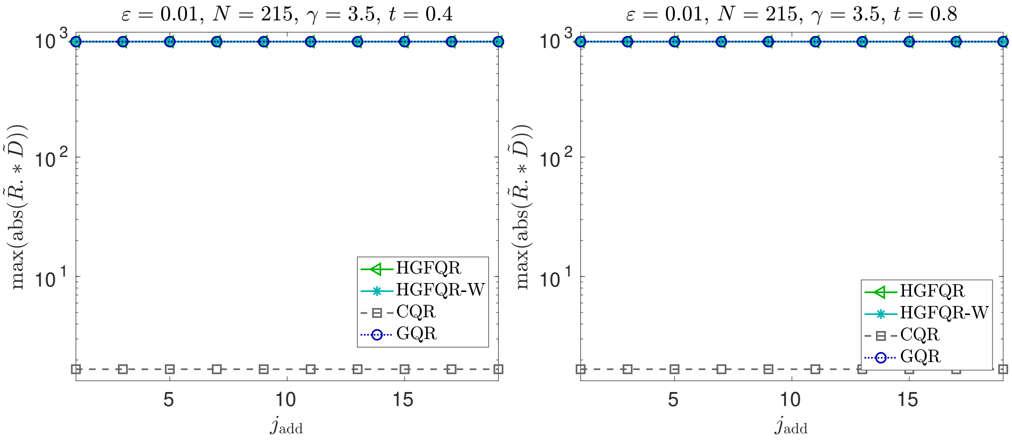
as the smallest cut-off degree $j_{\max}$ and then add additional degrees $j_{\text{add}} \in [1, 20]$ to see the impact of the newly added basis functions.

One can see in the Figure 8a that for $\varepsilon = 0.01$ for all methods the maximum value increases with $j_{\max}$ and then stagnates at a certain point. The values for the HermiteGF-based methods are stagnating slower than the reference methods. On the other hand, we observe in the Figure 8b that for a fixed number of expansion functions with $j_{\text{add}} = 10$, for the HermiteGF-based methods the maximum value of $\tilde{R}$ does not change with ε whereas for the Gauss-QR method, it starts to decay with the increase of ε. Slight changes can be also observed in the Chebyshev-QR method. This is due to the fact, that in the HermiteGF case the part containing ε cancels out in the isotropic case (see remark 8.3.1) whereas for other methods it might not be the case.



(a) The maximum of $\tilde{R}$ with the change of j_{add}. (b) The maximum of $\tilde{R}$ with the change of ε.

Figure 8 For all methods the values of $\tilde{R}$ grow with j_{add}. At the same time, for HermiteGF methods $\tilde{R}$ stays constant with the change of ε. In Chebyshev-QR we can observe some minor changes and for Gauss-QR the maximum starts to decay with the increase of ε.



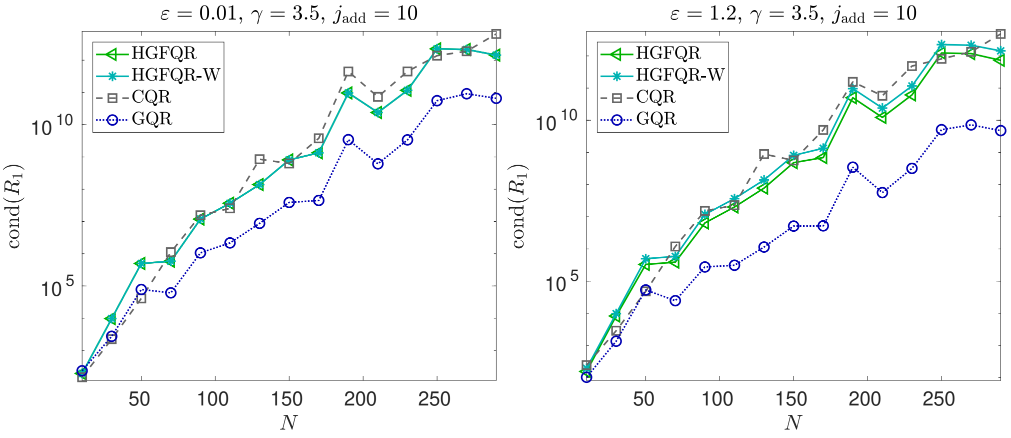
(a) The maximum of $\tilde{R}$ with the change of j_{add}. (b) The maximum of $\tilde{R}$ with the change of j_{add}.

Figure 9 For all methods the values of $\tilde{R}.*\tilde{D}$ stay constant with the change of j_{add}. The impact of t is negligible for the HermiteGF-based methods.

Combining $\tilde{R}$ with the corresponding $\tilde{D}$ yields a constant result with the change of j_{add} for all methods (see figures 9a and 9b). This indicates that the growth in $\tilde{R}$ is compensated by the decay in D. Note that even though in HermiteGF-QR methods the matrix D depends on t, it does not have a big impact on the magnitude of the correction term.

Observation 4. *Even though the maximum value of the matrix $\tilde{R}$ grows with j_{add} for all methods, it is compensated by the elements of $\tilde{D}$ when forming the matrix R_D. Indeed, for all methods, the maximum value of the matrix R_D, which is the Hadamard product of $\tilde{R}$ and $\tilde{D}$, was constant with the change of j_{add}.*

Finally, we take a look at the conditioning of the matrix R_1 that we invert. We are particularly interested in its relation to the Vandermonde splitting counterpart



(a) The maximum of $\tilde{R}$ with the change of j_{add}.

(b) The maximum of $\tilde{R}$ with the change of ε.

Figure 10 For the HermiteGF-QR and Chebyshev-QR methods, the conditioning is comparable with the Vandermonde HermiteGF alternative. For the GaussQR method, the conditioning is better.

W_1. One can see in the Figure 10 that the conditioning of R_1 is comparable to W_1 for the HermiteGF-QR method for both smaller and larger values of ε. Moreover, it is the same order of magnitude as the R_1 of the Chebyshev-QR method. Gauss-QR method, however, yields better conditioning.

10. Cut-off of the expansion in HermiteGF-QR method

To make the RBF-QR methods usable for numerical computations, one has to cut the expansion (6.3) at a certain polynomial degree $j_{\max} \in \mathbb{N}$ which in 1D also corresponds to the number of stabilizing basis functions M. In the multivariate setting, the number of basis functions $M_{j_{\max}}$ equals

$$M_{j_{\max}} = \binom{j_{\max} + d}{d}.$$

However, choosing an efficient cut-off degree $j_{\max}$ is not a trivial task. As we have mentioned in section 5.1 describing the general flow of RBF-QR algorithms, one of the common ways to determine the cut-off degree $j_{\max}$ is based on deducing it from the values of the matrix $\tilde{D}$. Once the values of $\tilde{D}$ become less than machine precision, we can stop.

For the HermiteGF-QR case, we first derive the specific formula for this criterion. However, for our case, it turns out to be inefficient, i.e. it overestimates the number $j_{\max}$. We then derive a new cut-off criterion based on the theoretical framework presented in the previous sections. This new criterion allows us to directly control the approximation error of the stable basis which is more efficient while still being effective.

10.1. State of the art criterion

Let us now derive the specific formulation of the standard cut-off criterion based on $\tilde{D}$ for the HermiteGF-QR case. Recall that the matrix $\tilde{D}$ that contains the effects of D_1^{-1} and D_2 can be written as

$$\tilde{D}_{ij} = \frac{D_{j+N,j+N}}{D_{ii}} = \frac{\mathbf{d}_{\text{vec}}^{\mathbf{j}}\sqrt{2^{|\mathbf{j}|}}}{\sqrt{t^{|\mathbf{j}|}\mathbf{j}!}} \cdot \frac{\sqrt{t^{|\mathbf{i}|}\mathbf{i}!}}{\mathbf{d}_{\text{vec}}^{\mathbf{i}}\sqrt{2^{|\mathbf{i}|}}}$$

with $i \in \{1, \ldots, N\}$, $j \geq 1$. Adding one more polynomial degree, i.e. increasing $j_{\max}$ by one, means adding $\binom{j_{\max} + d + 1}{d} - \binom{j_{\max} + d}{d}$ columns to $\tilde{D}$. As in the general case, we stop once all elements of the new block of $\tilde{D}$ are below machine precision. Namely, when

$$\max_{\substack{i=1\ldots N,\\ |\mathbf{j}|\geq j_{\max}+1}} \tilde{D}_{i\mathbf{j}} < \varepsilon_{\text{mach}}. \tag{10.1}$$

Let us now derive an explicit formula for the expression above. However, we first prove some important properties of the matrix D.

Lemma 10.1.1: Properties of the matrix D

If $\mathrm{Diag}_{ii} < \sqrt{\frac{t}{2}}$ for all $i = 1 \dots d$. Then, for all $j \geq 0$

$$\min_{0 \leq |\boldsymbol{\ell}| \leq j} \{D_{\boldsymbol{\ell}\boldsymbol{\ell}}\} = \min_{|\boldsymbol{\ell}|=j} \{D_{\boldsymbol{\ell}\boldsymbol{\ell}}\} \quad \text{and} \quad \max_{|\boldsymbol{\ell}|>j} \{D_{\boldsymbol{\ell}\boldsymbol{\ell}}\} = \max_{|\boldsymbol{\ell}|=j+1} \{D_{\boldsymbol{\ell}\boldsymbol{\ell}}\}.$$

Proof.

1. Denote $d_{\min} = \min_{i=1\dots d}\{\mathrm{Diag}_{ii}\}$. Then,

$$\min_{|\boldsymbol{\ell}|=j} \{D_{\ell\ell}\} = d_{\min}^{j} \sqrt{\frac{2^j}{t^j j!}} \quad \forall j \geq 0.$$

Indeed, all other elements of the corresponding level of the matrix Diag would involve a bigger numerator and a smaller denominator, since

$$d_{\min}^{j} \leq \prod_{i=1}^{d} \mathrm{Diag}_{ii}^{\ell_i} \quad \text{and} \quad j! > \boldsymbol{\ell}! \quad \text{for all} \quad \boldsymbol{\ell} \quad \text{with} \quad |\boldsymbol{\ell}| = j.$$

We look if the minimum on the level $j+1$ is less than the one at the level j. Indeed,

$$d_{\min}^{j+1} \sqrt{\frac{2^{j+1}}{t^{j+1}(j+1)!}} < d_{\min}^{j} \sqrt{\frac{2^j}{t^j j!}} \quad \Leftrightarrow \quad d_{\min} < \sqrt{\frac{t(j+1)}{2}},$$

which holds for all $j \geq 0$ in case $d_{\min} < \sqrt{\frac{t}{2}}$.

2. We want to prove that maximum on the level $j+1$ is less than the one at the level j. It is enough to prove

$$\forall \boldsymbol{\ell} : |\boldsymbol{\ell}| = j+1 \quad \exists \boldsymbol{\ell}' : |\boldsymbol{\ell}'| = j \quad \text{such that} \quad D_{\ell\ell} < D_{\ell'\ell'}.$$

Consider $\boldsymbol{\ell}' = \begin{pmatrix} \boldsymbol{\ell}_1 & \dots & \boldsymbol{\ell}_{i'} - 1 & \dots & \boldsymbol{\ell}_d \end{pmatrix}$. Then, using the fact that $\mathrm{Diag}_{ii} < \sqrt{\frac{t}{2}}$, we get

$$D_{\ell'\ell'} = \sqrt{\frac{2^j}{t^j \boldsymbol{\ell}'!}} \prod_{i=1}^{d} \mathrm{Diag}_{ii}^{\ell_i'} = \sqrt{\frac{2^j \boldsymbol{\ell}_{i'}}{t^j \boldsymbol{\ell}!}} \frac{\prod_{i=1}^{d} \mathrm{Diag}_{ii}^{\ell_i}}{\mathrm{Diag}_{i'i'}} = \sqrt{\frac{\boldsymbol{\ell}_{i'} t}{2}} \frac{D_{\ell\ell}}{\mathrm{Diag}_{i'i'}} > D_{\ell\ell}. \quad \square$$

With the help of the lemma, we note that

$$\max_{\substack{|\mathbf{j}| \geq j_{\max}+1, \\ i=1\dots N}} \tilde{D}_{ij} = \frac{\max\limits_{|\mathbf{j}| \geq j_{\max}+1} D_{jj}}{\min\limits_{i=1\dots N} D_{ii}} = \frac{\max\limits_{|\mathbf{j}| = j_{\max}+1} D_{jj}}{\min\limits_{i=1\dots N} D_{ii}} < \frac{\max\limits_{|\mathbf{j}| = j_{\max}+1} D_{jj}}{\min\limits_{|\mathbf{i}|=P} D_{ii}},$$

where $P = \min\left\{ K \in \mathbb{N}_0 \,\middle|\, N \leq \binom{K+d}{d} \right\}$. We can now formulate our cut-off criterion for the case when $\mathrm{Diag}_{ii} < \sqrt{\frac{t}{2}}$ for all $i = 1 \dots d$.

State-of-the-art criterion

$$\frac{\max\limits_{|\boldsymbol{\ell}|=j_{\max}+1}\{D_{\ell\ell}\}}{\min\limits_{|\boldsymbol{\ell}|=P}\{D_{\ell\ell}\}} < \varepsilon_{\mathrm{mach}}. \tag{10.2}$$

The criterion (10.2) guarantees that all additional columns that could be added to D (or, in particular, D_2) would yield elements in $\tilde{D}$ that are below machine precision.

Let us now recall our observations regarding the behavior of the elements of the matrix D. One can see in Figure 7 that for small values of ε the behavior is very similar for all three methods. If we consider a larger $\varepsilon = 1.2$, even though the condition on the Diag is broken, numerically we can see that in this case the decay of maximums and minimums is still fulfilled. However, the rate of decay in D is particularly bad in our formulation. This criterion also neglects the matrix $\tilde{R}$ and the effect of the polynomial vector $H^{G,E,t}(x-x_0)$. In particular, we know from section 6.4 that the tail of the polynomial vector also has some decay.

10.2. HermiteGF cut-off criterion

In this section, we derive a more holistic criterion for the cut-off in the HermiteGF expansion. We use the Vandermonde formulation of the method since it provides an explicit expression for the elements of all matrices which simplifies the analytic study of the method. We cut the polynomial vector as

$$H^{G,E,t}(\mathbf{x}-\mathbf{x}_0) = \begin{pmatrix}\hat{H}^{G,E,t}(\mathbf{x}-\mathbf{x}_0) & H^{G,E,t}_{\infty}(\mathbf{x}-\mathbf{x}_0)\end{pmatrix}$$

with $\hat{H}^{G,E,t} \in \mathbb{R}^{1\times M}$, where the number of basis functions used is larger than the number of collocation points, that is, $M \geq N$. Analogously we cut the $N \times \infty$ Vandermonde matrix W_2 and the infinite diagonal matrix D_2,

$$W_2 = \begin{pmatrix}\hat{W}_2 & W_\infty\end{pmatrix} \quad \text{and} \quad D_2 = \begin{pmatrix}\hat{D}_2 & 0\\ 0 & D_\infty\end{pmatrix}$$

with $\hat{W}_2 \in \mathbb{R}^{N\times(M-N)}$ and $\hat{D}_2 \in \mathbb{R}^{(M-N)\times(M-N)}$. We note that the infinite matrix W_∞ contains the columns of the full Vandermonde matrix W from column $M+1$ onward, while the infinite matrix D_∞ contains the diagonal entries of the full diagonal matrix D starting from the entry $M+1$. We then rewrite the formulation of the method (see section 8.3) after the cut-off,

$$\hat{\Psi}(\mathbf{x})^T = \begin{pmatrix}\mathrm{Id}_{N\times N} & D_1^{-1}W_1^{-1}\hat{W}_2\hat{D}_2\end{pmatrix}\hat{H}^{G,E,t}(\mathbf{x}-\mathbf{x}_0)^T.$$

We want to make sure that

$$\delta\Psi(\mathbf{x}) = \Psi(\mathbf{x}) - \hat{\Psi}(\mathbf{x})$$

is small for all collocation points by choosing a sufficiently large but not too large truncation parameter $j_{\max}$. For estimating $\delta\Psi(\mathbf{x})$, we will need the result of Lemma 7.2.1 stating that for $\mathbf{y} \in \mathbb{R}^d$ with $y_i \geq 0$ for all $i = 1, \dots, d$ and $j_{\max} \in \mathbb{N}$

$$\sum_{|\boldsymbol{\ell}| \geq j_{\max}} \frac{\mathbf{y}^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} \leq \sum_{|\boldsymbol{\ell}| = j_{\max}} \exp(\|\mathbf{y}\|_1) \frac{\mathbf{y}^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!},$$

where $\|\mathbf{y}\|_1 = |y_1| + \ldots + |y_d|$.

Before proceeding to the estimation of the truncation error $\|\delta\Psi(\mathbf{x})\|_2$, we recall the definition of the vectors

$$\mathbf{d}_{\text{vec}} = \operatorname{diag}(\mathrm{Diag}) \quad \text{and} \quad \mathbf{v}_k = (\mathrm{Id} + \mathrm{Diag}^{-1}\mathrm{Rem})\boldsymbol{\Delta}_k$$

for $k = 1, \dots, N$. They will contribute to the upper bound of the following estimate. In the isotropic case, they have the particularly simple form

$$\mathbf{d}_{\text{vec}} = \gamma^{-1}\varepsilon^2(1, \dots, 1) \quad \text{and} \quad \mathbf{v}_k = \boldsymbol{\Delta}_k \text{ for all } k = 1, \dots, N.$$

Theorem 10.2.1: Truncation estimate

For $k = 1, \dots, N$ we set

$$\omega_k = \sum_{i=1}^{N} (W_1^{-1})_{ki}^2 > 0 \quad \text{and} \quad \mathbf{y}_k = \mathrm{Diag}\, \mathbf{v}_k \in \mathbb{R}^d,$$

where W_1 is the upper left $N \times N$ block of the infinite Vandermonde matrix $W = (\mathbf{v}_k^{\boldsymbol{\ell}})$. For $j_{\max} \in \mathbb{N}$ we denote

$$\mathrm{const}_{j_{\max}} := \left(\sum_{k=1}^{N} \frac{\omega_k\, \mathbf{k}!\, (t/2)^{|\mathbf{k}|-(j_{\max}+1)}}{\mathbf{d}_{\text{vec}}^{2\mathbf{k}} (j_{\max}+1)!} \right) \cdot \left(\sum_{i=1}^{N} \exp\left(\frac{2}{t} \|\mathbf{y}_i\|_2^2 \right) \|\mathbf{y}_i\|_2^{2(j_{\max}+1)} \right).$$

Then, the truncation error $\delta\Psi$ satisfies for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\delta\Psi(\mathbf{x})\|_2^2 \;\leq\; \mathrm{const}_{j_{\max}} \cdot \left(H_{\lim}^{G,E,t}(\mathbf{x} - \mathbf{x}_0) - \sum_{|\boldsymbol{\ell}| \leq j_{\max}} H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x} - \mathbf{x}_0)^2 \right), \quad (10.3)$$

where $H_{\lim}^{G,E,t}(\mathbf{x} - \mathbf{x}_0)$ can be evaluated via (6.4).

Proof. We start by observing that

$$D_1^{-1} W_1^{-1} W_2 D_2 = D_1^{-1} W_1^{-1} \begin{pmatrix} \hat{W}_2 \hat{D}_2 & W_\infty D_\infty \end{pmatrix}.$$

Hence, it holds that

$$\delta\Psi(\mathbf{x})^T = D_1^{-1} W_1^{-1} W_\infty D_\infty H_\infty^{G,E,t}(\mathbf{x} - \mathbf{x}_0)^T,$$

and due to compatibility of the Frobenius norm and the 2-norm

$$\|\delta\Psi(\mathbf{x})\|_2^2 \leq \left\| D_1^{-1} W_1^{-1} W_\infty D_\infty \right\|_F^2 \cdot \left\| H_\infty^{G,E,t}(\mathbf{x} - \mathbf{x}_0) \right\|_2^2.$$

We further consider the two norms on the right-hand side separately. We first take a look at the Frobenius norm. Recall that in the RBF-QR method we evaluate the effect of the impact of $D_1^{-1} \dots D_2$ analytically. We can do the same here:

$$D_1^{-1} W_1^{-1} W_\infty D_\infty = \tilde{D}_\infty \,.*\, (W_1^{-1} W_\infty),$$

where $.*$ denotes the Hadamard product and $\tilde{D}_\infty$ is constructed analogously to (8.3). We write the Frobenius norm as

$$\|D_1^{-1} W_1^{-1} W_\infty D_\infty\|_F^2 = \sum_{k=1}^{N} \sum_{\ell > M} \left(\tilde{D}_{k\ell}^2 \cdot \left(\sum_{i=1}^{N} (W_1^{-1})_{ki} W_{i\ell} \right)^2 \right)$$

and estimate with the help of the Cauchy-Schwarz inequality,

$$\tilde{D}_{k\ell}^2 \left(\sum_{i=1}^{N} (W_1^{-1})_{ki} W_{i\ell} \right)^2 \le \frac{\omega_k}{D_{kk}^2} \, D_{\ell\ell}^2 \sum_{i=1}^{N} \mathbf{v}_i^{2\boldsymbol{\ell}},$$

where $\boldsymbol{\ell} \in \mathbb{N}^d$ is the ℓ-th multi-index corresponding to our basis enumeration. We used the explicit expression of $D_{\ell\ell}$ as defined in (8.2) and write

$$\frac{\omega_k}{D_{kk}^2} \, D_{\ell\ell}^2 \sum_{i=1}^{N} \mathbf{v}_i^{2\boldsymbol{\ell}} = \frac{\omega_k \, \mathbf{k}! \, t^{|\mathbf{k}|}}{\mathbf{d}_{\text{vec}}^{2\mathbf{k}} \, 2^{|\mathbf{k}|}} \, \frac{2^{|\boldsymbol{\ell}|}}{\boldsymbol{\ell}! \, t^{|\boldsymbol{\ell}|}} \sum_{i=1}^{N} (\operatorname{Diag} \mathbf{v}_i)^{2\boldsymbol{\ell}}.$$

We denote $\tilde{\mathbf{y}}_i = \begin{pmatrix} \frac{2}{t}(\mathbf{y}_i)_1^2 & \dots & \frac{2}{t}(\mathbf{y}_i)_d^2 \end{pmatrix}$. Then, by Lemma 7.2.1 we get

$$\begin{aligned} \|D_1^{-1} W_1^{-1} W_\infty D_\infty\|_F^2 &\le \sum_{k=1}^{N} \frac{\omega_k \, \mathbf{k}! \, t^{|\mathbf{k}|}}{\mathbf{d}_{\text{vec}}^{2\mathbf{k}} \, 2^{|\mathbf{k}|}} \sum_{i=1}^{N} \sum_{|\boldsymbol{\ell}| > j_{\max}} \frac{\tilde{\mathbf{y}}_i^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!} \\ &\le \sum_{k=1}^{N} \frac{\omega_k \, \mathbf{k}! \, t^{|\mathbf{k}|}}{\mathbf{d}_{\text{vec}}^{2\mathbf{k}} \, 2^{|\mathbf{k}|}} \sum_{i=1}^{N} \sum_{|\boldsymbol{\ell}| = j_{\max}+1} \exp\left(\|\tilde{\mathbf{y}}_i\|_1\right) \frac{\tilde{\mathbf{y}}_i^{\boldsymbol{\ell}}}{\boldsymbol{\ell}!}. \end{aligned}$$

Using the multinomial theorem and the fact that

$$\|H_\infty^{G,E,t}(\mathbf{x} - \mathbf{x}_0)\|_2^2 = H_{\lim}^{G,E,t}(\mathbf{x} - \mathbf{x}_0) - \sum_{|\boldsymbol{\ell}| \le j_{\max}} H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x} - \mathbf{x}_0)^2.$$

we arrive to estimate (10.3). □

The denominator $\mathbf{d}_{\text{vec}}^{2\mathbf{k}}$ in expression (10.3) can take extremely small values that can lead to underflow. To avoid this, it can be combined with $\|\mathbf{y}_i\|_2^{2(j_{\max}+1)}$. For this, we define the d-dimensional index $\mathbf{j}_d = \left(\frac{1}{d}, \dots, \frac{1}{d}\right)$ and use the following transformation

$$\|\mathbf{y}_i\|_2 = \mathbf{d}_{\text{vec}}^{\mathbf{j}_d} \|\mathbf{y}_i./(\mathbf{d}_{\text{vec}}^{\mathbf{j}_d})\|_2 =: \mathbf{d}_{\text{vec}}^{\mathbf{j}_d} \|\mathbf{y}_i^D\|_2,$$

where $./$ denotes component-wise division. Pulling out $\mathbf{d}_{\text{vec}}^{\mathbf{j}_d}$, the constant of the estimate (10.3) can be rewritten as

$$\text{const}_{j_{\max}} = \left(\sum_{k=1}^{N} \frac{\omega_k \mathbf{k}! \left(\frac{2}{t} \mathbf{d}_{\text{vec}}^2\right)^{-\mathbf{k} + (j_{\max}+1)\mathbf{j}_d}}{(j_{\max}+1)!} \right) \left(\sum_{i=1}^{N} \exp\left(\frac{2}{t} \|\mathbf{y}_i\|_2^2 \right) \|\mathbf{y}_i^D\|_2^{2(j_{\max}+1)} \right).$$

Note, that in the isotropic case, one can simplify $\mathbf{d}_{\text{vec}}^{-2\mathbf{k}+2(j_{\max}+1)\mathbf{j}_d} = (\varepsilon^4/\gamma^2)^{j_{\max}+1-|\mathbf{k}|}$. We are now ready to formulate our cut-off criterion.

HermiteGF cut-off criterion

We choose $j_{\max}$ for the HermiteGF-QR method such that

$$\max_{k=1\dots N} \frac{\|\delta\Psi(\mathbf{x}_k)\|_2}{\|\hat{\Psi}(\mathbf{x}_k)\|_2} \leq \mathit{TOL}, \tag{10.4}$$

where $\{\mathbf{x}_k\}_{k=1}^N$ are the collocation points.

Since we are looking at the relative error, the tolerance `TOL` need not be machine precision. The crucial difference to the state-of-the-art criterion eq. (10.1) is that now the `TOL` directly controls the accuracy of the stable basis Ψ. Depending on the desired accuracy, the tolerance can be adjusted for the specific problem.

Remark 10.2.1. *The state-of-the-art criterion that truncates diagonal elements below machine precision does not provide an error bound on the interpolant. On the one hand, the new criterion requires more computations for determining the cut-off degree. On the other hand, it allows to reduce the polynomial degree* $j_{\max}$ *while still guaranteeing a given truncation error. This in turn reduces the computational cost of the interpolation step.*

10.2.1. Automatic detection of t

One can use the criterion above also for determining the value of the parameter t. We scan the whole spectrum of the values of t and detect the one that yields the minimum amount of basis functions

$$\arg\min_{t\in(0,1)} j_{\max}(t) = t_{\text{auto}}.$$

Note that very small values of t can cause cancellations and should be excluded (see subsection 12.1.1). Even though this introduces additional computational cost in the determination of the suitable expansion, it could be profitable for the cases where the basis is used multiple times after having fixed the number $j_{\max}$ as e.g. in a time loop.

11. Implementation

We have implemented the HermiteGF interpolation in `MATLAB`. The code is available at `https://gitlab.mpcdf.mpg.de/clapp/hermiteGF`. Even though the described approach allows to reduce the ill-conditioning of the collocation and evaluation matrices, the algorithm involves formulas with factorials and polynomials of increasing power, so care has to be taken when implementing to avoid overflow and cancellation. Therefore, it is crucial to have a stability-aware implementation. Below we discuss the implementation of the parts of the code that proved to be crucial for stability and performance. This chapter originates from the section 5.1 of the preprint “Stable evaluation of Gaussian radial basis functions using Hermite polynomial” [YK17] by Yurova & Kormann. However, compared to the corresponding section, in this chapter, a lot of additional details are provided.

11.1. Evaluation of the basis

The HermiteGF-based matrices become increasingly ill-conditioned for growing number of basis functions. On the other hand, the product of the evaluation matrix $\Psi(X^{\text{eval}})$ and the inverse of the collocation matrix $\Psi(X^{\text{col}})$ is still well-conditioned. For this reason, it is crucial to take special care when building these matrices and inverting the collocation matrix. First of all, we evaluate the basis in a stable way by using the formulation through d-dimensional Hermite functions (3.1)

$$H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}) = \pi^{d/4} t^{|\boldsymbol{\ell}|/2} \psi_{\boldsymbol{\ell}}(G^T\mathbf{x}) \exp\left(\mathbf{x}^T \left(-E^T E + \frac{1}{2}GG^T\right)\mathbf{x}\right).$$

Since Hermite functions can be evaluated through the three-term recurrence, we completely avoid computing the polynomials. Moreover, it is enough to operate with only 1D Hermite functions and assemble the multidimensional structure only at the end. Alternatively, one can use directly the three-term recurrence (3.8) for the HermiteGF basis functions. In the current version of the code, the first approach is used and proved sufficient.

Another improvement that can be made, is computing $\Psi(X_\ell^{\text{eval}})\Psi(X_\ell^{\text{col}})^{-1}$ together which allows to cancel out some ill-conditioning. Using the built-in operator / for the inversion yields good results in `MATLAB`.

We now proceed to the implementation of the computation of other crucial parts of the HermiteGF-QR method.

11.2. The computation of the effect of $D_1^{-1} \dots D_2$

The effect of $D_1^{-1} \dots D_2$ can be computed analytically, however, it still has to

ALGORITHM 2 Compute $\tilde{D}$

```
Require: factorial_part1 – $N \times 1$ array of precomputed factorials.
Initialize factorial_part2 of size $1 \times M - N$ with ones.
for $j = N + 1 \dots M$ do
    for dim = $1 \dots d$ do
        factorial_part2(j - N + 1) /= $\mathbf{j}_{\texttt{dim}}!$
    end for
end for
$\tilde{D} = \sqrt{\texttt{factorial\_part1*factorial\_part2}}$
for dim = $1 \dots d$ do
    Deg $\leftarrow ((\boldsymbol{\ell}_1)_{\mathrm{dim}} - (\boldsymbol{\ell}_2)_{\mathrm{dim}})_{\ell_1=1\dots N, \ell_2=N+1\dots M}$
    $\tilde{D}$ = $\tilde{D}$ .* $(\sqrt{2/t}\,\mathbf{d}_{\mathrm{vec}})_{\texttt{dim}}$ .^ Deg.
end for
```

be implemented with care. Let ℓ be the index of the HermiteGF basis functions and $\boldsymbol{\ell}$ be corresponding multivariate indices. Recall that

$$\tilde{D}_{\ell_1,\ell_2-N} = \frac{D_{\ell_2\ell_2}}{D_{\ell_1\ell_1}} = \frac{\mathbf{d}_{\mathrm{vec}}^{\boldsymbol{\ell}_2}\sqrt{2^{|\boldsymbol{\ell}_2|}}}{\sqrt{t^{|\boldsymbol{\ell}_2|}\boldsymbol{\ell}_2!}} \cdot \frac{\sqrt{t^{|\boldsymbol{\ell}_1|}\boldsymbol{\ell}_1!}}{\mathbf{d}_{\mathrm{vec}}^{\boldsymbol{\ell}_1}\sqrt{2^{|\boldsymbol{\ell}_1|}}} = \underbrace{\left(\sqrt{\frac{2}{t}}\mathbf{d}_{\mathrm{vec}}\right)^{\boldsymbol{\ell}_2-\boldsymbol{\ell}_1}}_{\text{power part}} \underbrace{\sqrt{\frac{\boldsymbol{\ell}_1!}{\boldsymbol{\ell}_2!}}}_{\text{factorial part}},$$

where ℓ_1 are indices corresponding to D_1 and ℓ_2 are indices corresponding to D_2. Since the elements of $\mathbf{d}_{\mathrm{vec}}$ are usually small and $t < 1$, it is advantageous for the stability to first combine $\sqrt{2/t}$ with $\mathbf{d}_{\mathrm{vec}}$ before the exponentiation. This way, we make the base larger which allows us to prevent underflow. Let us now consider some optimizations that are required for a fast code.

11.2.1. Optimizations

Even though a straight-forward element-wise computation of the matrix $\tilde{D}$ via a double `for`-loop provides a numerically accurate result, the runtime of the $\tilde{D}$ computation becomes very large with the growth of $j_{\max}$ or N. With the corresponding function taking more time than all other computations, it turned out to be one of the major performance bottlenecks. To optimize the code we do the following:

1. Minimize the amount of factorial computations.

2. Vectorize the power computations.

As for the factorial part, it proved to be one of the most computationally extensive computations. For that reason, we employ the following representation

$$\text{factorial part} = \sqrt{(\boldsymbol{\ell}_1!)_{\ell_1 \in 1\dots N} \otimes (1/\boldsymbol{\ell}_2!)_{\ell_2 \in N+1\dots M}}.$$

We also note that the factorials of the indices corresponding to the first N basis functions are always required, independently from $j_{\max}$. That is why we precompute them at the very beginning of the computation and store them in an $N \times 1$ array factorial_part1. By the time we arrive to the function, where $\tilde{D}$ is evaluated, we already have the first N factorials. Keeping in mind the Kronecker product representation, we just have to compute the remaining $M - N$ factorials. In particular, we need to compute the inverse of the factorials corresponding to the D_2. We store the newly computed values in the $1 \times M - N$ array factorial_part2. Then, the factorial part of $\tilde{D}$ can be computed as a tensor product

$$\text{factorial part} = \underbrace{\texttt{factorial\_part2}}_{N\times 1} \star \underbrace{\texttt{factorial\_part1}}_{1\times M-N}.$$

This way the information required for assembling the factorial part of $\tilde{D}$ is stored in two vectors with cumulative size M until we need to form the full matrix. Moreover, the first N factorials can be saved at the very beginning and used later on in the computation of $j_{\max}$ (see section 11.3).

As for the power part, to achieve a good performance in MATLAB, we recall that MATLAB linear algebra routines are highly optimized. Therefore, it is beneficial for the performance to vectorize the code as much as possible (i.e. reduce the number of for-loops and replace them with matrix operations). In the isotropic case, such implementation would be rather straightforward. Indeed,

$$\sqrt{\frac{2}{t}}\mathbf{d}_{\text{vec}} = \left(\sqrt{\frac{2}{t}\frac{\varepsilon^2}{\gamma}}, \dots, \sqrt{\frac{2}{t}\frac{\varepsilon^2}{\gamma}}\right).$$

Therefore

$$\left(\sqrt{\frac{2}{t}}\mathbf{d}_{\text{vec}}\right)^{\boldsymbol{\ell}_2-\boldsymbol{\ell}_1} = \left(\sqrt{\frac{2}{t}\frac{\varepsilon^2}{\gamma}}\right)^{|\boldsymbol{\ell}_2|-|\boldsymbol{\ell}_1|}$$

After forming the matrix Deg of the degrees corresponding to $\boldsymbol{\ell}_2 - \boldsymbol{\ell}_1$, we can compute the desired powers in one line of MATLAB code:

$$\left(\sqrt{\frac{2}{t}}\mathbf{d}_{\text{vec}}\right)^{\boldsymbol{\ell}_2-\boldsymbol{\ell}_1} = \left(\sqrt{\frac{2}{t}\frac{\varepsilon^2}{\gamma}}\right)\texttt{.\^{}Deg},$$

where .^ denotes element-wise power operator. In the anisotropic case, it is a bit more tricky since the elements of $\mathbf{d}_{\text{vec}}$ might differ. In this case, we compute the required matrix iteratively, looping through dimensions (see Algorithm 2).

11.3. The computation of $j_{\max}$

In this section, we take a look at the implementation of the cut-off criterion (10.4):

$$j_{\max} = \min\left\{ j_{\max} \geq P \;\middle|\; \max_{k=1\dots N} \frac{\|\delta\Psi(\mathbf{x}_k)\|_2}{\|\hat{\Psi}(\mathbf{x}_k)\|_2} \leq \texttt{TOL} \right\},$$

where P is the largest polynomial degree in the first N basis functions. Recall the expression for the estimate of $\|\delta\Psi\|_2$:

$$\|\delta\Psi(\mathbf{x})\|_2^2 \leq \underbrace{\left(\sum_{k=1}^{N} \frac{\omega_k \mathbf{k}! \left(\frac{2}{t}\mathbf{d}_{\text{vec}}^2\right)^{-\mathbf{k}+(j_{\max}+1)\mathbf{j}_d}}{(j_{\max}+1)!}\right)}_{\mathbf{d}_{\text{vec}} \text{ part}} \underbrace{\left(\sum_{i=1}^{N} \exp\left(\frac{2\|\mathbf{y}_i\|_2^2}{t}\right) \|\mathbf{y}_i^D\|_2^{2(j_{\max}+1)}\right)}_{\mathbf{y} \text{ part}}$$

$$\cdot \underbrace{\left(H_{\lim}^{G,E,t}(\mathbf{x}) - \sum_{|\boldsymbol{\ell}|\leq j_{\max}} H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0)^2\right)}_{\text{Hermite part}}.$$

Let us first mention two implementation details which make the computations more stable:

1. The inverse W_1^{-1} can be stably computed using Moore–Penrose pseudoinverse [Pen55] (`pinv` function in `MATLAB`, see `https://de.mathworks.com/help/matlab/ref/pinv.html` for a reference).

2. The evaluation of the sum $\sum_{|\boldsymbol{\ell}|\leq j_{\max}} H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0)^2$ is more stable if we combine the summands by degree. In particular, we evaluate the sum as

$$\sum_{|\boldsymbol{\ell}|\leq j_{\max}} H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0)^2 = \sum_{\texttt{deg}=0}^{j_{\max}} \sum_{|\boldsymbol{\ell}|=\texttt{deg}} H_{\boldsymbol{\ell}}^{G,E,t}(\mathbf{x}-\mathbf{x}_0)^2.$$

   This formulation is also advantageous for the performance. If the $j_{\max}$ is increased by one, we just need to add the sum of the basis functions with the indices of the cumulative degree $j_{\max}+1$ to the existing sum.

Apart from the points mentioned above, similarly to the computation of the effect of $D_1^{-1}\dots D_2$, it is crucial to compute only cumulative powers of $\frac{2}{t}\mathbf{d}_{\text{vec}}^2$.

Remark 11.3.1. *There are special algorithms for computing the inverse of Vandermonde matrix (see, for example, [BP70, Tra66]). However, these algorithms have not been employed in the current version of the code.*

The function finding the suitable value of $j_{\max}$ is one of the most computationally intensive functions of the method. Below we discuss the optimizations that can be done in order to have an efficient implementation.

11.3.1. Optimization details

We compute the $j_{\max}$ in a `while`-loop

```
while $\max_{k=1\dots N} \frac{\|\delta\Psi(\mathbf{x}_k)\|_2}{\|\hat{\Psi}(\mathbf{x}_k)\|_2}$ <TOL do
  $j_{\max} = j_{\max} + 1$
  Update $\max_{k=1\dots N} \frac{\|\delta\Psi(\mathbf{x}_k)\|_2}{\|\hat{\Psi}(\mathbf{x}_k)\|_2}$
end while
```

Since the search of the final $j_{\max}$ takes several iterations, it is advantageous to precompute all terms that are not related to $j_{\max}$ and can be used within the `while`-loop. In particular, we precompute the following values

$$\omega_k,\ \exp\left(\frac{2\|\mathbf{y}_i\|_2^2}{t}\right),\ \|\mathbf{y}_i^D\|_2^2.$$

Additionally, as already mentioned in the section 11.2, by the time we arrive to the function computing $j_{\max}$, the factorials $\mathbf{k}!$ are already computed and stored in the array `factorial_part1`. As mentioned above, even though $(\frac{2}{t}\mathbf{d}^2_{\text{vec}})^{-\mathbf{k}}$ does not change over time, we do not compute it separately from the power $(j_{\max}+1)\mathbf{j}_d$ in order to avoid cancellations and overflow.

Recall that in the cut-off criterion we are looking at the error relative to $\|\hat{\Psi}(\mathbf{x})\|_2^2$ with

$$\hat{\Psi}(\mathbf{x})^T = \left(\mathrm{Id}_{N\times N} \quad D_1^{-1}W_1^{-1}\hat{W}_2\hat{D}_2\right)\hat{H}^{G,E,t}(\mathbf{x}-\mathbf{x}_0)^T.$$

From the estimate of $\|\delta\Psi\|_2^2$ we already have W_1^{-1} and the values of the basis functions $\hat{H}^{G,E,t}(\mathbf{x}-\mathbf{x}_0)$. We are left with computing the effect of $D_1^{-1}\ldots D_2$ and the remaining part of the Vandermonde matrix W_2. However, we would not want to compute the whole W_2 and D_2 for every $j_{\max}$. Due to the Vandermonde formulation, not only W_1 does not change with increase of $j_{\max}$, but also already computed columns of W_2. The same holds for D_2. Therefore, in this case, we can update $\Psi(\mathbf{x}_k)$ iteratively, adding the influence of only the new block at each loop iteration. In particular,

$$\hat{\Psi}^{j_{\max}+1}(\mathbf{x})^T = (\hat{\Psi}^{j_{\max}})^T + D_1^{-1}W_1^{-1}W_2^{j_{\max}+1}D_2^{j_{\max}+1}H^{G,E,t}_{|\boldsymbol{\ell}|=j_{\max}+1}(\mathbf{x}-\mathbf{x}_0)^T,$$

where $H^{G,E,t}_{|\boldsymbol{\ell}|=j_{\max}+1}(\mathbf{x}-\mathbf{x}_0)$ denotes a vector containing the values of the HermiteGF basis functions with cumulative polynomial degree $j_{\max}+1$, $W_2^{j_{\max}+1} \in \mathbb{R}^{N\times(M_{j_{\max}+1}-M_{j_{\max}})}$, $D_2^{j_{\max}+1} \in \mathbb{R}^{(M_{j_{\max}+1}-M_{j_{\max}})\times(M_{j_{\max}+1}-M_{j_{\max}})}$ with $M_{j_{\max}+1}-M_{j_{\max}}$ being the number of basis functions on the level $j_{\max}+1$.

The same approach applies to the sum $\sum_{|\boldsymbol{\ell}|\leq j_{\max}} H^{G,E,t}_{\boldsymbol{\ell}}(\mathbf{x}-\mathbf{x}_0)^2$. Indeed, in every iteration, we then need to add only the sum of the basis functions from the new block.

12. Numerical results

We compare the isotropic HermiteGF-based algorithm with the existing stabilization methods, the Chebyshev-QR method[1] and the Gauss-QR method[2]. We evaluate the influence of different parameters, such as ε, γ, number of collocation points N on the quality of the interpolation. For the Gauss-QR method, we take the free parameter α to be equal to our value of γ, i.e., $\alpha_{\mathrm{Gauss-QR}} = \gamma$. Since there are no stabilization methods available for fully anisotropic interpolation, we test the anisotropic HermiteGF-QR only against the direct algorithm to verify the correctness. To determine the cut-off degree in the HermiteGF method, we use the HermiteGF cut-off criterion with $\texttt{TOL} = 10^{-6}$, unless stated otherwise. For this tolerance, the HermiteGF-QR method provides results that match the results from Chebyshev-QR and Gauss-QR. The parameter t is detected automatically. In all tests we evaluate the interpolant at a set of evaluation points $\{\mathbf{z}_k\}_{k=1}^{N_{\mathrm{ev}}}$ and look at the average error of the form [FM12, § 5.1, Expr. (5.2)]:

$$\text{error} = \frac{1}{N_{\mathrm{ev}}} \sqrt{\sum_{k=1}^{N_{\mathrm{ev}}} \left(\frac{f(\mathbf{z}_k) - s(\mathbf{z}_k)}{f(\mathbf{z}_k)} \right)^2}.$$

12.1. 2D isotropic interpolation

In this section, we take a look at the two-dimensional isotropic interpolation with the HermiteGF-QR method. We take multiples of the identity for both E and G. We look at a hyperbolic domain (see Figure 11) defined by the inequality

$$0.04 \leq (x + 1.2)^2 - 4y^2 \leq 1 \tag{12.1}$$

with a boundary condition $x^2 + y^2 \leq 1$. The hyperbola of type (12.1) can then be parameterized as

$$(x, y) = r(t) = (c\cosh(t) - 1.2, 0.5c\sinh(t)),\ t \in \mathbb{R},\ c \in [0.2, 1].$$

We run the tests for the following function (f_4 from [FLF11, § 6]):

$$f_{\mathrm{h}}(x, y) = \sin(x^2 + 2y^2) - \sin(2x^2 + (y - 0.5)^2).$$

We investigate the behavior of the performance of the interpolation with respect to the parameters γ, ε, and number of functions N. We use $\gamma = 3.5$, $\varepsilon = 0.05$ and optimize t from the set $\mathrm{tvec} = $ `linspace(0.1, 0.99, 10)`, unless stated otherwise.

[1] Code downloaded from `http://www.it.uu.se/research/scientific_computing/software/rbf_qr` on September 10, 2018.

[2] Code downloaded from `http://math.iit.edu/~mccomic/gaussqr/` on September 5, 2018.

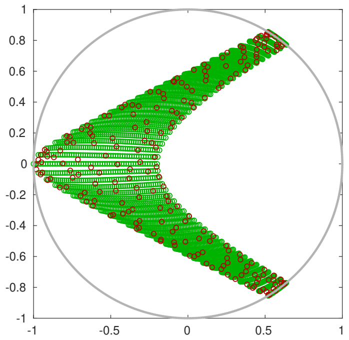
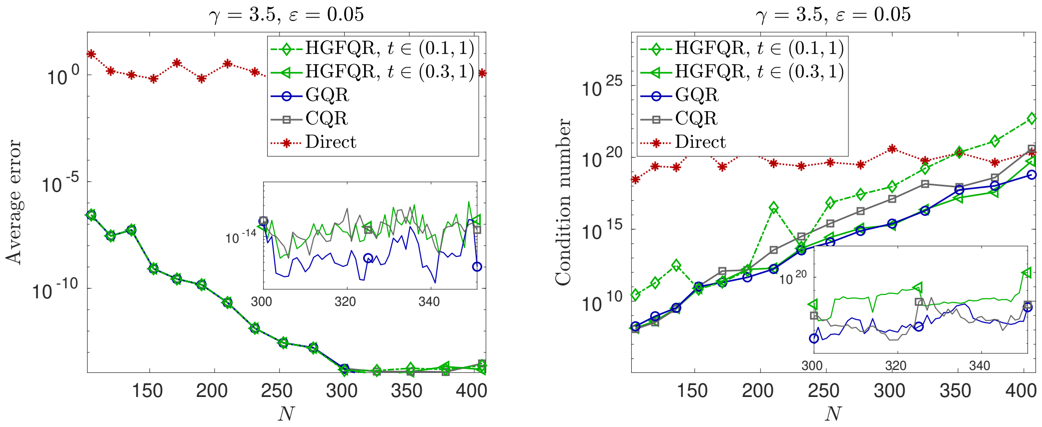


Figure 11 Hyperbolic domain. Evaluation grid (green) and $N = 210$ clustered Halton node points (red).

ALGORITHM 3 Generation of collocation nodes on hyperbolic domain

```
for $k = 1 \ldots N$ do
    Sample a 2D Halton point $(c_k, t_k)$ on a square $[0.2, 1] \times [-1, 1]$.
    Push the point to the boundary [FLF11, $\S 6.1.1$]:
        $$(c_k, t_k) = \left(\sin\left(\frac{\pi c_k}{2}\right), \sin\left(\frac{\pi t_k}{2}\right)\right)$$
    Compute the $t_{\max}$ for which the positive branch of the hyperbola intersects
    with the boundary circle on the right:
        $$t_{\max}: \quad (c_k \cosh(t_{\max}) - 1.2)^2 + 0.25 c_k^2 \sinh(t_{\max})^2 - 1 = 0.$$
    Most of the non-linear solvers require an initial guess. Since we know that
    our circle intersects the hyperbola at a point with $x \leq 1$, i.e. $c_k \cosh(t_{\max}) \leq$
    $2.2$, we can use the inverse of $\cosh$ in order to compute $t$ corresponding to
    that value:
        $$t_{\text{max\_guess}} = \left|\log\left(\frac{2.2}{c_k} + \sqrt{\frac{(2.2)^2}{c_k^2} - 1}\right)\right|.$$
    Scale $t_k$ with respect to $t_{\max}$:
        $$t = t_k t_{\max}.$$
    Assign $(x_k, y_k) = (c \cosh(t), 0.5 c \sinh(t))$.
end for
$\{(x_k - 1.2, y_k)\}$ are the generated points on the hyperbolic domain.
```

We sample the collocation points from Halton points that are clustered near the boundary to improve the conditioning of the polynomial interpolation and are then mapped to the hyperbolic domain (see Algorithm 3). For all tests, we use $N_{\text{ev}} = 53^2$ evaluation points that are sampled similarly to the collocation points,



(a) Average error.

(b) Condition number.

Figure 12 For the two-dimensional isotropic test case, the interpolation quality is the same for all three stabilization methods. The conditioning is slightly worse for small values of t. There is small noise for all methods when the number of radial basis functions N does not correspond to a number of all polynomials of a degree $\leq P$ for a certain P.

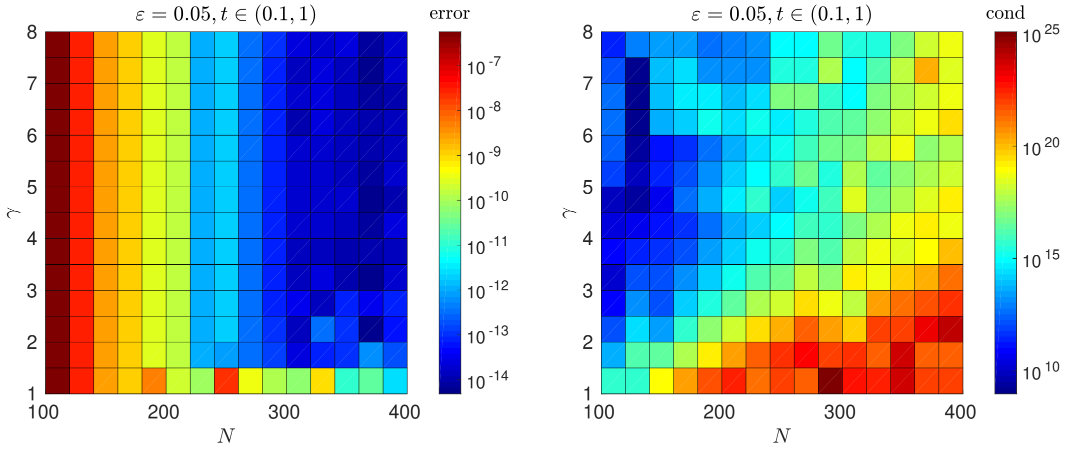
but based on a uniform grid and without clustering. The nodes distribution is depicted in Figure 11. This domain and sampling strategy choice was inspired by [FLF11, § 6.1.2].

12.1.1. The number of nodes N

Let us first look at the behavior of the method for different numbers of nodes, N. We take the values of $N \in [100, 410]$ of the form $\binom{P+2}{2}$ for some integer P, such that there are no same powers of ε present in both D_1 and D_2. In Figure 12, we see that the error consistently decays for all the tested methods. Choosing the truncation parameter t in the interval $t \in (0.1, 1)$, the conditioning of the HermiteGF-QR method is slightly worse than for the other methods, since big powers of smaller values of t yield cancellations. Indeed, limiting the range of t to $t \in (0.3, 1)$ brings the conditioning to the level of the other methods. Using all integers in the interval $[100, 410]$ also provided consistent results for all three methods, however, the picture gets noisy. A snippet of that behavior can be seen in the zoomed regions in Figure 12. This can be related to the fact that for the values of N of the form above the limit of the RBF interpolant in the flat limit $\varepsilon \to 0$ is a *unique* polynomial of degree P [LF05, §4 Theorem 4.1] whereas for other values the uniqueness is not guaranteed.

12.1.2. Sensitivity to γ

Let us take a look at the influence of the parameter γ on the interpolation quality. We see in Figure 13 that for small N the interpolation quality is not sensitive to the value of γ. However, for larger N the parameter γ has to be chosen with care. One can see in the Figure 13b that the conditioning is worse for small γ. However, one should be careful while increasing γ since it also increases the evaluation domain of the Hermite polynomials, which take very large values on large domains which can lead to overflow. This effect becomes more pronounced as the degree of the Hermite polynomials increases. The optimal balance depends on the particular function and the number of basis functions.



(a) Error.

(b) Conditioning.

Figure 13 Average error and condition number for the isotropic two-dimensional test case. For small and moderate N the interpolation quality is not sensitive to the value of γ, whereas for the larger N one should carefully choose the value of γ.

12.1.3. Cut-off degree $j_{\max}$

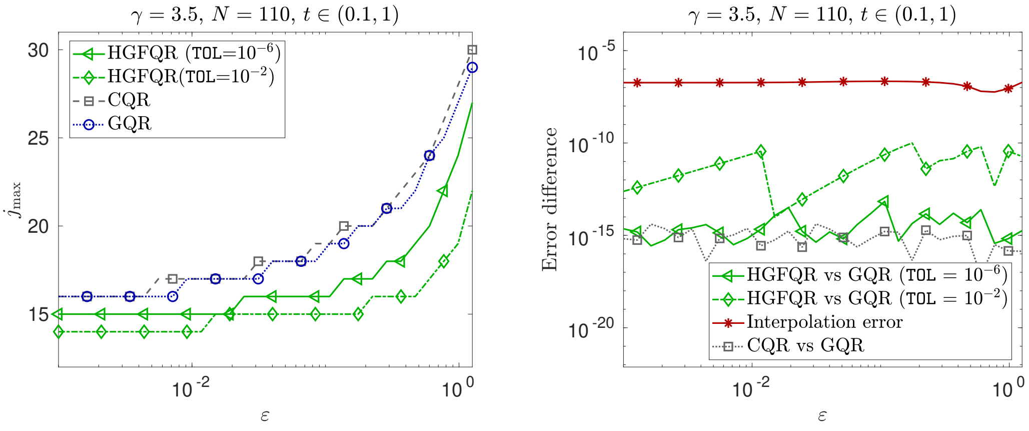
Next, we look at the influence of the value of $\texttt{TOL}$ on the quality of the interpolation. We compare the error only to the Gauss-QR method since the difference between the Chebyshev-QR and Gauss-QR results is down to machine precision. One can see in Figure 14 that for $\texttt{TOL} = 10^{-6}$ the difference HermiteGF-QR and Gauss-QR is also down to machine precision. If we relax the tolerance to 10^{-2}, the error is still small compared to the magnitude of the interpolation error, while having smaller $j_{\max}$, which yields an improved computational efficiency. Also, the figure shows a general trend that the expansion decays rather fast for small ε while an increasing number of basis functions is needed for ε close to 1.

12.2. 2D anisotropic interpolation

To test the performance of the HermiteGF expansion for anisotropic basis functions, we consider the function:

$$f_{\mathrm{a}}(x,y) = \frac{1}{x^2 + xy + y^2} + 2, \quad x, y \in [-1,1].$$

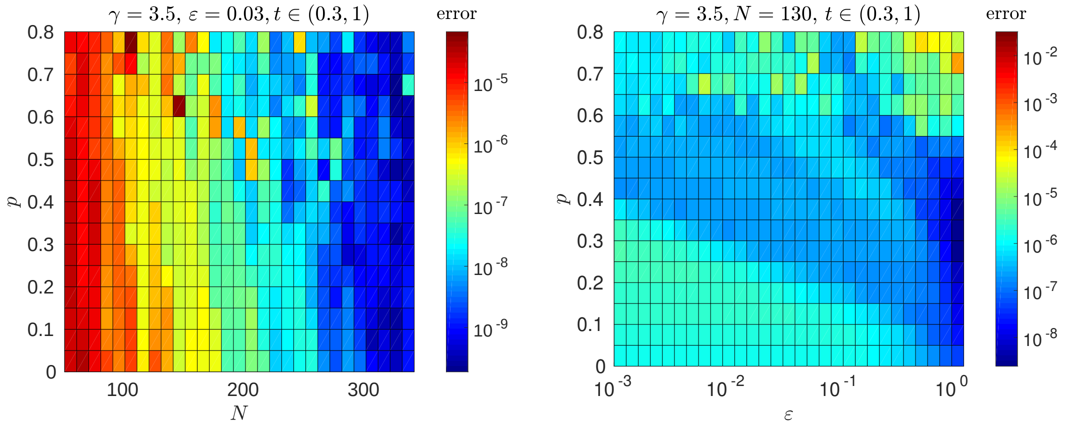
As collocation points, we use Halton points clustered toward the boundaries. For the evaluation grid we use 53×53 uniformly distributed points. As for the shape



(a) Dependence of $j_{\max}$ on ε.

(b) Absolute difference to reference errors.

Figure 14 Optimized truncation value and error differences for the two-dimensional isotropic test case. For the coarser TOL$=10^{-2}$ we get fewer basis functions, with truncation error still much below the interpolation error.



(a) Dependence on N.

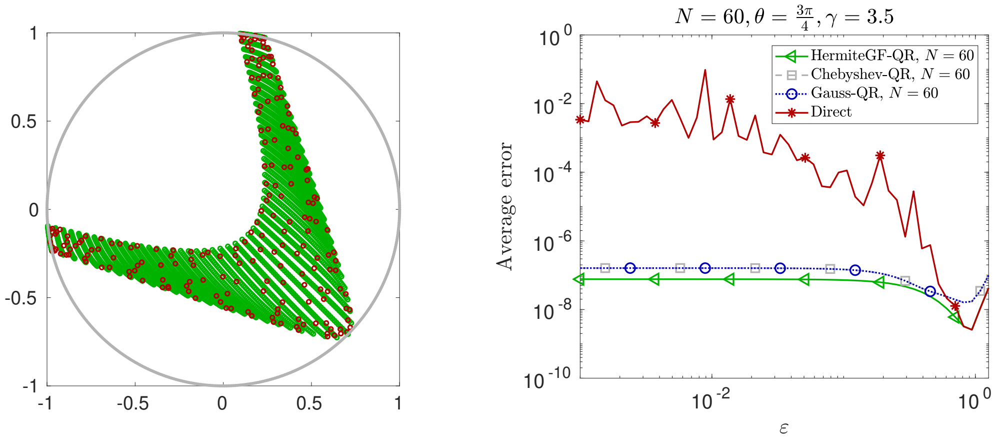
(b) Dependence on ε.

Figure 15 Two-dimensional anisotropic interpolation. For a fixed value of p, the error generally decreases with the growth of N as expected. Choosing an anisotropic shape matrix E ($p \neq 0$) often improves the interpolation quality.

matrix E, we check whether the off-diagonal elements influence the quality of the results. We choose a non-diagonal matrix G of arbitrary pattern to demonstrate the robustness of the method. We fix E and G to be of the following form:

$$E = \varepsilon \begin{pmatrix} 1 & p \\ p & 1 \end{pmatrix} \text{ with } p \in [0, 0.8], \quad G = \gamma \begin{pmatrix} 1 & 0.3 \\ 0.1 & 1.3 \end{pmatrix} \text{ with } \gamma = 3.5.$$

We restrict t to the interval $\mathrm{tvec} =$ `linspace(0.3, 1, 10)` in order to improve the stability of the computations. Let us take a look on how much the off-diagonal elements of the matrix E influence the error. For our scan, we take 30 logarithmically distributed values of $\varepsilon \in [10^{-3}, 10^{0.1}]$. One can see from Figure 15 that certain choices of off-diagonal elements can improve the quality of the interpolation compared to purely diagonal shape matrices. However, for larger p slight instabilities occur which might be explained since E becomes singular for $p \to 1$.



(a) Hyperbolic domain turned by $\theta = \frac{3\pi}{4}$.

(b) Average error.

Figure 16 HermiteGF-QR provides a considerable improvement for small ε compared to the direct anisotropic solver. For larger ε the average error of the anisotropic HermiteGF-QR method corresponds to the one of the direct method. That validates the accuracy of the HermiteGF-QR method. Compared to the reference QR methods, a slight improvement is observed.

12.3. Arbitrary domain

The method can also be successfully applied to arbitrary domains. We consider the hyperbolic domain from the 12.1 turned by $\theta = \frac{3\pi}{4}$ (see Figure 16a) and the same function f_{a}. We use the same sampling strategy as in the section 12.1, and turn the domain afterward. We set $\gamma = 3.5$, and the matrix G of the form

$$G = \begin{pmatrix} \gamma & 0.3\gamma \\ 0.1\gamma & 1.3\gamma \end{pmatrix}$$

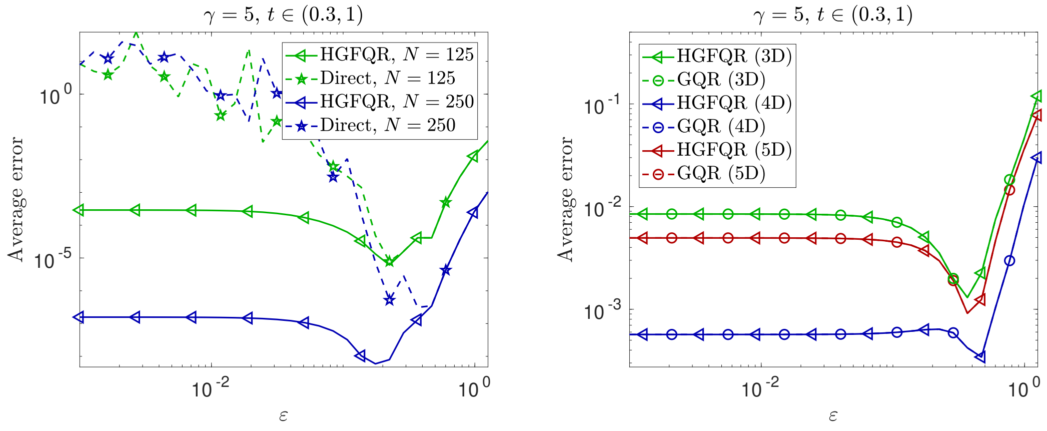
We use $N = 60$ interpolation points and $p = 0.25$ for the off-diagonal scaling of E. For the evaluation grid we use 53×53 points. We compare the accuracy of the interpolation to the isotropic HermiteGF-QR and Gauss-QR methods as well as to the direct anisotropic solver. One can see in the Figure 16b that the error is slightly better for the anisotropic HermiteGF-QR for most values of ε in comparison to the reference methods with the same value of ε. Moreover, for larger values of ε the HermiteGF-QR solution converges to the one of the direct anisotropic interpolation. For smaller values of ε HermiteGF-QR provides a significant improvement in accuracy in comparison to the direct solver.

12.4. Multivariate interpolation

In this section, we consider an example of the usage of HermiteGF-QR in higher dimension. For all tests, we use the function

$$f(\mathbf{x}) = \cos(|\mathbf{x}|), \quad \mathbf{x} \in [-1, 1]^d,$$

where $|\mathbf{x}| = \sum_{i=1}^{d} x_i$. We use Halton collocation points and 1000 Halton points, excluding the ones used for collocation, for the evaluation grid. We fix $G =$



(a) Anisotropic interpolation in 3D.

(b) Isotropic interpolation in 3-5D.

Figure 17 Interpolation in dimensions 3-5D: Both the anisotropic and the isotropic HermiteGF-QR method match the average error of the Gauss-QR method.

$5\mathrm{Id}_d$, $\varepsilon = \texttt{logspace(-3, 0.1, 30)}$, and we again optimize the parameter t over the set $\mathrm{tvec} = \texttt{linspace(0.3, 1, 10)}$. As before, we choose the tolerance $\texttt{TOL} = 10^{-6}$. We first look whether the anisotropic HermiteGF-QR method converges to the results of the direct interpolation as ε increases. We choose an arbitrary pattern for E in order to verify that the stabilization works for a truly anisotropic interpolation,

$$E_{\mathrm{a}} = \varepsilon \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.15 \\ 0.1 & 0.3 & 1 \end{pmatrix}.$$

In Figure 17a, we can see that HermiteGF-QR interpolation works stably even for very small values of ε for different N. On the other hand, for larger values of ε the result matches the direct anisotropic interpolation.

In order to validate the HermiteGF-QR method in higher dimensions against the existing methods, we compare the isotropic HermiteGF-QR method with the Gauss-QR method in 3-5D. For that, we fix the shape matrix E and the number of interpolation points N as

$$E = \mathrm{Id}_d \quad \text{and} \quad N = 4^d,$$

where d is the dimensionality. We choose the tolerance $\texttt{TOL} = 10^{-2}$ since it is enough to meet the overall accuracy of the method. One can see in Figure 17b that the HermiteGF-QR method matches the reference Gauss-QR method in 3-5D.

12.5. Leave-one-out cross-validation

Cross-validation is a technique for model selection based on a given data set [WR06, § 5.3]. It is widely used in statistics, in particular for statistical data fitting of big amounts of data in order to avoid overfitting. In the context of the isotropic RBF interpolation, we consider the shape parameter ε of the basis functions as the model parameter we try to find. Fasshauer & McCourt discussed in [FM15, § 14.2] how the RBF-QR methods can be also employed for stabilizing the cross-validation procedure in the range of small ε. Let us briefly recap the cross-validation algorithm for the RBF interpolation based on [FM15, § 14.2].

A general idea of the cross-validation is to split the existing data into a training set $\mathcal{T}$ and a validation set $\mathcal{V}$. The first set $\mathcal{T}$ is then used to build an interpolant. The validation set $\mathcal{V}$ is used to check how well the interpolant fits the points outside of the training set. We then accumulate the error in the validation set and use it as an indicator of the optimality of the shape parameter ε. If the split is performed only once and the validation set is small, the result might have a large variance. That is why one of the common choices is to use k-fold cross-validation. In this case, the data is split into k disjoint parts

$$\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_k, \quad \mathcal{X}_i \cap \mathcal{X}_j = \varnothing \quad \text{for} \quad i, j = 1 \ldots k.$$

Then, one subset is used as the validation set and the rest of the subsets are used as the training set

$$\mathcal{V}^{(i)} = \mathcal{X}_i, \quad \mathcal{T}^{(i)} = \mathcal{X} \setminus \mathcal{X}_i, \quad \text{for} \quad i = 1 \ldots k.$$

The procedure is repeated k times, so that each subset appeared as the validation set and all data were used in training, and the average error is used as the final result.

One of the common choices of the partitioning of the data is to use all points, except from one, as the training set and the remaining point as the validation set. The procedure is repeated N times, where N is the number of data points, so that every point in the set acts as a validation point once. This corresponds to the k-fold cross-validation with $k = 1$ and is called *leave-one-out cross-validation*, or LOO-CV. Let us now formalize the leave-one-out cross-validation procedure.

Recall the initial interpolation problem from section 4.3. Given a set $\{\phi_k(\cdot)\}_{k=1}^N$ of basis functions and the values $\{f_i\}$ of the function f at points $\{\mathbf{x}_i^{\text{col}}\}_{i=1}^N$ we seek to find an interpolant of the following form,

$$s(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j \phi_j(\mathbf{x}).$$

such that it satisfies the N collocation conditions,

$$s(\mathbf{x}_i^{\text{col}}) = f_i \quad \text{for} \quad i = 1, \ldots, N,$$

where we used center points of the Gaussians as collocation points. The straightforward approach is to find the coefficients $\{\alpha_j\}$ as a solution of the linear system,

$$\Phi^{\text{col}}\boldsymbol{\alpha} = \boldsymbol{f}, \quad \text{with} \quad \Phi^{\text{col}}_{ij} = \phi_j(\mathbf{x}^{\text{col}}_i).$$

The matrix Φ^{col} is called *collocation matrix*. Denote

$$\mathcal{X} = \{\mathbf{x}^{\text{col}}_i\}_{i=1}^N.$$

These are the data sites in which we have the data, or the value of the unknown function f. For the leave-one-out cross-validation, we set

$$\mathcal{V}^{(k)} = \mathbf{x}_k, \quad k = 1 \dots N$$

and the corresponding training sets

$$\mathcal{T}^{(k)} = \mathcal{X} \setminus \{\mathbf{x}_k\}.$$

Denote the $N-1 \times N-1$ matrix containing the values of the basis functions computed at $\mathcal{T}^{(k)}$ as Φ_k so that

$$(\Phi_k)_{ij} = \phi_j(\mathcal{T}^{(k)}_i).$$

Then, the leave-one-out cross-validation value can be computed as a sum of residuals at the validation set

$$C_{\text{CV}}(\varepsilon, \mathcal{V}) = \frac{1}{N}\sum_{k=1}^N \|f_k - \Phi_{\mathcal{T}^{(k)}}(\mathbf{x}_k)\Phi_k^{-1} f_{\mathcal{T}^{(k)}}\|_2,$$

where $\Phi_{\mathcal{T}^{(k)}}(\mathbf{x}_k) = (\phi_1(\mathbf{x}_k), \dots, \phi_{k-1}(\mathbf{x}_k), \phi_{k+1}(\mathbf{x}_k), \dots \phi_N(\mathbf{x}_k))$ and the vector $f_{\mathcal{T}^{(k)}}$ contains the values of f at the points from $\mathcal{T}^{(k)}$. Here we use a 2-norm, however in case of LOO-CV, it is equivalent to taking an absolute value of the difference since we only have one validation point. However, other norms can also be considered. It is clear that for small values of ε, the matrices Φ_k will be ill-conditioned. We recall from section 8.2 that after the HermiteGF-QR stabilization, the interpolant s can be computed via the stable basis Ψ as

$$s(\mathbf{x}) = \Psi(\mathbf{x})(\Psi^{\text{col}})^{-1}\boldsymbol{f} \quad \text{with} \quad \Psi^{\text{col}}_{ij} = \psi_j(\mathbf{x}^{\text{col}}_i).$$

Therefore, if we employ the HermiteGF-QR stabilization, our formula for the cross-validation transforms into

$$C_{\text{CV}}(\varepsilon, \mathcal{V}) = \sum_{k=1}^N \|f_k - \Psi_{\mathcal{T}^{(k)}}(\mathbf{x}_k)\Psi_k^{-1} f_{\mathcal{T}^{(k)}}\|_2,$$

where the matrices Ψ_k have been formed from the corresponding matrices Φ_k and $\Psi_{\mathcal{T}^{(k)}}(\mathbf{x}_k)$ is formed analogously to $\Phi_{\mathcal{T}^{(k)}}(\mathbf{x}_k)$. We seek to minimize the cumulative error in order to find the optimal ε.

We now check numerically how the HermiteGF-QR method performs in the context of cross-validation of a small artificial example. We build a data set sampling values from the function

$$f_{\text{c}}(x, y) = \cos(3\pi(x + y)).$$

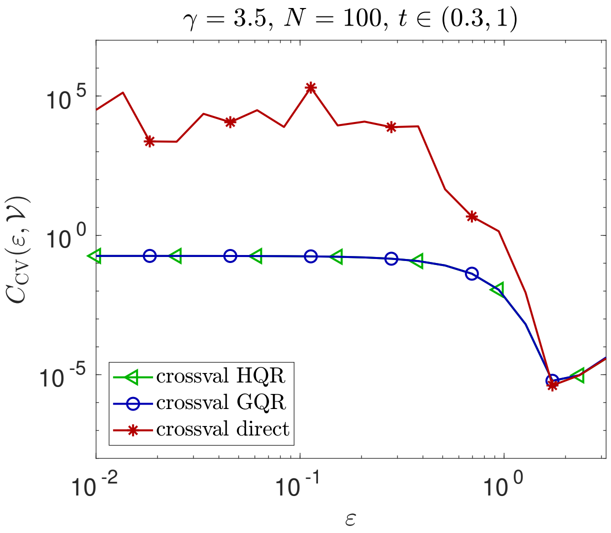


Figure 18 Leave-one-out cross-validation. HermiteGF-QR method performs equally well as the GaussQR method for stabilizing the cross-validation computations.

The cross-validation has been implemented in the framework of the GaussQR code from `http://math.iit.edu/~mccomic/gaussqr/` for a general interpolation procedure. We plug in our isotropic HermiteGF-QR method (with $E = \varepsilon \mathrm{Id}_2$) and look if the resulting cross-validation results agree with the ones produced by the GaussQR method and the RBF-Direct. One can see in Figure 18 that indeed, the values of $C_{\mathrm{CV}}(\varepsilon, \mathcal{V})$ agree for the HermiteGF-QR and Gauss-QR methods. We can also see that for larger ε, the values are also in tact with the ones obtained from the RBF-Direct. Even though in this case, the optimal value of ε lies within a range that is also reachable for the RBF-Direct, it does not have to be the case. With the help of the stabilization methods, we can access also small values of ε and extend the range of potential shape parameter values for cross-validation.

Part III

Generalized Fourier–Hermite method for the Vlasov equation

13. Vlasov–Poisson model

The Vlasov–Maxwell system describes the time evolution of charged particles in an electromagnetic field. An accurate and efficient numerical solution of the Vlasov–Maxwell system is one of the major problems in numerical plasma physics (see [FS03] for the overview of available methods). One of the topics of recent interest in the plasma physics community are *spectral* methods where the distribution function is expanded in series of basis functions of choice, and then, the resulting differential equations for the expansion coefficients are solved. Compared to the widely used *Particle-In-Cell (or PIC)* methods [BL04, HE88], which employ macroparticles that move throughout the computational mesh according Newton's equations, spectral methods are not prone to statistical noise. Spectral methods naturally introduce increased computational costs in exchange for improved accuracy. That is why, even though the first advancements in this area were made as early as 1963 [EFMO63], these methods became popular only recently due to the rapid increase of the computational power available. For most of the spectral methods, the Fourier basis is used in space due to its spectral nature and periodicity. For the velocity discretization, one of the most common choices are Hermite-type basis functions. Alternatively, Fourier [KF94], Chebyshev [SK74] or Legendre [MDVM16] bases could be used in velocity. However, in this thesis we focus on the Hermite-based methods.

Many test cases are electrostatic in nature and can be described by the simplified Vlasov–Poisson model. This model is frequently used for the derivation of new methods since it is slightly simpler in structure than the Vlasov–Maxwell one. In this part of the thesis, we derive a generalization of the existing Hermite-based spectral methods for the 1d1v Vlasov–Poisson system. We expect, however, that a similar approach could be used for a derivation of the method for the full Vlasov–Maxwell system, but it requires further investigation. Let us first set up the theoretical basis for the Vlasov–Poisson system that is relevant for the future discretization.

13.1. The model

The dimensionless 1d1v Vlasov–Poisson system for the distribution function $f : \mathbb{R}^2 \times \mathbb{R}^+ \to \mathbb{R}$ for electrons in neutralizing background reads as:

$$\begin{gathered}\frac{\partial f(x,v,t)}{\partial t} + v \cdot \nabla_x f(x,v,t) - E(x,t) \cdot \nabla_v f(x,v,t) = 0, \qquad (13.1)\\ -\Delta\phi(x,t) = 1 - \rho(x,t), \quad E(x,t) = -\nabla\phi(x,t),\\ x \in [0, L_x], L_x \in \mathbb{R}, v \in \mathbb{R}, t \in \mathbb{R}_+,\end{gathered}$$

where the electric field $E(x,t)$ is computed via the Poisson equation and the density is given by

$$\rho(x,t) = \int_{\mathbb{R}} f(x,v,t)\mathrm{d}v.$$

We consider the system on a periodic domain $[0, L_x]$ in x and the whole space $\mathbb{R}$ in velocity. The initial condition is given by:

$$f(x,v,0) = f_0(x,v), \qquad \int_{\mathbb{R}} \int_0^{L_x} f_0(x,v)\mathrm{d}x\mathrm{d}v = L_x.$$

Denote

$$L(x,v,t) = \begin{pmatrix} v \\ -E \end{pmatrix}.$$

Then the equation (13.1) reads as follows

$$\frac{\partial f}{\partial t} + L \cdot \nabla_{(x,v)} f = 0. \tag{13.2}$$

In the next sections we summarize a few properties of the Vlasov–Poisson model.

13.2. Characteristics

We start with defining the *characteristics* which are an essential tool for the proofs of the properties of the Vlasov equation.

Definition 13.2.1: Characteristics

Consider a differential system for a given $s \in \mathbb{R}_+$, $\mathbf{z} \in \mathbb{R}^2$

$$\frac{\mathrm{d}\mathbf{Z}}{\mathrm{d}t} = L(\mathbf{Z},t) \tag{13.3}$$
$$\mathbf{Z}(s) = \mathbf{z}.$$

Its solutions $\mathbf{Z}(\,\cdot\,; s, \mathbf{z})$ are called *characteristics* of the advection equation (13.2).

The following theorem formulates the main property of *characteristics* that we will use later [Son13, § 3.1, Theorem 2].

Theorem 13.2.1: Properties of characteristics

Consider the advection equation (13.2) with an initial distribution function $f(\mathbf{z},0) = f_0(\mathbf{z}) \in C^1(\mathbb{R}^2)$. Assume that $L \in C^{k-1}(\mathbb{R}^2 \times [0,T])$, $\nabla L \in C^{k-1}(\mathbb{R}^2 \times [0,T])$ for $k \geq 1$, and

$$|L(\mathbf{z},t)| \leq \kappa(1 + |\mathbf{z}|) \quad \forall t \in [0,T] \quad \forall \mathbf{z} \in \mathbb{R}^2.$$

Then there exists a unique solution $f(\mathbf{z}, t)$ of (13.1) and it is given by:

$$f(\mathbf{z}, t) = f_0(\mathbf{Z}(0; t, \mathbf{z})), \tag{13.4}$$

where Z corresponds to the characteristics associated to L.

Proof. See [Son13, § 3.1, Theorem 2]. □

The Theorem 13.2.1 implies that in order to find $f(\mathbf{z}, t)$, it is enough to solve the characteristics system (13.3) with an initial condition $\mathbf{Z}(t) = \mathbf{z}$ backwards in time and then evaluate $f_0(\mathbf{Z}(0; t, \mathbf{z}))$. We now look at another useful property of the characteristics.

Lemma 13.2.1

Let $J(t; s, 1) = \det(\nabla_z \mathbf{Z}(t; s, \mathbf{z}))$. Let L be an operator fulfilling the conditions of the previous theorem. Then

$$\frac{\partial J}{\partial t} = (\nabla \cdot L)(t; , \mathbf{Z}(t; s, \mathbf{z}))J \quad \text{and} \quad J > 0. \tag{13.5}$$

In case $\nabla \cdot L = 0$, then

$$J(t; s, 1) = J(s; s, 1) = 1. \tag{13.6}$$

Proof. See [BGP00, § 1.1, Proposition 1.1] □

13.3. Conservation properties

An important part of the simulation validation is monitoring the quantities that should be conserved over time. Although there are many quantities that are conserved for the Vlasov–Poisson system (see, for example, [Son13, § 3.2.2]), we are going to focus on the ones that we will employ for the simulation validation.

Theorem 13.3.1: Conservation properties of the Vlasov–Poisson model

The following conservation properties, among others, hold for the Vlasov–Poisson system:

1. Maximum principle:

$$0 \leq f(x, v, t) \leq \max_{(x,v)}\{f_0(x, v)\}.$$

2. Conservation of the L^p norm for $1 \leq p \leq \infty$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}} \int_0^{L_x} (f(x, v, t))^p \mathrm{d}x\mathrm{d}v = 0.$$

3. Conservation of volume:
$$\int_{\mathrm{Vol}}\int_0^{L_x} f(x,v,t)\mathrm{d}x\mathrm{d}v = \int_{F^{-1}(\mathrm{Vol})} f_0(y,u)\mathrm{d}y\mathrm{d}u.$$

4. Conservation of momentum:
$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}}\int_0^{L_x} vf(x,v,t)\mathrm{d}x\mathrm{d}v = 0.$$

5. Conservation of energy:
$$\frac{\mathrm{d}}{\mathrm{d}t}\left(W^{\mathrm{K}}(t)+W^{\mathrm{E}}(t)\right)=0,$$
where $W^{\mathrm{K}}(t)$ and $W^{\mathrm{E}}(t)$ are the kinetic and potential energy
$$W^{\mathrm{K}}(t)=\frac{1}{2}\int_{\mathbb{R}}\int_0^{L_x} v^2 f(x,v,t)\mathrm{d}x\mathrm{d}v,\quad W^{E}(t)=\frac{1}{2}\int_0^{L_x} E(x,t)^2\mathrm{d}x$$

Proof. Here we adapt the proof from [Son13, § 3.2.2, Proposition 5].

1. *Maximum principle.*

According to the Theorem 13.2.1, the solution of the Vlasov equation at time t can be written as follows:
$$f(x,v,t)=f_0(X(0;x,v,t),V(0;x,v,t)),$$
where $X(t;x,v,s),V(t;x,v,s)$ are characteristics associated with the Vlasov equation. Since f_0 is non-negative:
$$0\le f(x,v,t)\le \max_{(x,v)}\{f_0(x,v)\}.$$

2. *Conservation of L^p-norm.*
$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}}\int_0^{L_x}(f(x,v,t))^p\mathrm{d}x\mathrm{d}v=\int_{\mathbb{R}}\int_0^{L_x} pf^{p-1}(x,v,t)\frac{\partial f(x,v,t)}{\partial t}\mathrm{d}x\mathrm{d}v$$
$$=\int_{\mathbb{R}}\int_0^{L_x} pf^{p-1}(x,v,t)\left(E(x,t)\cdot\nabla_v f(x,v,t)-v\cdot\nabla_x f(x,v,t)\right)\mathrm{d}x\mathrm{d}v.$$
Integrating by parts in v we note that
$$\int_{\mathbb{R}} f^{p-1}(x,v,t)\left(E(x,t)\cdot\nabla_v f(x,v,t)\right)\mathrm{d}v$$
$$=-E(x,t)\int_{\mathbb{R}} f(x,v,t)(p-1)f(x,v,t)^{p-2}\nabla_v f(x,v,t)\mathrm{d}v.$$
Therefore,
$$\int_{\mathbb{R}} pf^{p-1}(x,v,t)\left(E(x,t)\cdot\nabla_v f(x,v,t)\right)\mathrm{d}v=0.$$

Analogously, integrating by parts in x gives

$$\int_0^{L_x} f^{p-1}(x,v,t)v{\cdot}\nabla_x f(x,v,t)\mathrm{d}x$$
$$= -v\int_0^{L_x} f(x,v,t)(p-1)f(x,v,t)^{p-2}\nabla_x f(x,v,t)\mathrm{d}x,$$

where we used the periodicity in x. Hence,

$$\int_{\mathbb{R}} pf^{p-1}(x,v,t)\left(v\cdot\nabla_x f(x,v,t)\right)\mathrm{d}x = 0$$

which proves (2) for $1 \leq p < \infty$. The result for $p = \infty$ follows from the maximum principle.

3. *Conservation of volume.*

   Using the Theorem 13.2.1 we get:

   $$\int_{\mathrm{Vol}} f(x,v,t)\mathrm{d}x\mathrm{d}v = \int_{\mathrm{Vol}} f_0(X(0;x,v,t),V(0;x,v,t))\mathrm{d}x\mathrm{d}v.$$

   Let us make a change of variable:

   $$\begin{pmatrix} y \\ u \end{pmatrix} = F(x,v) = \begin{pmatrix} X(0;t,x,v) \\ V(0;t,x,v) \end{pmatrix}.$$

   Since

   $$\nabla_{(x,v)}\cdot\begin{pmatrix} v \\ -E \end{pmatrix} = 0,$$

   according to the Lemma 13.2.1

   $$\det\left(\nabla_{(x,v)}\begin{pmatrix} X(0;x,v,t) \\ V(0;x,v,t) \end{pmatrix}\right) = 1.$$

   Therefore,

   $$\int_{\mathrm{Vol}} f(x,v,t)\mathrm{d}x\mathrm{d}v = \int_{F^{-1}(\mathrm{Vol})} f_0(y,u)\mathrm{d}y\mathrm{d}u.$$

4. The proofs of the conservation of momentum and energy are a bit more lengthy, so we will not reproduce them here. One can find the full proofs in [Son13, § 3.2.2, Proposition 5].

□

14. Generally weighted Hermite method

Hermite-type basis functions that are used for the discretization of the Vlasov equation are functions of the type Hermite polynomial times Gaussian. These functions have been of interest already in early 1d1v Vlasov–Poisson simulations [Arm67, AM67, GF67, JKM71]. Due to the lack of computational resources at the time, these ideas stayed on a rather theoretical level until the late 1990s.

The reignited interest in Hermite-based methods started in 1996, when Holloway proposed in [Hol96] two possible velocity discretizations of the Vlasov equation based on Hermite polynomials. The first approach, that seems natural, is to use Hermite functions

$$\psi_\ell(v) = \frac{\pi^{-1/4}}{\sqrt{2^\ell \ell!}} h_\ell(v) \mathrm{e}^{-v^2/2},$$

as the basis in velocity. Then, the standard Galerkin method was used with Hermite functions as the test functions. This method is called *symmetrically-weighted (SW) Hermite method*. However, it turned out that for this method, mass and momentum cannot be conserved simultaneously. To overcome that pitfall, an alternative basis was proposed. In particular, so-called *asymmetrically-weighted Hermite basis*, or AW basis

$$\psi^{\mathrm{a}}(v) = \frac{\pi^{-1/2}}{\sqrt{2^\ell \ell!}} h_\ell(v) \mathrm{e}^{-v^2}.$$

In this case, in order to preserve orthogonality, another set of functions was used as test functions. In particular, scaled Hermite polynomials were utilized

$$\psi^{\mathrm{a}}_{\mathrm{test}}(v) = \frac{1}{\sqrt{2^\ell \ell!}} h_\ell(v).$$

For this method, it turned out to be possible to conserve mass, momentum, and energy exactly. For both methods scaling of the argument of the basis functions was considered. Certain choices of the scaling parameter proved to provide significant improvements in the quality of the results. It is consistent with the result of Boyd [Boy84] that the scaling of the series of Hermite functions is beneficial for the accuracy. From then on, in most of the practical applications the scaling was included in the bases in order to gain additional accuracy. It is often implied that the scaling is included, when AW or SW method is considered. In this thesis, we also use this general notion and always include the scaling in the AW and SW basis by default.

In the follow up work [SH98], Schumer & Holloway carried out a thorough numerical study of both methods which indicated even though the AW method preserves mass, momentum and energy, the SW method is more robust and better suited for long-term simulations. Despite that, most of the further developments were focusing on the AW method. In [LBDVJ06, LB07] both spectral Galerkin and spectral collocation methods for the AW Fourier–Hermite discretization have been considered. Around the same time, the Hermite-based solution of the linearized Vlasov–Maxwell model has been investigated in [CDLD06]. A multi-dimensional spectral Vlasov–Maxwell solver based on the AW discretization has been proposed by Delzanno in 2015 in [Del15]. The work [CDBM16] by Camporeale, Delzanno, Bergen, & Moulton demonstrated that for certain test cases the AW Fourier–Hermite method can be significantly more accurate than the PIC method. The spectral solver has been further enhanced by adding an adaptive strategy for regulating the number of basis functions [VDJ$^+$15] and yielded the `SpectralPlasmaSolver` code [VDM$^+$16]. At the same time, the AW Fourier–Hermite method was also considered in the gyrokinetics framework [PD15].

Even though the SW method is considerably less popular, Gibelli & Shizgal studied the convergence of the expansion of the distribution functions via Hermite functions in [GS06]. In 2017, convergence theory for the SW Fourier–Hermite method was provided in [MFD17]. This study is, however, limited to the finite velocity interval.

In this thesis, we derive a generalization of the above described two methods. Consider the following basis functions

$$H_\ell^{\gamma,\varepsilon}(v) = \frac{1}{\sqrt{2^\ell \ell!}} h_\ell(\gamma v) \mathrm{e}^{-\varepsilon^2 v^2}.$$

These are the generalized anisotropic Hermite functions, that we introduced in part I, with the truncation parameter $t_{\mathrm{HGF}} = 1$. Here we use t_{HGF} instead of the truncation parameter t of the anisotropic Hermite basis to avoid the confusion with the time variable t. For the rest of this thesis we will only consider

$$t_{\mathrm{HGF}} = 1.$$

We note that both SW and AW bases are special cases of the basis $\{H_\ell^{\gamma,\varepsilon}\}_{\ell\in\mathbb{N}_0}$ with

$$\begin{cases} \gamma = \varepsilon\sqrt{2} & \text{for the SW method} \\ \gamma = \varepsilon & \text{for the AW method} \end{cases}$$

up to normalization constants $\pi^{-1/4}$, $\pi^{-1/2}$. In this chapter, we derive the general method for arbitrary γ and ε. The first steps in this direction were already made when considering the scaling of the argument of the SW and AW bases. However, in our case, the scaling of the exponent and of the Hermite polynomial are decoupled. This yields a family of Hermite-based spectral methods, or *generally weighted (GW) methods*, that includes AW and SW methods as special cases. Moreover, we later consider mass, momentum, and energy for the general method. We identify the cases when the observables are conserved. We also derive explicit formulas for the deviation of observables in other cases.

14.1. Generalized Hermite basis revisited

In this section, we briefly review the properties of the generalized anisotropic Hermite basis that are relevant for the 1d1v Vlasov–Poisson setup. We also look into the properties of the averages of the basis functions which will be useful later on, when we take a look at the observables. We start by recalling that the basis functions read for all $\ell \in \mathbb{N}_0$ as

$$H_\ell^{\gamma,\varepsilon}(v) = \frac{1}{\sqrt{2^\ell \ell!}} h_\ell(\gamma v) \mathrm{e}^{-\varepsilon^2 v^2}.$$

In this case, the associated weight takes the form

$$\omega(v) = \pi^{-1/2} \gamma \mathrm{e}^{(2\varepsilon^2 - \gamma^2) v^2}$$

and the basis functions are *orthonormal* in the corresponding weighted L^2 space. Indeed, writing down the result of the Lemma 3.2.1, we get

$$\langle H_{\ell_1}^{\gamma,\varepsilon,t_{\mathrm{HGF}}}, H_{\ell_2}^{\gamma,\varepsilon,t_{\mathrm{HGF}}} \rangle_\omega = t_{\mathrm{HGF}}^{(\ell_1+\ell_2)} \delta_{\ell_1,\ell_2} \overset{t_{\mathrm{HGF}}=1}{=} \delta_{\ell_1,\ell_2} = \langle H_{\ell_1}^{\gamma,\varepsilon}, H_{\ell_2}^{\gamma,\varepsilon} \rangle_\omega.$$

Recall the algebraic properties of the generalized anisotropic Hermite basis. In particular, properties (3.6), (3.7) in 1D take the form

Recursion relations of 1D generalized Hermite functions

1. $$\frac{\partial H_\ell^{\gamma,\varepsilon}(v)}{\partial v} = \frac{\sqrt{2}}{\gamma} \left((\gamma^2 - \varepsilon^2) \sqrt{\ell} H_{\ell-1}^{\gamma,\varepsilon}(v) - \varepsilon^2 \sqrt{\ell+1} H_{\ell+1}^{\gamma,\varepsilon}(v) \right). \qquad (14.1)$$

2. $$v H_\ell^{\gamma,\varepsilon}(v) = \frac{1}{\sqrt{2}\gamma} \left(\sqrt{\ell} H_{\ell-1}^{\gamma,\varepsilon}(v) + \sqrt{\ell+1} H_{\ell+1}^{\gamma,\varepsilon}(v) \right). \qquad (14.2)$$

Let us now derive a couple of other properties that involve integrals of the basis function over $\mathbb{R}$. These formulas will be useful for the computation of the observables. Analogous formulas for the SW and AW cases were considered in [Hol96], [SH98]. We now derive them for the generalized setup.

Lemma 14.1.1

Denote

$$I_\ell = \int_{\mathbb{R}} H_\ell^{\gamma,\varepsilon}(v) \mathrm{d}v, \quad J_\ell = \int_{\mathbb{R}} v H_\ell^{\gamma,\varepsilon}(v) \mathrm{d}v, \quad \bar{I}_\ell = \int_{\mathbb{R}} v^2 H_\ell^{\gamma,\varepsilon}(v) \mathrm{d}v.$$

Then,

1. The following relation holds for even integer $\ell \geq 2$:

$$\frac{I_{\ell+2}}{I_\ell} = \sqrt{\frac{\ell+1}{\ell+2}} \left(\frac{\gamma^2}{\varepsilon^2} - 1 \right) \qquad (14.3)$$

with $I_0 = \frac{\sqrt{\pi}}{\varepsilon}$. Moreover, $I_\ell = 0$ for odd $\ell \in \mathbb{N}_0$.

2. The following relation holds for odd $\ell \geq 3$:

$$J_{\ell+1} = \frac{\gamma}{\varepsilon^2}\sqrt{\frac{\ell+1}{2}} I_\ell = \sqrt{\frac{\ell+1}{\ell}} \left(\frac{\gamma^2}{\varepsilon^2} - 1\right) J_{\ell-1} \tag{14.4}$$

and $J_1 = \frac{\gamma}{\varepsilon^3}\sqrt{\frac{\pi}{2}} = \frac{\gamma}{\varepsilon^2}\sqrt{\frac{1}{2}} I_0$. Moreover, $J_\ell = 0$ for even ℓ.

3. The following relation holds for even $\ell \geq 2$:

$$\bar{I}_\ell = \frac{1}{\gamma\sqrt{2}} \left(\frac{\ell\varepsilon^2 + (\ell+1)(\gamma^2 - \varepsilon^2)}{\varepsilon^2\sqrt{\ell}}\right) J_{\ell-1}. \tag{14.5}$$

and $\bar{I}_0 = \frac{1}{2}\frac{\sqrt{\pi}}{\varepsilon^3}$. Moreover, $\bar{I}_\ell = 0$ for odd ℓ.

Proof. We first prove the relation (14.3). To start, let us compute I_0 and I_1.

$$I_0 = \int_{\mathbb{R}} H_0^{\gamma,\varepsilon}(v)\mathrm{d}v = \int_{\mathbb{R}} \mathrm{e}^{-\varepsilon^2 v^2}\mathrm{d}v = \frac{\sqrt{\pi}}{\varepsilon}.$$

As for the I_1, we get

$$I_1 = \int_{\mathbb{R}} \sqrt{2}\gamma v \mathrm{e}^{-\varepsilon^2 v^2}\mathrm{d}v = 0.$$

The same reasoning can be applied to all odd $\ell \in \mathbb{R}$ since odd degree Hermite polynomials are formed from odd degree monomials. We now move on to compute I_ℓ for even $\ell \in \mathbb{N}_0$. The following relation holds for the product of an even Hermite polynomial and a Gaussian [AS64, Expr. 22.13.17]:

$$\int_{\mathbb{R}} \mathrm{e}^{-a^2} H_{2m}(ax)\mathrm{d}a = \sqrt{\pi}\frac{(2m)!}{m!}(x^2-1)^m \quad \text{for all} \quad x \in \mathbb{R},\ m \in \mathbb{N}_0.$$

Then,

$$\begin{aligned} I_{2\ell} &= \frac{1}{\sqrt{2^{2\ell}(2\ell)!}} \int_{\mathbb{R}} h_{2\ell}(\gamma v)\mathrm{e}^{-\varepsilon^2 v^2}\mathrm{d}v \overset{\tilde{v}=\varepsilon v}{=} \frac{1}{\varepsilon\sqrt{2^{2\ell}(2\ell)!}} \int_{\mathbb{R}} h_{2\ell}\left(\frac{\gamma}{\varepsilon}\tilde{v}\right)\mathrm{e}^{-\tilde{v}^2}\mathrm{d}\tilde{v} \\ &= \frac{\sqrt{\pi}}{\varepsilon\sqrt{2^{2\ell}(2\ell)!}}\frac{(2\ell)!}{\ell!}\left(\frac{\gamma^2}{\varepsilon^2}-1\right)^\ell. \end{aligned}$$

Analogously, we compute $I_{2\ell+2}$

$$I_{2\ell+2} = \frac{\sqrt{\pi}}{\varepsilon\sqrt{2^{(2\ell+2)}(2\ell+2)!}}\frac{(2\ell+2)!}{(\ell+1)!}\left(\frac{\gamma^2}{\varepsilon^2}-1\right)^{\ell+1}.$$

Therefore,

$$\frac{I_{2\ell+2}}{I_{2\ell}} = \sqrt{\frac{2\ell+1}{2\ell+2}}\left(\frac{\gamma^2}{\varepsilon^2}-1\right)$$

which proves (14.3).

We start proving the relation (14.4) by observing that $J_0 = 0$ since

$$J_0 = \int_{\mathbb{R}} vH_0^{\gamma,\varepsilon,t}(v)\mathrm{d}v = \int_{\mathbb{R}} v\mathrm{e}^{-\varepsilon^2 v^2} = 0.$$

Let us now compute J_1. Using [GR14, 3.381, Expr. 11], we get

$$J_1 = \int_{\mathbb{R}} \frac{v}{\sqrt{2}} H_1^{\gamma,\varepsilon,t}(v)\mathrm{e}^{-\varepsilon^2 v^2}\mathrm{d}v = \sqrt{2}\int_{\mathbb{R}} \gamma v^2 \mathrm{e}^{-\varepsilon^2 v^2}\mathrm{d}v = \frac{\gamma}{\varepsilon^3}\sqrt{\frac{\pi}{2}} = \frac{\gamma}{\varepsilon^2\sqrt{2}} I_0.$$

To prove the first equality in the expression (14.4), it is enough to use the property (14.2) of generalized Hermite functions and the expression (14.3). Indeed,

$$J_{\ell+1} = \int_{\mathbb{R}} v H_\ell^{\gamma,\varepsilon}(v) \mathrm{d}v \overset{(14.2)}{=} \frac{1}{\gamma}\left(\sqrt{\frac{\ell+1}{2}} I_\ell + \sqrt{\frac{\ell+2}{2}} I_{\ell+2}\right)$$
$$\overset{(14.3)}{=} \frac{1}{\gamma}\left(\sqrt{\frac{\ell+1}{2}} I_\ell + \sqrt{\frac{\ell+2}{2}}\sqrt{\frac{\ell+1}{\ell+2}}\left(\frac{\gamma^2}{\varepsilon^2}-1\right) I_\ell\right) = \frac{\gamma}{\varepsilon^2}\sqrt{\frac{\ell+1}{2}} I_\ell.$$

Using this formula, we observe that, for $\ell \geq 2$

$$J_{\ell-1} = \frac{\gamma}{\varepsilon^2}\sqrt{\frac{\ell-1}{2}} I_{\ell-2} \overset{(14.3)}{=} \frac{\gamma}{\varepsilon^2}\sqrt{\frac{\ell-1}{2}}\frac{\varepsilon^2}{\gamma^2-\varepsilon^2}\sqrt{\frac{\ell}{\ell-1}} I_\ell = \frac{\gamma}{\gamma^2-\varepsilon^2}\sqrt{\frac{\ell}{2}} I_\ell.$$

Therefore,

$$J_{\ell+1} = \sqrt{\frac{\ell+1}{\ell}}\left(\frac{\gamma^2}{\varepsilon^2}-1\right) J_{\ell-1},$$

which completes the proof of the second equality in (14.4). This expression, together with the property (14.2) of generalized Hermite functions yields the final relation (14.5) for $\ell \neq 0$

$$\bar{I}_\ell = \int_{\mathbb{R}} v^2 H_\ell^{\gamma,\varepsilon}(v)\mathrm{d}v \overset{(14.2)}{=} \frac{1}{\gamma\sqrt{2}}\int_{\mathbb{R}} v\left(\sqrt{\ell} H_{\ell-1}^{\gamma,\varepsilon}(v) + \sqrt{\ell+1} H_{\ell+1}^{\gamma,\varepsilon}(v)\right)\mathrm{d}v$$
$$= \frac{1}{\gamma\sqrt{2}}\left(\sqrt{\ell} J_{\ell-1} + \sqrt{\ell+1} J_{\ell+1}\right) \overset{(14.4)}{=} \frac{1}{\gamma\sqrt{2}}\left(\sqrt{\ell} J_{\ell-1} + \frac{\ell+1}{\sqrt{\ell}}\left(\frac{\gamma^2}{\varepsilon^2}-1\right) J_{\ell-1}\right)$$
$$= \frac{1}{\gamma\sqrt{2}}\left(\frac{\ell\varepsilon^2 + (\ell+1)(\gamma^2-\varepsilon^2)}{\varepsilon^2\sqrt{\ell}}\right) J_{\ell-1}.$$

For $\ell = 0$ we use [GR14, 3.381, Expr. 11] and get

$$\bar{I}_0 = \int_{\mathbb{R}} v^2 \varepsilon^{-\varepsilon^2 v^2}\mathrm{d}v = \frac{1}{2}\frac{\sqrt{\pi}}{\varepsilon^3}. \qquad \square$$

14.2. Discretization in velocity

We now proceed to the discretization of the Vlasov equation in velocity. We look for an approximation f_{N_v} of the distribution function f in an approximation space

$$V_{N_v} = \text{span}\{H_{\ell_1}^{\gamma,\varepsilon}, H_{\ell_2}^{\gamma,\varepsilon}, \ldots, H_{\ell_{N_v-1}}^{\gamma,\varepsilon}\}$$

of generalized Hermite functions

$$f_{N_v}(x,v,t) = \sum_{\ell=0}^{N_v-1} c_\ell(x,t) H_\ell^{\gamma,\varepsilon}(v). \tag{14.6}$$

In order to treat the coefficients with the indexes outside of the range $0,\ldots,N_v-1$, we need a closure scheme. The most obvious choice is to assign them to zero

$$c_\ell(x,t) = 0, \quad \text{for all,} \quad \ell \notin 0,\ldots,N_v-1,\ x \in \mathbb{R},\ t \in \mathbb{R}_+.$$

This is the most common closure scheme but the alternatives can e.g. be found in [EFMO63].

We use the Galerkin method in the weighted space $L^2_\omega(\mathbb{R})$. In particular, we demand that at each time t and fixed point x in space, the approximate distribution function $f_{N_v}(x,v,t) \in V_{N_v}$ must satisfy:

$$\frac{\partial f_{N_v}}{\partial t} \in V_{N_v} \quad \text{and} \quad \left\langle \frac{\partial f_{N_v}}{\partial t} + v \cdot \nabla_x f_{N_v} - E(x,t) \cdot \nabla_v f_{N_v}, \phi \right\rangle_\omega = 0 \quad \forall \phi \in V_{N_v}. \tag{14.7}$$

Since $\{H_\ell^{\gamma,\varepsilon}\}_{\ell=0}^{N_v-1}$ is a basis of V_{N_v}, it is enough to demand that (14.7) is fulfilled for basis functions $\{H_\ell^{\gamma,\varepsilon}\}_{\ell=0}^{N_v}$.

We first insert the expression (14.6) into the Vlasov equation (13.1):

$$\sum_{\ell=0}^{N_v-1} \frac{\partial c_\ell(x,t)}{\partial t} H_\ell^{\gamma,\varepsilon}(v) + v \sum_{\ell=0}^{N_v-1} \frac{\partial c_\ell(x,t)}{\partial x} H_\ell^{\gamma,\varepsilon}(v) - E(x,t) \sum_{\ell=0}^{N_v-1} c_\ell(x,t) \frac{\partial H_\ell^{\gamma,\varepsilon}(v)}{\partial v} = 0.$$

Using the relations (14.1) and (14.2) we get:

$$\begin{aligned}\sum_{\ell=0}^{N_v-1} \frac{\partial c_\ell(x,t)}{\partial t} H_\ell^{\gamma,\varepsilon}(v) + \sum_{\ell=0}^{N_v-1} \frac{\partial c_\ell(x,t)}{\partial x} \frac{1}{\sqrt{2}\gamma} \left(\sqrt{\ell} H_{\ell-1}^{\gamma,\varepsilon}(v) + \sqrt{\ell+1} H_{\ell+1}^{\gamma,\varepsilon}(v) \right) \qquad (14.8)\\ - E(x,t) \sum_{\ell=0}^{N_v-1} c_\ell(x,t) \frac{\sqrt{2}}{\gamma} \left((\gamma^2 - \varepsilon^2) \sqrt{\ell} H_{\ell-1}^{\gamma,\varepsilon}(v) - \varepsilon^2 \sqrt{\ell+1} H_{\ell+1}^{\gamma,\varepsilon}(v) \right) = 0\end{aligned}$$

With the use of the expression (14.8), the Galerkin condition (14.7) transforms into:

Generalized Hermite velocity discretization

Using the orthonormality of the generalized anisotropic Hermite functions basis with $t_{\mathrm{HGF}} = 1$, we get the following discretization in velocity

$$\begin{aligned}\frac{\partial c_\ell(x,t)}{\partial t} + \frac{1}{\gamma\sqrt{2}} \left(\sqrt{\ell+1} \frac{\partial c_{\ell+1}(x,t)}{\partial x} + \sqrt{\ell} \frac{\partial c_{\ell-1}(x,t)}{\partial x} \right) \qquad (14.9)\\ - \frac{\sqrt{2}E(x,t)}{\gamma} \left((\gamma^2 - \varepsilon^2) \sqrt{\ell+1} c_{\ell+1}(x,t) - \varepsilon^2 \sqrt{\ell} c_{\ell-1}(x,t) \right) = 0,\end{aligned}$$

where $\ell = 0, \dots, N_v - 1$.

14.3. Discretization in space

We now discretize our system in space via the Fourier basis, which is a standard choice for Vlasov spectral methods. In particular, for all $\ell = 0 \dots N_v - 1$, we take the following ansatz:

$$c_\ell(x,t) = \sum_{k=-N_x}^{N_x} c_\ell^k(t) \exp\left(\frac{2\pi k i x}{L_x} \right). \tag{14.10}$$

For the convenience of notation, we consider only the odd number of Fourier modes. For the even number of modes, when $k = -N_x, \dots, N_x + 1$, the coefficients corresponding to the indexes 0 and $N_x + 1$ have to be tackled separately to preserve the symmetricity.

The electric field E can be represented in the Fourier basis as

$$E(x,t) = \sum_{m=-N_x}^{N_x} E_m(t) \exp\left(\frac{2\pi m\mathrm{i}x}{L_x}\right).$$

We assume for now that the coefficients E_m are already known. Before we proceed to the discretization of the Vlasov equation in space, let us compute the product $E(x,t)c_\ell(x,t)$ in terms of the corresponding Fourier coefficients. We get

$$\begin{aligned} E(x,t)c_\ell(x,t) &= \sum_{m=-N_x}^{N_x} E_m(t) \exp\left(\frac{2\pi m\mathrm{i}x}{L_x}\right) \sum_{k=-N_x}^{N_x} c_\ell^k(t) \exp\left(\frac{2\pi k\mathrm{i}x}{L_x}\right) \\ &= \sum_{m=-N_x}^{N_x} \sum_{k=-N_x}^{N_x} E_m(t)c_\ell^k(t) \exp\left(\frac{2\pi(m+k)\mathrm{i}x}{L_x}\right) \\ &= \sum_{p=-2N_x}^{2N_x} \exp\left(\frac{2\pi p\mathrm{i}x}{L_x}\right) \sum_{k=-N_x}^{N_x} E_{p-k}(t)c_\ell^k(t) \\ &= \sum_{p=-2N_x}^{2N_x} \exp\left(\frac{2\pi p\mathrm{i}x}{L_x}\right) [\mathbf{E}(t) * \mathbf{c}_\ell(t)][p], \end{aligned} \tag{14.11}$$

where $\mathbf{E}(t)$ is the vector of the coefficients $\{E_m(t)\}_{m=-N_x}^{N_x}$, $\mathbf{c}_\ell(t)$ is the vector of coefficients $\{c_\ell^k(t)\}_{k=-N_x}^{N_x}$ and the vector $\mathbf{E}(t)$ is additionally padded with zeros for all other indexes. Here $*$ denotes the convolution

$$[\mathbf{E}(t) * \mathbf{c}_\ell(t)][p] = \sum_{j=-N_x}^{N_x} E_{p-j}(t)c_\ell^j(t).$$

We are now ready to proceed to the discretization of the Vlasov equation in space. We plug in the ansatz (14.10) in the velocity discretization (14.9) and use the Galerkin method. Multiplying both sides by $\exp\left(-\frac{2\pi k'\mathrm{i}x}{L_x}\right)$ and integrating from 0 to L_x, we get the Galerkin condition in space

Generalized Fourier–Hermite (GW) discretization

$$\begin{aligned} &\frac{\partial c_\ell^k(t)}{\partial t} + \frac{1}{\gamma\sqrt{2}} \frac{2\pi k\mathrm{i}}{L_x} \left(\sqrt{\ell+1}c_{\ell+1}^k(t) + \sqrt{\ell}c_{\ell-1}^k(t)\right) \\ &- \frac{\sqrt{2}}{\gamma}\left((\gamma^2-\varepsilon^2)\sqrt{\ell+1}[\mathbf{E}(t) * \mathbf{c}_{\ell+1}(t)][k] - \varepsilon^2\sqrt{\ell}[\mathbf{E}(t) * \mathbf{c}_{\ell-1}(t)][k]\right) = 0 \end{aligned} \tag{14.12}$$

for all $\gamma, \varepsilon > 0$, $k = -N_x \dots N_x$, $\ell = 0 \dots N_v - 1$, $t \in \mathbb{R}_+$.

We will later refer to the discretization (14.12) as the *generalized Fourier–Hermite discretization* or the *generally weighted (GW) Hermite method*.

For the shortness of notation, denote

$$\beta_\ell^k(t) = [\mathbf{E}(t) * \mathbf{c}_\ell(t)][k]. \tag{14.13}$$

14.4. Computation of the electric field

To complete the formulation of the method, it is now left to compute the coefficients $\{E_m(t)\}_{m=-N_x}^{N_x}$ of the electric field E in the Fourier basis. We now proceed to the computation of the electric field $E = -\nabla\phi$ from the Poisson equation for the potential

$$-\Delta\phi(x,t) = 1 - \rho(x,t) \quad \text{with} \quad \rho(x,t) = \int_{\mathbb{R}} f(x,v,t)\mathrm{d}v \tag{14.14}$$

with $x \in [0, L_x]$, $v \in \mathbb{R}$, $t \in \mathbb{R}_+$.

Let us first compute the density ρ. Plugging in the discretization (14.6) of f in velocity, we get

$$\rho(x,t) = \int_{\mathbb{R}} f(x,v,t)\mathrm{d}v \approx \int_{\mathbb{R}} \sum_{\ell=0}^{N_v-1} c_\ell(x,t) H_\ell^{\gamma,\varepsilon}(v)\mathrm{d}v = \sum_{\ell=0}^{N_v-1} c_\ell(x,t) I_\ell,$$

where I_ℓ is computed by the recursion (14.3). We now consider the Fourier ansatz in space for both coefficients $\{c_\ell(x,t)\}_{\ell=0}^{N_v-1}$ and $\phi(x,t)$:

$$c_\ell(x,t) = \sum_{k=-N_x}^{N_x} c_\ell^k(t) \exp\left(\frac{2\pi k \mathrm{i} x}{L_x}\right), \quad \phi(x,t) = \sum_{k=-N_x}^{N_x} \phi_k(t) \exp\left(\frac{2\pi k \mathrm{i} x}{L_x}\right).$$

Inserting the ansatz into the Poisson equation (14.14) and imposing Galerkin conditions by multiplying both sides by $\exp\left(\frac{-2\pi k' \mathrm{i} x}{L_x}\right)$ and integrating from 0 to L_x, we get for all $k' \neq 0$

$$\frac{4\pi^2 k'^2}{L_x^2} \phi_{k'}(t) = -\sum_{\ell=0}^{N_v-1} c_\ell^{k'}(t) I_\ell.$$

Therefore, for $k' \neq 0$ the coefficients can be computed as

$$\phi_{k'}(t) = -\frac{L_x^2}{4\pi^2 k'^2} \sum_{\ell=0}^{N_v-1} c_\ell^{k'}(t) I_\ell.$$

For $k' = 0$, we get

$$0 = L_x - L_x \sum_{\ell=0}^{N_v-1} c_\ell^0(t) I_\ell.$$

This corresponds to the mass conservation, however, it leaves ϕ_0 undefined. Therefore, let us impose

$$\phi_0(t) = 0.$$

We can now compute the coefficients of $E(x,t)$.

Electric field computation

$$E_k(t) = \frac{\mathrm{i} L_x}{2\pi k} \sum_{\ell=0}^{N_v-1} c_\ell^k(t) I_\ell \quad \text{for} \quad k \neq 0; \quad E_0 = 0. \tag{14.15}$$

For the complete setup of the method, we need to discuss the representation of the initial distribution function in the generalized Fourier–Hermite basis.

14.5. Initial distribution representation in the Fourier–Hermite basis

A large number of initial distributions typically considered for the Vlasov model has the following form

$$f_0(x,v) = (1+\alpha\cos(mk_0x))\sum_{n=1}^{N_\mathrm{G}} \frac{a_n}{\sigma_n\sqrt{2\pi}}\mathrm{e}^{-\frac{(v-v_n)^2}{2\sigma_n^2}}, \quad x\in\left[0,\frac{2\pi}{k_0}\right], \, k_0,\alpha\in\mathbb{R},\, m\in\mathbb{N},$$

where N_G is the number of the Gaussians, $a_n\in\mathbb{R}$ are constants, $\sigma_n\in\mathbb{R}$ control the width of the Gaussians, $\{v_n\}_{n=1}^{N_\mathrm{G}}$ are centers of the Gaussians. Since the variables x and v are decoupled for this type of functions, we can compute the representation of the trigonometric functions in Fourier basis separately from the representation of the sum of Gaussians in the generalized Hermite basis. Once we have the corresponding coefficients, we can get the coefficients of the full initial distribution representation as a tensor product of the two sets.

Representation of trigonometric functions in Fourier basis is a trivial task and can be usually done exactly. Indeed, we need to find $\{c^k\}$ such that

$$(1+\alpha\cos(mk_0x)) = \sum_{k=-N_x}^{N_x} c^k\exp\left(k_0k\mathrm{i}x\right) = \sum_{k=-N_x}^{N_x} c^k\left(\cos(k_0kx)+\mathrm{i}\sin(k_0kx)\right)$$

If we take

$$c^0=1, \quad c^m=c^{-m}=\frac{\alpha}{2}$$

and set all other indexes to zero, we will get the exact initial representation of the function $(1+\alpha\cos(mk_0x))$ in the Fourier basis. For the case when the trigonometric function consists from a sum of multiple $\sin$ and $\cos$ functions, one can also proceed in similar fashion.

Representation of the sum of Gaussians in the generalized Hermite basis can be straightforward as well, in case $N_\mathrm{G}=1$ and $v_1=0$. In this case, there is only one Gaussian in the sum, that is centered at zero. Therefore, it is possible to represent it exactly by adjusting the width ε of the Gaussian in the generalized Hermite basis accordingly. Indeed, if we take

$$\varepsilon = \frac{1}{\sigma_1\sqrt{2}}$$

in the generalized Fourier–Hermite basis, then

$$H_0^{\gamma,\varepsilon}(v) = \mathrm{e}^{-\frac{v^2}{2\sigma_1^2}}$$

and we just need to take

$$c_0 = a_1,$$

and set all other coefficients to 0 to get an exact representation of the initial distribution in velocity. This is, for example, the case for the Landau damping test case (see Section 18.1).

In other cases, one needs to do a Galerkin projection of the initial distribution onto the corresponding $L^2_\omega(\mathbb{R})$ space. In particular, we need to find the coefficients $\{c_\ell\}_{\ell=0}^{N_v-1}$ such that

$$\left\langle f_0^v - \sum_{\ell=0}^{N_v-1} c_\ell H_\ell^{\gamma,\varepsilon}, H_\ell^{\gamma,\varepsilon} \right\rangle_\omega = 0, \quad \text{for all} \quad \ell = 0 \dots N_v - 1,$$

where

$$f_0^v(v) = \sum_{n=1}^{N_\mathrm{G}} \frac{a_n}{\sigma_n \sqrt{2\pi}} \mathrm{e}^{-\frac{(v-v_n)^2}{2\sigma_n^2}}.$$

Therefore, we have to compute

$$c_\ell = \langle f_0^v, H_\ell^{\gamma,\varepsilon} \rangle_\omega, \quad \text{for all} \quad \ell = 0 \dots N_v - 1.$$

It turns out that in case

$$\sigma_1 = \sigma_2 = \dots = \sigma_{N_\mathrm{G}} = \sigma,$$

which is true, for example, for the two stream instability test case (see Section 18.1), it is possible to utilize the theory from part II if we decide to fix the shape parameter for our basis to the width of the Gaussians in the initial condition. Indeed, for this setup, the sum of the Gaussians in f_0 takes the form of the Gaussian RBF interpolant (4.1). If we take

$$\varepsilon = \frac{1}{\sigma\sqrt{2}},$$

then the coefficients $\{c_\ell\}_{\ell=0}^{N_v-1}$ correspond to the coefficients of the HermiteGF expansion (6.2) of the interpolant and

$$f_0^v(v) = \sum_{n=1}^{N_\mathrm{G}} a_n \mathrm{e}^{-\varepsilon^2(v-v_n)^2} = \sum_{n=1}^{N_\mathrm{G}} a_n \exp\left(\varepsilon^2 v_n^2 \left(\frac{\varepsilon^2}{\gamma^2} - 1\right)\right) \sum_{\ell \geq 0} \frac{\varepsilon^{2\ell}\sqrt{2^\ell}}{\gamma^\ell \sqrt{\ell!}} v_n^\ell H_\ell^{\gamma,\varepsilon}(v).$$

Indeed, we know from Proposition 7.1.1 that after the truncation of the basis, the coefficients of the HermiteGF expansion correspond to the generalized Fourier coefficients $\langle f_{N_v}^0, H_\ell^{\gamma,\varepsilon} \rangle_\omega$. In particular,

$$\langle f_0^v, H_\ell^{\gamma,\varepsilon} \rangle_\omega = \sum_{n=1}^{N_\mathrm{G}} a_n \exp\left(\varepsilon^2 v_n^2 \left(\frac{\varepsilon^2}{\gamma^2} - 1\right)\right) \frac{\varepsilon^{2\ell}\sqrt{2^\ell}}{\gamma^\ell \sqrt{\ell!}} v_n^\ell = c_\ell.$$

For other cases, we derive an analytic formula for the scalar products $\langle f_0, H_\ell^{\gamma,\varepsilon} \rangle_\omega$ in the next section.

14.5.1. Galerkin projection of the sum of multiple Gaussians

Consider a generic initial distribution

$$f_0(x,v) = (1+\alpha\cos(mk_0x))\sum_{n=1}^{N_{\mathrm{G}}}\frac{a_n}{\sigma_n\sqrt{2\pi}}\mathrm{e}^{-\frac{(v-v_n)^2}{2\sigma_n^2}},\ x\in\left[0,\frac{2\pi}{k_0}\right],\ k_0,\alpha\in\mathbb{R},\ m\in\mathbb{N}.$$

The linear perturbation $(1+\alpha\cos(mk_0x))$ in x can be exactly represented via the Fourier basis. Therefore, let us focus on the representation of the part corresponding to the velocity

$$f_0^v(v) = \sum_{n=1}^{N_{\mathrm{G}}}\frac{a_n}{\sigma_n\sqrt{2\pi}}\mathrm{e}^{-\frac{(v-v_n)^2}{2\sigma_n^2}}$$

in the GW basis. For that, we need to find the projection of f_0^v onto the GW basis. In particular, we need to find

$$c_\ell = \langle f_0^v, H_\ell^{\gamma,\varepsilon}\rangle_\omega, \quad \text{for} \quad \ell = 0\dots N_v-1.$$

Then, the approximation of f_0^v in the GW basis reads as

$$f_0^v(v) \approx \sum_{\ell=0}^{N_v-1} c_\ell H_\ell^{\gamma,\varepsilon}(v).$$

It turns out that for f_0^v it is possible to compute those coefficients analytically.

Proposition 14.5.1: Initial distribution approximation

The initial distribution f_0^v in velocity can be approximated in the GW basis as

$$f_0^v(v) \approx \sum_{\ell=0}^{N_v-1}\langle f_0^v, H_\ell^{\gamma,\varepsilon}\rangle_\omega H_\ell^{\gamma,\varepsilon}(v),$$

where the generalized Fourier coefficients of f_0^v can be evaluated as

$$\langle f_0^v, H_\ell^{\gamma,\varepsilon}\rangle_\omega = \sum_{n=1}^{N_G}\frac{\alpha_n}{\sigma_n\sqrt{2\pi}}\frac{\gamma\sigma_n\sqrt{2}}{\tilde{\varepsilon}_n}\exp\left(\frac{\varepsilon^2v_n^2}{\tilde{\varepsilon}_n^2}\right)\left(\frac{1-2\sigma_n^2\varepsilon^2}{\tilde{\varepsilon}_n^2}\right)^{\ell/2} H_\ell^{\frac{\gamma}{\tilde{\varepsilon}_n},\bar{\gamma}}(v_n) \tag{14.16}$$

where

$$\tilde{\varepsilon}_n = \sqrt{1+2\gamma^2\sigma_n^2-2\varepsilon^2\sigma_n^2} \quad \text{and} \quad \bar{\gamma}_n = \frac{\gamma}{\tilde{\varepsilon}_n\sqrt{1-2\sigma_n^2\varepsilon^2}}.$$

Proof. Let us first compute an integral of a single Gaussian

$$I^{\sigma_n,v_n} = \int_{\mathbb{R}}\mathrm{e}^{-\frac{(v-v_n)^2}{2\sigma_n^2}}H_\ell^{\gamma,\varepsilon}(v)\omega(v)\mathrm{d}v = \pi^{-1/2}\gamma\int_{\mathbb{R}}\frac{1}{\sqrt{2^\ell\ell!}}h_\ell(\gamma v)\mathrm{e}^{-\varepsilon^2v^2}e^{-\frac{(v-v_n)^2}{2\sigma_n^2}}\mathrm{e}^{2\varepsilon^2-\gamma^2}\mathrm{d}v$$

$$= \pi^{-1/2}\gamma\int_{\mathbb{R}}\frac{1}{\sqrt{2^\ell\ell!}}h_\ell(\gamma v)\mathrm{e}^{(\varepsilon^2-\gamma^2)v^2-\frac{(v-v_n)^2}{2\sigma_n^2}}\mathrm{d}v.$$

Denote

$$\tilde{\varepsilon}_n = \sqrt{1+2\gamma^2\sigma_n^2-2\varepsilon^2\sigma_n^2}.$$

We observe that

$$(\varepsilon^2 - \gamma^2)v^2 - \frac{(v - v_n)^2}{2\sigma_n^2} = \frac{1}{2\sigma_n^2}\left(-\tilde{\varepsilon}_n^2 v^2 + 2vv_n - v_n^2\right)$$
$$= -\left(\frac{\tilde{\varepsilon}_n v}{\sigma_n\sqrt{2}} - \frac{v_n}{\tilde{\varepsilon}_n\sigma_n\sqrt{2}}\right)^2 + v_n^2\left(\frac{-\gamma^2 + \varepsilon^2}{\tilde{\varepsilon}_n^2}\right)$$

Denote

$$\bar{v} = \frac{\tilde{\varepsilon}_n v}{\sigma_n\sqrt{2}}.$$

Then,

$$I^{\sigma_n,v_n} = \frac{\pi^{-1/2}\gamma\sigma_n\sqrt{2}}{\tilde{\varepsilon}_n\sqrt{2^\ell \ell!}}\mathrm{e}^{v_n^2\left(\frac{-\gamma^2+\varepsilon^2}{\tilde{\varepsilon}_n^2}\right)}\int_{\mathbb{R}} h_\ell\left(\frac{\sigma_n\gamma\sqrt{2}}{\tilde{\varepsilon}_n}\bar{v}\right)\exp\left(-\left(\bar{v}^2 - \frac{v_n}{\tilde{\varepsilon}_n\sigma_n\sqrt{2}}\right)^2\right)\mathrm{d}\bar{v}.$$

We recall the analytic formula [GR14, $\S 7.374$, Expr. 8] for this type of integral

$$\int h_\ell(\alpha x)\mathrm{e}^{(x-y)^2}\mathrm{d}x = \pi^{1/2}(1-\alpha^2)^{\ell/2}h_\ell\left(\frac{\alpha y}{\sqrt{1-\alpha^2}}\right)$$

for all real $\alpha \neq 1$. Then, we take

$$\alpha = \frac{\sigma_n\gamma\sqrt{2}}{\tilde{\varepsilon}_n}, \quad \text{and} \quad y = \frac{v_n}{\tilde{\varepsilon}_n\sigma_n\sqrt{2}}$$

Finally, we arrive to the final expression

$$I^{\sigma_2,v_n} = \frac{\gamma\sigma_n\sqrt{2}}{\tilde{\varepsilon}_n\sqrt{2^\ell \ell!}}\mathrm{e}^{v_n^2\left(\frac{-\gamma^2+\varepsilon^2}{\tilde{\varepsilon}_n^2}\right)}\left(\frac{1-2\sigma_n^2\varepsilon^2}{\tilde{\varepsilon}_n^2}\right)^{\ell/2}h_\ell\left(\frac{\gamma v_n}{\tilde{\varepsilon}_n\sqrt{1-2\sigma_n^2\varepsilon^2}}\right).$$

To simplify the expression above, we denote

$$\bar{\gamma}_n = \frac{\gamma}{\tilde{\varepsilon}_n\sqrt{1-2\sigma_n^2\varepsilon^2}}.$$

Then,

$$I^{\sigma_n,v_n} = \frac{\gamma\sigma_n\sqrt{2}}{\tilde{\varepsilon}_n}\exp\left(\frac{\varepsilon^2 v_n^2}{\tilde{\varepsilon}_n^2}\right)\left(\frac{1-2\sigma_n^2\varepsilon^2}{\tilde{\varepsilon}_n^2}\right)^{\ell/2}H_\ell^{\frac{\gamma}{\tilde{\varepsilon}_n},\bar{\gamma}_n}(v_n).$$

Taking a linear combination of the coefficients I^{σ_2,v_n} corresponding to single Gaussians, we arrive at (14.16). □

It turns out that one can get a simplified expression for Gaussians that are centered at zero.

Corollary 14.5.1. *The coefficients of the Galerkin projection of a Gaussian centered at zero can be computed as*

$$\langle \mathrm{e}^{-\frac{v^2}{2\sigma_n^2}}, H_\ell^{\gamma,\varepsilon}\rangle_\omega = \pi^{-1/2}\gamma I_{\ell,\bar{\varepsilon}_n} \quad \textit{with} \quad \bar{\varepsilon}_n = \frac{\tilde{\varepsilon}_n}{\sigma_n\sqrt{2}},$$

where, as before, $\tilde{\varepsilon}_n = \sqrt{1 + 2\gamma^2\sigma_n^2 - 2\varepsilon^2\sigma_n^2}$ *and the values of* $I_{\ell,\bar{\varepsilon}_n}$ *can be computed from* (14.3) *via recurrence for even* ℓ

$$\frac{I_{\ell+2,\bar{\varepsilon}_n}}{I_{\ell,\bar{\varepsilon}_n}} = \sqrt{\frac{\ell+1}{\ell+2}}\left(\frac{\gamma^2}{\bar{\varepsilon}_n^2} - 1\right)$$

with $I_{0,\bar{\varepsilon}_n} = \frac{\sqrt{\pi}}{\bar{\varepsilon}_n}$*. And* $I_{\ell,\bar{\varepsilon}_n} = 0$ *for odd* $\ell \in \mathbb{N}_0$*.*

Proof. Using the result (14.16) with $v_n = 0$, we get

$$\langle \mathrm{e}^{-\frac{v^2}{2\sigma^2}}, H_\ell^{\gamma,\varepsilon} \rangle_\omega = \frac{\gamma\sigma_n\sqrt{2}}{\tilde{\varepsilon}_n\sqrt{2^\ell \ell!}} \left(\frac{1 - 2\sigma_n^2\varepsilon^2}{\tilde{\varepsilon}_n^2} \right)^{\ell/2} h_\ell(0).$$

Recall that according to [GR14, $\S 8.956$], for odd indexes Hermite polynomials turn to zero when the argument is zero

$$h_{2\ell+1}(0) = 0 \quad \forall\, \ell \in \mathbb{N}_0.$$

For the even ones, the following relation holds

$$h_{2\ell}(0) = (-1)^\ell 2^\ell (2\ell-1)!! = (-1)^\ell \frac{(2\ell)!}{\ell!}$$

For $\ell = 0$, we recover

$$\langle \mathrm{e}^{-\frac{v^2}{2\sigma^2}}, H_0^{\gamma,\varepsilon} \rangle_\omega = \frac{\gamma\sigma_n\sqrt{2}}{\tilde{\varepsilon}_n} = \frac{\gamma}{\bar{\varepsilon}_n} = \pi^{-1/2}\gamma I_{0,\bar{\varepsilon}_n}.$$

For an arbitrary even index, we get

$$\begin{aligned} \frac{I_{2(\ell+1),\bar{\varepsilon}_n}}{I_{2\ell,\bar{\varepsilon}_n}} &= -\frac{1}{2\sqrt{(2\ell+1)(2\ell+2)}} \left(\frac{1 - 2\sigma_n^2\varepsilon^2}{\tilde{\varepsilon}_n^2} \right) \frac{(2\ell+1)(2\ell+2)}{(\ell+1)} \\ &= \sqrt{\frac{2\ell+1}{2\ell+2}} \left(\frac{2\gamma^2\sigma_n^2}{\tilde{\varepsilon}_n^2} - 1 \right) = \sqrt{\frac{2\ell+1}{2\ell+2}} \left(\frac{\gamma^2}{\tilde{\varepsilon}_n^2} - 1 \right). \end{aligned}$$ □

Note that the recurrence for $I_{\ell,\bar{\varepsilon}_n}$ is of the same form as the one considered Lemma 14.1.1 in for

$$I_\ell = \int_{\mathbb{R}} H_\ell^{\gamma,\varepsilon}(v)\mathrm{d}v.$$

This is not a coincidence, since the integral $\langle \mathrm{e}^{-\frac{v^2}{2\sigma^2}}, H_\ell^{\gamma,\varepsilon} \rangle_\omega$ can also be expressed through the integral of a GW basis function with certain parameters ε, γ that differ from the original ones.

15. Computation of the observables and conservation properties

Let us take a look whether the generalized Fourier–Hermite discretization preserves the physical conservation properties of the Vlasov equation that we have discussed in Theorem 13.3.1. We focus on mass, momentum, and energy. First, we consider a few relevant telescopic sums that will be useful for the computation of the observables. After that, we compute the explicit expressions of the time derivatives of the observables. We then investigate in which cases these derivatives are equal to zero, which implies the conservation of the observables over time. It turns out, that mass and energy can be conserved exactly for an odd number of basis functions N_v in velocity, whereas the momentum can only be conserved for an even number of basis functions. For the L^2 norm, we provide an analytic formula for its computation and postpone the discussion of its conservation to the next chapter.

15.1. Telescopic sums

Let us prove some technical results which will serve as the basis of the later proofs of the conservation properties. We compute the sums involving the integrals with the GW Hermite basis functions computed in Lemma 14.1.1 in combination with the terms resembling parts of the discretization (14.12). Note, that even though the structure of the sums is based on the discretization (14.12), the results are independent of the specific coefficients $\{c_\ell^k(t)\}$.

The idea of employing telescopic sums involving the integrals of the basis functions for the investigation of the conservation properties was mentioned in [Hol96] for SW and AW cases. We now prove these relations for the generalized setup.

Lemma 15.1.1: Telescopic sums

Consider a sequence $\{a_\ell\}_{\ell=0}^{N_v}$ of real numbers. Denote

$$S_I = \sum_{\ell=0}^{N_v-1} I_\ell \left((\gamma^2 - \varepsilon^2)\sqrt{\ell+1}a_{\ell+1} - \varepsilon^2\sqrt{\ell}a_{\ell-1} \right).$$

We set $a_{-1} = 0$. Let S_J and $S_{\bar{I}}$ be the same sums as above but with the terms J_ℓ or $\bar{I}_\ell$ instead of I_ℓ correspondingly. Then, following expressions hold for these sums

1.
$$S_I = \frac{\varepsilon^2\sqrt{2}(\gamma^2-\varepsilon^2)}{\gamma}\begin{cases} J_{N_v}a_{N_v} & \text{for} \quad N_v \quad \text{odd.} \\ J_{N_v-1}a_{N_v-1} & \text{for} \quad N_v \quad \text{even.}\end{cases} \tag{15.1}$$

2.
$$S_J = \frac{\gamma}{\sqrt{2}}\begin{cases} N_v I_{N_v-1}a_{N_v-1} - \sum_{\ell=0}^{N_v-1} a_\ell I_\ell & \text{for} \quad N_v \quad \text{odd,} \\ N_v I_{N_v}a_{N_v} - \sum_{\ell=0}^{N_v-1} a_\ell I_\ell & \text{for} \quad N_v \quad \text{even.}\end{cases} \tag{15.2}$$

3.
$$S_{\bar{I}} = \begin{cases} \frac{\gamma^2(N_v-1)+\gamma^2-\varepsilon^2}{\gamma\sqrt{2}} J_{N_v}a_{N_v} - \gamma\sqrt{2}\sum_{\ell=0}^{N_v-1} J_\ell a_\ell & \text{for} \quad N_v \quad \text{odd,} \\ \frac{\gamma^2 N_v+\gamma^2-\varepsilon^2}{\gamma\sqrt{2}} J_{N_v-1}a_{N_v-1} - \gamma\sqrt{2}\sum_{\ell=0}^{N_v-1} J_\ell a_\ell & \text{for} \quad N_v \quad \text{even.}\end{cases} \tag{15.3}$$

Proof. Let us first prove the expression for S_I. Using the expression (14.4) of I_ℓ through $J_{\ell+1}$ and $J_{\ell-1}$, we get

$$\begin{aligned} S_I &= I_0(\gamma^2-\varepsilon^2)a_1 + \sum_{\ell=2}^{N_v-1} I_\ell\left((\gamma^2-\varepsilon^2)\sqrt{\ell+1}a_{\ell+1} - \varepsilon^2\sqrt{\ell}a_{\ell-1}\right) \\ &\overset{(14.4)}{=} I_0(\gamma^2-\varepsilon^2)a_1 + \sum_{\ell=2}^{N_v-1}\left(\frac{\varepsilon^2\sqrt{2}(\gamma^2-\varepsilon^2)}{\gamma}J_{\ell+1}a_{\ell+1} - \frac{\varepsilon^2\sqrt{2}(\gamma^2-\varepsilon^2)}{\gamma}a_{\ell-1}J_{\ell-1}\right) \\ &= I_0(\gamma^2-\varepsilon^2)a_1 + \frac{\varepsilon^2\sqrt{2}(\gamma^2-\varepsilon^2)}{\gamma}\sum_{\ell=2}^{N_v-1}(J_{\ell+1}a_{\ell+1} - J_{\ell-1}a_{\ell-1}). \end{aligned}$$

Then, telescoping the sum above and noting that

$$\frac{\varepsilon^2\sqrt{2}(\gamma^2-\varepsilon^2)}{\gamma}J_1a_1 = \sqrt{\pi}\frac{(\gamma^2-\varepsilon^2)}{\varepsilon}a_1 = I_0(\gamma^2-\varepsilon^2)a_1$$

we arrive to the formula (15.1).

The formula for S_J can be proved in similar fashion, but this time we express J_ℓ through $I_{\ell-1}$ with the help of (14.4).

$$\begin{aligned} S_J &= \sum_{\ell=0}^{N_v-1} J_\ell\left((\gamma^2-\varepsilon^2)\sqrt{\ell+1}a_{\ell+1} - \varepsilon^2\sqrt{\ell}a_{\ell-1}\right) \\ &\overset{(14.4)}{=} \sum_{\ell=1}^{N_v-1}\frac{\gamma}{\varepsilon^2}\sqrt{\frac{\ell}{2}}I_{\ell-1}\left((\gamma^2-\varepsilon^2)\sqrt{\ell+1}a_{\ell+1} - \varepsilon^2\sqrt{\ell}a_{\ell-1}\right) \\ &\overset{(14.3)}{=} \frac{\gamma}{\sqrt{2}}\sum_{\ell=1}^{N_v-1}((\ell+1)I_{\ell+1}a_{\ell+1} - \ell I_{\ell-1}a_{\ell-1}) \\ &= \frac{\gamma}{\sqrt{2}}\sum_{\ell=1}^{N_v-1}((\ell+1)I_{\ell+1}a_{\ell+1} - (\ell-1)I_{\ell-1}a_{\ell-1}) - \frac{\gamma}{\sqrt{2}}\sum_{\ell=0}^{N_v-2}a_\ell I_\ell. \end{aligned}$$

We note that the first term is a telescopic sum and can be computed as

$$\sum_{\ell=1}^{N_v-1} ((\ell+1)I_{\ell+1}a_{\ell+1} - (\ell-1)I_{\ell-1}a_{\ell-1}) = \begin{cases} (N_v-1)I_{N_v-1}a_{N_v-1} & \text{for} \quad N_v \quad \text{odd.} \\ N_v I_{N_v} a_{N_v} & \text{for} \quad N_v \quad \text{even,} \end{cases}$$

We observe that $I_{N_v-1} = 0$ for even N_v, therefore $\sum_{\ell=0}^{N_v-2} a_\ell I_\ell = \sum_{\ell=0}^{N_v-1} a_\ell I_\ell$ in this case. On the other hand, for the odd N_v, we can borrow the missing term $-I_{N_v-1}a_{N_v-1}$ from the result of the telescopic sum. Putting it all together we arrive at the expression (15.2).

For the last sum $S_{\bar{I}}$ let us first use the representation (14.5) of $\bar{I}_\ell$ through $J_{\ell-1}$. We get

$$\begin{aligned} S_{\bar{I}} &= \bar{I}_0(\gamma^2-\varepsilon^2)a_1 + \sum_{\ell=2}^{N_v-1} \bar{I}_\ell \left((\gamma^2-\varepsilon^2)\sqrt{\ell+1}a_{\ell+1} - \varepsilon^2\sqrt{\ell}a_{\ell-1}\right) \\ &= \bar{I}_0(\gamma^2-\varepsilon^2)a_1 + \frac{1}{\gamma\sqrt{2}} \sum_{\ell=2}^{N_v-1} \left(\frac{\ell\varepsilon^2 + (\ell+1)(\gamma^2-\varepsilon^2)}{\varepsilon^2\sqrt{\ell}}\right)(\gamma^2-\varepsilon^2)\sqrt{\ell+1}J_{\ell-1}a_{\ell+1} \\ &\quad - \frac{1}{\gamma\sqrt{2}} \sum_{\ell=2}^{N_v-1} (\ell\varepsilon^2 + (\ell+1)(\gamma^2-\varepsilon^2))J_{\ell-1}a_{\ell-1}. \end{aligned}$$

We now apply (14.4) to the first sum to obtain

$$\begin{aligned} S_{\bar{I}} &= \bar{I}_0(\gamma^2-\varepsilon^2)a_1 + \frac{1}{\gamma\sqrt{2}} \sum_{\ell=2}^{N_v-1} (\ell\varepsilon^2 + (\ell+1)(\gamma^2-\varepsilon^2))J_{\ell+1}a_{\ell+1} \\ &\quad - \frac{1}{\gamma\sqrt{2}} \sum_{\ell=2}^{N_v-1} (\ell\varepsilon^2 + (\ell+1)(\gamma^2-\varepsilon^2))J_{\ell-1}a_{\ell-1} \\ &= \bar{I}_0(\gamma^2-\varepsilon^2)a_1 + \frac{1}{\gamma\sqrt{2}} \sum_{\ell=2}^{N_v-1} (\ell\gamma^2 + \gamma^2 - \varepsilon^2)(J_{\ell+1}a_{\ell+1} - J_{\ell-1}a_{\ell-1}). \end{aligned}$$

Splitting the terms of the sum containing ℓ from the other, we get

$$\begin{aligned} S_{\bar{I}} &= \bar{I}_0(\gamma^2-\varepsilon^2)a_1 + \frac{\gamma}{\sqrt{2}} \sum_{\ell=2}^{N_v-1} \ell(J_{\ell+1}a_{\ell+1} - J_{\ell-1}a_{\ell-1}) \\ &\quad + \frac{\gamma^2-\varepsilon^2}{\gamma\sqrt{2}} \sum_{\ell=2}^{N_v-1} (J_{\ell+1}a_{\ell+1} - J_{\ell-1}a_{\ell-1}) \\ &= \bar{I}_0(\gamma^2-\varepsilon^2)a_1 + \frac{\gamma}{\sqrt{2}} \sum_{\ell=2}^{N_v-1} ((\ell+1)J_{\ell+1}a_{\ell+1} - (\ell-1)J_{\ell-1}a_{\ell-1}) - \frac{\gamma}{\sqrt{2}} \sum_{\ell=3}^{N_v} J_\ell a_\ell \\ &\quad - \frac{\gamma}{\sqrt{2}} \sum_{\ell=1}^{N_v-2} J_\ell a_\ell + \frac{\gamma^2-\varepsilon^2}{\gamma\sqrt{2}} \sum_{\ell=2}^{N_v-1} (J_{\ell+1}a_{\ell+1} - J_{\ell-1}a_{\ell-1}). \end{aligned}$$

Next, we use that

$$\frac{\gamma^2-\varepsilon^2}{\gamma\sqrt{2}} J_1 a_1 = \frac{\sqrt{\pi}(\gamma^2-\varepsilon^2)}{2\varepsilon^3} a_1 = \bar{I}_0(\gamma^2-\varepsilon^2)a_1.$$

and $J_2 = 0$ and telescope the sums to find

$$\begin{aligned}
S_{\bar{I}} &= \bar{I}_0(\gamma^2 - \varepsilon^2)a_1 + \frac{\gamma}{\sqrt{2}}\left(N_v J_{N_v} a_{N_v} + (N_v - 1)J_{N_v-1}a_{N_v-1} - 2J_2a_2 - J_1a_1\right) \\
&\quad - \frac{\gamma}{\sqrt{2}}\sum_{\ell=3}^{N_v} J_\ell a_\ell - \frac{\gamma}{\sqrt{2}}\sum_{\ell=1}^{N_v-2} J_\ell a_\ell + \frac{\gamma^2 - \varepsilon^2}{\gamma\sqrt{2}}\left(J_{N_v}a_{N_v} + J_{N_v-1}a_{N_v-1} - J_2a_2 - J_1a_1\right) \\
&= -\gamma\sqrt{2}\sum_{\ell=1}^{N_v-1} J_\ell a_\ell + \frac{\gamma}{\sqrt{2}}(N_v - 1)J_{N_v}a_{N_v} + \frac{\gamma}{\sqrt{2}}N_v J_{N_v-1}a_{N_v-1} \\
&\quad + \frac{\gamma^2 - \varepsilon^2}{\gamma\sqrt{2}}(J_{N_v}a_{N_v} + J_{N_v-1}a_{N_v-1}).
\end{aligned}$$

Using that $J_{N_v-1} = 0$ for odd N_v and $J_{N_v} = 0$ for even N_v, together with the fact that $J_0 = 0$, we arrive to (15.3). □

Now, when everything is set up, we move on to the investigation of conservation properties of the generalized Fourier–Hermite discretization.

15.2. Mass

We first take a look at the evolution of mass with time when the generalized Fourier–Hermite discretization is used. The mass is defined as follows:

$$M(t) = \int_{\mathbb{R}}\int_0^{L_x} f(x, v, t)\mathrm{d}x\mathrm{d}v.$$

According to the properties of the Vlasov equation (see Theorem 13.3.1), the mass should be preserved over time:

$$\frac{\mathrm{d}M(t)}{\mathrm{d}t} = 0.$$

We now check if this condition is fulfilled for the generalized Fourier–Hermite discretization. Numerically, the mass of the approximated solution can be computed as:

$$M_{N_v,N_x}(t) = \sum_{\ell=0}^{N_v-1} I_\ell \int_0^{L_x} \sum_{k=-N_x}^{N_x} c_\ell^k(t)\exp\left(\frac{2\pi\mathrm{i}kx}{L_x}\right)\mathrm{d}x = L_x\sum_{\ell=0}^{N_v-1} I_\ell c_\ell^0(t). \tag{15.4}$$

Theorem 15.2.1: Fourier–Hermite mass evolution

The mass is preserved for an odd number of basis functions N_v. Moreover,

$$\frac{\mathrm{d}M_{N_v,N_x}(t)}{\mathrm{d}t} = \begin{cases} 0 & \text{for} \quad N_v \quad \text{odd.} \\ L_x\frac{2\varepsilon^2}{\gamma^2}(\gamma^2 - \varepsilon^2)J_{N_v-1}\beta^0_{N_v-1}(t) & \text{for} \quad N_v \quad \text{even,} \end{cases} \tag{15.5}$$

where

$$\beta^0_{N_v-1}(t) = [\mathbf{E}(t) * \mathbf{c}_{N_v-1}(t)][0].$$

Proof. Using the representation (15.4) together with expression (14.12), we get:

$$\frac{\mathrm{d}M_{N_v,N_x}(t)}{\mathrm{d}t} = \frac{\sqrt{2}L_x}{\gamma} \sum_{\ell=0}^{N_v-1} I_\ell \left((\gamma^2-\varepsilon^2)\sqrt{\ell+1}\beta^0_{\ell+1}(t) - \varepsilon^2\sqrt{\ell}\beta^0_{\ell-1}(t) \right).$$

Using expression (15.1) for the sum above with $\{\beta^0_\ell(t)\}$ in place of the sequence $\{a_\ell\}$, we get

$$\frac{\mathrm{d}M_{N_v,N_x}(t)}{\mathrm{d}t} = \frac{2\varepsilon^2 L_x}{\gamma^2}(\gamma^2-\varepsilon^2) \begin{cases} J_{N_v}\beta^0_{N_v}(t) & \text{for} \quad N_v \quad \text{odd.} \\ J_{N_v-1}\beta^0_{N_v-1}(t) & \text{for} \quad N_v \quad \text{even.} \end{cases}$$

Keeping in mind that the closure scheme implies $c^0_{N_v}(t) = 0$ for all $t \in \mathbb{R}_+$, and hence $\beta^0_{N_v}(t) = 0$, for all t, we arrive to the expression (15.5). □

Remark 15.2.1. *Note that a similar expression also holds on the semi-discrete level in phase space, i.e. after the velocity discretization while the spacial variable is left continuous.*

15.3. Momentum

The second observable that we will take a look at is the momentum. The momentum is defined as follows:

$$P(t) = \int_{\mathbb{R}} \int_0^{L_x} v f(x,v,t) \mathrm{d}x \mathrm{d}v.$$

According to the properties of the Vlasov equation (see Theorem 13.3.1), the momentum should be preserved over time:

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = 0.$$

We now check if this condition is fulfilled for the generalized Fourier–Hermite discretization. Numerically, the momentum of the approximated solution can be computed as

$$P_{N_v,N_x} = \sum_{\ell=0}^{N_v-1} J_\ell \int_0^{L_x} \sum_{k=-N_x}^{N_x} c^k_\ell(t) \exp\left(\frac{2\pi \mathrm{i} k x}{L_x}\right) \mathrm{d}x = L_x \sum_{\ell=0}^{N_v-1} J_\ell c^0_\ell(t). \tag{15.6}$$

Theorem 15.3.1: Fourier–Hermite momentum evolution

The momentum is preserved for the even number of basis functions. Moreover,

$$\frac{\mathrm{d}P_{N_v,N_x}(t)}{\mathrm{d}t} = \begin{cases} L_x N_v I_{N_v-1}\beta^0_{N_v-1}(t) & \text{for} \quad N_v \quad \text{odd,} \\ 0 & \text{for} \quad N_v \quad \text{even,} \end{cases} \tag{15.7}$$

where

$$\beta^0_{N_v-1}(t) = [\mathbf{E}(t) * \mathbf{c}_{N_v-1}(t)][0].$$

Proof. Using the representation (15.6) together with the expression (14.12), we get:

$$\frac{\mathrm{d}P_{N_v,N_x}(t)}{\mathrm{d}t} = \frac{\sqrt{2}L_x}{\gamma}\sum_{\ell=0}^{N_v-1} J_\ell\left((\gamma^2-\varepsilon^2)\sqrt{\ell+1}\beta^0_{\ell+1}(t) - \varepsilon^2\sqrt{\ell}\beta^0_{\ell-1}(t)\right).$$

Using (15.2) for the sum above with $\{\beta^0_\ell(t)\}$ in place of the sequence $\{a_\ell\}$, we get

$$\frac{\mathrm{d}P_{N_v,N_x}(t)}{\mathrm{d}t} = L_x\begin{cases} N_v I_{N_v-1}\beta^0_{N_v-1}(t) - \sum_{\ell=0}^{N_v-1}\beta^0_\ell(t)I_\ell & \text{for} \quad N_v \quad \text{odd,}\\ N_v I_{N_v}\beta^0_{N_v}(t) - \sum_{\ell=0}^{N_v-1}\beta^0_\ell(t)I_\ell & \text{for} \quad N_v \quad \text{even.}\end{cases}$$

Since the closure scheme implies $c^0_{N_v}(t)=0$, and hence $\beta^0_{N_v}(t)=0$, for all t, for even N_v, the term $\beta^0_{N_v}(t)$ is zero. Let us now consider the sum $\sum_{\ell=0}^{N_v-1}\beta^0_\ell(t)I_\ell$. Recall that, by definition (14.13),

$$\beta^0_\ell(t) = [\mathbf{E}(t) * \mathbf{c}_\ell(t)][0].$$

Then, using the explicit formula (14.15) for the Fourier coefficients $E_k(t)$ of the electric field through the coefficients $\{c^k_\ell(t)\}$, we get

$$\begin{aligned}\sum_{\ell=0}^{N_v-1}\beta^0_\ell(t)I_\ell &= \sum_{\ell=0}^{N_v-1} I_\ell \sum_{k=-N_x}^{N_x} E_k(t)c^{-k}_\ell(t) = \sum_{k=-N_x}^{N_x} E_k(t)\sum_{\ell=0}^{N_v-1} I_\ell c^{-k}_\ell(t)\\ &= \sum_{\substack{k=-N_x\\k\neq0}}^{N_x}\frac{iL_x}{2\pi k}\left(\sum_{\ell=0}^{N_v-1}c^k_\ell(t)I_\ell\right)\left(\sum_{\ell=0}^{N_v-1}c^{-k}_\ell(t)I_\ell\right) = 0,\end{aligned}$$

since we have fixed $E_0 = 0$. □

15.4. Energy

As we know from the chapter 13, the energy of the system consists of the two components: kinetic energy and potential energy

$$W(t) = W^{\mathrm{K}}(t) + W^{\mathrm{E}}(t).$$

For the Vlasov–Maxwell model the potential energy consists from the electric and magnetic parts. However, for the Vlasov–Poisson model, only the electric energy is present since the magnetic field is zero and therefore, we use W^E for its notation. According to the properties of the Vlasov equation (see Theorem 13.3.1), the full energy should be preserved over time:

$$\frac{\mathrm{d}W(t)}{\mathrm{d}t} = 0.$$

Recall that in the continuous case, the kinetic energy is computed as

$$W^{\mathrm{K}}(t) = \int_{\mathbb{R}}\int_0^{L_x} v^2 f(x,v,t)\mathrm{d}x\mathrm{d}v.$$

Similarly to the other observables, the kinetic energy of the approximated solution can be computed as

$$W^{\mathrm{K}}_{N_v,N_x}(t) = \frac{1}{2}\sum_{\ell=0}^{N_v-1} \bar{I}_\ell \int_0^{L_x} \sum_{k=-N_x}^{N_x} c^k_\ell(t)\exp\left(\frac{2\pi \mathrm{i} k x}{L_x}\right)\mathrm{d}x = \frac{L_x}{2}\sum_{\ell=0}^{N_v-1} \bar{I}_\ell c^0_\ell(t). \qquad (15.8)$$

The electric energy in the continuous case writes as

$$W^E(t) = \frac{1}{2}\int_0^{L_x} E(x,t)^2 \mathrm{d}x.$$

After the Fourier discretization of E in space, we get

$$W^E_{N_x}(t) = \frac{1}{2}\int_0^{L_x}\left(\sum_{k=-N_x}^{N_x} E_k(t)\exp\left(\frac{2\pi k \mathrm{i} x}{L_x}\right)\right)^2 \mathrm{d}x = \frac{L_x}{2}\sum_{k=-N_x}^{N_x}|E_k(t)|^2.$$

Theorem 15.4.1: Fourier–Hermite energy evolution

For odd N_v, the energy is conserved, i.e.

$$\frac{\mathrm{d}W_{N_v,N_x}(t)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(W^E_{N_x}(t) + W^K_{N_v,N_x}(t)) = 0.$$

For the even N_v, the time derivative of the energy can be computed as

$$\begin{aligned}\frac{\mathrm{d}W_{N_v,N_x}(t)}{\mathrm{d}t} &= \frac{\mathrm{i}L_x^2}{2\pi}\sum_{k=-N_x}^{N_x}\frac{\Delta M^k_{N_v}(t)}{k}E_{-k}(t) \qquad (15.9)\\ &\quad + L_x\frac{\gamma^2 N_v + \varepsilon^2 - \gamma^2}{2\gamma^2}J_{N_v-1}\beta^0_{N_v-1}(t)\end{aligned}$$

with $\Delta M^k_{N_v}(t) = L_x\frac{2\varepsilon^2}{\gamma^2}(\gamma^2-\varepsilon^2)J_{N_v-1}\beta^k_{N_v-1}(t)$ and $\beta^k_{N_v-1}(t) = [\mathbf{E}(t)*\mathbf{c}_{N_v-1}(t)][k]$.

Proof. We first compute the derivative of the electric energy.

$$\frac{\mathrm{d}W^E_{N_x}}{\mathrm{d}t} = \frac{L_x}{2}\sum_{k=-N_x}^{N_x}\frac{\mathrm{d}|E_k(t)|^2}{\mathrm{d}t} = L_x\sum_{k=-N_x}^{N_x}\frac{\mathrm{d}E_k(t)}{\mathrm{d}t}E_{-k}(t).$$

Let us now compute the time derivative of the Fourier coefficients E_k of the electric field E. Using the expression (14.15) for the coefficients $\{E_k\}$ and the equation (14.12) for the coefficients $c^k_\ell(t)$, we get, for $k \neq 0$

$$\begin{aligned}\frac{\mathrm{d}E_k(t)}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{i}L_x}{2\pi k}\sum_{\ell=0}^{N_v-1}c^k_\ell(t)I_\ell\right)\\ &\overset{(14.12)}{=} \frac{1}{\gamma\sqrt{2}}\sum_{\ell=0}^{N_v-1}I_\ell(\sqrt{\ell+1}c^k_{\ell+1}(t) + \sqrt{\ell}c^k_{\ell-1}(t))\\ &\quad + \frac{\sqrt{2}\mathrm{i}L_x}{2\pi k\gamma}\sum_{\ell=0}^{N_v-1}I_\ell\Big((\gamma^2-\varepsilon^2)\sqrt{\ell+1}\beta^k_{\ell+1}(t) - \varepsilon^2\sqrt{\ell}\beta^k_{\ell-1}(t)\Big)\end{aligned}$$

Using (15.1) for the second sum, we get, for $k \neq 0$

$$\frac{\sqrt{2}\mathrm{i}L_x}{2\pi k\gamma}\sum_{\ell=0}^{N_v-1} I_\ell\Big((\gamma^2-\varepsilon^2)\sqrt{\ell+1}\beta^k_{\ell+1}(t)-\varepsilon^2\sqrt{\ell}\beta^k_{\ell-1}(t)\Big)$$
$$=\begin{cases} 0 & \text{for} \quad N_v \quad \text{odd.} \\ \frac{\mathrm{i}L_x}{2\pi k}\frac{2\varepsilon^2}{\gamma^2}(\gamma^2-\varepsilon^2)J_{N_v-1}\beta^k_{N_v-1}(t) & \text{for} \quad N_v \quad \text{even,}\end{cases}$$

where we used the fact that according to the closure scheme $c^k_{N_v}(t) = 0$ for all $t > 0$, $k = -N_x, \dots, N_x$. As for the first sum, using the expression (14.4) of I_ℓ through $J_{\ell-1}$ and $J_{\ell+1}$, we get

$$\frac{1}{\gamma\sqrt{2}}\sum_{\ell=0}^{N_v-1} I_\ell(\sqrt{\ell+1}c^k_{\ell+1}(t)+\sqrt{\ell}c^k_{\ell-1}(t))$$
$$= I_0\frac{c^k_1(t)}{\gamma\sqrt{2}}+\frac{\varepsilon^2}{\gamma^2}\sum_{\ell=2}^{N_v-1}(J_{\ell+1}c^k_{\ell+1}(t)-J_{\ell-1}c^k_{\ell-1}(t))+\sum_{\ell=2}^{N_v-1}J_{\ell-1}c^k_{\ell-1}(t).$$
$$=\begin{cases} \sum_{\ell=0}^{N_v-1} J_\ell c^k_\ell(t) & \text{for} \quad N_v \quad \text{odd,} \\ \frac{\varepsilon^2}{\gamma^2}J_{N_v-1}c^k_{N_v-1}(t)+\sum_{\ell=0}^{N_v-2}J_\ell c^k_\ell(t) & \text{for} \quad N_v \quad \text{even,}\end{cases}$$

where we used that

$$I_0\frac{c^k_1(t)}{\gamma\sqrt{2}} = \frac{\sqrt{\pi}c^k_1(t)}{\varepsilon\gamma\sqrt{2}} = \frac{\varepsilon^2}{\gamma^2}J_1c^k_1(t)$$

and we added the term $J_0c^k_0(t)$ to the sum for both odd and even N_v since $J_0 = 0$. The term $J_{N_v-1}c^k_{N_v-1}(t)$ we added for the odd N_v only, since $J_{N_v-1} \neq 0$ for the even N_v. Denote

$$\Delta M^k_{N_v}(t) = L_x\frac{2\varepsilon^2}{\gamma^2}(\gamma^2-\varepsilon^2)J_{N_v-1}\beta^k_{N_v-1}(t).$$

Then

$$\frac{\mathrm{d}E_k(t)}{\mathrm{d}t} = \begin{cases} \sum_{\ell=0}^{N_v-1} J_\ell c^k_\ell(t) & \text{for} \quad N_v \quad \text{odd,} \\ \frac{\mathrm{i}L_x}{2\pi k}\Delta M_{N_v}(t)+\frac{\varepsilon^2}{\gamma^2}J_{N_v-1}c^k_{N_v-1}(t)+\sum_{\ell=0}^{N_v-2}J_\ell c^k_\ell(t) & \text{for} \quad N_v \quad \text{even.}\end{cases}$$

Therefore, the derivative of the electric energy is given by

$$\frac{\mathrm{d}W^E_{N_x}}{\mathrm{d}t} = L_x\sum_{k=-N_x}^{N_x}\frac{\mathrm{d}E_k(t)}{\mathrm{d}t}E_{-k}(t),$$

where $\frac{\mathrm{d}E_k(t)}{\mathrm{d}t}$ is computed as described above.

Let us now consider the kinetic energy. Using the representation (15.8), expression (14.12), we get

$$\frac{\mathrm{d}W^{\mathrm{K}}_{N_v,N_x}(t)}{\mathrm{d}t} = \frac{L_x}{\gamma\sqrt{2}}\sum_{\ell=0}^{N_v-1}\bar{I}_\ell\left((\gamma^2-\varepsilon^2)\sqrt{\ell+1}\beta^0_{\ell+1}(t)-\varepsilon^2\sqrt{\ell}\beta^0_{\ell-1}(t)\right).$$

Using the expression (15.3) for the sum above with $\{\beta^0_\ell(t)\}$ in place of the sequence $\{a_\ell\}$, we get

$$\frac{\mathrm{d}W^{\mathrm{K}}_{N_v,N_x}(t)}{\mathrm{d}t} = L_x \begin{cases} -\sum_{\ell=0}^{N_v-1} J_\ell \beta^0_\ell(t) & \text{for} \quad N_v \quad \text{odd,} \\ \frac{\gamma^2 N_v+\gamma^2-\varepsilon^2}{2\gamma^2} J_{N_v-1}\beta^0_{N_v-1}(t) - \sum_{\ell=0}^{N_v-1} J_\ell \beta^0_\ell(t) & \text{for} \quad N_v \quad \text{even.} \end{cases}$$

where we used again that $\beta^0_{N_v}(t) = 0$ for all t to eliminate the first term of the telescopic sum for odd N_v. We can now combine the expressions for the derivatives of the kinetic and electric energy to obtain the full energy evolution.

We recall that, by definition (14.13),

$$\beta^0_\ell(t) = [\mathbf{E}(t) * \mathbf{c}_\ell(t)][0] = \sum_{k=-N_x}^{N_x} c^k_\ell(t) E_{-k}(t).$$

Let us first compute the derivative of the full energy for the case of the odd N_v. Using the results above together with the explicit expression for $\beta^0_\ell(t)$, we get

$$\frac{\mathrm{d}W^{\mathrm{E}}_{N_x}}{\mathrm{d}t} + \frac{\mathrm{d}W^{\mathrm{K}}_{N_v,N_x}}{\mathrm{d}t} = L_x \sum_{k=-N_x}^{N_x} E_{-k}(t) \sum_{\ell=0}^{N_v-1} J_\ell c^k_\ell(t) - L_x \sum_{\ell=0}^{N_v-1} J_\ell \sum_{k=-N_x}^{N_x} c^k_\ell(t) E_{-k}(t) = 0.$$

As for the case of the even N_v, we get the following full expression for the electric energy derivative

$$\begin{aligned} \frac{\mathrm{d}W^{E}_{N_x}(t)}{\mathrm{d}t} &= L_x \sum_{k=-N_x}^{N_x} \left(\frac{\mathrm{i}L_x}{2\pi k} \Delta M_{N_v}(t) + \frac{\varepsilon^2}{\gamma^2} J_{N_v-1} c^k_{N_v-1}(t) + \sum_{\ell=0}^{N_v-2} J_\ell c^k_\ell(t) \right) E_{-k}(t) \\ &= L_x \sum_{k=-N_x}^{N_x} \left(\frac{iL_x}{2\pi k} \Delta M^k_{N_v}(t) + \frac{\varepsilon^2-\gamma^2}{\gamma^2} J_{N_v-1} c^k_{N_v-1}(t) \right) E_{-k}(t) \\ &\quad + L_x \sum_{k=-N_x}^{N_x} \sum_{\ell=0}^{N_v-1} J_\ell c^k_\ell(t) E_{-k}(t). \end{aligned}$$

At the same time, the kinetic energy is given by

$$\begin{aligned} \frac{\mathrm{d}W^{K}_{N_x,N_v}(t)}{\mathrm{d}t} &= L_x \frac{\gamma^2 N_v + \gamma^2 - \varepsilon^2}{2\gamma^2} J_{N_v-1} \beta^0_{N_v-1}(t) - \ L_x \sum_{\ell=0}^{N_v-1} J_\ell \beta^0_\ell(t) \\ = L_x \frac{\gamma^2 N_v + \gamma^2 - \varepsilon^2}{2\gamma^2} & J_{N_v-1} \sum_{k=-N_x}^{N_x} c^k_{N_v-1}(t) E_{-k}(t) - L_x \sum_{\ell=0}^{N_v-1} J_\ell \sum_{k=-N_x}^{N_x} c^k_\ell(t) E_{-k}(t) \end{aligned}$$

Therefore, the derivative of the full energy is given by

$$\begin{aligned} \frac{\mathrm{d}W_{N_x,N_v}(t)}{\mathrm{d}t} &= \frac{\mathrm{d}W^{\mathrm{K}}_{N_v,N_x}(t)}{\mathrm{d}t} + \frac{\mathrm{d}W^{\mathrm{E}}_{N_x}(t)}{\mathrm{d}t} \\ &= \frac{\mathrm{i}L_x^2}{2\pi} \sum_{k=-N_x}^{N_x} \frac{\Delta M^k_{N_v}(t)}{k} E_{-k}(t) + L_x \frac{\gamma^2 N_v + \varepsilon^2 - \gamma^2}{2\gamma^2} J_{N_v-1} \beta^0_{N_v-1}(t). \square \end{aligned}$$

15.5. L^2 norm

The only observable left to compute is the L^2 norm. Contrary to other observables, we cannot write an analytic expression for the L^2 norm right away. Let us first derive the formulas for the computation of the L^2 norm through the expansion coefficients.

Proposition 15.5.1: L^2 norm computation

Consider a function

$$f_{N_v,N_x}(x,v,t) = \sum_{k=-N_x}^{N_x} \sum_{\ell=0}^{N_v-1} c_\ell^k(t) \exp\left(\frac{2\pi \mathrm{i} k x}{L_x}\right) H_\ell^{\gamma,\varepsilon}(v),$$

with $x \in [0, L_x],\ v \in \mathbb{R},\ t \in \mathbb{R}_+$.

Then, if $\gamma = \varepsilon\sqrt{2}$, corresponding to the SW basis, the L^2 norm of the function f_{N_v,N_x} can be computed as

$$\|f_{N_v,N_x}\|_{L^2}^2 = \frac{L_x\sqrt{\pi}}{\varepsilon\sqrt{2}} \sum_{k=-N_x}^{N_x} \sum_{\ell=0}^{N_v-1} |c_\ell^k(t)|^2.$$

For all other cases, the expression for the L^2 norm writes as

$$\|f_{N_v,N_x}\|_{L^2}^2 = L_x \sum_{k=-N_x}^{N_x} \sum_{\substack{\ell_1,\ell_2=0 \\ \ell_1+\ell_2 \text{ is even}}}^{N_v-1} c_{\ell_1}^k(t) c_{\ell_2}^k(t) \langle H_{\ell_1}^{\gamma,\varepsilon}, H_{\ell_2}^{\gamma,\varepsilon}\rangle_{L^2} \tag{15.10}$$

where

$$\langle H_{\ell_1}^{\gamma,\varepsilon}, H_{\ell_2}^{\gamma,\varepsilon}\rangle_{L^2} = \frac{1}{\varepsilon\sqrt{2}} \frac{\sqrt{2^{\ell_1+\ell_2}}}{\ell_1!\ell_2!} \Gamma\left(\frac{\ell_1+\ell_2+1}{2}\right) \left(\frac{\gamma}{\varepsilon\sqrt{2}}\right)^{2\ell_2} \left(\frac{\gamma^2}{2\varepsilon^2} - 1\right)^{\frac{\ell_1-\ell_2}{2}}$$
$$\cdot\ {}_2F_1\left(-\frac{1}{2}\ell_2, \frac{1}{2}(1-\ell_2); \frac{1}{2}(1-\ell_1-\ell_2); \frac{4\varepsilon^2}{\gamma^2} - \frac{4\varepsilon^4}{\gamma^4}\right),$$

with Γ and ${}_2F_1$ being the gamma and hypergeometric functions respectively.

Proof. Consider

$$f_{N_v,N_x}(x,v,t) = \sum_{k=-N_x}^{N_x} \sum_{\ell=0}^{N_v-1} c_\ell^k(t) \exp\left(\frac{2\pi \mathrm{i} k x}{L_x}\right) H_\ell^{\gamma,\varepsilon}(v).$$

Denote the k-the Fourier basis function as

$$\phi_k(x) = \exp\left(\frac{2\pi \mathrm{i} k x}{L_x}\right).$$

Then, the L^2 norm at time t can be computed as

$$\begin{aligned}\|f_{N_v,N_x}\|_{L^2}^2 &= \left\langle \sum_{k_1=-N_x}^{N_x} \sum_{\ell_1=0}^{N_v-1} c_{\ell_1}^{k_1}(t)\phi_{k_1} H_{\ell_1}^{\gamma,\varepsilon}, \sum_{k_2=-N_x}^{N_x} \sum_{\ell_2=0}^{N_v-1} c_{\ell_2}^{k_2}(t)\phi_{k_2} H_{\ell_2}^{\gamma,\varepsilon} \right\rangle_{L^2} \\ &= \sum_{k_1,k_2=-N_x}^{N_x} \left\langle \phi_{k_1} \sum_{\ell_1=0}^{N_v-1} c_{\ell_1}^{k_1}(t) H_{\ell_1}^{\gamma,\varepsilon}, \phi_{k_2} \sum_{\ell_2=0}^{N_v-1} c_{\ell_2}^{k_2}(t) H_{\ell_2}^{\gamma,\varepsilon} \right\rangle_{L^2}\end{aligned}$$

We note that for the Fourier basis

$$\langle \phi_{k_1}, \phi_{k_2}\rangle_{L^2} = \delta_{k_1+k_2,0} L_x.$$

Therefore,

$$\|f_{N_v,N_x}\|_{L^2}^2 = L_x \sum_{k=-N_x}^{N_x} \left\langle \sum_{\ell_1=0}^{N_v-1} c_{\ell_1}^k(t) H_{\ell_1}^{\gamma,\varepsilon}, \sum_{\ell_2=0}^{N_v-1} c_{\ell_2}^{-k}(t) H_{\ell_2}^{\gamma,\varepsilon} \right\rangle_{L^2}$$
$$= L_x \sum_{k=-N_x}^{N_x} \sum_{\ell_1,\ell_2=0}^{N_v-1} c_{\ell_1}^k(t) c_{\ell_2}^{-k}(t) \langle H_{\ell_1}^{\gamma,\varepsilon}, H_{\ell_2}^{\gamma,\varepsilon}\rangle_{L^2}.$$

We are left to compute the products $\langle H_{\ell_1}^{\gamma,\varepsilon}, H_{\ell_2}^{\gamma,\varepsilon}\rangle_{L^2}$. For the SW case, i.e. when $\gamma = \varepsilon\sqrt{2}$, we get

$$\langle H_{\ell_1}^{\gamma,\varepsilon}, H_{\ell_2}^{\gamma,\varepsilon}\rangle_{L^2} = \int_{\mathbb{R}} \frac{1}{\sqrt{2^{\ell_1}\ell_1!}} \frac{1}{\sqrt{2^{\ell_2}\ell_2!}} h_{\ell_1}(\varepsilon\sqrt{2}v) h_{\ell_2}(\varepsilon\sqrt{2}v) \mathrm{e}^{-2\varepsilon^2 v^2} \mathrm{d}v$$
$$= \frac{\sqrt{\pi}}{\varepsilon\sqrt{2}} \int_{\mathbb{R}} \psi_{\ell_1}(v)\psi_{\ell_2}(v) \mathrm{d}v = \frac{\sqrt{\pi}}{\varepsilon\sqrt{2}} \delta_{\ell_1,\ell_2},$$

where ψ_ℓ denote Hermite functions. Then, for the SW setup, the L^2 norm takes the form

$$\|f_{N_v,N_x}\|_{L^2}^2 = \frac{L_x\sqrt{\pi}}{\varepsilon\sqrt{2}} \sum_{k=-N_x}^{N_x} \sum_{\ell=0}^{N_v-1} |c_\ell^k(t)|^2.$$

For the other cases, we get

$$\langle H_{\ell_1}^{\gamma,\varepsilon}, H_{\ell_2}^{\gamma,\varepsilon}\rangle_{L^2} = \frac{1}{\sqrt{2^{\ell_1+\ell_2}\ell_1!\ell_2!}} \int_{\mathbb{R}} h_{\ell_1}(\gamma v) h_{\ell_2}(\gamma v) \mathrm{e}^{-2\varepsilon^2 v^2} \mathrm{d}v$$
$$\overset{\bar{v}=\varepsilon\sqrt{2}}{=} \frac{1}{\varepsilon\sqrt{2}} \frac{1}{\sqrt{2^{\ell_1+\ell_2}\ell_1!\ell_2!}} \int_{\mathbb{R}} h_{\ell_1}\left(\frac{\gamma}{\varepsilon\sqrt{2}}\bar{v}\right) h_{\ell_2}\left(\frac{\gamma}{\varepsilon\sqrt{2}}\bar{v}\right) \mathrm{e}^{-\bar{v}^2} \mathrm{d}\bar{v}.$$

The integral of the type above has been considered in [Bai48], where multiple analytic formulas for the these integrals were derived. For the case of odd $\ell_1 + \ell_2$ the function $h_{\ell_1}(\gamma v) h_{\ell_2}(\gamma v)\mathrm{e}^{-2\varepsilon^2 v^2}$ is odd, and, therefore, its integral over $\mathbb{R}$ is zero. This means that

$$\langle H_{\ell_1}^{\gamma,\varepsilon}, H_{\ell_2}^{\gamma,\varepsilon}\rangle_{L^2} = 0 \quad \text{if} \quad \ell_1 + \ell_2 \text{ is odd.}$$

For the case when $\ell_1 + \ell_2$ is even, we use the expression (1.3) from [Bai48] with $a = b = \frac{\gamma}{\varepsilon\sqrt{2}}$ to compute the remaining integral

$$\langle H_{\ell_1}^{\gamma,\varepsilon}, H_{\ell_2}^{\gamma,\varepsilon}\rangle_{L^2} = \frac{1}{\varepsilon\sqrt{2}} \frac{\sqrt{2^{\ell_1+\ell_2}}}{\sqrt{\ell_1!\ell_2!}} \Gamma\left(\frac{\ell_1+\ell_2+1}{2}\right) \left(\frac{\gamma}{\varepsilon\sqrt{2}}\right)^{2\ell_2} \left(\frac{\gamma^2}{2\varepsilon^2} - 1\right)^{\frac{\ell_1-\ell_2}{2}}$$
$$\cdot \; {}_2F_1\left(-\frac{1}{2}\ell_2, \frac{1}{2}(1-\ell_2); \frac{1}{2}(1-\ell_1-\ell_2); \frac{4\varepsilon^2}{\gamma^2} - \frac{4\varepsilon^4}{\gamma^4}\right),$$

where the notation Γ and ${}_2F_1$ denote the gamma and hypergeometric functions, respectively. □

Even though the formula above looks massive for the non-SW case, it can be explicitly computed using built-in `MATLAB` functions `gamma` and `hypergeom`. Note that scalar products have to be computed only once at the beginning since they are not time-dependent. Moreover, even though the formula above for scalar products $\langle H^{\gamma,\varepsilon}_{\ell_1}, H^{\gamma,\varepsilon}_{\ell_2}\rangle$ is not written in a symmetric way, it still exhibits a symmetric behavior with respect to indexes ℓ_1, ℓ_2 since

$$\langle H^{\gamma,\varepsilon}_{\ell_1}, H^{\gamma,\varepsilon}_{\ell_2}\rangle_{L^2} = \langle H^{\gamma,\varepsilon}_{\ell_2}, H^{\gamma,\varepsilon}_{\ell_1}\rangle_{L^2}.$$

Therefore, we only need to compute a half of the matrix. It appears to be more robust to compute the lower triangular part of the matrix, when $\ell_1 > \ell_2$. This becomes crucial when γ approaches $\varepsilon\sqrt{2}$ (which also corresponds to the SW setup). In this case, the term $(\gamma^2/2\varepsilon^2 - 1)$ approaches zero and taking negative powers of it would yield numerical issues.

As for the conservation of the L^2 norm, it has been mentioned in [Hol96] that one can guarantee the conservation only for the SW case, where the projection operator is self-adjoint. One could derive the L^2 norm evolution based on (15.10). However, we take another approach to emphasize the structural difference between the SW method and all other GW setups. Along with other considerations, we will reproduce the proof of the conservation of the L^2 norm for the SW case and point out the crucial differences to the general case in chapter 16.

16. Error estimate of the SW solution and pitfalls of the basis asymmetry

In this chapter, we discuss the issues that arise from the fact that while the Vlasov equation is considered in the $L^2(\mathbb{R})$ space, the Galerkin projection is happening in the weighted space $L^2_\omega(\mathbb{R})$. For the conservation of mass, momentum, and energy we were still able to consider the observables in the $L^2(\mathbb{R})$ space anyway. However, some of the other results do not translate as seamlessly. For example, another important observable is the L^2 norm of the solution. If the norm is conserved, that guarantees that the numerical solution will not "explode" and produce extremely large values. It was already mentioned in [Hol96, § 3.1] that we can only guarantee the L^2 norm conservation for the SW method. One potential alternative could be to consider the weighted L^2 norm. However, due to the properties of the differential operator L of the Vlasov equation, the weighted L^2 norm need not be conserved and, therefore, should not be used as an observable. We reproduce the proof of the L^2 norm conservation for the SW case and discuss the discrepancies caused by the asymmetry in the basis in section 16.1.

Another illustration of the impact of the weight on the projection and the final result is presented in section 16.2. There, we derive an error estimate for the Hermite functions (or SW) solution of the advection equation based on the ladder operator theory. Even though we have derived the analogous theory for the case of the weighted basis in sections 3.4 and 3.5, it turns out to not be possible to derive the analogous error estimate. The reason for that is that certain properties of the original differential operator do not hold in the weighted space, while holding in the $L^2(\mathbb{R})$ space.

16.1. L^2 norm conservation

The conservation of the L^2 norm is crucial for the numerical stability of the method. Note that in theory, the Vlasov–Poisson system preserves all L^p norms. The SW basis functions correspond to the GW basis functions with $\gamma = \varepsilon\sqrt{2}$ and are orthogonal in $L^2(\mathbb{R})$ where the Vlasov equation is considered. The corresponding weight then takes the form

$$\omega_{\mathrm{SW}}(v) = \pi^{-1/2}\varepsilon\sqrt{2}.$$

We note that the scalar product with this weighting corresponds to the scaled scalar product in the $L^2(\mathbb{R})$ space. Indeed, for all $f, g \in L^2_{\omega_{\mathrm{SW}}}(\mathbb{R})$, we get

$$\langle f, g\rangle_{\omega_{\mathrm{SW}}} = \int_{\mathbb{R}} f(v)g(v)\pi^{-1/2}\varepsilon\sqrt{2}\mathrm{d}v = \pi^{-1/2}\varepsilon\sqrt{2}\langle f, g\rangle_{L^2}.$$

It is also clear that if $f \in L^2_{\omega_{\mathrm{SW}}}(\mathbb{R})$, then $f \in L^2(\mathbb{R})$. Let us now prove that the L^2 norm is conserved for the SW case.

Proposition 16.1.1: L^2 norm conservation for SW discretization

Let $f_{N_v}(x, v, t)$ be a Galerkin approximation of the distribution function f in the SW basis. In particular,

$$f_{N_v}(x, v, t) = \sum_{\ell=0}^{N_v-1} c_\ell(x, t) H_\ell^{\varepsilon\sqrt{2},\varepsilon}(v),$$

and the coefficients $\{c_\ell(x, t)\}$ satisfy equations (14.9). Then,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}} \int_0^{L_x} f_{N_v}(x, v, t)^2 \mathrm{d}x\mathrm{d}v = 0.$$

This implies that the Galerkin discretization (14.9) preserves the L^2 norm.

Proof. The Vlasov equation can be written in the following form:

$$\frac{\partial f(x, v, t)}{\partial t} = L(x, v, t) f(x, t),$$

where L is a differential operator:

$$L = -v\frac{\partial}{\partial x} + E(x, t)\frac{\partial}{\partial v}.$$

Let us prove via integration by parts that the operator L has the following property:

$$\langle f, Lf \rangle_{L_2} = 0,$$

for all f periodic in x and square-integrable in v. We first write the explicit expression for $\langle f, Lf \rangle_{L_2}$. We get

$$\langle f, Lf \rangle_{L^2} = \int_{\mathbb{R}} \int_0^{L_x} f(x, v, t) \left(-v\frac{\partial f(x, v, t)}{\partial x} + E(x, t)\frac{\partial f(x, v, t)}{\partial v} \right) \mathrm{d}x\mathrm{d}v.$$

We first consider the second term of the sum. Due to the square integrability in v, we get

$$\int_{\mathbb{R}} f(x, v, t)\frac{\partial f(x, v, t)}{\partial v}\mathrm{d}v = -\int_{\mathbb{R}} \frac{\partial f(x, v, t)}{\partial v} f(x, v, t)\mathrm{d}v.$$

Hence

$$\int_{\mathbb{R}} f(x, v, t)\frac{\partial f(x, v, t)}{\partial v}\mathrm{d}v = 0.$$

In the same fashion one can prove that the first term also integrates to zero in x using the fact that f is periodic in x.

The Galerkin approximation of the equation of such form can be written as [GO77, § 2, Eq. (2.5)]:

$$\frac{\partial f_{N_v}}{\partial t} = L_{N_v} f_{N_v} \quad \text{with} \quad L_{N_v} = P_{N_v} L P_{N_v},$$

where P_{N_v} is the Galerkin projection. Using the fact that the Galerkin projection operator is self-adjoint in the corresponding space [GO77, § 2], we write:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}}\int_0^{L_x} f_{N_v}(x,v,t)^2\mathrm{d}x\mathrm{d}v = \frac{\pi^{1/2}}{\varepsilon\sqrt{2}}\frac{\mathrm{d}}{\mathrm{d}t}\langle f_{N_v}, f_{N_v}\rangle_{\omega_{\mathrm{SW}}}$$
$$= \frac{\pi^{1/2}}{\varepsilon\sqrt{2}}\langle f_{N_v}, P_{N_v}LP_{N_v}f_{N_v}\rangle_{\omega_{\mathrm{SW}}} = \frac{\pi^{1/2}}{\varepsilon\sqrt{2}}\langle P_{N_v}f_{N_v}, LP_{N_v}f_{N_v}\rangle_{\omega_{\mathrm{SW}}}$$
$$= \langle P_{N_v}f_{N_v}, LP_{N_v}f_{N_v}\rangle_{L^2} = 0. \qquad \square$$

The proof above relies on the two major properties:

$$\langle f, Lf\rangle_{L^2} = 0 \quad \text{and} \quad \langle P_{N_v}f, g\rangle_{\omega_{\mathrm{SW}}} = \langle f, P_{N_v}g\rangle_{\omega_{\mathrm{SW}}}.$$

If we take the weighted Galerkin projection with a non-constant weight instead, which is the case for all setups except from the SW one, we cannot transfer from the L^2 norm of f_{N_v} to the weighted one by means of multiplication with a constant. On the other hand, if we stay in the $L^2(\mathbb{R})$ space, the adjointness of the Galerkin operator will be lost.

If we consider the weighted L^2 norm of f_{N_v} instead, the relevant property of the operator L is lost. In particular,

$$\langle f, Lf\rangle_\omega = \int_{\mathbb{R}}\int_0^{L_x} f(x,v,t)\left(-v\frac{\partial f(x,v,t)}{\partial x} + E(x,t)\frac{\partial f(x,v,t)}{\partial v}\right)\omega(v)\mathrm{d}x\mathrm{d}v \neq 0.$$

Indeed, even though the first term of the sum will not be affected by the addition of the weight and will still integrate to zero due to the periodicity considerations, the second part loses its symmetry property even for $f \in L^2_\omega(\mathbb{R})$

$$\int_{\mathbb{R}} f(x,v,t)\frac{\partial f(x,v,t)}{\partial v}\omega(v)\mathrm{d}v = -\int_{\mathbb{R}} f(x,v,t)\frac{\partial(f(x,v,t)\omega(v))}{\partial v}\mathrm{d}v$$
$$= -\int_{\mathbb{R}}\left(f(x,v,t)\frac{\partial f(x,v,t)}{\partial v}\omega(v) + f(x,v,t)^2\frac{\partial\omega(v)}{\partial v}\right)\mathrm{d}v,$$

where the boundary term vanishes for $f \in L^2_\omega(\mathbb{R})$. Therefore, we cannot guarantee neither the conservation of the standard L^2 norm, nor the weighted one.

16.2. Error estimate for the SW solution of the advection equation

In this section, we derive a Galerkin error estimate for the pure velocity approximation. We simplify our model and consider a simple 1D advection equation instead

$$\frac{\partial f(v,t)}{\partial t} - a\frac{\partial f(v,t)}{\partial v} = 0, \quad a \in \mathbb{R}. \tag{16.1}$$

to investigate the error introduced by the velocity discretization only. We first prove an error estimate of the Galerkin solution of the advection equation for the Hermite functions discretization. For that, we employ the theory of ladder operators of Hermite functions (see section 2.3). We use the pure Hermite functions, corresponding to the SW basis without scaling, to be consistent with the result [Lub08, § III.1.1, Theorem 1.3] we base our derivation on. However, it can be easily included in the proof by adding appropriate scaling factors. After deriving the error estimate for the Hermite functions case, we discuss the reasons why the analogous theory for the weighted space (see section 3.5) does not yield the desired outcome.

Let us first note that the equation (16.1) can be written in terms of differentiation operators as

$$\frac{\partial f}{\partial t} = Lf,$$

where $L = a\frac{\partial}{\partial v}$. We recall from section 16.1, that the following property holds for this operator

$$\langle f, Lf\rangle_{L^2} = 0 \tag{16.2}$$

for square-integrable functions f. As in the proof of the L^2-error conservation, this property will play a major role in the derivation of the Galerkin error.

Lemma 16.2.1: SW solution error estimate for the advection equation

Consider the advection equation (16.1) and the numerical solution f_{N_v} obtained via Galerkin method with Hermite functions basis $\{\psi_0, \ldots, \psi_{N_v-1}\}$. Then, if $f_{N_v}(0) = P_{N_v} f(0)$ and the exact solution $f \in D(A^{s+1})$ for some $s \leq N_v - 1$, the error can be estimated as

$$\|f - f_{N_v}\|_{L^2} \leq \frac{\left(1 + at\sqrt{N_v/2}\right)}{\sqrt{N_v(N_v-1)\ldots(N_v-s)}} \max_{0\leq\tau\leq t} \|A^{s+1} f(\tau)\|_{L^2}, \tag{16.3}$$

where $\|\cdot\|_{L^2}$ denotes the $L^2(\mathbb{R})$ norm and A is the lowering operator (2.11) for Hermite functions.

Proof. We follow the general structure of the proof [Lub08, § III.1.1, Theorem 1.3] of a similar estimate for the Schrödinger equation. As before, the Galerkin approximation of the advection equation can be written as

$$\frac{\partial f_{N_v}}{\partial t} = P_{N_v} L P_{N_v} f_{N_v},$$

where $L = a\frac{\partial}{\partial v}$. Applying P_{N_v} to the advection equation (16.1), we get

$$P_{N_v}\frac{\partial f}{\partial t} = P_{N_v} L P_{N_v}(P_{N_v} f + P_{N_v}^{\perp} f) = P_{N_v} L P_{N_v} P_{N_v} f + P_{N_v} L P_{N_v}^{\perp} f,$$

where $P_{N_v}^{\perp} = I - P_{N_v}$ with I being the identity operator. Hence,

$$\frac{\partial}{\partial t}(f_{N_v} - P_{N_v} f) = P_{N_v} L P_{N_v}(f_{N_v} - P_{N_v} f) - P_{N_v} L P_{N_v}^{\perp} f.$$

and

$$
\begin{aligned}
&\left\langle \frac{\partial}{\partial t}(f_{N_v} - P_{N_v} f), f_{N_v} - P_{N_v} f \right\rangle_{L^2} &(16.4)\\
&= \left\langle P_{N_v} L P_{N_v}(f_{N_v} - P_{N_v} f) - P_{N_v} L P_{N_v}^{\perp} f, f_{N_v} - P_{N_v} f \right\rangle_{L^2}\\
&= \langle L P_{N_v}(f_{N_v} - P_{N_v} f), P_{N_v} f_{N_v} - P_{N_v} f \rangle_{L^2} - \left\langle P_{N_v} L P_{N_v}^{\perp} f, f_{N_v} - P_{N_v} f \right\rangle_{L^2}\\
&\le \|f_{N_v} - P_{N_v} f\|_{L^2} \cdot \|P_{N_v} L P_{N_v}^{\perp} f\|_{L^2}, &(16.5)
\end{aligned}
$$

where we used the property (16.2) for the operator L.

On the other hand,

$$
\begin{aligned}
\left\langle \frac{\partial}{\partial t}(f_{N_v} - P_{N_v} f), f_{N_v} - P_{N_v} f \right\rangle_{L^2} &= \frac{1}{2}\frac{\partial}{\partial t}\|f_{N_v} - P_{N_v} f\|_{L^2}^2\\
&= \|f_{N_v} - P_{N_v} f\|_{L^2} \frac{\partial}{\partial t}\|f_{N_v} - P_{N_v} f\|_{L^2}.
\end{aligned}
$$

Therefore,

$$
\frac{\partial}{\partial t}\|f_{N_v} - P_{N_v} f\|_{L^2} \le \|P_{N_v} L P_{N_v}^{T} f\|_{L^2}
$$

and

$$
\|f_{N_v} - P_{N_v} f\|_{L^2} \le \int_0^t \|P_{N_v} L P_{N_v}^{\perp} f\|_{L^2} \mathrm{d}t + \|f_{N_v}(0) - P_{N_v} f(0)\|_{L^2}.
$$

Let us now estimate $\|P_{N_v} L P_{N_v}^{\perp} f\|$. Denote the coefficients of the function f in the Hermite basis before the cut-off as $\{c_\ell\}_{\ell=0}^{\infty}$. Then,

$$
P_{N_v}^{\perp} f = \sum_{\ell=N_v}^{\infty} c_\ell \psi_\ell.
$$

Then, using the property (3.6) of Hermite functions, we get

$$
L P_{N_v}^{\perp} f = a \frac{\partial}{\partial v} \sum_{\ell=N_v}^{\infty} c_\ell \psi_\ell = a \sum_{\ell=N_v}^{\infty} c_\ell \left(\sqrt{\frac{\ell}{2}} \psi_{\ell-1} - \sqrt{\frac{\ell+1}{2}} \psi_{\ell+1} \right).
$$

Therefore,

$$
P_{N_v} L P_{N_v}^{\perp} f = a c_{N_v} \sqrt{\frac{N_v}{2}} \psi_{N_v - 1}.
$$

To estimate the coefficient c_{N_v}, we employ the same strategy as in Theorem 2.3.1

$$
\begin{aligned}
c_{N_v} = \langle f, \psi_{N_v} \rangle_{L^2} &= \frac{\langle f, (A^\dagger)^{s+1} \psi_{\ell-s-1} \rangle_{L^2}}{\sqrt{N_v (N_v - 1) \dots (N_v - s)}} = \frac{\langle A^{s+1} f, \psi_{\ell-s-1} \rangle_{L^2}}{\sqrt{N_v (N_v - 1) \dots (N_v - s)}}\\
&\le \frac{1}{\sqrt{N_v (N_v - 1) \dots (N_v - s)}} \|A^{s+1} f\|_{L^2},
\end{aligned}
$$

where $A^\dagger$, A are Hermite ladder operators (see section 2.3). At the same time, Theorem 2.3.1 provides an estimate for $\|f - P_{N_v} f\|_{L^2}$

$$
\|f - P_{N_v} f\|_{L^2} \le \frac{1}{\sqrt{N_v (N_v - 1) \dots (N_v - s)}} \|A^{s+1} f\|_{L^2}
$$

Therefore,

$$\begin{aligned}\|f - f_{N_v}\|_{L^2} &\leq \|f - P_{N_v} f\|_{L^2} + \|P_{N_v} f - f_{N_v}\|_{L^2} \\ &\leq \frac{\left(1 + at\sqrt{N_v/2}\right)}{\sqrt{N_v(N_v - 1)\dots(N_v - s)}} \max_{0\leq\tau\leq t} \|A^{s+1} f(\tau)\|_{L^2}. \qquad \square\end{aligned}$$

As in the case of the L^2 norm conservation, analogous proof is impossible for the case of a weighted basis. Indeed, if we consider the scalar products in L^2 space, then the Galerkin projection becomes non-adjoint, making the second equality of (16.5) invalid. If we consider the $L^2_\omega(\mathbb{R})$ norm instead, then the third inequality of (16.5) does not hold, since

$$\langle f, Lf \rangle_\omega \neq 0.$$

This proof illustrates once again that asymmetry in the basis introduces additional factors to the structural properties of the method and should be treated carefully. One of the possible ways to deal with this discrepancy could be to use variation of constants method, also referred to as the Duhamel's principle. However, this is more complicated and requires further investigation.

Manzini, Funaro, & Delzanno provide in [MFD17] an analysis of the L^2 norm conservation for the SW method for the Vlasov–Poisson system for the case of the bounded velocity domain. The authors of [MFD17] mention that some of the convergence results could be transferred to the AW case, however, to our best knowledge, that has not yet been developed.

17. Symmetrically and asymmetrically weighted Fourier–Hermite methods

Now, when we have derived and analyzed the GW Hermite method, or the generalized Fourier–Hermite method, let us come back to the original two methods: symmetrically weighted (SW) Hermite method and asymmetrically weighted (AW) Hermite method. We write again the general formula for the generalized anisotropic Hermite basis used for the GW Hermite method

$$H^{\gamma,\varepsilon}(v) = \frac{1}{\sqrt{2^\ell \ell!}} h_\ell(\gamma v) \mathrm{e}^{-\varepsilon^2 v^2}.$$

The SW Hermite method then corresponds to

$$\gamma = \varepsilon\sqrt{2}.$$

for the GW method. It is a discretization with standard Hermite functions and the weight is constant in this case

$$\omega(v) = \gamma \pi^{-1/2} = \pi^{-1/2}\varepsilon\sqrt{2}.$$

The only difference of the GW method with these parameters to the standard SW Hermite method is the constant $\pi^{-1/4}$ in the basis. However, it does not play a major role in the Galerkin setup. Note that the scaling of the argument of Hermite functions, which form the SW basis, is already included in this representation since we only consider the ratio of the γ and ε and not the precise values.

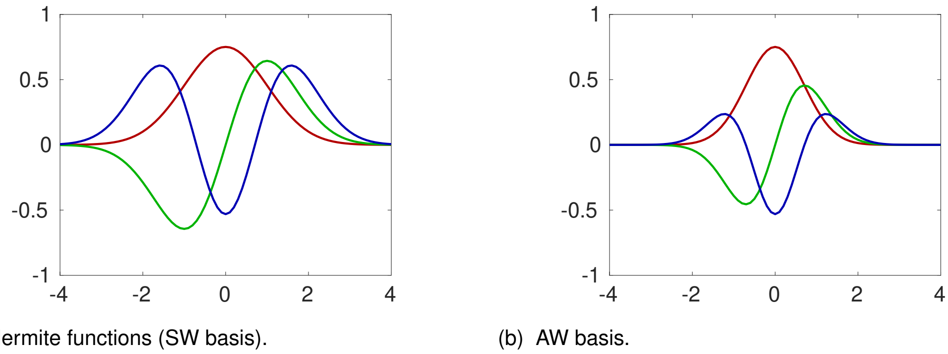
Similarly, the AW Hermite method in its general form, with the scaling of the argument included, can be obtained by setting

$$\gamma = \varepsilon.$$

In this case, the weight takes the form

$$\omega(v) = \gamma \pi^{-1/2} \mathrm{e}^{\varepsilon^2 v^2} = \gamma \pi^{-1/2} \mathrm{e}^{\gamma^2 v^2}.$$

A special property of this particular case of the generalized anisotropic Hermite



(a) Hermite functions (SW basis).

(b) AW basis.

Figure 19 One can see that the AW basis implies way more decay than regular Hermite functions.

basis is that, according to our formulas (14.3), (14.5)

$$I_\ell = \bar{I}_\ell = 0 \quad \text{for} \quad \ell \geq 1.$$

and, due to (14.4)

$$J_\ell = 0 \quad \text{for} \quad \ell \geq 2.$$

This means that in this case mass, momentum, and energy are always exactly conserved. However, one can see on Figure 19 that the AW basis implies way more decay than the standard Hermite functions. Some adjustments could be done via scaling. However, the width of the Gaussian ε is usually determined by minimizing the error in the initial distribution representation. Therefore, in practice, the scaling is usually fixed for a specific problem. Then, the question remains what the optimal decay of the functions is. As mentioned in [Hol96], the SW basis is more stable, since it always preserves the L^2 norm. However, maybe there exists another combination of parameters ε and γ that allows to have better conservation than the SW basis, but more stability than the AW basis. With that in mind, we now move to numerical experiments to investigate intermediate setups.

18. Numerical results

The generally weighted (GW) Hermite method has been implemented in `MATLAB`. The parameter ε of the generalized Hermite basis, corresponding to the width of the Gaussians, is chosen based on the initial conditions of the corresponding test cases. The parameter γ, however, is left free and its influence is studied. From chapter 17 we know that in this case the special cases of the AW and SW methods correspond to

$$\gamma_{\mathsf{AW}} = \varepsilon, \quad \gamma_{\mathsf{SW}} = \varepsilon\sqrt{2}.$$

For the time integration we have used the explicit Runge-Kutta method of order 4 with the time step

$$\Delta t_{\mathsf{RK}} = 0.001,$$

which proved to be sufficient to investigate the conservation properties of our new spatial discretization.

A general observation for all test cases is that the method gets less stable the further we go away from the SW case. However, the conservation properties are better closer to the AW case. It turns out, that an intermediate value of γ might be a good solution, providing an improved conservation, compared to the SW method, and better stability as opposed to the AW case.

18.1. Test cases

Let us first review the three testing setups that we will use for our experiments. In particular, we consider Landau damping, two stream instability, and bump-on-tail initial distributions. In the following sections we discuss how we initialize the simulation for these scenarios.

18.1.1. Landau damping

The initial distribution for this case takes the following form

$$f_0(x,v) = \frac{1}{\sqrt{2\pi}}(1 + \alpha\cos(k_0 x))\mathrm{e}^{-v^2/2}, \quad x \in \left[0, \frac{2\pi}{k_0}\right], \quad v \in \mathbb{R}.$$

For our numerical experiments, we take $\alpha = 0.05$ and $k = 0.5$, as this setup has also been used in [CDBM16, § 4.2]. As for the other parameters, we set

- $N_v = 33$ or $N_v = 32$,
- $N_x = 32$ (total number of 65 modes).

For the verification of the results, we compare the damping rate in the electric energy to the one provided by linear theory [Son17, § 4.4.2]. In particular, for $k = 0.5$, the damping rate of the electric field is -0.1533 and, therefore, the electric energy is damped by the factor of -0.3066.

We know from section 14.5 that in this case, the initial distribution can be reproduced exactly in the Fourier–Hermite basis. For that, we set

$$\varepsilon = \frac{1}{\sqrt{2}}$$

for the Hermite basis in order to match the width of the Gaussian. Then, the initial values of the coefficients are given by

$$c_0^0(0) = \frac{1}{\sqrt{2\pi}}, \quad c_0^1(0) = c_0^{-1}(0) = \frac{0.025}{\sqrt{2\pi}}.$$

18.1.2. Two stream instability

We now consider another standard test case, the two stream instability, corresponding to the following initial distribution

$$f_0(x,v) = \frac{1}{2\sqrt{2\pi}}(1+\alpha\cos(k_0 x))\left(\mathrm{e}^{-\frac{(v-v_0)^2}{2}} + \mathrm{e}^{-\frac{(v+v_0)^2}{2}}\right), \quad x \in \left[0, \frac{2\pi}{k_0}\right], \quad v \in \mathbb{R},$$

with $\alpha = 0.001$, $k = 0.2$ and $v_0 = 3$. As for the other parameters, we set

- $N_v = 65$ or $N_v = 64$,
- $N_x = 32$ (total number of 65 modes).

In this case, the linear theory predicts the growth rate of 0.569 (see [Son17, § 4.4.2]). In order to set up the simulation, we need to find the representation of f_0 in the Fourier–Hermite basis. We set

$$\varepsilon = \frac{1}{\sqrt{2}}$$

that matches the width of the Gaussians. As in section 14.5, we observed that f_0 is of the form of the Gaussian RBF interpolant (4.1) that we considered in part II. Therefore, we have the initial approximation

$$\mathrm{e}^{-\frac{(v-v_0)^2}{2}} + \mathrm{e}^{-\frac{(v+v_0)^2}{2}} \approx \sum_{\ell=0}^{N_v-1} c_\ell H^{\gamma,\varepsilon}(v)$$

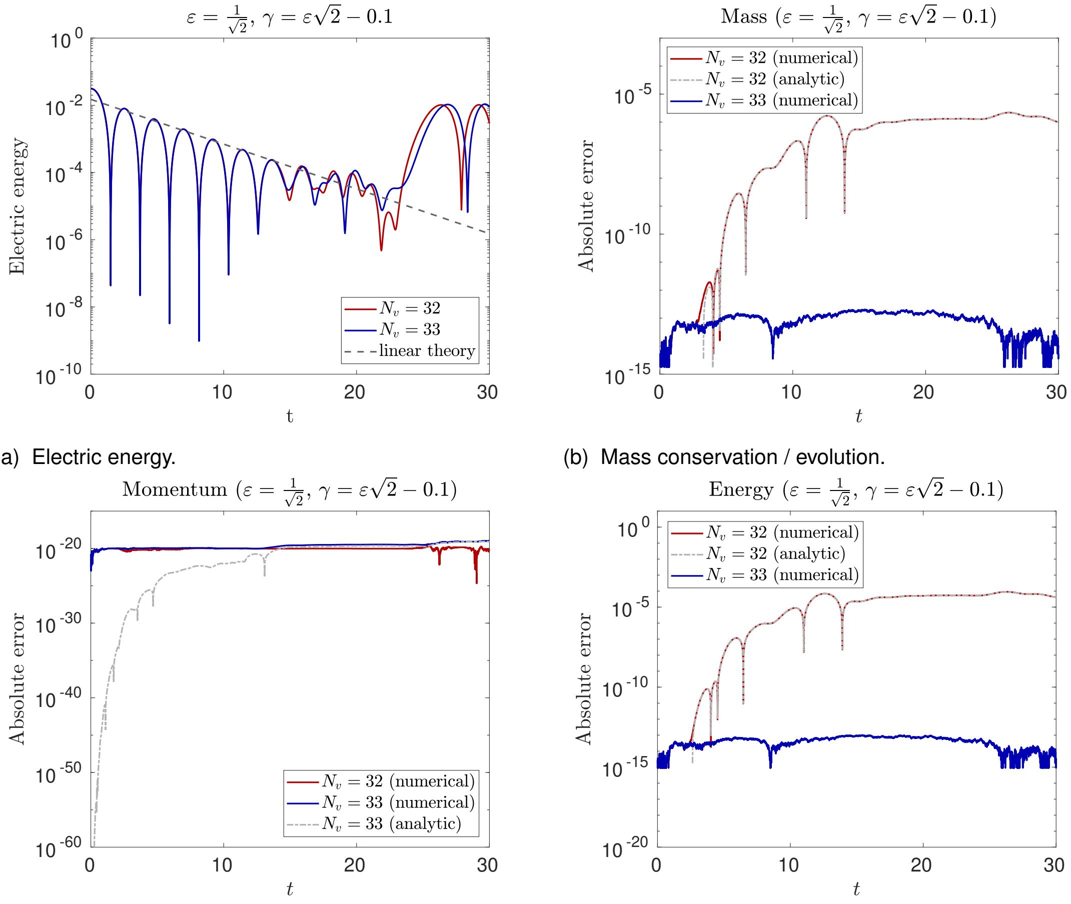
with

$$c_\ell = \begin{cases} 2\exp\left(\varepsilon^2 v_0^2\left(\frac{\varepsilon^2}{\gamma^2}-1\right)\right)\frac{\varepsilon^{2\ell}\sqrt{2^\ell}}{\gamma^\ell\sqrt{\ell!}}v_0^\ell & \text{for even } \ell \\ 0 & \text{for odd } \ell. \end{cases}$$

In the x direction, the perturbation $(1+\alpha\cos(kx))$ can be represented in the Fourier basis as before, with coefficients

$$c^0 = 1, \quad c^1 = c^{-1} = \frac{\alpha}{2}.$$

In the code, we take the tensor product of the coefficients $\{c_\ell\}$ and $\{c^k\}$ scaled with the factor of $2\sqrt{2\pi}$ as the initial representation of f_0 in our basis.



(a) Electric energy.

(b) Mass conservation / evolution.

(c) Momentum conservation / evolution.

(d) Energy conservation / evolution.

Figure 20 Landau damping. The GW method recovers the damping rate predicted by linear theory both with even and odd number of basis functions in velocity. Mass and energy are preserved for the odd N_v, whereas the momentum for the even N_v. The deviation in the non-preserving cases matches the analytic prediction.

18.1.3. Bump-on-tail instability

Finally, we look at the, so-called, “bump-on-tail” test case corresponding to the following initial distribution function

$$f_0(x,v) = \frac{1}{\sqrt{2\pi}}(1+\alpha\cos(kx))\left(\frac{\delta}{\sigma_1}\mathrm{e}^{-\frac{v^2}{2\sigma_1^2}} + \frac{1-\delta}{\sigma_2}\mathrm{e}^{-\frac{(v-v_b)^2}{2\sigma_2^2}}\right)$$

with

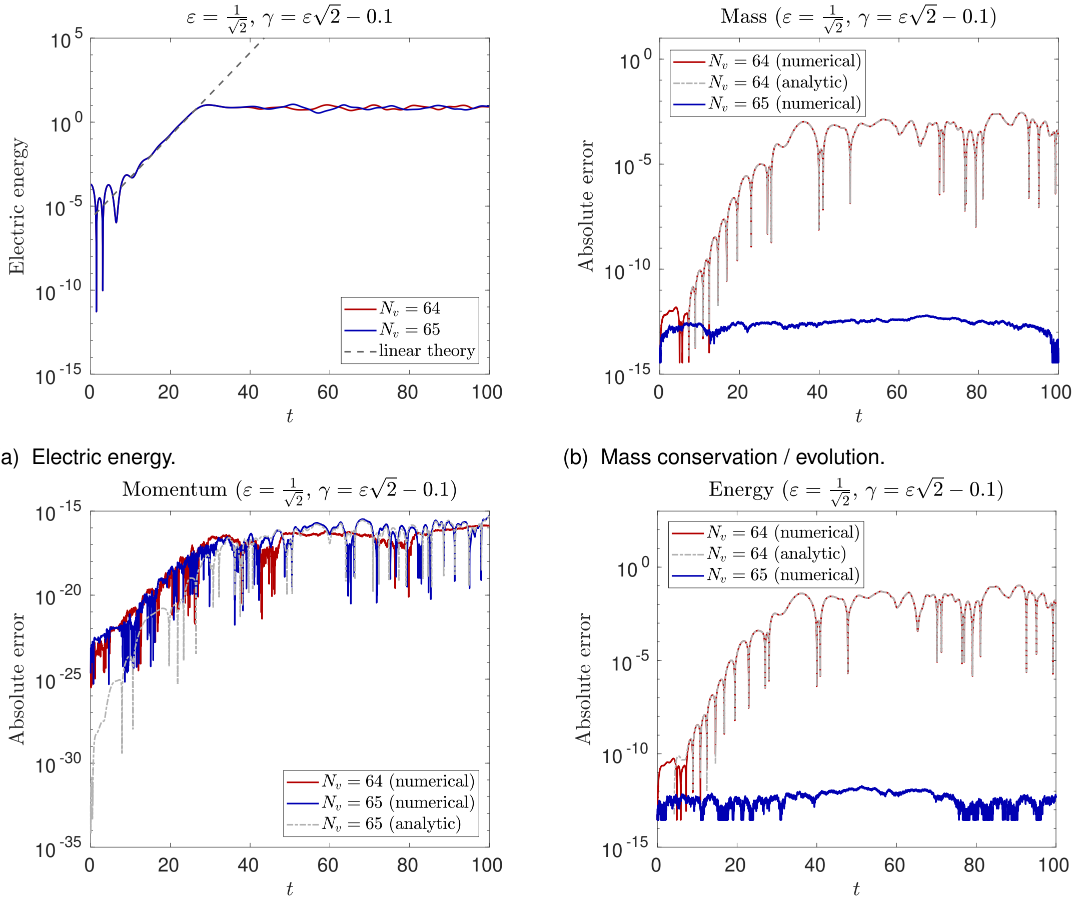
$$k = 0.3, \quad \alpha = 0.03, \quad \delta = \frac{9}{10}, \quad \sigma_1 = 1, \quad \sigma_2 = \frac{1}{\sqrt{2}}, \quad v_b = 4.5.$$

As for the other parameters, we set

- $N_v = 65$ or $N_v = 64$,
- $N_x = 32$ (total number of 65 modes).

For the initialization we use the result of Proposition 14.5.1 with the corresponding parameters.

As before, in the x direction, the perturbation $(1+\alpha\cos(kx))$ can be repre-



(a) Electric energy.

(b) Mass conservation / evolution.

(c) Momentum conservation / evolution.

(d) Energy conservation / evolution.

Figure 21 Two stream instability. The GW method recovers the growth rate predicted by linear theory both with even and odd number of basis functions in velocity. Mass and energy are preserved for the odd N_v, whereas the momentum for the even N_v. The deviation in the non-preserving cases matches the analytic prediction.

sented in the Fourier basis, with coefficients

$$c^0 = 1, \quad c^1 = c^{-1} = \frac{\alpha}{2}.$$

Once the coefficients in velocity are computed, we can get the initial representation of f_0 by taking a tensor product with the coefficients in x.

We set the width of the Gaussian to

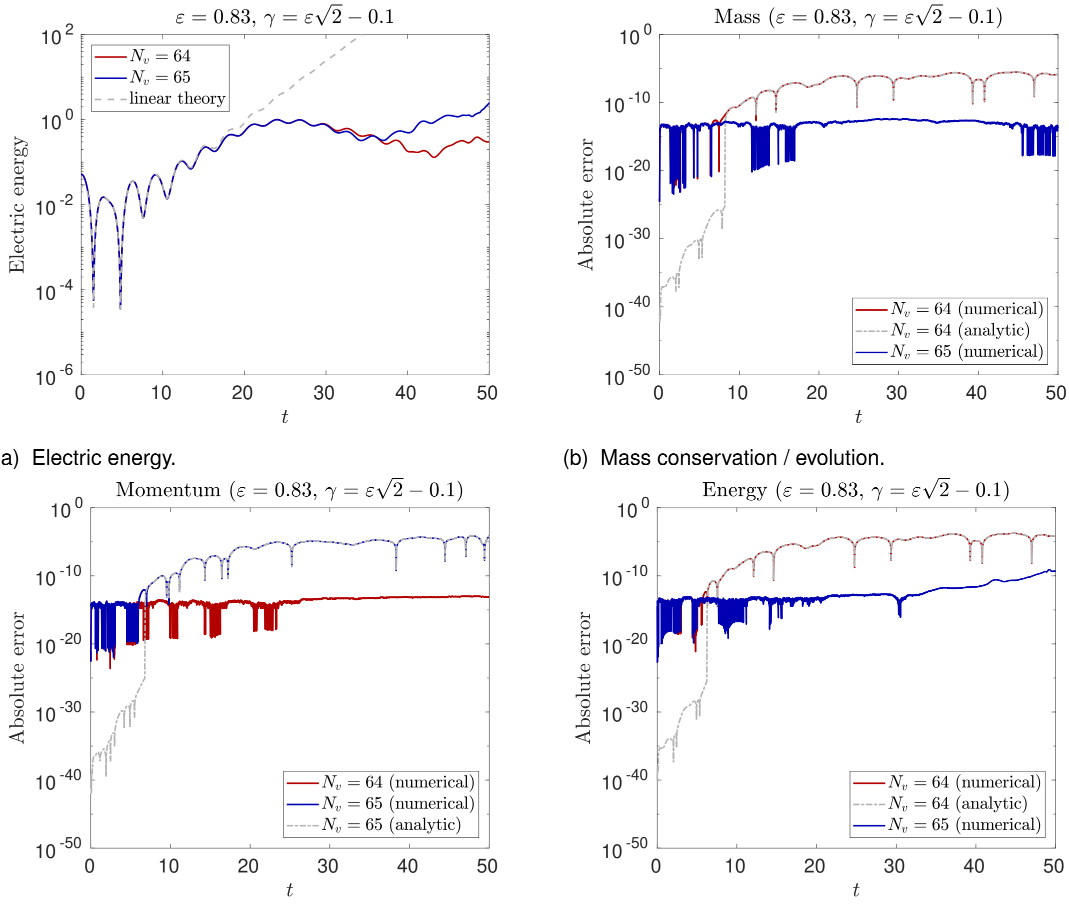
$$\varepsilon = 0.83,$$

which proved to give good results in this case.

18.2. Validation of the method for an intermediate γ

For our first test we consider an intermediate between the SW and AW setups value of

$$\gamma = \sqrt{2}\varepsilon - 0.1$$



(a) Electric energy.

(b) Mass conservation / evolution.

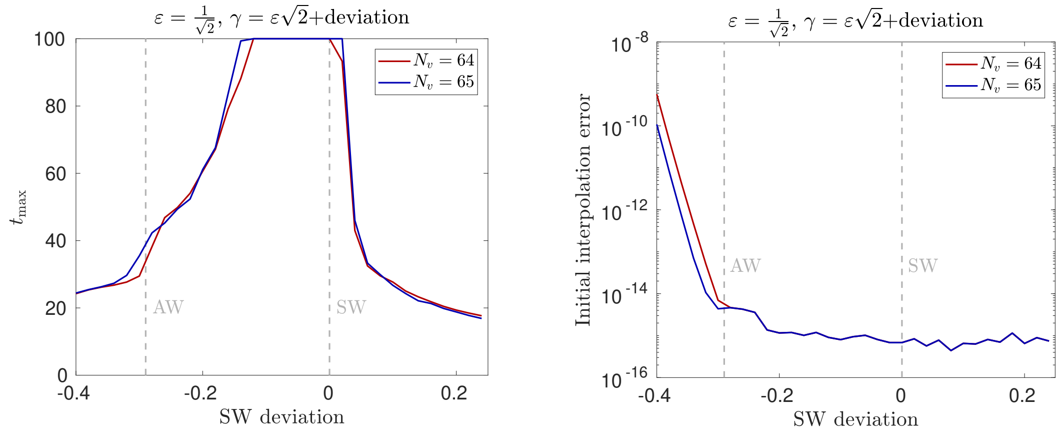
(c) Momentum conservation / evolution.

(d) Energy conservation / evolution.

Figure 22 Bump-on-tail instability. Mass and energy are preserved for the odd N_v, whereas the momentum for the even N_v. The deviation in the non-preserving cases matches the analytic prediction.

in order to check whether the generalized method works as expected. We verify that the electric energy matches growth (or decay) rates obtained from the linear theory. Moreover, we check if the error in the conservation of observables matches the analytic formulas derived in chapter 15. To obtain the predicted values, we evolved in time the corresponding explicit expressions (15.5), (15.7) and (15.9) with the 4th order Runge-Kutta method with the same time step $\Delta t = 0.001$.

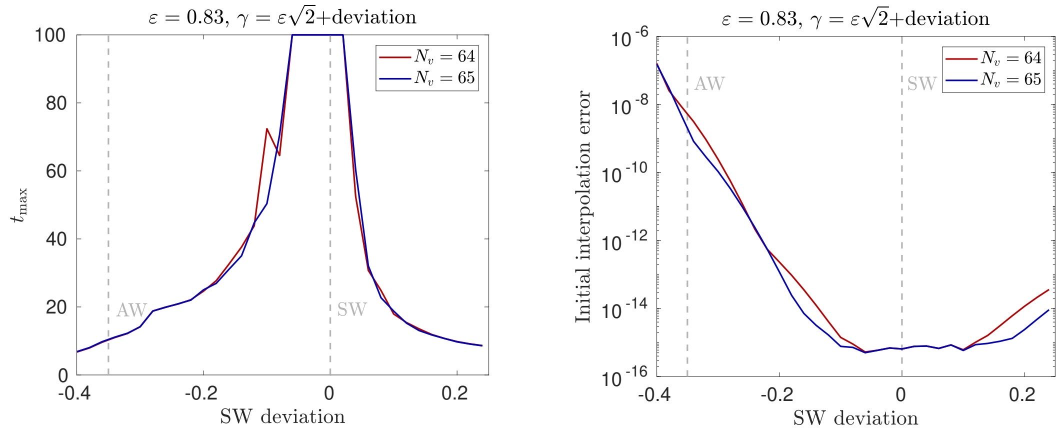
For the Landau damping and two stream instability test cases, one can see in Figure 20a that both odd an even number of basis functions in velocity allow to reproduce the damping rate provided by the linear theory. However, the conservation properties vary significantly between the two. As predicted by the theory, mass and energy are preserved for the case of odd number of basis functions (see Figure 20b, Figure 20d). As for the momentum, it is preserved slightly better for the even number of basis functions (see Figure 20c). However, in these particular test cases the magnitude of the momentum deviation are extremely small even for the case of odd number of basis functions. For the bump-on-tail test



(a) Maximum time reached.

(b) Initial interpolation error.

Figure 23 Two stream instability. One can see that the setups close to the SW case demonstrate better stability. Also, there is no direct correlation between the instability an the error in the initial representation.



(a) Maximum time reached.

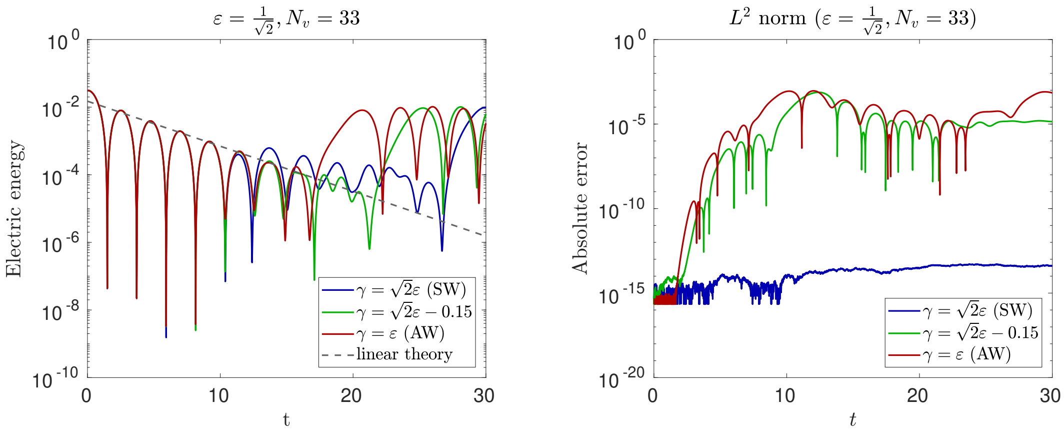
(b) Initial interpolation error.

Figure 24 Bump-on-tail instability. One can see that the setups close to the SW case demonstrate better stability. Also, there is no direct correlation between the instability an the error in the initial representation.

case, the phase of the linear growth is quite short, so we compare the electric energy obtained in our code to the electric energy obtained from a linear dispersion analysis (see [Mur18]). One can see in Figure 22a that our electric energy matches the one provided by the linear theory. Another difference of the bump-on-tail test case is that the error in the momentum conservation goes away from zero for the odd N_v. The error in the conservation of the observables obtained from the numerical simulation matches the analytic results.

18.3. L^2 stability and quality of the solution

We now investigate how the parameter γ influences the method. We first take a look at the Landau damping test case. The recurrence appearing in the electric field is a purely numerical effect and affects most of the grid-based numerical methods. However, in Figure 20a it appears rather early in time. This effect has



(a) Electric energy for different γ.

(b) L^2 norm conservation / evolution.

Figure 25 Landau damping. The choice of γ corresponding to the SW method yields exact L^2 norm conservation and delays the recurrence in the electric field the most. For the AW setup and the intermediate case, the L^2 norm is not conserved. Moreover, the further the setup is from the SW one, the earlier the recurrence appears.

been also mentioned in [CDBM16, § 4.2] where it was proposed to introduce an artificial collision term to overcome this problem. However, since the GW method allows for the variation of γ, we now investigate whether the recurrence can be delayed by changing γ. In Figure 25 we see that the choice of the SW basis delays the recurrence the most, however a slight deviation of the electric field from the linear prediction appears earlier. The choice of the AW basis, even though in this case the conservation of mass, momentum, and energy are guaranteed, yields the earliest recurrence. The intermediate case of $\gamma = \varepsilon\sqrt{2} - 0.15$ preserves the exact damping rate the longest with the recurrence appearing in between of the SW and AW cases. This indicates that in certain situations the choice of the intermediate γ might be advantageous. As for the L^2 norm conservation, it is significantly worse for the cases other than SW, where it is always fulfilled.

We now consider the two stream and bump-on-tail instabilities and look how long a simulation can run without serious stability issues depending on the value of γ. As we have proved in section 16.1, the L^2 norm is preserved for the SW case, i.e. when

$$\gamma = \varepsilon\sqrt{2}.$$

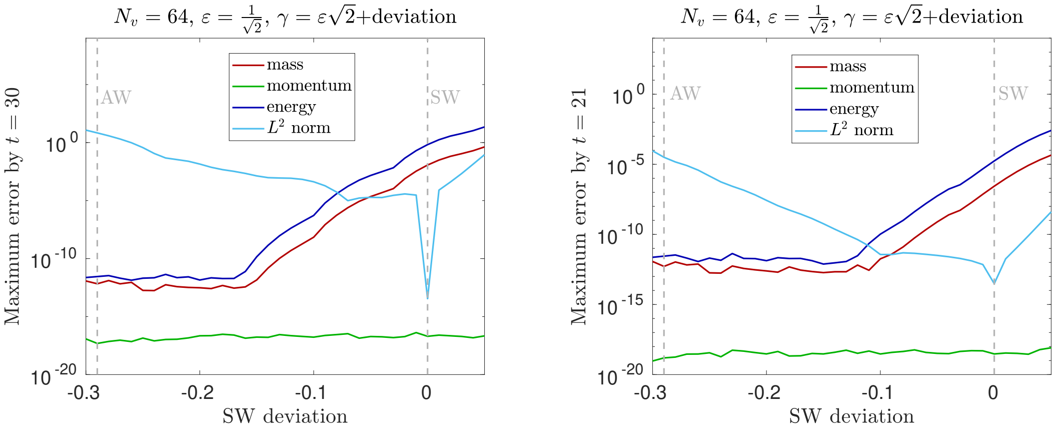
We vary the deviation from the SW case and take a look at

$$\gamma \in [\varepsilon\sqrt{2} - 0.4, \varepsilon\sqrt{2} + 0.25].$$

Recall that the AW method implies

$$\gamma = \varepsilon.$$

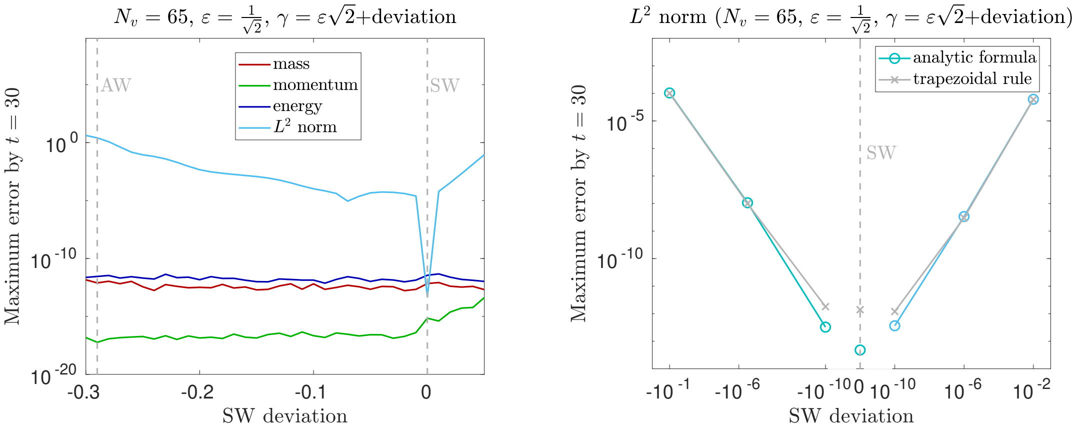
For the two stream instability, when $\varepsilon = 1/\sqrt{2}$, the AW case corresponds to the



(a) $t = 30$.

(b) $t = 21$.

Figure 26 Two stream instability. One can see that the maximum error in mass and energy grows with the deviation. L^2 error, in turn, decays with the increase of the deviation and then takes an abrupt sink when deviation approaches 0 which corresponds to the SW setup.



(a) Maximum errors in mass, momentum, energy and L^2 norm.

(b) Zoomed decay of the L^2 norm close to the SW setup in logarithmic scale.

Figure 27 Two stream instability. One can see that in this case, the SW setup is the most advantageous since it allows to preserve all observables. Moreover, we see that the SW setup is special when it comes to the L^2 norm conservation. The convergence of the L^2 is very slow near the zero deviation.

deviation of approximately -0.29. For the bump-on-tail instability, with $\varepsilon = 0.83$, the AW method corresponds to the deviation of -0.35.

In our simulation we iterate over the deviation range $[-0.4, 0.25]$ with the step of 0.02. For each value of the deviation we observe the maximum time $t_{\max}$ that the simulation can reach without "blowing up". In our code, we consider that to be the case once the error in the L^2 norm exceeds 10^1. If the simulation has reached the time $t = 100$, it is considered stable and we do not proceed further. One can see in figures 23a, 24a that the closer we get to the SW case (with zero deviation), the further we can run the simulation without stability issues for both even and odd number of basis functions. For the two stream instability, the range of the values of the deviation where the simulation has reached the time

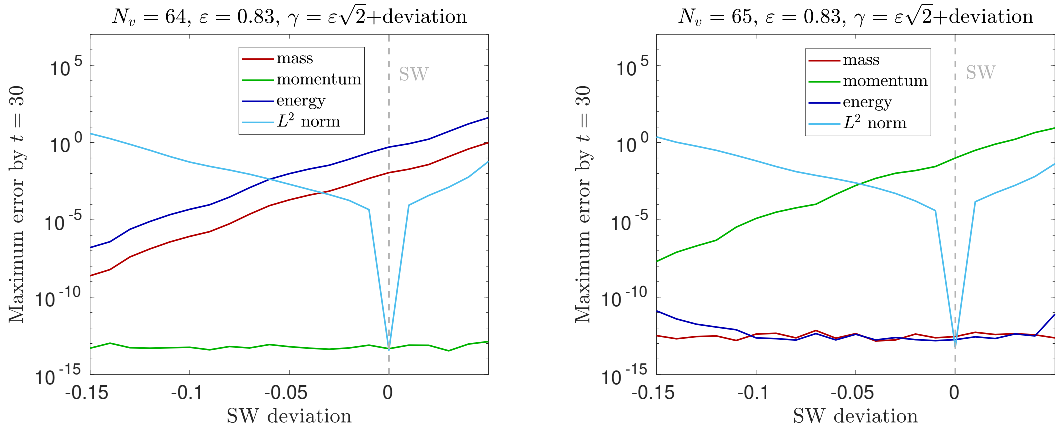
100 is slightly wider for $N_v = 65$. However, in Figure 24a we see that for the bump-on-tail test case, this range is the same for $N_v = 64$ and $N_v = 65$.

Of course, the initial interpolation quality also varies with the parameter γ. However, from figures 23b and 24b, it is clear that there is no direct correlation with $t_{\max}$. For both test cases, after the deviation surpasses the zero value the interpolation error stays around machine precision, whereas the maximum time drops significantly.

18.4. Conservation of the observables

In order to gain more insight in the conservation properties of the method, we set up another experiment. We limit the deviation range to $[-0.3, 0.05]$ in the two stream instability test case and to $[-0.15, 0.05]$ in the bump-on-tail test case. We set the step in deviation values to 0.01. We run the code only until the time $t = 30$. Note that for the bump-on-tail test case the deviation of -0.35, corresponding to the AW setup, did not pass our threshold of having an L^2 norm error less than 10^1 by the time 30. From figures 23a, 24a we know, that the simulation should not "blow up" until this point within this parameter range. We then observe the maximum error over the time $[0, 30]$ in mass, momentum, energy and the L^2 norm. We first take the even number of basis functions $N_v = 64$. As expected we observe in figures 26a, 28a the error in momentum stays around machine precision for all values of the deviation. However, mass and energy error are growing with the growth of the deviation. We know that for the AW case, all observables should be conserved all the time. We observe that indeed, for the initial deviation of -0.29, corresponding to the AW method for the two stream instability test case, the error in mass and momentum is very small. However, we can also see that it corresponds to the largest L^2 error. This indicates that an intermediate value of the deviation can be advantageous. One can also observe an interesting effect that an L^2 norm error slowly decays as the deviation increases but then abruptly sinks all the way to machine precision for the SW setup.

The behavior of the instabilities has two clear phases: the *linear* phase, when the electric energy grows and the *non-linear* one, where the electric energy flattens out. The non-linear phase of the simulation is generally harder to capture than the linear one. At $t = 30$ in the two-stream instability test case, the simulation has already reached the non-linear phase. Let us see if the rapid sink of the L^2 norm is specific for this phase. If we take a look at the same experiment, but with the final time $t = 21$, which is in the middle of the linear phase, we see in Figure 26b a similar picture. Even though in this case the L^2 norm error decays more steeply, we can still see an abrupt fall all the way to machine precision for the SW case. In Figure 27b we plot on a logarithmic scale the behavior of the



(a) $N_v = 64$.

(b) $N_v = 65$.

Figure 28 Bump-on-tail instability. One can see that the maximum error in mass and energy grows with the deviation. L^2 error, in turn, decays with the increase of the deviation and then takes an abrupt sink when deviation approaches 0 which corresponds to the SW setup.

L^2 norm near the zero deviation. In order to check whether it is an artifact of the analytic formula for the L^2 norm computation derived in Proposition 15.5.1, we use trapezoidal rule with 1000 grid points as a reference. We break the "x" axis near zero to emphasize that the zero value itself is undefined on the log scale. However, we can still compute the corresponding value of the L^2 norm. One can see in Figure 27b the the L^2 error decays logarithmically as the deviation approaches zero. Moreover, that the result of the trapezoidal rule matches the analytic formula. This means that the SW case is indeed a special case when it comes to the L^2 norm conservation.

Finally, we set $N_v = 65$ and look at the behavior of the observables in this case. It turns out, that for the two stream instability test case the momentum keeps being conserved also for the case of the odd number of basis functions (see Figure 27a). Meanwhile, the L^2 norm exhibits the same behavior as in the case of the even N_v. Therefore, for this case the clear choice would be to use the SW method with odd number of basis functions.

For the bump-on-tail test case, however, the situation is not as clear. One can see in Figure 28 that for the odd number of basis functions, the error in momentum grows, while mass and energy keep being conserved. Also, the abrupt sink of the L^2 error for the SW setup is present as before for both odd and even N_v. We know from the theory that for the deviation of -0.35, we would be able to gain all conservation properties. However, as discussed before, in this case the simulation is not stable. Therefore, it could be advantageous to use a γ corresponding to an intermediate setup between AW and SW.

18.5. Discussion

From the experiments above we can see the convenience of having two parameters in the basis. The parameter ε, corresponding to the width of the Gaussian in the basis, is usually fixed at the beginning to get the optimal representation of the initial distribution. The parameter γ, however, allows to vary the scaling of the argument of the Hermite polynomials in the basis, which in practice means that it controls the ratio between the scaling of the exponent and the Hermite polynomial in the basis. The AW and SW methods are just cases of specific values of γ. Overall, the tests have shown that the method works correctly also for intermediate values of γ, not corresponding to the SW and AW setups.

18.5.1. Conservation properties

As for the conservation properties, it has been numerically confirmed that, in the general case, we cannot guarantee simultaneous conservation of mass, momentum, and energy. However, for some specific test cases it might be the case for a wide range of γ, as in Figure 27a. For other cases, as predicted by the theory in chapter 15, mass and energy are preserved for the odd number of basis functions, whereas the momentum is preserved for the even number of basis functions. In general, however, if the conservation of mass, momentum, and energy is needed simultaneously, one needs to go closer to the AW setup. It is worth noticing that one need not hit the AW setup exactly to get the error in the mass, momentum, and energy conservation be below a certain threshold. From Figure 26a we see that it is true for a certain interval of the values of γ. Unfortunately, we can also see in Figure 26a that for this range of γ the error in the L^2 norm is larger. In order to guarantee the L^2 stability we would need the L^2 norm conservation, which, in theory, is only provided for the SW setup. We see a general trend in all test cases, that the simulations are more stable the closer we get to the SW setup. When it comes to the precise L^2 norm conservation, however, it is necessary to hit the exact SW setup (see Figure 27b).

18.5.2. Summary

All in all, enforcing the simultaneous conservation of mass, momentum, and energy yields instabilities, whereas enforcing the precise L^2 norm conservation confines us to the SW basis. However, the precise L^2 norm conservation is not necessary to run the simulation long enough. With the help of γ we could sacrifice some conservation in momentum (or mass/energy) to get more stability. In particular, taking an intermediate γ between the AW and SW setups provides a more stable simulation than the AW one, and smaller errors in momentum (or mass/energy) than the SW one.

Acknowledgments

Several people have participated in my PhD journey. First of all, I would like to thank my academic advisor, Caroline Lasser, for guiding me through my research and assisting with a profound scientific advice. I am deeply grateful to my supervisor Katharina Kormann, who provided me a tremendous support on a day-to-day basis and who always found time to answer my questions no matter how busy she was. I would like to thank Eric Sonnendrücker who gave me the opportunity to be a part of the Numerical Methods in Plasma Physics Department at Max Planck Institute for Plasma Physics. I would also like to thank Elisabeth Larsson who kindly agreed to be an examiner of this thesis. Last, but not least, I would like to express my gratitude to Peter Manz and Peter Kurz for always offering a personal advice and support throughout the HEPP graduate school.

I would like to acknowledge my colleagues from IPP who contributed to my PhD life. I am deeply grateful to Jalal Lakhlili for making my time at IPP more fun and for documenting it with funny pictures and gifs. It was a pleasure sharing the office with you! I would also like to thank Edoardo Zoni for always being on top of everything PhD students have to do and making my life easier by freeing me from figuring it out all by myself. I would also like to thank Laura Fahrner who always assisted me with all administrative work and also advice in complicated situations.

On a personal level, I would like to thank my boyfriend Philipp, who has been there for me every step of the way. I cannot imagine how I would be able to complete this PhD without having you by my side. At last, I would like to express my deep gratitude to my parents who made it possible for me to study in Germany in the first place. I would especially like to thank my mother Elena Yurova whose experience of working at a university in Russia has been invaluable for all my studies.

I gratefully acknowledge the support of the program “Global Challenges for Women in Math Science”, which provided me with an opportunity to present my research results at the ICOSAHOM 2018 conference. I would like to express my gratitude to the “SIAM Student Travel Award” for providing a financial support for my trip to the SIAM Conference on Computational Science and Engineering 2019 to present my research results.

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014–2018 and 2019–2020 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

List of publications

Prior publications of parts of the thesis “Generalized anisotropic Hermite functions and their applications” by Anna Yurova:

1. Katharina Kormann, Caroline Lasser, Anna Yurova, *”Stable interpolation with isotropic and anisotropic Gaussians using Hermite generating function”*. To appear in SIAM Journal on Scientific Computing. Accepted for publication on 18.09.2019, preprint available at `https://arxiv.org/abs/1905.09542`. Appears in the bibliography as [KLY19].

2. Anna Yurova, Katharina Kormann, *“Stable evaluation of Gaussian radial basis functions using Hermite polynomials”*, preprint available at `https://arxiv.org/abs/1709.02164`, 2017. Appears in the bibliography as [YK17].

Bibliography

[AM67] Thomas P. Armstrong and David Montgomery. Asymptotic state of the two-stream instability. *Journal of Plasma Physics*, 1(4):425–433, 1967.

[Arm67] Thomas P. Armstrong. Numerical studies of the nonlinear Vlasov equation. *The Physics of Fluids*, 10(6):1269–1280, 1967.

[AS64] Milton Abramowitz and Irene A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

[Bai48] W.N. Bailey. Some integrals involving Hermite polynomials. *Journal of the London Mathematical Society*, 1(4):291–297, 1948.

[BDL10] Rick Beatson, Oleg Davydov, and Jeremy Levesley. Error bounds for anisotropic rbf interpolation. *Journal of Approximation Theory*, 162(3):512 – 527, 2010.

[BGP00] François Bouchut, François Golse, and Mario Pulvirenti. *Kinetic equations and asymptotic theory*. 2000.

[BL04] Charles K. Birdsall and A. Bruce Langdon. *Plasma Physics via Computer Simulation*. Series in Plasma Physics and Fluid Dynamics. Taylor & Francis, 2004.

[Boy84] John P. Boyd. Asymptotic coefficients of Hermite function series. *Journal of Computational Physics*, 54(3):382–410, 1984.

[BP70] Ake Björck and Victor Pereyra. Solution of Vandermonde systems of equations. *Mathematics of computation*, 24(112):893–903, 1970.

[CDBM16] Enrico Camporeale, Gian Luca Delzanno, Benjamin K. Bergen, and J. David Moulton. On the velocity space discretization for the Vlasov–Poisson system: Comparison between implicit Hermite spectral and particle-in-cell methods. *Computer Physics Communications*, 198:47–58, 2016.

[CDLD06] Enrico Camporeale, Gian Luca Delzanno, Giovanni Lapenta, and William Daughton. New approach for the study of linear Vlasov stability of inhomogeneous systems. *Physics of plasmas*, 13(9):092110, 2006.

[Del15] Gian Luca Delzanno. Multi-dimensional, fully-implicit, spectral method for the Vlasov–Maxwell equations with exact conservation laws in discrete form. *Journal of Computational Physics*, 301:338–356, 2015.

[DKT17] Helge Dietert, Johannes Keller, and Stephanie Troppmann. An invariant class of wave packets for the Wigner transform. *Journal of Mathematical Analysis and Applications*, 450(2):1317–1332, 2017.

[DMS13] Stefano De Marchi and Gabrieles Santin. A new stable basis for radial basis function interpolation. *Journal of Computational and Applied Mathematics*, 253:1–13, 2013.

[EFMO63] Folker Engelmann, Marc R. Feix, Ettore Minardi, and Joachim Oxenius. Nonlinear effects from Vlasov's equation. *The Physics of Fluids*, 6(2):266–275, 1963.

[FF15] Bengt Fornberg and Natasha Flyer. Solving PDEs with radial basis functions. *Acta Numerica*, 24:215–258, 2015.

[FHW12] Gregory E. Fasshauer, Fred J. Hickernell, and Henryk Woźniakowski. On dimension-independent rates of convergence for function approximation with Gaussian kernels. *SIAM Journal on Numerical Analysis*, 50(1):247–271, 2012.

[FLF11] Bengt Fornberg, Elisabeth Larsson, and Natasha Flyer. Stable computations with Gaussian radial basis functions. *SIAM Journal on Scientific Computing*, 33(2):869–892, 2011.

[FLP13] Bengt Fornberg, Erik Lehto, and Collin Powell. Stable calculation of Gaussian-based RBF-FD stencils. *Computers & Mathematics with Applications*, 65(4):627–637, 2013.

[FM12] Gregory E. Fasshauer and Michael J. McCourt. Stable evaluation of Gaussian radial basis function interpolants. *SIAM Journal on Scientific Computing*, 34(2):A737–A762, 2012.

[FM15] Gregory E. Fasshauer and Michael McCourt. *Kernel-based approximation methods using Matlab*, volume 19. World Scientific Publishing Company, 2015.

[Fol89] Gerald B. Folland. *Harmonic analysis in phase space*. Princeton university press, 1989.

[Fol09] Gerald B. Folland. *Fourier analysis and its applications*, volume 4. American Mathematical Soc., 2009.

[FP07] Bengt Fornberg and Cécile Piret. A stable algorithm for flat radial basis functions on a sphere. *SIAM Journal on Scientific Computing*, 30(1):60–80, 2007.

[FS03] Francis Filbet and Eric Sonnendrücker. Numerical methods for the Vlasov equation. In *Numerical mathematics and advanced applications*, pages 459–468. Springer, 2003.

[FW04] Bengt Fornberg and Grady B. Wright. Stable computation of multiquadric interpolants for all values of the shape parameter. *Computers*

& Mathematics with Applications, 48(5):853 – 867, 2004.

[FZ07] Bengt Fornberg and Julia Zuev. The Runge phenomenon and spatially variable shape parameters in RBF interpolation. *Computers & Mathematics with Applications*, 54(3):379–398, 2007.

[GF67] Frederick C. Grant and Marc R. Feix. Fourier–Hermite solutions of the Vlasov equations in the linearized limit. *The Physics of Fluids*, 10(4):696–702, 1967.

[GO77] David Gottlieb and Steven A. Orszag. *Numerical analysis of spectral methods: theory and applications*, volume 26. SIAM, 1977.

[GR14] Izrail S. Gradshteyn and Iosif M. Ryzhik. *Table of integrals, series, and products*. Academic press, 2014.

[Gra49] Harold Grad. Note on n-dimensional Hermite polynomials. *Communications on Pure and Applied Mathematics*, 2(4):325–330, 1949.

[GRZ15] Michael Griebel, Christian Rieger, and Barbara Zwicknagl. Multiscale approximation and reproducing kernel Hilbert space methods. *SIAM Journal on Numerical Analysis*, 53(2):852–873, 2015.

[GS06] Livio Gibelli and Bernie D. Shizgal. Spectral convergence of the Hermite basis function solution of the Vlasov equation. *Journal of Computational Physics*, 219(2):477–488, 2006.

[Hag15] George A. Hagedorn. Generating function and a Rodrigues formula for the polynomials in d-dimensional semiclassical wave packets. *Annals of Physics*, 362:603–608, 2015.

[HE88] Roger W. Hockney and James W. Eastwood. *Computer simulation using particles*. crc Press, 1988.

[Hol96] James Paul Holloway. Spectral velocity discretizations for the Vlasov–Maxwell equations. *Transport theory and statistical physics*, 25(1):1–32, 1996.

[Hör90] Lars Hörmander. *The analysis of linear partial differential operators I*. Grundlehren der mathematischen Wissenschaften. Springer, 1990.

[JKM71] Glenn Joyce, Georg Knorr, and Homer K. Meier. Numerical integration methods of the Vlasov equation. *Journal of Computational Physics*, 8(1):53–63, 1971.

[KF94] Alexander J. Klimas and William M. Farrell. A splitting algorithm for Vlasov simulation with filamentation filtration. *Journal of computational physics*, 110(1):150–163, 1994.

[KLY19] Katharina Kormann, Caroline Lasser, and Anna Yurova. Stable interpolation with isotropic and anisotropic Gaussians using Hermite generating function. *To appear at SIAM Journal on Scientific Com-*

puting, accepted for publication on 18.09.2019, 2019.

[LB07] Soléne Le Bourdiec. *Méthodes déterministes de résolution des équations de Vlasov–Maxwell relativistes en vue du calcul de la dynamique des ceintures de Van Allen*. PhD thesis, 2007.

[LBDVJ06] Soléne Le Bourdiec, Florian De Vuyst, and Laurent Jacquet. Numerical solution of the Vlasov–Poisson system using generalized Hermite functions. *Computer physics communications*, 175(8):528–544, 2006.

[LF05] Elisabeth Larsson and Bengt Fornberg. Theoretical and computational aspects of multivariate interpolation with increasingly flat radial basis functions. *Computers & Mathematics with Applications*, 49(1):103–130, 2005.

[LLHF13] Elisabeth Larsson, Erik Lehto, Alfa Heryudono, and Bengt Fornberg. Stable computation of differentiation matrices and scattered node stencils based on Gaussian radial basis functions. *SIAM Journal on Scientific Computing*, 35(4):A2096–A2119, 2013.

[LSH17] Elisabeth Larsson, Victor Shcherbakov, and Alfa Heryudono. A least squares radial basis function partition of unity method for solving PDEs. *SIAM Journal on Scientific Computing*, 39(6):A2538–A2563, 2017.

[Lub08] Christian Lubich. *From quantum to classical molecular dynamics: reduced models and numerical analysis*. European Mathematical Society, 2008.

[MDVM16] Gianmarco Manzini, Gian Luca Delzanno, Juris Vencels, and Stefano Markidis. A Legendre–Fourier spectral method with exact conservation laws for the Vlasov–Poisson system. *Journal of Computational Physics*, 317:82–107, 2016.

[MF17] Michael McCourt and Gregory E. Fasshauer. Stable likelihood computation for Gaussian random fields. In *Recent Applications of Harmonic Analysis to Function Spaces, Differential Equations, and Data Science*, pages 917–943. Springer, 2017.

[MFD17] Gianmarco Manzini, Daniele Funaro, and Gian Luca Delzanno. Convergence of spectral discretizations of the Vlasov–Poisson system. *SIAM Journal on Numerical Analysis*, 55(5):2312–2335, 2017.

[Mur18] Moahan Murugappan. Unsicherheitsquantifizierung für die Vlasov-Poisson-Gleichung basierend auf Hierarchischen-Tucker-Tensoren. Master's thesis, Technische Universität München, 2018.

[PD15] Joseph T. Parker and Paul J. Dellar. Fourier–Hermite spectral representation for the Vlasov–Poisson system in the weakly collisional limit. *Journal of Plasma Physics*, 81(2), 2015.

[Pen55] Roger Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

[Rai71] Earl D. Rainville. *Special Functions*. AMS Chelsea Publishing Series. Chelsea Publishing Company, 1971.

[Ras03] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[RFK16] Jalil Rashidinia, Gregory E. Fasshauer, and Manoochehr Khasi. A stable method for the evaluation of Gaussian radial basis function solutions of interpolation and collocation problems. *Computers & Mathematics with Applications*, 72(1):178–193, 2016.

[Rud91] Walter Rudin. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1991.

[SH98] Joseph W. Schumer and James Paul Holloway. Vlasov simulations using velocity-scaled Hermite representations. *Journal of Computational Physics*, 144(2):626–661, 1998.

[SK74] Magdi Shoucri and Georg Knorr. Numerical integration of the Vlasov equation. *Journal of Computational Physics*, 14(1):84–92, 1974.

[Son13] Eric Sonnendrücker. Lecture notes in numerical methods for Vlasov equations, 2013.

[Son17] Eric Sonnendrücker. Numerical methods for the Vlasov–Maxwell equations. In preparation, 2017.

[Tar85] Audry Ellen Tarwater. Parameter study of Hardy's multiquadric method for scattered data interpolation. Technical Report UCRL-53670, Lawrence Livermore National Lab., CA (USA), 1985.

[Tha93] Sundaram Thangavelu. *Lectures on Hermite and Laguerre expansions*, volume 42. Princeton University Press, 1993.

[Tra66] Joseph F. Traub. Associated polynomials and uniform methods for the solution of linear problems. *SIAM Review*, 8(3):277–301, 1966.

[VDJ$^+$15] Juris Vencels, Gian Luca Delzanno, Alec Johnson, Ivy Bo Peng, Erwin Laure, and Stefano Markidis. Spectral solver for multi-scale plasma physics simulations with dynamically adaptive number of moments. *Procedia Computer Science*, 51:1148–1157, 2015.

[VDM$^+$16] Juris Vencels, Gian Luca Delzanno, Gianmarco Manzini, Stefano Markidis, Ivy Bo Peng, and Vadim Roytershteyn. Spectralplasmasolver: a spectral code for multiscale simulations of collisionless, magnetized plasmas. In *Journal of Physics: Conference Series*, volume 719, page 012022. IOP Publishing, 2016.

[Wat33] G.N. Watson. Notes on generating functions of polynomials:(2) Hermite polynomials. *Journal of the London Mathematical Society*, 1(3):194–199, 1933.

[WF17] Grady B. Wright and Bengt Fornberg. Stable computations with flat radial basis functions using vector-valued rational approximations. *Journal of Computational Physics*, 331:137–156, 2017.

[Won99] Man Wah Wong. *An Introduction to Pseudo-differential Operators*. World Scientific, 1999.

[WR06] Christopher K.I. Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[Wü15] Alfred Wünsche. Generating functions for products of special laguerre 2d and Hermite 2d polynomials. *Applied Mathematics*, 06:2142–2168, 01 2015.

[YK17] Anna Yurova and Katharina Kormann. Stable evaluation of Gaussian radial basis functions using Hermite polynomials. *E-print available at* `https://arxiv.org/abs/1709.02164`, 2017.

[Yse10] Harry Yserentant. *Regularity and approximability of electronic wave functions*. Springer, 2010.