ANOVA-boosting for Random Fourier Features

Daniel Potts
Faculty of Mathematics
Chemnitz University of Technology
D-09107 Chemnitz, Germany
daniel.potts@math.tu-chemnitz.de

Laura Weidensager
Faculty of Mathematics
Chemnitz University of Technology
D-09107 Chemnitz, Germany
laura.weidensager@math.tu-chemnitz.de

April 3, 2024

ABSTRACT

We propose two algorithms for boosting random Fourier feature models for approximating high-dimensional functions. These methods utilize the classical and generalized analysis of variance (ANOVA) decomposition to learn low-order functions, where there are few interactions between the variables. Our algorithms are able to find an index set of important input variables and variable interactions reliably.

Furthermore, we generalize already existing random Fourier feature models to an ANOVA setting, where terms of different order can be used. Our algorithms have the advantage of interpretability, meaning that the influence of every input variable is known in the learned model, even for dependent input variables. We give theoretical as well as numerical results that our algorithms perform well for sensitivity analysis. The ANOVA-boosting step reduces the approximation error of existing methods significantly.

Keywords ANOVA decomposition · global sensitivity analysis · random Fourier features · high-dimensional approximation

1 Introduction

Developing predictive models based on empirical data is a current field of research with diverse applications. The continuous growth in data collection leads to complex datasets, necessitating the handling of regression or classification tasks in high-dimensional spaces. Traditional machine learning techniques such as support vector machines, neural networks, and decision trees are commonly used to address these challenges. However, a crucial concern alongside prediction accuracy is the interpretability of these models, which is essential for understanding the underlying reasoning behind predictions.

Many current approaches, although effective with smaller or moderate number of input variables, stop working when confronted with high-dimensional challenges. The main problem to practical computability is often related to high dimension of the multivariate integration or interpolation problem, known as the curse of dimensionality. A well-known foundation of a dimensional decomposition is the analysis of variance (ANOVA) decomposition, first presented by Hoeffding in the 1940s. Since then, the ANOVA decomposition has been studied a lot in the literature, see for example [3, 27, 19, 13, 9, 6, 24].

However, the classical ANOVA decomposition is only available for independent, product-type probability measures of the input density. In practice, there could be notable correlations or dependencies among input variables. Therefore, the classical decomposition must be generalized for an arbitrary, non product type probability measure. Achieving this will require modifying the original orthogonality conditions. Indeed, inspired by Stone [37] and employing a set of weakened annihilating conditions, Hooker [10] provided an existential proof of a unique ANOVA decomposition for dependent variables, referred to as the generalized ANOVA decomposition in this paper, subject to a mild restriction on the probability measure. Afterwards, different approaches for calculating the component functions are studied for example in [14, 29].

The methodology presented in this paper offers an alternative to traditional machine learning methods by proposing an initial ANOVA boosting step for random feature methods, but also provides a natural means to assess the importance and influence of attributes on the predicted outcomes. The generalization of the classical ANOVA decomposition, e.g. in [10, 29] forms the basis for our algorithm, which can be applied to possibly dependent input variables. We calculate an approximation by a least squares regression, which penalizes the non-orthogonality between ANOVA terms.

Consider the following standard supervised learning setup. Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\} \subset \mathbb{R}^d$ be a set of discrete samples. We consider the problem of reconstructing a multivariate function $f : \mathbb{R}^d \to \mathbb{C}$ from discrete function samples on the set of nodes $\mathcal{X}$, which are sampled from the density $\mu \colon \mathbb{R}^d \to \mathbb{R}_+$. We study the scattered-data problem, i.e. we have given as labels the (possibly noisy) function values $\boldsymbol{f} = (f(\boldsymbol{x}) + \epsilon_{\boldsymbol{x}})_{\boldsymbol{x} \in \mathcal{X}}$. In contrast, in [30] the authors give explicit advice on the location of good sampling points for approximating high-dimensional functions as finite sums of lower-dimensional functions.

It is natural to express the model output $f(\boldsymbol{x})$ as a finite hierarchical expansion in terms of the input variables,

$$f(\boldsymbol{x}) = f_\varnothing + \sum_{i=1}^d f_i(x_i) + \sum_{1 \le i < j \le d} f_{\{i,j\}}(x_i, x_j) + \ldots + f_{\{1,\ldots,d\}}(x_1, x_2, \ldots, x_d), \tag{1.1}$$

where the zero-th order component function $f_\varnothing$ is a constant representing the mean of $f(\boldsymbol{x})$, the first order component function $f_i(x_i)$ gives the independent contribution to $f(\boldsymbol{x})$ by the i-th input variable acting alone, the second order component function $f_{\{i,j\}}(x_i, x_j)$ gives the pair cooperative contribution to $f(\boldsymbol{x})$ by the input variables x_i and x_j, etc. The last term $f_{\{1,\ldots,d\}}(x_1, x_2, \ldots, x_d)$ contains any residual d-th order cooperative contribution of all the input variables. The classical ANOVA decomposition, [3, 19, 9], of a function is a tool for capturing high-dimensional behaviour by demanding orthogonality with respect to the measure μ between the terms in (1.1) for functions $f \in L_2(\mathbb{R}^d, \mu)$.

In many settings functions may arise naturally as sums of functions, each with a limited variable interaction. Such low-order structure may also be used to reduce the curse of dimensionality, [3, 38, 34]. Another approach to capture low-dimensional structures by Gaussian mixtures was done in [8]. In this regard, two classes of problems arise: either all of the input variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ are independent or at least some portion of the variables in $\boldsymbol{x}$ are correlated. Standard formulations of the ANOVA deal with the case of independent variables. We, on the other hand, use the extension of [14] to also treat correlated variables.

Kernel-based approaches have been extensively used in high-dimensional function approximation since they often perform well in practice. The random feature model [28] is a popular technique for approximating the kernel (and thus the minimizer of kernel regression problems) using a randomized basis that can avoid the cost of full kernel methods. An alternative perspective to view the random feature model is as a two-layer network with a randomized but fixed single hidden layer, [28, 18]. The random feature model takes the form

$$f^\#(\boldsymbol{x}) = \sum_{k=1}^N a_k \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_k, \boldsymbol{x} \rangle} = \boldsymbol{a}^\top \mathrm{e}^{\mathrm{i}\langle \boldsymbol{W}, \boldsymbol{x} \rangle}, \quad \boldsymbol{\omega_k} \in \mathbb{R}^d,$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the input data, $\boldsymbol{W} \in \mathbb{R}^{d \times N}$ is a random weight matrix, and $\boldsymbol{a} \in \mathbb{C}^N$ is the final weight layer. The entries of the matrix $\boldsymbol{W}$ are independent and identically distributed (i.i.d.) random variables generated by the (user defined) probability density function $\rho(\boldsymbol{\omega})$. We construct the feature matrix

$$\boldsymbol{A} = (\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle})_{\boldsymbol{\omega} \in \mathcal{I}, \boldsymbol{x} \in \mathcal{X}}.$$

Given the collection of $M = |\mathcal{X}|$ measurements, $\boldsymbol{f} = (f(\boldsymbol{x}) + \epsilon_{\boldsymbol{x}})_{\boldsymbol{x} \in \mathcal{X}}$, the random feature regression problem becomes training $\boldsymbol{a}$ by optimizing

$$\min_{\boldsymbol{a} \in \mathbb{C}^N} \|\boldsymbol{f} - \boldsymbol{A}\boldsymbol{a}\|_2^2 + \mathcal{R}(\boldsymbol{a})$$

with some penalty function $\mathcal{R} \colon \mathbb{C}^N \to \mathbb{R}$. The most common choice for $\mathcal{R}$ is the ridge penalty $\mathcal{R}(\boldsymbol{a}) = \lambda \|\boldsymbol{a}\|_2^2$, which leads to the random feature ridge regression problem [31, 15, 21]. We will use another penalty function $\mathcal{R}$, which incorporates the ANOVA decomposition of the function f. We propose a random Fourier feature-framework where we explain the ANOVA decomposition within the random Fourier feature (RFF) structure. This specification allows us to associate variances and covariances to input variables, leading to interpretability that is not present in existing RFF models. Our algorithm aims to boost existing RFF algorithms like SHRIMP [39] (uses iterative magnitude pruning to select features) or HARFE [32] (uses hard thresholding pursuit to select features) by introducing a first approximation step, which is demonstrated by numerical examples.

This paper is organized as follows. In Section 2 we introduce the well-known ANOVA decomposition for independent input variables and relate this decomposition to the Fourier transform of the function f. Furthermore, we introduce

functions of lower order and show that they occur naturally in function spaces of mixed smoothness. In Section 3 we generalize the ANOVA decomposition to possibly dependent input variables. We summarize the idea of random Fourier feature algorithms in Section 4 and apply them to the ANOVA setting. The resulting boosting algorithms are summarized in Section 5, where we show how to do sensitivity analysis on an approximation with random Fourier features. The theoretical analysis in Section 6 generalizes the theory in [7, 39] for random Fourier features and finally in Section 7 we show with numerical examples the power of our boosting algorithms.

Our main contributions are as follows:

- We propose an ANOVA boosting, which extends and further develops the sparse random feature approximation from [7, 39, 32] to arbitrary index-sets $U \subset \mathcal{P}([d])$ and give a new connection between the Fourier transform of a function and the ANOVA terms. This leads to random features, which are adapted to the function.

- Generalizing the theory of sparse random Fourier features: In many cases, for example in the target case of functions of low order, the Fourier transform only exists in distributional sense. The norm on which the existing literature is based on, contains a maximum norm of the Fourier transform, which has to be generalized to the setting of tempered distributions.

- Introducing and analysing a first approximation step which calculates the important ANOVA terms for independent or dependent input variables.

- We improve the interpretability of previous random feature models by reducing the importance of variables that are only correlated with other variables and do not influence the function.

We will distinguish the two cases where the input variables are independent or dependent. In the first case, we will use the classical ANOVA decomposition, whereas in the latter case we have to generalize this decomposition.

Related work

- The authors from [20] propose the notion of neural decomposition, which integrates the classical ANOVA and deep neural networks for dimensionality reduction and variance decomposition. Similar to our approach for dependent input variables, they show that identifiability for independent input variables can be achieved by training models subject to constraints on the marginal properties of the decoder networks.

- The D-MORPH algorithm [14] uses orthogonal basis with respect to sampling density μ, which requires knowledge about the sampling density. Our approach is applicable independent of the sampling density μ. Furthermore, instead of calculating the solution of the minimization problem directly by using an SVD, we solve the problem by an iterative algorithm. This work was followed, among other, by [2], where the procedure was generalized to sampling from mixture densities, but also in this case an orthogonal basis is necessary.
  In [29] the generalized ANOVA decomposition is constructed by a constructive method by employing multivariate orthogonal polynomials as bases and calculating the expansion coefficients involved from the solution of linear algebraic equations.

- The authors in [4] also study indices measuring the sensitivity of the output with respect to dependent input variables, but they are restricted to independent pairs of dependent input variables.

- We generalize the approximation with random Fourier features, for already existing algorithms see for example [7, 39, 32]. See also [18] for a nice overview of random features for kernel approximation.

Definitions and Notation

In this paper we denote by $[d]$ the set $\{1, \ldots, d\}$ and its power set by $\mathcal{P}([d])$. The d-dimensional input variable of the function f is $\boldsymbol{x}$, where we denote the subset-vector by $\boldsymbol{x_u} = (x_i)_{i \in \boldsymbol{u}}$ for a subset $\boldsymbol{u} \subseteq [d]$. The complement of those subsets is always with respect to $[d]$, i.e., $\boldsymbol{u}^c = [d] \backslash \boldsymbol{u}$. For an index set $\boldsymbol{u} \subseteq [d]$ we define $|\boldsymbol{u}|$ as the number of elements in $\boldsymbol{u}$. Define the Fourier transform by

$$\hat{f}(\boldsymbol{\omega}) := \int_{\mathbb{R}^d} f(\boldsymbol{x}) \, \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle} \, \mathrm{d}\boldsymbol{x} \quad \text{for } \boldsymbol{w} \in \mathbb{R}^d. \tag{1.2}$$

If $f \in L_2(\mathbb{R}^d)$ with $\hat{f} \in L_1(\mathbb{R}^d)$, the Fourier inversion formula

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\omega}) \, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle} \, \mathrm{d}\boldsymbol{\omega}$$

holds true for almost all $\boldsymbol{x} \in \mathbb{R}^d$. For $f \notin L_2(\mathbb{R}^d)$ the Fourier transform $\hat{f}$ is defined only in distributional sense. Let $\mathcal{S}(\mathbb{R}^d)$ be the Schwartz space of rapidly decreasing functions on $\mathbb{R}^d$. Then, for slowly increasing function f, we can formulate the functional $T_f \colon \mathcal{S}(\mathbb{R}^d) \to \mathbb{C}$,

$$\langle T_f, \varphi \rangle = \int_{\mathbb{R}^d} f(\boldsymbol{x})\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad \varphi \in \mathcal{S}(\mathbb{R}^d).$$

The Dirac distribution δ is defined by $\langle \delta, \varphi \rangle = \varphi(\boldsymbol{0})$ for all $\varphi \in \mathcal{S}(\mathbb{R}^d)$. The Fourier transform is a linear functional on the Schwartz space,

$$\langle \hat{T}_f, \varphi \rangle = \langle T_f, \hat{\varphi} \rangle.$$

We denote for $\gamma > 0$ some frequently used densities by

$$\mu_{\mathcal{N}}(\boldsymbol{x}) = \frac{1}{(2\pi\gamma^2)^{d/2}}\mathrm{e}^{-\frac{\|\boldsymbol{x}\|^2}{2\gamma}} \quad \text{Gaussian,}$$

$$\mu_{\mathcal{C}}(\boldsymbol{x}) = \prod_{i=1}^{d} \frac{1}{\pi\sigma(1 + x_i^2/\gamma^2)} \quad \text{Cauchy.}$$

Let $s > 0$. Then we define Sobolev spaces of dominating mixed smoothness by

$$H_{\mathrm{mix}}^s(\mathbb{R}^d) := \left\{ f : \mathbb{R}^d \to \mathbb{C} \mid \|f\|_{H_{\mathrm{mix}}^s(\mathbb{R}^d)} < \infty \right\},$$

where the norm is defined by

$$\|f\|_{H_{\mathrm{mix}}^s(\mathbb{R}^d)}^2 = \int_{\mathbb{R}^d} |\hat{f}(\boldsymbol{\omega})|^2 \prod_{i=1}^{d} (1 + |\omega_i|^2)^s \,\mathrm{d}\boldsymbol{\omega}.$$

2 The ANOVA decomposition for independent input variables

In this section we study the case of independent input variables, which coincides with the density μ having tensor product structure. For periodic functions there is a connection between the Fourier coefficients of the ANOVA terms, which is used to construct approximation algorithms for high-dimensional functions with low effective dimension in an efficient and fast manner, see [26]. This was the motivation to study the more general setting, seeking a connection between the Fourier transform $\hat{f}$ and the ANOVA terms, which we will do in the following.

The curse of dimensionality comes into play when analysing data in high-dimensional spaces. A frequently used concept is the following, [3, 19, 9]. See also [22, Chapter 8.4] or [24, Appendix] for a general introduction to functional decompositions.

Definition 2.1. Let f be in $L_2(\mathbb{R}^d, \mu)$. For a tensor product density

$$\mu(\boldsymbol{x}) = \prod_{i=1}^{d} \mu_i(\boldsymbol{x}_i) \tag{2.1}$$

and for a subset $\boldsymbol{u} \subseteq [d]$ we define the **ANOVA (Analysis of variance) terms** recursively by

$$f_\varnothing = \int_{\mathbb{R}^d} f(\boldsymbol{x})\mu(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) = \int_{\mathbb{R}^{d-|\boldsymbol{u}|}} f(\boldsymbol{x})\mu(\boldsymbol{x}_{\boldsymbol{u}^c}) \,\mathrm{d}\boldsymbol{x}_{\boldsymbol{u}^c} - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}). \tag{2.2}$$

The **ANOVA decomposition** with respect to μ of a function $f \colon \mathbb{R}^d \to \mathbb{C}$ is then given by

$$f(\boldsymbol{x}) = f_\varnothing + \sum_{i=1}^{d} f_{\{i\}}(x_i) + \sum_{i \neq j=1}^{d} f_{\{i,j\}}(x_i, x_j) + \cdots + f_{[d]}(\boldsymbol{x}) = \sum_{\boldsymbol{u} \subseteq [d]} f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}). \tag{2.3}$$

The terms (2.2) are the unique decomposition (2.3), such that

$$\langle f_{\boldsymbol{u}}, f_{\boldsymbol{v}} \rangle_\mu := \int_{\mathbb{R}^d} f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}})f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{u}})\mu(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0 \quad \text{for } \boldsymbol{v} \neq \boldsymbol{u} \subseteq [d] \tag{2.4}$$

$$\int_{\mathbb{R}} f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}})\mu_j(x_j)\mathrm{d}x_j = 0 \quad \text{for } j \in \boldsymbol{u}.$$

4

Note that in general, $f_{\boldsymbol{u}} \notin L_2(\mathbb{R}^{|\boldsymbol{u}|})$, but $f_{\boldsymbol{u}} \in L_2(\mathbb{R}^{|\boldsymbol{u}|}, \mu_{\boldsymbol{u}})$. Furthermore, every density has the property that $\mu \in L_1(\mathbb{R}^d)$, which implies that $\mu_i \in L_1(\mathbb{R})$, such that the one-dimensional Fourier transforms $\hat{\mu}_i$ exists. For a better readability of the following proofs, we introduce the notation

$$E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \boldsymbol{u}) := \prod_{i \in \boldsymbol{u}} \left(\mathrm{e}^{\mathrm{i}\omega_i x_i} - \hat{\mu}_i(-\omega_i) \right) \prod_{i \in \boldsymbol{u}^c} \hat{\mu}_i(-\omega_i). \tag{2.5}$$

In terms of the Fourier transform, the ANOVA terms (2.2) can then be described by the following.

Lemma 2.2. *Let the sampling distribution μ have a product structure* (2.1). *Then the ANOVA decomposition* (2.2) *of the function $f \in L_2(\mathbb{R}^d, \mu)$ is given by*

$$f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \boldsymbol{u}) \, \mathrm{d}\boldsymbol{\omega}. \tag{2.6}$$

Proof. The Fourier transform of the tensor product density μ can be decomposed as

$$\hat{\mu}(\boldsymbol{\omega}) = \prod_{i \in [d]} \hat{\mu}_i(\omega_i).$$

We prove (2.6) inductively over $|\boldsymbol{u}|$. First, observe that

$$f_{\varnothing} = \int_{\mathbb{R}^d} f(\boldsymbol{x}) \mu(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) \, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle} \, \mathrm{d}\boldsymbol{\omega} \, \mu(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle} \mu(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{\omega} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) \hat{\mu}(-\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega}$$

$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \varnothing) \, \mathrm{d}\boldsymbol{\omega},$$

where $E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \varnothing) = \prod_{i \in [d]} \hat{\mu}_i(-\omega_i)$ does not depend on $\boldsymbol{x}$. The induction step follows using Definition 2.1,

$$f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) = \int_{\mathbb{R}^{d-|\boldsymbol{u}|}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) \, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle} \, \mathrm{d}\boldsymbol{\omega} \mu(\boldsymbol{x}_{\boldsymbol{u}^c}) \, \mathrm{d}\boldsymbol{x}_{\boldsymbol{u}^c} - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}})$$

$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{u}} \rangle} \int_{\mathbb{R}^{d-|\boldsymbol{u}|}} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_{\boldsymbol{u}^c}, \boldsymbol{x}_{\boldsymbol{u}^c} \rangle} \mu(\boldsymbol{x}_{\boldsymbol{u}^c}) \, \mathrm{d}\boldsymbol{x}_{\boldsymbol{u}^c} \, \mathrm{d}\boldsymbol{\omega} - \frac{1}{(2\pi)^d} \sum_{\boldsymbol{v} \subset \boldsymbol{u}} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \boldsymbol{v}) \, \mathrm{d}\boldsymbol{\omega}$$

$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{u}} \rangle} \hat{\mu}_{\boldsymbol{u}^c}(-\boldsymbol{\omega}_{\boldsymbol{u}^c}) \, \mathrm{d}\boldsymbol{\omega} - \frac{1}{(2\pi)^d} \sum_{\boldsymbol{v} \subset \boldsymbol{u}} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) \prod_{i \in \boldsymbol{v}} \left(\mathrm{e}^{\mathrm{i}\omega_i x_i} - \hat{\mu}_i(-\omega_i) \right) \prod_{i \in \boldsymbol{v}^c} \hat{\mu}_i(-\omega_i) \, \mathrm{d}\boldsymbol{\omega}$$

$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) \hat{\mu}_{\boldsymbol{u}^c}(-\boldsymbol{\omega}_{\boldsymbol{u}^c}) \left(\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{u}} \rangle} - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} \prod_{i \in \boldsymbol{v}} \left(\mathrm{e}^{\mathrm{i}\omega_i x_i} - \hat{\mu}_i(-\omega_i) \right) \prod_{i \in \boldsymbol{u} \setminus \boldsymbol{v}} \hat{\mu}_i(-\omega_i) \right) \mathrm{d}\boldsymbol{\omega}$$

$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) \prod_{i \in \boldsymbol{u}} \left(\mathrm{e}^{\mathrm{i}\omega_i x_i} - \hat{\mu}_i(-\omega_i) \right) \prod_{i \in \boldsymbol{u}^c} \hat{\mu}_i(-\omega_i) \, \mathrm{d}\boldsymbol{\omega}$$

$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \boldsymbol{u}) \, \mathrm{d}\boldsymbol{\omega}.$$

This shows (2.6). ∎

Example 2.3. Suppose

$$f \colon \mathbb{R}^2 \to \mathbb{R} \qquad f(\boldsymbol{x}) = g_1(x_1) \, g_2(x_2) = \frac{|x_1|}{(1+x_1^2)^2} \cdot \max\left(1 - |x_2|, 0\right) \in H_{\mathrm{mix}}^{3/2-\epsilon}(\mathbb{R}^2, \mu). \tag{2.7}$$

This is a function of tensor product structure, which means that the ANOVA decomposition is

$$f_1(x_1) = \left(g_1(x_1) - \overline{f_1} \right) \cdot \overline{f_2}, \qquad\qquad \overline{f_1} := \int_{\mathbb{R}} g_1(x_1) \mu_1(x_1) \mathrm{d}x_1,$$

$$f_2(x_2) = \left(g_2(x_2) - \overline{f_2} \right) \cdot \overline{f_1}, \qquad\qquad \overline{f_2} := \int_{\mathbb{R}} g_2(x_2) \mu_2(x_2) \mathrm{d}x_2,$$

$$f_{1,2}(\boldsymbol{x}) = \left(g_1(x_1) - \overline{f_1} \right) \cdot \left(g_2(x_2) - \overline{f_2} \right),$$

where the constants with respect to standard Gaussian samples $\mu(\boldsymbol{x}) = \mu_{\mathcal{N}}(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}^d} \mathrm{e}^{-\|\boldsymbol{x}\|^2/2}$ and uniform samples on $[-1,1]^2$ are summarized here:

5

μ	Gaussian	Uniform
$f_\varnothing$	0.0792	0.125
$\overline{f_1}$	0.2148	0.25
$\overline{f_2}$	0.3687	0.5

The ANOVA decomposition is plotted in Figure 2.1 for these two cases. This example shows that tensor product type function can be easily decomposed in the ANOVA terms independent of the sampling density μ. Furthermore, the ANOVA decomposition depends on the sampling density μ. For more complicated functions this can have much more influence. $\qquad\square$

$f(x_1, x_2)$
$\downarrow$

Figure 2.1: The ANOVA decomposition of the function (2.7).

Sensitivity analysis

The ANOVA decomposition is the basis for sensitivity analysis, which is the study of how the uncertainty in the output of a mathematical model can be divided and allocated to different sources of uncertainty in its inputs. A measure describing the proportion of how much the variables $\boldsymbol{x_u}$ contribute to the variance of the function f itself are the variances

$$\sigma^2(f_{\boldsymbol{u}}) = \int_{\mathbb{R}^{|\boldsymbol{u}|}} |f_{\boldsymbol{u}}(\boldsymbol{x_u})|^2 \mu_{\boldsymbol{u}}(\boldsymbol{x_u}) \mathrm{d}\boldsymbol{x_u},$$

where $\sum_{\boldsymbol{u} \subseteq [d]} \sigma^2(f_{\boldsymbol{u}}) = \sigma^2(f)$. This provides an explanation for the name **analysis of variance** decomposition to this function decomposition. Sobol indices, first introduced by [33], are an often used tool for sensitivity analysis,

namely

$$S_{\boldsymbol{u}} = \frac{\sigma^2(f_{\boldsymbol{u}})}{\sigma^2(f)}. \tag{2.8}$$

Relation to the case of periodic functions

The ANOVA decomposition in terms of the Fourier transform $\hat{f}$ investigated in Lemma 2.2 is a generalization of the periodic case. To see this, let $\mu_i(x_i) = \frac{1}{2\pi}\, 1_{[-\pi,\pi]}$, the density belonging to the uniform density on the torus. For the Fourier transform of this density we calculate

$$\hat{\mu}_i(\omega_i) = \frac{\sin(\pi\omega_i)}{\pi\omega_i} = \begin{cases} 0 \text{ if } \omega_i \in \mathbb{Z}\backslash 0, \\ 1 \text{ if } \omega_i = 0. \end{cases} \tag{2.9}$$

For periodic functions the Fourier coefficients are defined by

$$c_{\boldsymbol{k}}(f) = \int_{-\pi}^{\pi} f(\boldsymbol{x}) \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{k}, \boldsymbol{x}\rangle}\, \mathrm{d}\boldsymbol{x}.$$

It is possible to extend the definition (1.2) to include periodic functions by viewing them as tempered distributions. This makes it possible to see a connection between the Fourier series and the Fourier transform for periodic functions that have a convergent Fourier series. If f is a periodic function with period 1, that has convergent Fourier series, then:

$$\hat{f}(\omega) = \sum_{k\in\mathbb{Z}} c_k(f)\delta\left(\omega - k\right),$$

where $c_k(f)$ are the Fourier coefficients of f and δ is the Dirac delta distribution. The corresponding multivariate case is

$$\hat{f}(\boldsymbol{\omega}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} c_{\boldsymbol{k}}(f)\delta\left(\boldsymbol{\omega} - \boldsymbol{k}\right),$$

Applying Lemma 2.2 to this setting, we have

$$f_{\boldsymbol{u}}(\boldsymbol{x_u}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\omega}) E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \boldsymbol{u})\, \mathrm{d}\boldsymbol{\omega} = \frac{1}{(2\pi)^d} \sum_{\boldsymbol{k}\in\mathbb{Z}^d} c_{\boldsymbol{k}}(f)\delta\left(\boldsymbol{\omega} - \boldsymbol{k}\right) E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \boldsymbol{u})$$

$$= \frac{1}{(2\pi)^d} \sum_{\boldsymbol{k}\in\mathbb{Z}^d} c_{\boldsymbol{k}}(f) E(\boldsymbol{x}, \boldsymbol{k}, \mu, \boldsymbol{u}) = \frac{1}{(2\pi)^d} \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{u}}} c_{\boldsymbol{k}}(f) \mathrm{e}^{\mathrm{i}\langle \boldsymbol{k_u}, \boldsymbol{x_u}\rangle},$$

where the index-set is $\mathcal{I}_{\boldsymbol{u}} = \{\boldsymbol{k} \in \mathbb{Z}^d \mid \operatorname{supp} \boldsymbol{k} = \boldsymbol{u}\}$. The last equality follows from (2.9) and the definition of the term E in (2.5). This connection was shown in [26] and was starting point for efficient algorithms.

2.1 Functions of low order

Often, high-dimensional functions that arise from important physical systems are of low order, meaning the function is dominated by a few terms, each depending on only a subset of the input variables, say q out of the d variables where $q \ll d$. For that reason, we formalize the notion of low order functions by extending the definition from [7].

Definition 2.4 (Functions of low order)**.** Fix $d, q \in \mathbb{N}$ with $q \leq d$. A function $f\colon \mathbb{R}^d \to \mathbb{C}$ is an **order-q** function, if

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{u}\in U_q} f_{\boldsymbol{u}}(\boldsymbol{x_u}) \quad \text{with} \quad U_q = \{\boldsymbol{u} \in [d] \mid |\boldsymbol{u}| \leq q\}.$$

Low order functions arise naturally in the physical world and are used as a form of the reduced complexity model for such systems.

The following gives a bound for the variances of the ANOVA terms, which is guided by a decomposition of the frequency domain of f. The proof can be found in Appendix A.

Lemma 2.5. *Let $f \in L_2(\mathbb{R}^d)$ with $\hat{f} \in L_1(\mathbb{R}^d)$, then the variances of the ANOVA terms $f_{\boldsymbol{u}}$ defined in (2.6) are bounded by*

$$\sigma^2(f_{\boldsymbol{u}}) \leq \frac{1}{(2\pi)^{2d}}\, \|\hat{\mu}\|_{L_1(\mathbb{R}^d)} \int_{\mathbb{R}^d} |\hat{f}(\boldsymbol{\omega})|^2 |E(\boldsymbol{0}, \boldsymbol{\omega}, \mu, \boldsymbol{u})|\, \mathrm{d}\boldsymbol{\omega}.$$

A summation of all inequalities for $\boldsymbol{u} \subseteq [d]$ from the previous result gives:

$$\sum_{\boldsymbol{u} \subseteq [d]} \sigma^2(f_{\boldsymbol{u}}) = \|f\|_{L_2(\mathbb{R}^d, \mu)}^2 = \int_{\mathbb{R}^d} |f(\boldsymbol{x})|^2 \mu(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq \frac{1}{(2\pi)^{2d}} \|\hat{\mu}\|_{L_1(\mathbb{R}^d)} \|\hat{f}\|_{L_2(\mathbb{R}^d)}^2.$$

Example 2.6. Consider again the standard Gaussian distribution $\mu = \mu_{\mathcal{N}}$, where $\hat{\mu}_{\mathcal{N}}(\boldsymbol{\omega}) = \mathrm{e}^{-\frac{\|\boldsymbol{\omega}\|^2}{2}}$. By the functions $|E(\boldsymbol{0}, \boldsymbol{\omega}, \mu, \boldsymbol{u})|$ appearing in Lemma 2.5 we decompose the frequency domain $\mathbb{R}^d$ concerning the different ANOVA indices $\boldsymbol{u} \subseteq [d]$. We plot the two-dimensional example in Figure 2.2. One can see that (this is in general the case, not only for this example)

$$\lim_{k \to \infty} |E(\boldsymbol{0}, k\boldsymbol{\omega}, \mu, \boldsymbol{u})| = \begin{cases} 1 & \text{if } \operatorname{supp} \boldsymbol{\omega} = \boldsymbol{u}, \\ 0 & \text{otherwise} . \end{cases}$$

The decomposition is the analogue of the discrete decomposition of the Fourier series into ANOVA terms, see [26, Fig.1]. □

Figure 2.2: 2-dimensional decomposition of the frequency domain into ANOVA terms for Gaussian distribution $\mu_{\mathcal{N}}$. Plotted are the functions $|E(\boldsymbol{0}, \boldsymbol{\omega}, \mu_{\mathcal{N}}, \boldsymbol{u})|$ for $\boldsymbol{u} \subseteq \{1, 2\}$.

To study functions of low-order q, define

$$\mathcal{T}_q f := \sum_{|\boldsymbol{u}| \leq q} f_{\boldsymbol{u}}.$$

In [34, Corollary 2.32] the author delivers error estimates for the truncation error $\|f - \mathcal{T}_q f\|_{L_2(\mathbb{R}^d, \mu)}$ for functions f in function spaces with product and order-dependent weights, which builds on a transformation of periodic functions and an ANOVA decomposition based on Fourier coefficients on the torus. However, the estimates there are related to the smoothness of the transformed function on the torus, see also [16] for details of the idea of transformations from $\mathbb{R}^d$ to the torus and the transformation of the smoothness thereby. Since the transformation can destroy the smoothness, in the following theorem we give a bound on the truncation error $\|f - \mathcal{T}_q f\|_{L_2(\mathbb{R}^d, \mu)}$ relative to the norm $\|f\|_{H_{\mathrm{mix}}^s(\mathbb{R}^d)}$. The proof can be found in Appendix A.

Theorem 2.7. *Let the measures μ_i be either symmetric and have positive Fourier transform or fulfill for fixed $s > \frac{1}{2}$ the mild condition*

$$\frac{|1 - \hat{\mu}_i(-\omega_i)| + |\hat{\mu}_i(-\omega_i)|}{(1 + |\omega_i|^2)^s} \leq 1 \tag{2.10}$$

for all $i \in [d]$ and all $\omega_i \in \mathbb{R}$. Define the constant

$$c_{\mu, s} = \max_{i \in [d]} \max_{\omega \in \mathbb{R}} \frac{1 - \hat{\mu}_i(-\omega)}{(1 + |\omega|^2)^s}.$$

Then for $f \in H_{\mathrm{mix}}^s(\mathbb{R}^d)$ the truncation error is bounded by

$$\|f - \mathcal{T}_q f\|_{L_2(\mathbb{R}^d, \mu)}^2 \leq \frac{c_{\mu, s}^q}{(2\pi)^{2d}} \|\hat{\mu}\|_{L_1(\mathbb{R}^d)} \|f\|_{H_{\mathrm{mix}}^s(\mathbb{R}^d)}^2.$$

Furthermore, if the Fourier transforms $\hat{\mu}_i$ are differentiable, we have

$$c_{\mu, s} = \max_{i \in [d]} \max_{\omega \in \mathbb{R}} \frac{-\hat{\mu}_i'(\omega)}{2s\omega (1 + \omega^2)^{s-1}}. \tag{2.11}$$

We specifically point out that in the previous result the truncation error $\|f - \mathcal{T}_q f\|^2_{L_2(\mathbb{R}^d, \mu)}$ is bounded by a constant $\mathcal{O}(2^{-q})$ and $\|f\|_{H^s_{\mathrm{mix}}(\mathbb{R}^d)}$, which is determined by the smoothness of f. In general, this norm can increase with increasing d, but on the other hand the factor $(2\pi)^{-2d} \|\hat{\mu}\|_{L_1(\mathbb{R}^d)}$ decays exponentially with increasing d.

Example 2.8. Let us have a look at Gaussian samples $\mu_i(x_i) = \mu_{\mathcal{N}}(x_i) = \frac{1}{\gamma\sqrt{2\pi}} e^{-\frac{x_i^2}{2\gamma^2}}$ with variance γ, which means $\hat{\mu}_i(\omega_i) = e^{-\gamma^2 \omega_i^2/2}$. Then

$$c_{\mu,s} = \frac{\gamma\omega_i}{2s\omega_i(1+\omega_i^2)^{s-1}} e^{-\gamma^2\omega_i^2/2} = \frac{\gamma}{2s(1+\omega_i^2)^{s-1}} e^{-\gamma^2\omega_i^2/2} \leq \frac{\gamma}{2s},$$

$$\|\hat{\mu}_i\|_{L_1(\mathbb{R})} = \frac{\sqrt{2\pi}}{\gamma},$$

such that

$$\|f - \mathcal{T}_q f\|^2_{L_2(\mathbb{R}^d, \mu_{\mathcal{N}})} \leq (2\pi)^{-\frac{3}{2}d}(2s)^{-(q+1)} \|f\|^2_{H^s_{\mathrm{mix}}(\mathbb{R}^d)}.$$

Let us have a look at Cauchy distributed samples $\mu_i(x_i) = \mu_{\mathcal{C}}(x_i) = \frac{1}{\pi\gamma(1+x_i^2/\gamma^2)}$ with variance γ, which means $\hat{\mu}_i(\omega_i) = e^{-\gamma|\omega_i|}$. Then

$$c_{\mu,s} = \sup_{\omega_i > 0} \frac{\gamma\omega_i}{2s\omega_i(1+\omega_i^2)^{s-1}} e^{-\gamma\omega_i} = \frac{\gamma}{2s(1+\omega_i^2)^{s-1}} e^{-\gamma^2\omega_i^2/2} \leq \frac{\gamma}{2s},$$

$$\|\hat{\mu}_i\|_{L_1(\mathbb{R})} = \frac{\sqrt{2}}{\gamma},$$

such that

$$\|f - \mathcal{T}_q f\|^2_{L_2(\mathbb{R}^d, \mu_{\mathcal{C}})} \leq 2^{-d}\pi^{-2d}(2s)^{-(q+1)} \|f\|^2_{H^s_{\mathrm{mix}}(\mathbb{R}^d)}.$$

$\square$

3 The generalized ANOVA decomposition for correlated input variables

The main assumption of the ANOVA decomposition is that the input parameters $x_i, i = 1, \ldots, d$, are independent. This is unrealistic in many cases. Clearly, the correlation structure of random variables heavily influences the composition of component functions as well as global sensitivity analysis.

However, when the dependence is present among variables, the variance contribution of an individual variable x_i consists of not only the contribution resulting from the variable itself, but also contains the dependent contribution resulting from the dependence between variable x_i and other variables. So far, the literature [14, 29] discusses the variance contributions with dependent variables, and makes a distinction between the independent contribution and dependent contribution of the variables.

We now consider possibly dependent input variables x_i, i.e. an arbitrary non-product type probability density function $\mu\colon \mathbb{R}^d \to \mathbb{R}$, that has marginal probability density functions

$$\mu_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) \coloneqq \int_{\mathbb{R}^{d-|\boldsymbol{u}|}} \mu(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}_{\boldsymbol{u}^c},$$

where $\varnothing \neq \boldsymbol{u} \subseteq [d]$.

In the case for independent variables, the ANOVA decomposition (2.3) is unique by demanding the condition (2.4). For dependent variables, it is in general not possible to find an orthogonal decomposition. First, there is a mild condition to the measure μ needed, to construct an ANOVA decomposition of the form (2.3): Assume that for every $\boldsymbol{u} \subseteq [d]$ the support of $\mu_{\boldsymbol{u}}$ is **grid-closed** [10]. The grid closure implies that for any point $\boldsymbol{x}_{\boldsymbol{u}} \in \operatorname{supp} \mu_{\boldsymbol{u}}$ we can move in each coordinate direction and find another point in the support of $\mu_{\boldsymbol{u}}$. This is a mild regularity requirement, which is fulfilled by common probability distributions. The grid closure excludes only degenerated distributions like $\mu_{\{1,2\}} = \mathbf{1}_{\{x_1=x_2\}}$, where it anyway is not possible to distinguish between the input variables x_1 and x_2.

Second, we have to replace the condition (2.4) of the classical ANOVA decomposition in the setting for independent variables to a milder condition, **hierarchical orthogonality** condition,

$$\int_{\mathbb{R}^d} f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) \mu(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0 \quad \text{for all } \boldsymbol{v} \subset \boldsymbol{u}, \tag{3.1}$$

see [10, 29] for more details. The **weak annihilating conditions**, that is

$$\int_{\mathbb{R}} f_{\boldsymbol{u}}(\boldsymbol{x_u})\mu_{\boldsymbol{u}}(\boldsymbol{x_u})\mathrm{d}x_i = 0 \quad \text{for all } i \in \boldsymbol{u}, \boldsymbol{u} \subseteq [d] \tag{3.2}$$

are equivalent to (3.1) and are appropriate for the generalized ANOVA decomposition. Every square-integrable multivariate function f with respect to the marginal probability measure $\mu_{\boldsymbol{u}}$ supported on $\mathbb{R}^{|\boldsymbol{u}|}$ admits a unique, finite, hierarchical expansion

$$f(\boldsymbol{x}) = \sum_{\varnothing \neq \boldsymbol{u} \subseteq [d]} f_{\boldsymbol{u}}(\boldsymbol{x_u}), \tag{3.3}$$

referred to as the **generalized ANOVA decomposition**. The existence and uniqueness of the decomposition in (3.3) under mild conditions has been proven in [4, 10, 37]. That is, for the existence the support of every $\mu_{\boldsymbol{u}}$ has to be grid-closed and the uniqueness follows by demanding (3.2). Note that the generalized ANOVA decomposition matches the classical ANOVA decomposition (2.3), if the input variables are independent.

There exist many methods for calculating the global sensitivity indices (2.8) of a function of independent variables. In contrast, only a few methods, such as those presented in [4, 10, 12, 14, 29] are available for models with dependent or correlated input. In all literature the Sobol indices (2.8) are generalized to the following.

Definition 3.1. *The Sobol indices for an ANOVA term $f_{\boldsymbol{u}}$ measuring the contribution of $\boldsymbol{x_u}$ into the model, denoted by $S_{\boldsymbol{u},\mathrm{var}}, S_{\boldsymbol{u},\mathrm{cor}}$ and $S_{\boldsymbol{u}}$ are given by*

$$S_{\boldsymbol{u},\mathrm{var}} = \frac{\sigma^2(f_{\boldsymbol{u}})}{\sigma^2(f)}$$

$$S_{\boldsymbol{u},\mathrm{cor}} = \frac{\sum_{\substack{\varnothing \neq \boldsymbol{v} \subseteq [d] \\ \boldsymbol{v} \cap \boldsymbol{u} \neq \varnothing, \boldsymbol{v} \not\subseteq \boldsymbol{u}}} \langle f_{\boldsymbol{u}}, f_{\boldsymbol{v}} \rangle_{\mu}}{\sigma^2(f)}$$

$$S_{\boldsymbol{u}} = S_{\boldsymbol{u},\mathrm{var}} + S_{\boldsymbol{u},\mathrm{cor}}.$$

*The first two indices $S_{\boldsymbol{u},\mathrm{var}}$ and $S_{\boldsymbol{u},\mathrm{cor}}$ represent the normalized versions of the **variance contributions** and **covariance contributions** from $f_{\boldsymbol{u}}$ to $\sigma^2(f)$. The third index, $S_{\boldsymbol{u}}$, referred to as the **total global sensitivity index** is the sum of variance and covariance contributions.*

When the random variables are independent, the covariance contributions to the total sensitivity index $S_{\boldsymbol{u},\mathrm{cor}}$ vanish for all $\boldsymbol{u} \subseteq [d]$, leaving only one sensitivity index for the classical ANOVA decomposition.

3.1 ANOVA decomposition of the frequency domain

In case of periodic functions and independent input variables there is a connection between the ANOVA terms $f_{\boldsymbol{u}}$ and the Fourier coefficients [26] or the wavelet coefficients [17]. This connection leads to efficient algorithms. In the following we study a similar connection for functions on $\mathbb{R}^d$ and the Fourier transform.

Let $f \in L_2(\mathbb{R}^d) \cap L_2(\mathbb{R}^d, \mu)$ have the generalized ANOVA decomposition $f = \sum_{\boldsymbol{u} \in U} f_{\boldsymbol{u}}(\boldsymbol{x_u})$. The functions $f_{\boldsymbol{u}}$ depend only on the variables $\boldsymbol{x_u}$, i.e.,

$$\hat{T}_f(\boldsymbol{\omega}) = \sum_{\boldsymbol{u} \in U} \delta_{\boldsymbol{\omega}_{\boldsymbol{u}^c}} \hat{T}_{f_{\boldsymbol{u}}}(\boldsymbol{\omega_u}).$$

Then the Fourier inversion formula yields,

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega}) \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle} \mathrm{d}\boldsymbol{\omega} = \sum_{\boldsymbol{u} \in U} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^{|\boldsymbol{u}|}} \hat{T}_{f_{\boldsymbol{u}}}(\boldsymbol{\omega_u}) \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega_u}, \boldsymbol{x_u} \rangle} \mathrm{d}\boldsymbol{\omega_u}$$

$$= \sum_{\boldsymbol{u} \in U} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{T}_{f_{\boldsymbol{u}}}(\boldsymbol{\omega_u}) \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega_u}, \boldsymbol{x_u} \rangle} \delta_{\boldsymbol{\omega}_{\boldsymbol{u}^c}} \mathrm{d}\boldsymbol{\omega}.$$

This somehow decomposes the frequency domain $\mathbb{R}^d$ into parts which belong to the different ANOVA-terms, in the sense that for the ANOVA term $\boldsymbol{u}$ the corresponding frequencies $\boldsymbol{\omega}$ have to fulfill $\boldsymbol{\omega}_{\boldsymbol{u}^c} = \boldsymbol{0}$. For that reason let us define the frequency decomposition

$$Q_{\boldsymbol{u}} := \{\boldsymbol{\omega} \in \mathbb{R}^d \mid \boldsymbol{\omega}_{\boldsymbol{u}^c} = \boldsymbol{0}, \omega_i \neq 0 \text{ for } i \in \boldsymbol{u}\}. \tag{3.4}$$

An illustration for the three-dimensional case can be found in Figure 3.1. In general $f_{\boldsymbol{u}} \notin L_2(\mathbb{R}^{|\boldsymbol{u}|})$, but $f_{\boldsymbol{u}} \in L_2(\mathbb{R}^{|\boldsymbol{u}|}, \mu_{\boldsymbol{u}})$.

Figure 3.1: Decomposition of the frequency domain $\mathbb{R}^3$ into the lower dimensional parts. The lower dimensional subsets $Q_{\boldsymbol{v}}$ are not part of the higher-dimensional $Q_{\boldsymbol{u}}$ for $\boldsymbol{v} \subset \boldsymbol{u}$.

4 Random Fourier features

Kernel-based approaches have been extensively used in data-based applications, including image classification and high-dimensional function approximations since they often perform well in practice. The random feature model is a popular technique for approximating the kernel using a randomized basis that can avoid the cost of full kernel methods. An alternative perspective is to view the random feature model as a non-linear randomized function approximation. The theoretical foundation of random Fourier features builds on Bochner's characterization of positive definite functions, see also the pioneering work for random features [28].

Theorem 4.1 (Bochner's theorem [1]). *A continous and shift-invariant function $\kappa \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if κ is the Fourier transform of a non-negative measure.*

If a shift-invariant kernel κ is properly scaled with $k(\boldsymbol{0}) = 1$, Bochner's theorem guarantees that its Fourier transform $\rho(\boldsymbol{\omega})$ is a proper probability distribution, which means that

$$k(\boldsymbol{x} - \boldsymbol{x}') = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} - \boldsymbol{x}' \rangle} \rho(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} = \mathbb{E}_{\boldsymbol{\omega} \sim \rho} \left(\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle} \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x}' \rangle} \right). \tag{4.1}$$

See also [25, Chapter 4.4] According to (4.1), the random Fourier feature model makes use of the standard Monte Carlo sampling scheme to approximate $\kappa(\boldsymbol{x}, \boldsymbol{x}')$. In particular, one uses the approximation

$$k(\boldsymbol{x} - \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{\omega} \sim \rho} \left(\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle} \cdot \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x}' \rangle} \right) \approx \boldsymbol{A}(\boldsymbol{x}) \cdot \boldsymbol{A}(\boldsymbol{x}')^*,$$

with the explicit feature mapping

$$\boldsymbol{A}(\boldsymbol{x}) = \left(\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_1, \boldsymbol{x} \rangle}, \cdots, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_N, \boldsymbol{x} \rangle} \right) \in \mathbb{C}^{|\mathcal{X}|, N},$$

where $\{\boldsymbol{\omega}_k\}_{k=1}^N$ are sampled from $\rho(\boldsymbol{\omega})$ independently of the training set $\mathcal{X}$. Consequently, the original kernel matrix $\boldsymbol{K} = (\kappa(\boldsymbol{x} - \boldsymbol{x}'))_{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{x}' \in \mathcal{X}}$ can be approximated by $\boldsymbol{K} \approx \frac{1}{N} \boldsymbol{A} \boldsymbol{A}^*$. In the case of $N \ll |\mathcal{X}|$, this is a low rank approximation of the kernel which is for a big amount of samples $|\mathcal{X}|$ computational more feasible than the kernel matrix $\boldsymbol{K}$ itself.

We will focus on two families of feature distributions, Gaussian features and the Sobolev-type features with i.i.d coordinates (associated to the Gaussian kernel and Laplace-type kernel, respectively):

$$\rho_{\mathcal{N}}^{\sigma}(\boldsymbol{\omega}) := \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{d} \exp\left(-\|\boldsymbol{\omega}\|^{2}/(2\sigma^{2})\right), \qquad \text{Gaussian density}, \sigma > 0$$

$$\rho_{\Pi}^{s,\sigma}(\boldsymbol{\omega}) := c_{\rho_{\Pi}^{s}} \prod_{i\in[d]} \frac{1}{\sigma\left(1+\omega_{i}^{2}/\sigma^{2}\right)^{s}}, \qquad \text{tensor-product density } (d \geq 2, s > \tfrac{1}{2}, \sigma > 0),$$

with the constant $c_{\rho_{\Pi}^{s}} := \left(\frac{\Gamma(s)}{\sqrt{\pi}\,\Gamma(s-\frac{1}{2})}\right)^{d}$ chosen to ensure the associated densities have unit mass. One special case of the tensor-product density, is the tensor-product **Cauchy** distribution

$$\rho_{C}^{\sigma}(\boldsymbol{\omega}) := \prod_{i=1}^{d} \frac{1}{\pi\sigma(1+w_{i}^{2}/\sigma^{2})}.$$

From the neural network point of view, this is a two-layer network with a randomized but fixed single hidden layer. Given an unknown function $f \colon \mathbb{R}^{d} \to \mathbb{C}$, the random Fourier feature model takes the form

$$f^{\#}(\boldsymbol{x}) = \sum_{j=1}^{N} a_{j}\mathrm{e}^{\mathrm{i}\langle\boldsymbol{\omega}_{j},\boldsymbol{x}\rangle},$$

where $\boldsymbol{x} \in \mathbb{R}^{d}$ is the input data, $(\boldsymbol{\omega}_{j})_{j=1}^{N}$ are the random weights and $\boldsymbol{a} = (a_{j})_{j=1}^{N} \in \mathbb{C}^{N}$ is the final weight layer.

Existing algorithms differ in how they select features $\boldsymbol{\omega}_{j}$ and weights a_{j}. In most cases, the features $\boldsymbol{\omega}_{j}$ are independent and identically distributed random variables generated by the (user defined) probability density function $\rho(\boldsymbol{\omega})$. Then, for the random Fourier feature model, the output layer $\boldsymbol{a}$ is trained (training data-dependent or independent), while the hidden layer (the weights $\boldsymbol{\omega}_{j}$) are fixed.

Suppose we are given a probability density ρ used to sample the entries of the random weights $\boldsymbol{\omega}$. Let us recall the definition for bounded $\mathcal{F}(\rho)$-norm functions, see also [7, 39, 32].

Definition 4.2. *Let $\rho \colon \mathbb{R}^{d} \to \mathbb{R}$ be a density function. A function $f \colon \mathbb{R}^{d} \to \mathbb{R}$ has finite $\mathcal{F}(\rho)$-norm with respect to $\mathrm{e}^{\langle\boldsymbol{\omega},\cdot\rangle}$ if it belongs to the class*

$$\mathcal{F}(\rho) := \left\{ f(\boldsymbol{x}) = \int_{\mathbb{R}^{d}} \hat{f}(\boldsymbol{\omega})\mathrm{e}^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}\rangle}\,\mathrm{d}\boldsymbol{\omega} \mid \|f\|_{\mathcal{F}(\rho)} := \sup_{\boldsymbol{\omega}\in\mathbb{R}^{d}} \left|\frac{\hat{f}(\boldsymbol{\omega})}{\rho(\boldsymbol{\omega})}\right| < \infty \right\}. \tag{4.2}$$

Choosing a Gaussian distribution ρ like in [39, 7] is a very restrictive condition to the function space $\mathcal{F}(\rho)$, since the Fourier transform $\hat{f}$ has to decay faster than exponentially. For instance, in Sobolev spaces the decay of the Fourier transform is polynomially. For functions in $\mathcal{F}(\rho)$ generalization error bounds for random feature ridge regression from [39, 7] achieve the rate $\mathcal{O}(M^{-1})$, provided the number of data samples grows with M and satisfies certain statistical assumptions.

When more information is known about the target function f, the rates and complexity bounds improve (especially with respect to the dimension). This helps mitigate issues with the approximation of functions in high-dimensions. Especially, if the function is of low order. This is what we want to study in this Section. If the function f is of low order, say q, the Fourier transform $\hat{f}$ is not defined as a function, but only in distributional sense, which makes the norm in Definition 4.2 infinite. This also concerns the ρ-norm for function of lower dimension, defined in [7] if the effective dimension of the function f does not equal q, the sparsity of the drawn random features. For that reason we introduce in the following ANOVA truncated random Fourier features.

4.1 Random Fourier Features and ANOVA

In this section we relate the generalized ANOVA decomposition (3.3) to the random feature approximation. If nothing is known about the function f, it might be appropriate to use random Fourier features from a d-dimensional density ρ. But if the function is of low order or sparse in the ANOVA-terms, it is useful to decompose the density ρ having the ANOVA decomposition of the function f in mind. Shortly, we draw $n_{\boldsymbol{u}}$ random Fourier features supported on $\boldsymbol{u}$ for every ANOVA term $\boldsymbol{u}$ in the set $U \subset \mathcal{P}([d])$.

Assume the function f has a sparse ANOVA decomposition, i.e. for small $\epsilon > 0$,

$$\|f - \mathcal{T}_U f\|_{L_2(\mathbb{R}^d, \mu)} \leq \epsilon, \qquad \text{where } \mathcal{T}_U f(\boldsymbol{x}) = \sum_{\boldsymbol{u} \in U} f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}).$$

For such functions it is reasonable to reduce the dimension of the random Fourier features, where the remaining main task is to find the index-set U. In this paper we aim to develop Algorithms 1 and 2 for finding this index set U.

Definition 4.3. *[ANOVA-truncated random Fourier features] Let $U \subset \mathcal{P}([d])$ be a set of ANOVA indices and the functions $\rho_{\boldsymbol{u}} \colon \mathbb{R}^{|\boldsymbol{u}|} \to \mathbb{R}$ be probability distributions. A collection of $N = \sum_{\boldsymbol{u} \in U} n_{\boldsymbol{u}}$ weight vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N$ is called a set of **ANOVA-truncated random Fourier features**, if it is generated as follows: For each index $\boldsymbol{u} \in U$ draw $n_{\boldsymbol{u}}$ realizations $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n_{\boldsymbol{u}}}$ from $\rho_{\boldsymbol{u}}$ and construct $|\boldsymbol{u}|$-sparse features $\boldsymbol{\omega}_k$ by setting $\mathrm{supp}(\boldsymbol{\omega}_k) = \boldsymbol{u}$ and $(\boldsymbol{\omega}_k)_{\boldsymbol{u}} = \boldsymbol{z}_k$. All random Fourier features are collected in the index-set $\mathcal{I}$, and define the notation $\mathcal{I}_{\boldsymbol{u}} = \{\boldsymbol{\omega} \in \mathcal{I} \mid \boldsymbol{\omega} \in Q_{\boldsymbol{u}}\}$, where $Q_{\boldsymbol{u}}$ is defined in (3.4).*

Note that the algorithms [7, 32, 39] are restricted to index-sets $U = \{\boldsymbol{u} \subseteq [d] \mid |\boldsymbol{u}| = q\}$. Concerning the interpretability of the results this is a disadvantage, since it can happen that non-important input variables gain significance, see the following example. For that reason, we use random Fourier features of different dimension up to order q, and not only random Fourier features of order q.

Example 4.4. Introduce a function of the form

$$f(x_1, \ldots, x_{20}) = f_{1,2}(x_1, x_2) + f_3(x_3) + f_4(x_4) + f_5(x_5),$$

this includes the Friedmann function considered in [32, Fig.4]. The Fourier transform $T_{\hat{f}}$ is supported on

$$Q_{\varnothing} \cup Q_{\{1\}} \cup Q_{\{2\}} \cup Q_{\{1,2\}} \cup Q_{\{3\}} \cup Q_{\{4\}} \cup Q_{\{5\}}.$$

Choosing q-sparse random Fourier features with $U = \{\boldsymbol{u} \subset [d] \mid |\boldsymbol{u}| = 1\}$ and without demanding some orthogonality to the ANOVA terms, means that for example the ANOVA term f_1 can be described by a sum $\sum_{k=2}^{20} \sum_{\boldsymbol{\omega} \in \mathcal{I}_{\{1,k\}}} \mathrm{e}^{\mathrm{i}(\omega_1 x_1 + \omega_k x_k)}$. This leads to the problem, that coefficients $a_{\boldsymbol{\omega}}$ for $\boldsymbol{\omega} \in \mathcal{I}_{\{1,k\}}$ for some $k \in \{2, \ldots, 20\}$ describe the ANOVA term f_1 and are non-zero, despite the fact that the variable x_k does not play a role in the function. Then, analyzing the histogram based on the occurrence rate (as a percentage) of the input variables obtained from the HARFE model like in [32, Fig.4], leads to non-zero weights for non-necessary variables. $\qquad \square$

In applications the function f is unknown, even if a formula for f is available, the component functions $f_{\boldsymbol{u}}$ can not be calculated analytically using for example the coupled equations like in [29]. In typical applications, the function f is only available by sampling points $\boldsymbol{x} \in \mathcal{X}$ from modelling or experiments. Therefore, a practical numerical method is needed to construct each unique component function. Similar to [14], we minimize the squared error under the hierarchical orthogonality condition (3.1). For a set $U \subseteq \mathcal{P}([d])$ and ANOVA-truncated random Fourier features $\boldsymbol{\omega} \in \mathcal{I}$ drawn according to Definition 4.3, we construct the random feature matrix

$$\boldsymbol{A} = (\boldsymbol{A}_{\boldsymbol{u}})_{\boldsymbol{u} \in U} \quad \text{with} \quad \boldsymbol{A}_{\boldsymbol{u}} = \left(\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{u}} \rangle}\right)_{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{\omega}_{\boldsymbol{u}} \in \mathcal{I}_{\boldsymbol{u}}}. \tag{4.3}$$

Employing the generalized ANOVA decomposition, every term $f_{\boldsymbol{u}}$ of the function f by a sum

$$f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) \approx \sum_{\boldsymbol{\omega}_{\boldsymbol{u}} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{\omega}} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{u}} \rangle},$$

where the related random Fourier features $\boldsymbol{\omega}$ are in $Q_{\boldsymbol{u}}$. To find a suitable vector $\boldsymbol{a} = (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega} \in \mathcal{I}}$, we use a regularization, which is similar to defining the cost function like the D-MORPH algorithm [14] does: A solution vector $\boldsymbol{a}$ should fulfill simultaneously $\|\boldsymbol{A}\boldsymbol{a} - \boldsymbol{f}\|_2 = 0$ and the hierarchical orthogonality (3.1). Since the regularization by forcing the integrals (3.2) to be zero is not numerically feasible, we use a discretization of the integrals $\langle f_{\boldsymbol{u}}, f_{\boldsymbol{v}} \rangle_{\mu}$ instead:

$$\langle f_{\boldsymbol{u}}, f_{\boldsymbol{v}} \rangle_{\mathcal{X}} := \frac{1}{M} \sum_{\boldsymbol{x} \in \mathcal{X}} f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) \overline{f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}})}. \tag{4.4}$$

In contrast to [14], we do not want to enforce the hierarchical orthogonality (3.1) by calculating an SVD, but rather by penalizing by using a regularization. The solution vector $\boldsymbol{a}^{\#}$ can then be obtained by minimizing

$$\boldsymbol{a}^{\#} = \operatorname*{argmin}_{\boldsymbol{a}} \|\boldsymbol{A}\boldsymbol{a} - \boldsymbol{f}\|^2 + \lambda \|\boldsymbol{a}\|_{\hat{\boldsymbol{W}}}^2,$$

$$= \operatorname*{argmin}_{\boldsymbol{a}} \left\| \begin{pmatrix} \boldsymbol{A} \\ \sqrt{\lambda \hat{\boldsymbol{W}}} \end{pmatrix} \boldsymbol{a} - \begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{0} \end{pmatrix} \right\|_2^2 \tag{4.5}$$

$$\text{where } \|\boldsymbol{a}\|_{\hat{\boldsymbol{W}}}^2 = \boldsymbol{a}^* \hat{\boldsymbol{W}} \boldsymbol{a} = \sum_{\boldsymbol{u} \in U} \boldsymbol{a}_{\boldsymbol{u}} \hat{\boldsymbol{W}}_{\boldsymbol{u}} \boldsymbol{a}_{\boldsymbol{u}}, \tag{4.6}$$

13

where we introduce the weight matrix $\hat{\boldsymbol{W}}$ by

$$\hat{\boldsymbol{W}} := \operatorname{diag}\left(\hat{\boldsymbol{W}}_{\boldsymbol{u}}\right)_{\boldsymbol{u}\in U},$$

$$\hat{\boldsymbol{W}}_{\boldsymbol{u}} = \frac{1}{M^2}\boldsymbol{A}_{\boldsymbol{u}}^*\left(\boldsymbol{A}_{\boldsymbol{v}_1}\boldsymbol{A}_{\boldsymbol{v}_2}\cdots\right)\begin{pmatrix}\boldsymbol{A}_{\boldsymbol{v}_1}^*\\\boldsymbol{A}_{\boldsymbol{v}_2}^*\\\cdots\end{pmatrix}\boldsymbol{A}_{\boldsymbol{u}} = \frac{1}{M^2}\boldsymbol{A}_{\boldsymbol{u}}^*\left(\sum_{\boldsymbol{v}\subset\boldsymbol{u}}\boldsymbol{A}_{\boldsymbol{v}}\boldsymbol{A}_{\boldsymbol{v}}^*\right)\boldsymbol{A}_{\boldsymbol{u}} \in \mathbb{C}^{n_{\boldsymbol{u}}\times n_{\boldsymbol{u}}}, \tag{4.7}$$

where $\boldsymbol{v}_i$ runs through all subsets $\boldsymbol{v}\subset\boldsymbol{u}$ and $\varnothing$ is also a subset $\boldsymbol{v}$ of $\boldsymbol{u}$. The following lemma shows that this regularization coincides with the hierarchical orthogonality (3.1).

Lemma 4.5. *The regularization term $\boldsymbol{a}_{\boldsymbol{u}}^*\hat{\boldsymbol{W}}_{\boldsymbol{u}}\boldsymbol{a}_{\boldsymbol{u}}$ in (4.5) with the weight matrices $\hat{\boldsymbol{W}}_{\boldsymbol{u}}$, defined in (4.7) penalizes the non-orthogonality of the terms $f_{\boldsymbol{u}}$ and $f_{\boldsymbol{v}}$ where $\boldsymbol{v}\subset\boldsymbol{u}$ with the discrete scalar product (4.4) in the sense that*

$$\boldsymbol{a}_{\boldsymbol{u}}^*\hat{\boldsymbol{W}}_{\boldsymbol{u}}\boldsymbol{a}_{\boldsymbol{u}} \geq \sum_{\boldsymbol{v}\subset\boldsymbol{u}}\frac{1}{\|\boldsymbol{a}_{\boldsymbol{v}}\|^2}|\langle f_{\boldsymbol{u}}, f_{\boldsymbol{v}}\rangle_{\mathcal{X}}|^2.$$

Proof. The Cauchy-Schwarz inequality gives for all vectors $\boldsymbol{b}, \boldsymbol{c}$ and matrix $\boldsymbol{A}$ with suitable size that

$$|\langle \boldsymbol{A}\boldsymbol{b}, \boldsymbol{c}\rangle|^2 \leq \|\boldsymbol{A}\boldsymbol{b}\|^2\|\boldsymbol{c}\|^2.$$

Applying this to our setting yields for

$$\begin{aligned}
|\langle f_{\boldsymbol{u}}, f_{\boldsymbol{v}}\rangle_{\mathcal{X}}|^2 &= \frac{1}{M^2}|\langle\boldsymbol{A}_{\boldsymbol{u}}\boldsymbol{a}_{\boldsymbol{u}}, \boldsymbol{A}_{\boldsymbol{v}}\boldsymbol{a}_{\boldsymbol{v}}\rangle|^2 = \frac{1}{M^2}|\langle\boldsymbol{A}_{\boldsymbol{v}}^*\boldsymbol{A}_{\boldsymbol{u}}\boldsymbol{a}_{\boldsymbol{u}}, \boldsymbol{a}_{\boldsymbol{v}}\rangle|^2 \\
&\leq \frac{1}{M^2}\|\boldsymbol{A}_{\boldsymbol{v}}^*\boldsymbol{A}_{\boldsymbol{u}}\boldsymbol{a}_{\boldsymbol{u}}\|^2\|\boldsymbol{a}_{\boldsymbol{v}}\|^2 = \frac{\|\boldsymbol{a}_{\boldsymbol{v}}\|^2}{M^2}|\langle\boldsymbol{A}_{\boldsymbol{v}}^*\boldsymbol{A}_{\boldsymbol{u}}\boldsymbol{a}_{\boldsymbol{u}}, \boldsymbol{A}_{\boldsymbol{v}}^*\boldsymbol{A}_{\boldsymbol{u}}\boldsymbol{a}_{\boldsymbol{u}}\rangle| \\
&= \frac{\|\boldsymbol{a}_{\boldsymbol{v}}\|^2}{M^2}\boldsymbol{a}_{\boldsymbol{u}}^*\boldsymbol{A}_{\boldsymbol{u}}^*\boldsymbol{A}_{\boldsymbol{v}}\boldsymbol{A}_{\boldsymbol{v}}^*\boldsymbol{A}_{\boldsymbol{u}}\boldsymbol{a}_{\boldsymbol{u}}.
\end{aligned}$$

Hence, the regularization term $\boldsymbol{a}_{\boldsymbol{u}}\hat{\boldsymbol{W}}_{\boldsymbol{u}}\boldsymbol{a}_{\boldsymbol{u}}$ has the following connection to the discrete orthogonality of the terms $f_{\boldsymbol{u}}$ and $f_{\boldsymbol{v}}$, if $\boldsymbol{a}_{\boldsymbol{v}}\neq\boldsymbol{0}$,

$$\boldsymbol{a}_{\boldsymbol{u}}^*\hat{\boldsymbol{W}}_{\boldsymbol{u}}\boldsymbol{a}_{\boldsymbol{u}} = \boldsymbol{a}_{\boldsymbol{u}}^*\left(\frac{1}{M^2}\boldsymbol{A}_{\boldsymbol{u}}^*\left(\sum_{\boldsymbol{v}\subset\boldsymbol{u}}\boldsymbol{A}_{\boldsymbol{v}}\boldsymbol{A}_{\boldsymbol{v}}^*\right)\boldsymbol{A}_{\boldsymbol{u}}\right)\boldsymbol{a}_{\boldsymbol{u}} \geq \sum_{\boldsymbol{v}\subset\boldsymbol{u}}\frac{1}{\|\boldsymbol{a}_{\boldsymbol{v}}\|^2}|\langle f_{\boldsymbol{u}}, f_{\boldsymbol{v}}\rangle_{\mathcal{X}}|^2.$$

This finishes the proof. ∎

For the one-dimensional terms $\boldsymbol{u} = \{i\}$ we obtain equality in the previous lemma, the weight matrix $\hat{\boldsymbol{W}}_{\boldsymbol{u}}$ contains only $\boldsymbol{A}_{\varnothing}$ and is in this case equal to

$$\hat{\boldsymbol{W}}_{\{i\}} = \boldsymbol{1}_{M\times M}$$

and

$$\boldsymbol{a}_{\{i\}}^*\hat{\boldsymbol{W}}_{\{i\}}\boldsymbol{a}_{\{i\}} = \frac{1}{M^2}\boldsymbol{a}_{\{i\}}^*\boldsymbol{A}_{\{i\}}^*\begin{pmatrix}1 & \cdots & 1\\ & \vdots & \\ 1 & \cdots & 1\end{pmatrix}\boldsymbol{A}_{\{i\}}\boldsymbol{a}_{\{i\}} = \left|\frac{1}{M}\sum_{\boldsymbol{x}\in\mathcal{X}}f_i(x_i)\right|^2 = \frac{1}{f_{\varnothing}^2}|\langle f_i, f_{\varnothing}\rangle_{\mathcal{X}}|^2 = \left|\sum_{\boldsymbol{x}\in\mathcal{X}}f_i(x_i)\right|^2.$$

For the two-dimensional case $\boldsymbol{u} = \{i, j\}$ we have

$$\hat{\boldsymbol{W}}_{\{i,j\}} = \boldsymbol{1}_{M\times M} + \boldsymbol{A}_{\{i\}}\boldsymbol{A}_{\{i\}}^* + \boldsymbol{A}_{\{j\}}\boldsymbol{A}_{\{j\}}^*.$$

To solve the regularized least squares problem (4.5), we have to construct the matrix $\boldsymbol{A}$ and the matrix $\hat{\boldsymbol{W}}$, which is a block diagonal matrix, with blocks belonging to every $\boldsymbol{u}\in U$, so the square root has to be calculated for every block separately only. The actual minimization problem is then solved by an iterative least squares algorithm.

5   Sensitivity analysis

The aim of sensitivity analysis is to study how the output of a mathematical model or system can be divided and allocated to different input variables. Or, speaking in the ANOVA setting, to compare the variances $\sigma^2(f_{\boldsymbol{u}})$ of the different ANOVA terms. This simplifies the model: Overly complex models may complicate analysing the inputs. By performing sensitivity analysis, users can better understand what factors don't actually matter and can be removed from the model.

Suppose a set of ANOVA indices $U \subseteq \mathcal{P}([d])$ and ANOVA-truncated random Fourier features $\boldsymbol{\omega} \in \mathcal{I}$ according to Definition 4.3. Let some approximation to f be a sum $f^{\#}$ of the form

$$f^{\#}(\boldsymbol{x}) = \sum_{\boldsymbol{u} \in U} \sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{\omega}}^{\#} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega_u}, \boldsymbol{x_u}\rangle}, \tag{5.1}$$

with some coefficients $a_{\boldsymbol{\omega}}^{\#}$ (which can be the output of some optimization algorithm). In the following we will show two methods for performing sensitivity analysis for a function of kind (5.1). The first approach in Section 5.1 can be applied for independent input variables and exploits the tensor product structure of the corresponding sampling density μ. We start with q-sparse random Fourier features and include successively needed ANOVA-terms of order smaller than q. In contrast to that, in Section 5.2 we will first start by incorporating all ANOVA-terms up to order q and omit these with low variance.

5.1   Sensitivity analysis for independent input variables

In the case of a sum $f^{\#}(\boldsymbol{x}) = \sum_j a_j e_j(x)$ with basis functions e_j being orthonormal in $L_2(\mathbb{R}^d, \mu)$, the variances of the ANOVA terms can be calculated easily from the coefficients a_j, see for example [34, 17] for the exponential basis or the wavelet basis on the torus, respectively. In this paper we want to study the problem of unknown sampling density μ, such that an orthogonal basis is not available.

However, analogously to Lemma 2.2 the recursive definition (2.2) splits the function $f^{\#}$ into the terms

$$f_{\varnothing}^{\#} = \int_{\mathbb{R}^d} \sum_{\boldsymbol{\omega} \in \mathcal{I}} a_{\boldsymbol{\omega}}^{\#} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x}\rangle} \mu(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \sum_{\boldsymbol{\omega} \in \mathcal{I}} a_{\boldsymbol{\omega}}^{\#} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x}\rangle} \mu(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \frac{1}{(2\pi)^d} \sum_{\boldsymbol{\omega} \in \mathcal{I}} a_{\boldsymbol{\omega}}^{\#} \hat{\mu}(-\boldsymbol{\omega})$$

$$f_{\boldsymbol{u}}^{\#}(\boldsymbol{x_u}) = \frac{1}{(2\pi)^d} \sum_{\boldsymbol{\omega} \in \mathcal{I}} a_{\boldsymbol{\omega}}^{\#} \prod_{i \in \boldsymbol{u}} \left(\mathrm{e}^{\mathrm{i}\omega_i x_i} - \hat{\mu}_i(-\omega_i)\right) \prod_{i \in \boldsymbol{u}^c} \hat{\mu}_i(-\omega_i) = \frac{1}{(2\pi)^d} \sum_{\boldsymbol{\omega} \in \mathcal{I}} a_{\boldsymbol{\omega}}^{\#} E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \boldsymbol{u}),$$

with the terms E defined in (2.5).

Instead of calculating the integrals for the ANOVA decomposition (2.2), we derive advantage from the fact that the points are sampled from the density μ. Then the Monte-Carlo approximation of the integrals can be calculated using the RFF matrix $\boldsymbol{A}$ and a coefficient vector $\boldsymbol{a}^{\#}$.

We consider the setting where the set of ANOVA indices U is completely arbitrary. We propose to start with $U = \{u \in [d] \mid |u| = q\}$ and refining this set iteratively, we will give more details later in this section. Since the functions $\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \cdot\rangle}$ are not orthogonal for random drawn frequencies $\boldsymbol{\omega}$, the decomposition (5.1) is not the unique ANOVA decomposition of $f^{\#}$. But we notice, that for $\boldsymbol{u} \in U$ with $\{\boldsymbol{v} \in U \mid \boldsymbol{u} \subset \boldsymbol{v}\} = \varnothing$, the ANOVA term $f_{\boldsymbol{u}}^{\#}$ is completely contained in the sum $\sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{\omega}}^{\#} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega_u}, \boldsymbol{x_u}\rangle}$. The main idea of our procedure is to calculate, if such an ANOVA term is really necessary or if the indices can be reduced to even lower dimensional sparse random Fourier features.

Lemma 5.1. *Let $f^{\#}$ from (5.1) be the output of a trained random feature model. Fix a subset $\boldsymbol{u} \in U$ with $\{\boldsymbol{v} \in U \mid \boldsymbol{u} \subset \boldsymbol{v}\} = \varnothing$ and denote $g(\boldsymbol{x_u}) := \sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{\omega}}^{\#} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega_u}, \boldsymbol{x_u}\rangle}$. Then the Monte-Carlo approximation of the ANOVA terms of the function g at the sample points $\mathcal{X}$ with $|\mathcal{X}| = M$ are*

$$g_{\varnothing}^{MC} = \frac{1}{M} \sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} \boldsymbol{a}_{\boldsymbol{\omega}} \sum_{\boldsymbol{x}^{(j)} \in \mathcal{X}} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{x}^{(j)}, \boldsymbol{w}\rangle},$$

$$g_{\boldsymbol{v}}^{MC}(\boldsymbol{x_v}) = \frac{1}{M} \sum_{\boldsymbol{\omega} \in \mathcal{I}} \boldsymbol{a}_{\boldsymbol{\omega}} \sum_{\boldsymbol{x}^{(j)} \in \mathcal{X}} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{x}_{\boldsymbol{u} \setminus \boldsymbol{v}}^{(j)}, \boldsymbol{w}_{\boldsymbol{u} \setminus \boldsymbol{v}}\rangle} \prod_{i \in \boldsymbol{v}} \left(\mathrm{e}^{\mathrm{i}x_i \omega_i} - \mathrm{e}^{\mathrm{i}x_i^{(j)} \omega_i}\right).$$

Proof. We use induction over $|\boldsymbol{v}|$, and begin with $\boldsymbol{v} = \varnothing$,

$$g_{\varnothing} = \int_{\mathbb{R}^d} f^{\#}(\boldsymbol{x}) \mu(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \approx \frac{1}{M} \sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{\boldsymbol{\omega} \in \mathcal{I}} a_{\boldsymbol{\omega}} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x}\rangle} =: g_{\varnothing}^{\mathrm{MC}}.$$

For the induction step we use the recursive definition (2.2) of the ANOVA-terms. Additionally, the tensor product structure of the density μ, (2.1), is the basis to approximate an $|\boldsymbol{v}|$-dimensional integral with respect to $\mu_{\boldsymbol{v}}$ by

$$\int_{\mathbb{R}^{|\boldsymbol{v}|}} g(\boldsymbol{x}_{\boldsymbol{v}}, \boldsymbol{x}_{\boldsymbol{v}^c}) \mu_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) \,\mathrm{d}\boldsymbol{x}_{\boldsymbol{v}} \approx \frac{1}{M} \sum_{\boldsymbol{x}^{(j)} \in \mathcal{X}} g(\boldsymbol{x}_{\boldsymbol{v}}^{(j)}, \boldsymbol{x}_{\boldsymbol{v}^c}),$$

which is a function that depends on the variables $\boldsymbol{x}_{\boldsymbol{v}^c}$ and not on the variables $\boldsymbol{x}_{\boldsymbol{v}}$. Applying this to the recursive definition (2.2) yields

$$\begin{aligned}
g_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) &= \int_{\mathbb{R}^{|\boldsymbol{u}|-|\boldsymbol{v}|}} g(\boldsymbol{x}_{\boldsymbol{u}}) \mu(\boldsymbol{x}_{\boldsymbol{v}^c}) \,\mathrm{d}\boldsymbol{x}_{\boldsymbol{v}^c} - \sum_{\boldsymbol{v}' \subset \boldsymbol{v}} g_{\boldsymbol{v}'}(\boldsymbol{x}_{\boldsymbol{v}'}) \\
&\approx \frac{1}{M} \sum_{\boldsymbol{x}^{(j)} \in \mathcal{X}} g(\boldsymbol{x}_{\boldsymbol{v}}, \boldsymbol{x}_{\boldsymbol{u} \setminus \boldsymbol{v}}^{(j)}) - \sum_{\boldsymbol{v}' \subset \boldsymbol{v}} g_{\boldsymbol{v}'}^{\mathrm{MC}}(\boldsymbol{x}_{\boldsymbol{v}}) \\
&= \frac{1}{M} \sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{w}} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{x}_{\boldsymbol{v}}, \boldsymbol{\omega}_{\boldsymbol{v}}\rangle} \sum_{\boldsymbol{x}^{(j)} \in \mathcal{X}} \left(\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_{\boldsymbol{u} \setminus \boldsymbol{v}}, \boldsymbol{x}_{\boldsymbol{u} \setminus \boldsymbol{v}}^{(j)}\rangle} \right) \\
&\quad - \frac{1}{M} \sum_{\boldsymbol{v}' \subset \boldsymbol{v}} \sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{w}} \sum_{\boldsymbol{x}^{(j)} \in \mathcal{X}} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{x}_{\boldsymbol{u} \setminus \boldsymbol{v}'}^{(j)}, \boldsymbol{\omega}_{\boldsymbol{u} \setminus \boldsymbol{v}'}\rangle} \prod_{i \in \boldsymbol{v}'} \left(\mathrm{e}^{\mathrm{i}x_i \omega_i} - \mathrm{e}^{\mathrm{i}x_i^{(j)}\omega_i} \right) \\
&= \frac{1}{M} \sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{w}} \sum_{\boldsymbol{x}^{(j)} \in \mathcal{X}} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_{\boldsymbol{u} \setminus \boldsymbol{v}}, \boldsymbol{x}_{\boldsymbol{u} \setminus \boldsymbol{v}}^{(j)}\rangle} \left(\mathrm{e}^{\mathrm{i}\langle \boldsymbol{x}_{\boldsymbol{v}}, \boldsymbol{\omega}_{\boldsymbol{v}}\rangle} - \sum_{\boldsymbol{v}' \subset \boldsymbol{v}} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{x}_{\boldsymbol{v} \setminus \boldsymbol{v}'}^{(j)}, \boldsymbol{\omega}_{\boldsymbol{v} \setminus \boldsymbol{v}'}\rangle} \prod_{i \in \boldsymbol{v}'} \left(\mathrm{e}^{\mathrm{i}x_i \omega_i} - \mathrm{e}^{\mathrm{i}x_i^{(j)}\omega_i} \right) \right) \\
&= \frac{1}{M} \sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{w}} \sum_{\boldsymbol{x}^{(j)} \in \mathcal{X}} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega}_{\boldsymbol{u} \setminus \boldsymbol{v}}, \boldsymbol{x}_{\boldsymbol{u} \setminus \boldsymbol{v}}^{(j)}\rangle} \left(\prod_{i \in \boldsymbol{v}} \left(\mathrm{e}^{\mathrm{i}x_i \omega_i} - \mathrm{e}^{\mathrm{i}x_i^{(j)}\omega_i} \right) \right)
\end{aligned}$$

This finishes the proof. ∎

Of special interest for our algorithm is the case where $\boldsymbol{v} = \boldsymbol{u}$ in the previous lemma. With the known ANOVA-terms $g_{\boldsymbol{u}}^{\mathrm{MC}}$ for every index $\boldsymbol{u}$ we can estimate the variance of the ANOVA term $\boldsymbol{u}$ in $f^{\#}$ by

$$\sigma_{\mathrm{MC}}^2(g_{\boldsymbol{u}}) = \frac{1}{M-1} \sum_{\boldsymbol{x}^{(j)} \in \mathcal{X}} \left(g_{\boldsymbol{u}}^{\mathrm{MC}}(\boldsymbol{x}^{(j)}) - \sum_{\boldsymbol{x}^{(j)} \in \mathcal{X}} g_{\boldsymbol{u}}^{\mathrm{MC}}(\boldsymbol{x}^{(j)}) \right)^2.$$

Remark 5.2. The case where $\boldsymbol{v} = \boldsymbol{u}$ in the previous Lemma 5.1 is of special interest. In this case we calculate,

$$\left(f_{\boldsymbol{u}}^{\#}(\boldsymbol{x}_{\boldsymbol{u}}) \right)_{\boldsymbol{x} \in \mathcal{X}} \approx \frac{1}{M} \sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{w}}^{\#} \sum_{\boldsymbol{x}^{(j)} \in \mathcal{X}} \left(\prod_{i \in \boldsymbol{u}} \left(\mathrm{e}^{\mathrm{i}x_i \omega_i} - \mathrm{e}^{\mathrm{i}x_i^{(j)}\omega_i} \right) \right)$$

$$\frac{\sigma^2(f_{\boldsymbol{u}}^{\#}(\boldsymbol{x}_{\boldsymbol{u}}))}{\sigma^2(f)} \approx \frac{\left\| f_{\boldsymbol{u}}^{\#}(\boldsymbol{x}_{\boldsymbol{u}}) \right\|_{\ell_2(\mathcal{X})}^2}{\sigma^2(\boldsymbol{f})}. \tag{5.2}$$

Introducing three-dimensional tensors $\mathrm{e}^{\mathrm{i}x_i^{(j)}\omega_i} - \mathrm{e}^{\mathrm{i}x_i^{(\tilde{j})}\omega_i} \in \mathbb{C}^{|\mathcal{X}| \times |\mathcal{X}| \times n_{\boldsymbol{u}}}$ this is calculated numerically by a tensor-vector multiplication, followed by a summation, point-wise squaring and a summation. We approximate the variance of f by the variance of the given data vector $\boldsymbol{f}$. A splitting of the given data into test data and validation data gives the possibility to use the prediction on the validation set to estimate the variances of the ANOVA terms on validation data. □

The variance of $f_{\boldsymbol{u}}^{\#}$ is a good approximation to the variance of $f_{\boldsymbol{u}}$ if the error $\left\| f_{\boldsymbol{u}} - f_{\boldsymbol{u}}^{\#} \right\|_{L_2(\mathbb{R}^{|\boldsymbol{u}|}, \mu_{\boldsymbol{u}})}$ is small, see

$$\begin{aligned}
|\sigma^2(f_{\boldsymbol{u}}) - \sigma^2(f_{\boldsymbol{u}}^{\#})| &= \left| \int_{\mathbb{R}^{|\boldsymbol{u}|}} \left(|f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}})|^2 - |f_{\boldsymbol{u}}^{\#}(\boldsymbol{x}_{\boldsymbol{u}})|^2 \right) \mu_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) \mathrm{d}\boldsymbol{x}_{\boldsymbol{u}} \right| \\
&= \left| \int_{\mathbb{R}^{|\boldsymbol{u}|}} \left(|f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}})| - |f_{\boldsymbol{u}}^{\#}(\boldsymbol{x}_{\boldsymbol{u}})| \right) \left(|f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}})| + |f_{\boldsymbol{u}}^{\#}(\boldsymbol{x}_{\boldsymbol{u}})| \right) \mu_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) \mathrm{d}\boldsymbol{x}_{\boldsymbol{u}} \right| \\
&\leq \left\| f_{\boldsymbol{u}} - f_{\boldsymbol{u}}^{\#} \right\|_{L_2(\mathbb{R}^{|\boldsymbol{u}|}, \mu_{\boldsymbol{u}})} \left\| f_{\boldsymbol{u}} + f_{\boldsymbol{u}}^{\#} \right\|_{L_2(\mathbb{R}^{|\boldsymbol{u}|}, \mu_{\boldsymbol{u}})} \\
&\leq \left\| f_{\boldsymbol{u}} - f_{\boldsymbol{u}}^{\#} \right\|_{L_2(\mathbb{R}^{|\boldsymbol{u}|}, \mu_{\boldsymbol{u}})} \left\| f_{\boldsymbol{u}} \right\|_{L_2(\mathbb{R}^{|\boldsymbol{u}|}, \mu_{\boldsymbol{u}})} \left(2 + \left\| f_{\boldsymbol{u}} - f_{\boldsymbol{u}}^{\#} \right\|_{L_2(\mathbb{R}^{|\boldsymbol{u}|}, \mu_{\boldsymbol{u}})} \right). \tag{5.3}
\end{aligned}$$

The algorithms in the exsiting literature [7, 39, 32] used q-sparse random Fourier features, but they numerically verified that choosing q equal to the real effective dimension of the function f leads to best approximation results. Furthermore, the norm $\mathcal{F}(\rho)$ from Definition 4.2 is not finite, the Fourier transform $\hat{f}$ is defined only in distributional sense.

Definition 5.3. *A set U is called **anti downward closed** if for every $\boldsymbol{u} \in U$ there is no index $\boldsymbol{v} \in U$, which is a subset of $\boldsymbol{u}$.*

We propose to choose an anti downward closed set U. Furthermore, we propose to start with q-sparse random Fourier features and customize the random features to the ANOVA decomposition of the function f iteratively. This works as follows. We start with q-sparse random Fourier features as proposed in the literature so far by choosing $U = \{\boldsymbol{u} \in \mathcal{P}([d]) \mid |\boldsymbol{u}| = q\}$. Then we draw in total N random Fourier features and learn a first approximation, described by the parameter vector $\boldsymbol{a}^{\#}$. This is Stage I of Algorithm 1.

Then, using the variance estimations (5.2), Algorithm 1 decides in Stage II for every $\boldsymbol{u} \in U$, if it keeps this ANOVA index or if it omits this ANOVA index and uses instead all indices $\boldsymbol{u}$ of order $q - 1$ which are contained in $\boldsymbol{u}$. This procedure is done q times, to reduce the ANOVA index-set U to the really necessary variable interactions. If the function has only a low amount of non-zero ANOVA-terms this leads to a huge decrease of non-necessary parameters in the model, which is the starting point for the iterative pruning steps. We summarize this in Algorithm 1.

Algorithm 1 ANOVA boosting for independent input variables

Input: $\mathcal{X} = (\boldsymbol{x}^{(i)})_{i=1}^M \in \mathbb{R}^d$ sampling nodes
  $\boldsymbol{f} = (f(x^{(i)}))_{i=1}^M$ function values at sampling nodes
  q maximal superposition dimension
  ε ANOVA threshold
  N number of total random Fourier features
  λ regularization parameter

Stage I: Initialization

1: $U = \{\boldsymbol{u} \subseteq [d] \mid |\boldsymbol{u}| = q\}$

2: $n = \text{floor}\left(\frac{N}{|U|}\right)$.

3: For every $\boldsymbol{u} \in U$ draw n q-sparse features $\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}$ and construct the matrix

$$\boldsymbol{A} = [\boldsymbol{A_u}]_{\boldsymbol{u} \in U} \in \mathbb{C}^{M \times N}, \quad \boldsymbol{A_u} = (\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega_u}, \boldsymbol{x_u} \rangle})_{\boldsymbol{x} \in \mathcal{X}_{\text{train}}, \boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} \in \mathbb{C}^{M \times N}.$$

  First approximation: $\boldsymbol{a} = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^* + \lambda \boldsymbol{I})^{-1}\boldsymbol{f}$.

Stage II: ANOVA boosting

4: **for** $t = 1, \ldots, q$ **do**

5:   For every $\boldsymbol{u} \in U$ calculate the variances $\sigma_{\text{MC}}^2(g_{\boldsymbol{u}})$ using (5.2).

6:   $U \leftarrow \{\boldsymbol{u} \in U \mid \sigma_{\text{MC}}^2(g_{\boldsymbol{u}}) \geq \epsilon\}$

7:   $U_t = \{\boldsymbol{v} \in [d] \mid |\boldsymbol{v}| = t - 1, \nexists \boldsymbol{u} \in U \text{ with } \boldsymbol{v} \subset \boldsymbol{u}\}$

8:   $U \leftarrow U \cup U_t$

9:   Draw $|\boldsymbol{u}|$-sparse features $\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}$ for every $\boldsymbol{u} \in U_t$

10:   Construct the matrix $\boldsymbol{A}$ and update the approximation $\boldsymbol{a} = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^* + \lambda \boldsymbol{I})^{-1}\boldsymbol{f}$.

11: **end for**

 Output: U

In Figure 5.1 we illustrate this procedure for an example function, which can be written in the form

$$f\colon \mathbb{R}^7 \to \mathbb{R}, \quad f(\boldsymbol{x}) = f_{\{1,2,3\}}(x_1, x_2, x_3) + f_1(x_1) + f_{\{1,3\}}(x_1, x_3)f_{\{5\}}(x_5) + f_{\{6,7\}}(x_6, x_7). \tag{5.4}$$

We start with all three-dimensional terms, i.e. $U = \{\boldsymbol{u} \subset [d] \mid |\boldsymbol{u}| = 3\}$. A sensitivity analysis of the first approximation $f^{\#}$ shows that only the three-dimensional term $\{1, 2, 3\}$ has variance bigger than some threshold ϵ. The other three-dimensional terms can be replaced by all two-dimensional terms, where the terms $\{1, 2\}$ and $\{2, 3\}$ are not needed, because they are contained in the term $\{1, 2, 3\}$. A second approximation shows that only the variances of $f_{\{1,2,3\}}^{\#}$ and $f_{\{6,7\}}^{\#}$ are bigger than a threshold. In the third approximation only the additional one-dimensional terms $\{4\}, \{5\}$ are needed, since the other ones are contained in the higher-dimensional terms. After the third approximation only the important terms in an anti downward closed set U remain for the next approximation step to reduce the over-parametrized model to an under-parametrized model, for example using SHRIMP or HARFE.

Figure 5.1: Example procedure of finding the ANOVA-sparse random Fourier features: Consider a 7-dimensional input function of the form (5.4), starting at $q = 3$. The terms with approximated variance $\sigma^2(f_{\boldsymbol{u}})$ bigger than the threshold ϵ are highlighted in magenta for each approximation step. The result is an anti downward closed set U. At the bottom we give the number of indices in the index-set U, which is used in the respective step.

5.2 Sensitivity analysis for correlated input variables

If we do not have information about the sample density μ, we do not have the tensor product structure as for independent input variables. We want to come back to the regularized least squares (4.5). As shown in Lemma 4.5, this regularization ensures the hierarchical orthogonality (3.1) of the ANOVA terms. In this setting we demand that the sum $\sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{\omega}}^{\#} \mathrm{e}^{\mathrm{i} \langle \boldsymbol{\omega}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{u}} \rangle}$ should approximate the ANOVA term $f_{\boldsymbol{u}}$ for every $\boldsymbol{u} \in U$. Therefore, we start with $U = \{\boldsymbol{u} \subseteq [d] \mid |\boldsymbol{u}| \leq q\}$. Then the solution vector $\boldsymbol{a}^{\#}$ of (4.5) seperates the ANOVA terms in the sense that

$$f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) \approx \sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{\omega}}^{\#} \mathrm{e}^{\mathrm{i} \langle \boldsymbol{\omega}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{u}} \rangle}.$$

Let the RFF matrix $\boldsymbol{A}$ be split like in (4.3) and denote the vectors $\boldsymbol{a_u} = (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{\omega}}}$. Then we approximate the Sobol indices, see Definition 3.1 by

$$S_{\boldsymbol{u}, \text{var}}^{\text{MC}} = \frac{\left\| \boldsymbol{A_u a_u^{\#}} \right\|_2^2}{\sigma^2(\boldsymbol{f})} \tag{5.5}$$

$$S_{\boldsymbol{u}, \text{cor}}^{\text{MC}} = \frac{\sum_{\substack{\varnothing \neq \boldsymbol{v} \subseteq [d] \\ \boldsymbol{v} \cap \boldsymbol{u} \neq \varnothing, \boldsymbol{v} \not\subseteq \boldsymbol{u}}} \langle \boldsymbol{A_v a_v^{\#}}, \boldsymbol{A_u a_u^{\#}} \rangle}{\sigma^2(\boldsymbol{f})} \tag{5.6}$$

$$S_{\boldsymbol{u}}^{\text{MC}} = S_{\boldsymbol{u}, \text{var}}^{\text{MC}} + S_{\boldsymbol{u}, \text{cor}}^{\text{MC}}. \tag{5.7}$$

The procedure is summarized in Algorithm 2. Note, that this procedure can also be applied to samples from tensor product sampling densities μ. In that case we expect the indices $S_{\boldsymbol{u}, \text{cor}}$ to be zero.

Algorithm 2 ANOVA boosting for possibly dependent input variables

---

Input: $\mathcal{X} = (\boldsymbol{x}^{(i)})_{i=1}^M \in \mathbb{R}^d$ sampling nodes
$\boldsymbol{f} = (f(x^{(i)}))_{i=1}^M$ function values at sampling nodes
q maximal superposition dimension
ε ANOVA threshold
N total number of random Fourier features
λ regularization parameter

Stage I: Initialization
1: $U = \{\boldsymbol{u} \subseteq [d] \mid |\boldsymbol{u}| \leq q\}$

Stage II: ANOVA boosting
2: **for** $t = q, \ldots, 1$ **do**
3:     $n = \text{floor}\left(\frac{N}{|U|}\right)$.
4:     For every $\boldsymbol{u} \in U$ draw n q-sparse features $\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}$ and construct the matrix

$$\boldsymbol{A} = [\boldsymbol{A_u}]_{\boldsymbol{u} \in U} \in \mathbb{C}^{M \times N}, \quad \boldsymbol{A_u} = (\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\omega_u}, \boldsymbol{x_u} \rangle})_{\boldsymbol{x} \in \mathcal{X}_{\text{train}}, \boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} \in \mathbb{C}^{M \times N}.$$

    The solution vector $\boldsymbol{a}$ is solution of minimization problem (4.5) by an iterative least squares algorithm.
5:     For every $\boldsymbol{u} \in U$ calculate the Sobol indices $S_{\boldsymbol{u}, \text{var}}^{\text{MC}}$, $S_{\boldsymbol{u}, \text{cor}}^{\text{MC}}$ and $S_{\boldsymbol{u}}^{\text{MC}}$ using (5.5) to (5.7).
6:     $U \leftarrow \{\boldsymbol{u} \in U \mid S_{\boldsymbol{u}, \text{var}}^{\text{MC}} > \epsilon \text{ or } |\boldsymbol{u}| < t\}$.
7: **end for**
8: make anti downward closed set U (see Definition 5.3):

$$U \leftarrow U \backslash \{\boldsymbol{u} \in U \mid \exists \boldsymbol{v} \in U \text{ with } \boldsymbol{u} \subset \boldsymbol{v}\}.$$

9: Draw $|\boldsymbol{u}|$-sparse features $\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}$ for every $\boldsymbol{u} \in U$
Output: U

A good choice for the index-set U

We want to discuss two procedures done in Algorithm 2: The first one is the for-loop, which is a similar proceeding as in Algorithm 1. Another possibility would be to do just one step of approximation and omit all indices $\boldsymbol{u} \in U$ with $S_{\boldsymbol{u}, \text{var}}^{\text{MC}}$ smaller than the threshold ϵ independent of the order $|\boldsymbol{u}|$. But it turned out in numerical tests, to be beneficial to use the loop, because otherwise the algorithm would not be able to detect the correct ANOVA terms. The variances of terms $f_{\boldsymbol{u}}$ of order less than q are not estimated well enough when using ANOVA-truncated random Fourier features belonging to all $|U| = \sum_{i=0}^q \binom{d}{i}$ terms of order smaller or equal q.

The second procedure is, that in line 8 of Algorithm 2 we shrink the index-set U, such that there are no two sets contained, which are subsets $\boldsymbol{u} \subset \boldsymbol{v}$, which is necessary to receive an anti downward closed set U, see Definition 5.3. This is the better choice, since the ANOVA terms $\boldsymbol{v} \subseteq \boldsymbol{u}$ are already contained in the sum $\sum_{\boldsymbol{\omega} \in \mathcal{I}_{\boldsymbol{u}}} a_{\boldsymbol{\omega}} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{x_u}, \boldsymbol{\omega_u} \rangle}$. This is made clearer by the following example. Assume a two-dimensional function $f = f_{\varnothing} + f_{\{1\}} + f_{\{2\}} + f_{\{1,2\}}$, which we approximate by the sum

$$f^{\#} = \sum_{\boldsymbol{\omega} \in \mathcal{I}_{\{1,2\}} \subset Q_{\{1,2\}}} a_{\boldsymbol{\omega}}^{\#} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{x_u}, \boldsymbol{\omega_u} \rangle}.$$

The approximation $f^{\#}$ has non-zero ANOVA terms $f_{\varnothing}^{\#}$, $f_{\{1\}}^{\#}$ and $f_{\{2\}}^{\#}$, since the weak annihilating condition (3.2) would require in the case $f^{\#} = f_{\{1,2\}}^{\#}$ that

$$0 = \int_{\mathbb{R}} f_{\{1,2\}}^{\#}(x_1, x_2)\mu_{\{1,2\}}(x_1, x_2)\,\mathrm{d}x_1 = \sum_{\boldsymbol{\omega}\in\mathcal{I}_{\{1,2\}}} a_{\boldsymbol{\omega}}^{\#} \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}(x_1\omega_1 + x_2\omega_2)}\mu(x_1, x_2)\,\mathrm{d}x_1$$

$$= \sum_{\boldsymbol{\omega}\in\mathcal{I}_{\{1,2\}}} a_{\boldsymbol{\omega}}^{\#}\mathrm{e}^{\mathrm{i}x_2\omega_2} \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}x_1\omega_1}\mu(x_1, x_2)\,\mathrm{d}x_1,$$

which can not be true for arbitrary density μ, and x_2 and is also not demanded in the minimization of the RFF algorithms. This also applies to larger dimension d and other index-sets $\boldsymbol{u}$, that the ANOVA-terms $f_{\boldsymbol{v}}^{\#}$ are non-zero for $\boldsymbol{v}\subseteq\boldsymbol{u}$, if we draw random Fourier features from the set $Q_{\boldsymbol{u}}$. Thus, it is beneficial to use an anti-downward closed subset U. Numerical tests also indicate that the RFF algorithms yield better results in this case.

6 Theoretical analysis

In this section, we improve the analysis for the generalization error for the approximation by sparse random Fourier features. Following [7, 39, 32], we go through the finite-sum approximation

$$f^{\star}(\boldsymbol{x}) = \sum_{\boldsymbol{u}\in U} f_{\boldsymbol{u}}^{\star}(\boldsymbol{x_u}) = \sum_{\boldsymbol{u}\in U}\sum_{\boldsymbol{\omega}\in\mathcal{I_u}} a_{\boldsymbol{\omega}}^{\star}\,\mathrm{e}^{\mathrm{i}\langle\boldsymbol{\omega_u},\boldsymbol{x_u}\rangle}, \qquad a_{\boldsymbol{\omega}}^{\star} = \frac{\hat{T}_{f_{\boldsymbol{u}}}(\boldsymbol{\omega_u})}{n_{\boldsymbol{u}}\,(2\pi)^d\rho_{\boldsymbol{u}}(\boldsymbol{\omega_u})}. \tag{6.1}$$

This is motivated by the Monte Carlo approximation of the integral

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \hat{T}_f(\boldsymbol{\omega})\mathrm{e}^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}\rangle}\mathrm{d}\boldsymbol{\omega} = \sum_{\boldsymbol{u}\in U}\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\hat{T}_{f_{\boldsymbol{u}}}(\boldsymbol{\omega_u})\mathrm{e}^{\mathrm{i}\langle\boldsymbol{\omega_u},\boldsymbol{x_u}\rangle}\delta_{\boldsymbol{\omega_{u^c}}}\,\mathrm{d}\boldsymbol{\omega}.$$

Note that $f^{\star}$ is not known in practice, because $\hat{T}_f$ is not known. Furthermore, $\mathbb{E}_{\boldsymbol{\omega}}\left[f_{\boldsymbol{u}}^{\star}(\boldsymbol{x_u})\right] = f_{\boldsymbol{u}}(\boldsymbol{x_u})$ for fixed $\boldsymbol{x}$, which means that $\mathbb{E}_{\boldsymbol{\omega}}\langle f_{\boldsymbol{u}}^{\star}, f_{\boldsymbol{v}}^{\star}\rangle_{\mathcal{X}} = \langle f_{\boldsymbol{u}}, f_{\boldsymbol{v}}\rangle_{\mathcal{X}}$.

The function $f^{\star}$ allows the following error splitting,

$$\left\|f - f^{\#}\right\|_{L_2(\mathbb{R}^d,\mu)} \le \|f - \mathcal{T}_U f\|_{L_2(\mathbb{R}^d,\mu)} + \|\mathcal{T}_U f - f^{\star}\|_{L_2(\mathbb{R}^d,\mu)} + \left\|f^{\star} - f^{\#}\right\|_{L_2(\mathbb{R}^d,\mu)}$$

The first error is bounded in Theorem 2.7. Furthermore, it is known that in many real world problems the underlying function is of low effective, see [3, 5, 13].

Our procedure is as follows:

- In Lemma 6.1 we generalize the error $\|\mathcal{T}_U f - f^{\star}\|_{L_2(\mathbb{R}^d,\mu)}$ to the ANOVA setting by using ANOVA-sparse random feature instead of only q-sparse random Fourier features.

- We want to perform sensitivity analysis to calculate an index-set U, which is adapted to the function f. In (5.3) we show that we have a good approximation to the variances, if the approximation error is small, so our procedure finds the important terms.

- Once we have fixed a good ANOVA index-set U, we use SHRIMP or HARFE, so the approximation bounds from there are applicable for the error $\left\|f^{\star} - f^{\#}\right\|_{L_2(\mathbb{R}^d,\mu)}$. The used norm $\mathcal{F}(\rho)$ from (4.2) can be replaced by our norm (6.2).

Let us define the $\mathcal{F}(\rho)$-norm by

$$\|\!|f|\!\|_{\mathcal{F}(\rho)}^2 = \sum_{\boldsymbol{u}\in U}\frac{N}{n_{\boldsymbol{u}}\,(2\pi)^d}\left(\sup_{\boldsymbol{\omega}\in Q_{\boldsymbol{u}}}\frac{|\hat{T}_{f_{\boldsymbol{u}}}(\boldsymbol{\omega_u})|}{\rho_{\boldsymbol{u}}(\boldsymbol{\omega_u})}\right)^2, \tag{6.2}$$

where $Q_{\boldsymbol{u}}$ is defined in (3.4). For the approximant $f^{\star}$ we have the following.

Lemma 6.1. *Fix $\delta,\epsilon > 0$. Consider the random feature approximation $f^{\star}$ from (6.1). If the total number of features $N = \sum_{\boldsymbol{u}\in U} n_{\boldsymbol{u}}$ satisfies the bound*

$$N \ge \frac{1}{\epsilon^2}\left(1 + \sqrt{2\log(1/\delta)}\right)^2,$$

then with probability at least $1 - \delta$ with respect to the draw of weights $\boldsymbol{\omega}_j$ the following holds

$$\|T_U f - f^{\star}\|_{L_2(\mathbb{R}^d,\mu)} \le \epsilon\|\!|f|\!\|_{\mathcal{F}(\rho)}.$$

Proof. The proof follows similar arguments like [7, Lemma 1], but applied to our setting of ANOVA truncated random Fourier features. The coefficients $a_{\boldsymbol{\omega}}^{\star}$, defined in (6.1) are bounded by

$$|a_{\boldsymbol{\omega}}^{\star}| \leq \frac{1}{n_{\boldsymbol{u}}\,(2\pi)^d}\,\sup_{\boldsymbol{\omega}\in Q_{\boldsymbol{u}}} \frac{\hat{T}_{f_{\boldsymbol{u}}}(\boldsymbol{\omega}_{\boldsymbol{u}})}{\rho_{\boldsymbol{u}}(\boldsymbol{\omega}_{\boldsymbol{u}})},$$

and for fixed $\boldsymbol{x}$, $\mathbb{E}_{\boldsymbol{\omega}}\left[f_{\boldsymbol{u}}^{\star}(\boldsymbol{x}_{\boldsymbol{u}})\right] = f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}})$. Define the random variable

$$v(\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_N) = \|f - f^{\star}\|_{L_2(\mathbb{R}^d,\mu)} = \left(\int_{\mathbb{R}^d} |\mathbb{E}_{\boldsymbol{\omega}}(f^{\star}(\boldsymbol{x})) - f^{\star}(\boldsymbol{x})|^2 \mu(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)^{1/2}.$$

To apply McDiarmid's inequality from Theorem A.2, we show that v is stable to perturbation. In particular, let $f^{\star}$ be the random feature approximation using random weights $(\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_k,\ldots,\boldsymbol{\omega}_N)$ and let $\tilde{f}^{\star}$ be the random feature approximation using random weights $(\boldsymbol{\omega}_1,\ldots,\tilde{\boldsymbol{\omega}}_k,\ldots,\boldsymbol{\omega}_N)$ with $\operatorname{supp}\boldsymbol{\omega}_k = \boldsymbol{u}$, then

$$\begin{aligned}
|v(\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_k,\ldots,\boldsymbol{\omega}_N) - v(\boldsymbol{\omega}_1,\ldots,\tilde{\boldsymbol{\omega}}_k,\ldots,\boldsymbol{\omega}_N)| &\leq \left\|f^{\star} - \tilde{f}^{\star}\right\|_{L_2(\mathbb{R}^d,\mu)} \\
&= \left\|a_{\boldsymbol{\omega}_k}^{\star}\mathrm{e}^{\mathrm{i}\langle(\boldsymbol{\omega}_k)_{\boldsymbol{u}},\boldsymbol{x}_{\boldsymbol{u}}\rangle} - \tilde{a}_{\boldsymbol{\omega}_k}^{\star}\mathrm{e}^{\mathrm{i}\langle(\tilde{\boldsymbol{\omega}}_k)_{\boldsymbol{u}},\boldsymbol{x}_{\boldsymbol{u}}\rangle}\right\|_{L_2(\mathbb{R}^d,\mu)} \\
&\leq \frac{2}{n_{\boldsymbol{u}}}\left(\sup_{\boldsymbol{\omega}\in Q_{\boldsymbol{u}}} \frac{\hat{T}_{f_{\boldsymbol{u}}}(\boldsymbol{\omega}_{\boldsymbol{u}})}{\rho_{\boldsymbol{u}}(\boldsymbol{\omega}_{\boldsymbol{u}})}\right) =: \Delta_k.
\end{aligned}$$

Summing over the Δ_k yields,

$$\begin{aligned}
\sum_{k=1}^{N}\Delta_k^2 &\leq \sum_{\boldsymbol{u}\in U}\frac{4}{n_{\boldsymbol{u}}^2}\left(\sup_{\boldsymbol{\omega}\in Q_{\boldsymbol{u}}}\frac{\hat{T}_{f_{\boldsymbol{u}}}(\boldsymbol{\omega}_{\boldsymbol{u}})}{\rho_{\boldsymbol{u}}(\boldsymbol{\omega}_{\boldsymbol{u}})}\right)^2 \\
&\leq \frac{4\|\!|f|\!\|_{\mathcal{F}(\rho)}^2}{N}.
\end{aligned}$$

To estimate the expectation of v, we bound the expectation of the second moment. By noting that the variance of an average of i.i.d. random variables is the average of the variances of each variable and by using the relation between the variance and the un-centered second moment, we have that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\omega}}(v^2) &= \mathbb{E}_{\boldsymbol{\omega}}\|\mathbb{E}_{\boldsymbol{\omega}}(f^{\star}) - f^{\star}\|_{L_2(\mathbb{R}^d,\mu)} \\
&= \mathbb{E}_{\boldsymbol{\omega}}\left\|\sum_{\boldsymbol{u}\in U}\sum_{\boldsymbol{\omega}\in\mathcal{I}_{\boldsymbol{u}}}a_{\boldsymbol{\omega}}^{\star}\mathrm{e}^{\mathrm{i}\langle\boldsymbol{\omega}_v,\boldsymbol{x}_v\rangle}\right\|_{L_2(\mathbb{R}^d,\mu)}^2 - \left\|\mathbb{E}_{\boldsymbol{\omega}}\left(\sum_{\boldsymbol{u}\in U}a_{\boldsymbol{\omega}}^{\star}\mathrm{e}^{\mathrm{i}\langle\boldsymbol{\omega}_v,\boldsymbol{x}_v\rangle}\right)\right\|_{L_2(\mathbb{R}^d,\mu)}^2 \\
&\leq \sum_{\boldsymbol{u}\in U}\frac{1}{n_{\boldsymbol{u}}}\left(\sup_{\boldsymbol{\omega}\in Q_{\boldsymbol{u}}}\frac{\hat{T}_{f_{\boldsymbol{u}}}(\boldsymbol{\omega}_{\boldsymbol{u}})}{\rho_{\boldsymbol{u}}(\boldsymbol{\omega}_{\boldsymbol{u}})}\right)^2 \leq \frac{\|\!|f|\!\|_{\mathcal{F}(\rho)}^2}{N}.
\end{aligned}$$

By Jensen's inequality, the expectation of v is bounded by

$$\mathbb{E}_{\boldsymbol{\omega}}(v) \leq \left(\mathbb{E}_{\boldsymbol{\omega}}(v^2)\right)^{1/2} \leq \frac{\|\!|f|\!\|_{\mathcal{F}(\rho)}}{\sqrt{N}}.$$

Applying McDiarmids inequality from Theorem A.2, yields

$$\mathbb{P}\left(v \geq \frac{\|\!|f|\!\|_{\mathcal{F}(\rho)}}{\sqrt{N}}\right) \leq \exp\left(-\frac{2t^2}{\sum_k \Delta_k^2}\right) = \exp\left(-\frac{2t^2 N}{4\|\!|f|\!\|_{\mathcal{F}(\rho)}^2}\right).$$

Setting t and N to

$$t = \|\!|f|\!\|_{\mathcal{F}(\rho)}\sqrt{\frac{2}{N}\log(1/\delta)}$$

$$N \geq \frac{1}{\epsilon^2}\left(1 + \sqrt{2\log(1/\delta)}\right)^2$$

enforces that $v \leq \epsilon\|\!|f|\!\|_{\mathcal{F}(\rho)}$ with probability at least $1 - \delta$. This completes the proof. ■

7   Numerical results

We illustrate in this section our sensitivity analysis on test examples. First, we summarize in Section 7.1 two random Fourier feature algorithms from the literature, which we then use for the numerical experiments. We suggest to use one of our algorithms to find a good index-set U and then to adapt a random Fourier feature algorithm to ANOVA sparse random Fourier features.

We present numerical results of Algorithm 1, where we use independent input variables. Further, we illustrate in Section 7.3 the approximation procedure of Algorithm 2 through several numerical applications.

7.1   Algorithms for RFF

In the literature there are several algorithms for approximating high-dimensional sparse additive functions. Here we will summarize two of them. Since we assume limited data availability, we wish to have a sparse representation of the function f by learning the coefficient vector $\boldsymbol{a}$ with a sparsity constraint.

- In [39] the non-linear **SHRIMP** algorithm was proposed. There the authors propose to use q-sparse frequencies, which means that $|\operatorname{supp}\boldsymbol{\omega}| = q$ for all random feature weights, where the non-zero components are sampled from the Gaussian distribution $\mathcal{N}(0, \frac{1}{\sqrt{q}})$. The algorithm begins with a strong over-parametrization $N \gg M$. The first solution vector $\boldsymbol{a}$ is calculated by

$$\boldsymbol{a} = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^* + \lambda\boldsymbol{I})^{-1}\boldsymbol{f}.$$

  Then, using iterative magnitude pruning (IMP) and selecting the best model via a validation set, the algorithm output is the solution vector $\boldsymbol{a}^{\#}$.

  Iterative Magnitude Pruning is used for compressing over-parametrized neural networks. The IMP procedure prunes features on their magnitude and then retrains the pruned sub-network in each pruning iteration. In every iteration the pruning rate is $p < 1$. After calculating the MSE error of a validation set in every iteration one can choose the final model by choosing the smallest validation error.

  We want to generalize this algorithm to ANOVA-sparse random Fourier features as defined in Definition 4.3. In fact this is a generalization of choosing a tensor product density or a q-sparse density to a density ρ, which can have an arbitrary ANOVA decomposition.

- In [32] the authors solve the sparse random feature regression problem by a greedy algorithm named hard-ridge random feature expansion (**HARFE**), which uses a hard thresholding pursuit (HTP) like algorithm to solve the random feature ridge regression problem. Specifically, they learn the vector $\boldsymbol{a}$ from the following minimization problem

$$\min_{\boldsymbol{a}} \|\boldsymbol{A}\boldsymbol{a} - \boldsymbol{f}\|_2^2 + \lambda\|\boldsymbol{a}\|_2^2 \quad \text{sucht that } \boldsymbol{a} \text{ is s-sparse.}$$

  The idea is to solve for the coefficients using a much smaller number of model terms. The subset S given by the indices of the s largest entries of one gradient descent step applied on the vector $\boldsymbol{a}$ is a good candidate for the support set of $\boldsymbol{a}$. The HTP algorithm iterates between these two steps and leads to a stable and robust reconstruction of sparse vectors depending on the restricted isometry property (RIP) constant of the matrix $\boldsymbol{A}$, which characterizes matrices which are nearly orthonormal, at least when operating on sparse vectors.

As the numerical test for the RFF suggest, choosing the sparsity q of the ANOVA terms equal to effective dimension of the function f, gives the lowest approximation errors. We suggest to first calculate a good ANOVA index-set U for the function f with Algorithm 1 or 2, use this to draw ANOVA-sparse random Fourier features adapted to the function f and apply afterwards an algorithm for sparse random features, for example SHRIMP or HARFE.
The best chances of improving the previous algorithms have functions with ANOVA terms of different orders: In this case drawing ANOVA random features adapted to the function f will decrease the approximation error significantly.

7.2   Numerical results for independent input variables

To test the performance of the ANOVA boosting for independent input variables, we test Algorithm 1 on synthetic functions:

$$f_{T1}(\boldsymbol{x}) = x_4^2 + x_2 x_3 + x_1 x_2 + x_4 \tag{7.1}$$

$$f_{T2}(\boldsymbol{x}) = \sin(x_1) + 7\sin^2(x_2) + 0.1x_3^4 \sin(x_1) \qquad \text{Ishigami function} \tag{7.2}$$

$$f_{T3}(\boldsymbol{x}) = \left(10\sin(\pi\, x_1 x_2) + 20\left(x_3 - \tfrac{1}{2}\right)^2 + 10\, x_4 + 5x_5\right) \qquad \text{Friedmann function} \tag{7.3}$$

We randomly draw M points from $\mathcal{N}(0, I_d)$ (functions f_{T1}, f_{T2}) or uniformly on $[0, 1]^d$ (function f_{T3}) and use $N = 5M$ random Fourier features in the initialization step. We additionally draw M test samples from the same distribution to validate the approximation error using the MSE,

$$\text{MSE} = \frac{1}{|\mathcal{X}_{\text{test}}|} \sum_{\boldsymbol{x} \in \mathcal{X}_{\text{test}}} |f(\boldsymbol{x}) - f^{\#}(\boldsymbol{x})|^2.$$

We compare the performance of the SHRIMP/HARFE algorithm and the ANOVA boosted SHRIMP/ HARFE and denote the resulting algorithms as **ANOVA-S** and **ANOVA-H**, respectively. The random Fourier features were distributed i.i.d. according to Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \frac{1}{q}\mathcal{I}_d)$ or according to Cauchy distribution $\sim \prod_{i \in [d]}(1 + w_i^2)^{-1}$ with variance $\sigma = \frac{1}{q}$. In every case we did the approximation 10 times and show the mean. In every case we have chosen the regularization parameter $\lambda = 10^{-6}$ and the cut-off parameter $\epsilon = 0.01$. In Table 7.1 we summarize the results. Note that for the HARFE algorithm we used the exponential function $\mathrm{e}^{-\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle}$ in contrast to the authors of the numerical tests in [32], who used the cosine function. Possibly further research could study why the numerical results are better with the cosine functions, despite the theoretical results are mostly stated for exponential functions.

For summarizing the results, our procedure detects the important ANOVA terms, if enough samples are available. The random feature algorithms benefit from the first ANOVA boosting in Algorithm 1, where we could improve the accuracy by factor up to 10^2. But in any case the approximation error is smaller for the ANOVA boosted algorithms or at least comparable. Furthermore, our procedure improves previous algorithms by being interpretable by showing clearly which ANOVA terms are zero and which input variables are necessary for the learned final model.

function	d	q	M	ρ	SHRIMP	ANOVA-S	HARFE	ANOVA-H
f_{T1}	5	2	300	$\rho_{\mathcal{N}}$	$1.8324 \cdot 10^{-6}$	$1.4060 \cdot 10^{-6}$	0.3615	0.3005
				$\rho_{\mathcal{C}}$	$2.0259 \cdot 10^{-5}$	$1.4160 \cdot 10^{-6}$	0.5826	0.9081
	10	3	500	$\rho_{\mathcal{N}}$	0.0005	$1.4581 \cdot 10^{-6}$	2.3772	0.2151
				$\rho_{\mathcal{C}}$	0.0152	$1.7514 \cdot 10^{-6}$	3.4952	0.8934
f_{T2}	5	2	500	$\rho_{\mathcal{N}}$	0.0082	$2.6587 \cdot 10^{-5}$	0.1378	0.6071
				$\rho_{\mathcal{C}}$	0.0025	0.0028	0.5712	0.3841
	10	2	1000	$\rho_{\mathcal{N}}$	0.0055	$2.6213 \cdot 10^{-5}$	0.6650	0.6910
				$\rho_{\mathcal{C}}$	0.1213	0.0063	1.1395	0.8896
f_{T3}	5	3	500	$\rho_{\mathcal{N}}$	0.0032	0.0027	5.3181	0.4766
				$\rho_{\mathcal{C}}$	0.0001	0.0002	3.6378	0.9084
	10	3	200	$\rho_{\mathcal{N}}$	0.3808	0.0098	5.8178	2.0540
				$\rho_{\mathcal{C}}$	0.5433	0.0133	3.9776	2.6890

Table 7.1: Approximation results: MSE on test data for different functions. We compare the performance of the SHRIMP and HARFE algorithm with the ANOVA boosted algorithms using Algorithm 1. The random Fourier features were distributed i.i.d. according to $\rho_{\mathcal{N}} = \mathcal{N}(\boldsymbol{0}, \frac{1}{q}\mathcal{I}_d)$ or according to $\rho_{\mathcal{C}} \sim \prod_{i \in [d]}(1 + w_i^2)^{-1}$ with variance $\sigma = \frac{1}{q}$. In every case we did the approximation 10 times and show the mean.

7.3 Numerical results for dependent input variables

We illustrate that even for dependent input variables, it is possible to find non-zero ANOVA terms of the unknown function f. In this example, the ANOVA boosting method from Algorithm 2 is tested on a slightly modified Friedmann function, which is used as benchmark example for certain approximation techniques,

$$f(x_1, \ldots, x_9) = 10 \sin(0.1 \pi x_1 x_2) + 20 (x_3 - \tfrac{1}{2})^2 + 10 x_4 + 5x_5. \tag{7.4}$$

In contrast to the literature we do not use uniform samples on $[0, 1]^d$, but we use (partly) dependent Gaussian samples. Due to this sampling we changed the original function slightly to have comparable variances of the non-zero ANOVA terms. Based on the example in [29], the samples $\boldsymbol{x} \in \mathcal{X}$ are Gaussian random vectors with mean $\mathbb{E}\boldsymbol{x} = \boldsymbol{0}$ and with

the covariance matrix being one of the following,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{I}_9, \qquad\qquad\qquad\qquad \text{uncorrelated}$$

$$\boldsymbol{\Sigma}_2 = \tfrac{4}{5}\boldsymbol{I}_9 + \tfrac{1}{5}\boldsymbol{1}_{9\times 9}, \qquad\qquad\qquad \text{equally correlated}$$

$$\boldsymbol{\Sigma}_3 = \boldsymbol{I}_3 \otimes \begin{pmatrix} 1 & -\tfrac{1}{5} & \tfrac{2}{5} \\ -\tfrac{1}{5} & 1 & -\tfrac{4}{5} \\ \tfrac{2}{5} & -\tfrac{4}{5} & 1 \end{pmatrix}, \qquad\qquad \text{mixed correlated}$$

where $M = 500$. Independent of the drawn samples $\boldsymbol{x} \in \mathcal{X}$, the function (7.4) has non-zero ANOVA terms only for the index-set

$$U = \{\varnothing, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{1, 2\}\}. \tag{7.5}$$

Furthermore, for independent input variables, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$ the Sobolev indices can be easily calculated analytically, which gives

$$S_{\{3\},\text{var}} \approx 0.2788 \qquad\qquad\qquad S_{\{4\},\text{var}} \approx 0.3718$$
$$S_{\{5\},\text{var}} \approx 0.0929 \qquad\qquad\qquad S_{\{1,2\},\text{var}} \approx 0.2564.$$

We use Algorithm 2 to calculate the indices $S_{\boldsymbol{u},\text{var}}^{\text{MC}}$, where we draw in total $N = 5000$ ANOVA-sparse random Fourier features with $q = 2$ and variance $\tfrac{1}{2}$, and choose the regularization parameter $\lambda = 1$. The results are plotted in Figure 7.1. Note that the Sobolev indices $S_{\boldsymbol{u},\text{var}}^{\text{MC}}$ can be bigger than one, since for dependent input variables the variances of the ANOVA terms $\sigma^2(f_{\boldsymbol{u}})$ do not sum up to the variance of the function $\sigma^2(f)$. In Figure 7.1 we normalized the indices $S_{\boldsymbol{u},\text{var}}^{\text{MC}}$ by the sum

$$\sum_{\boldsymbol{u} \in \{\boldsymbol{u} \,||\, \boldsymbol{u} | \leq 2\}} S_{\boldsymbol{u},\text{var}}^{\text{MC}},$$

which has in total $\binom{9}{1} + \binom{9}{2} = 45$ summands. It can be clearly seen, that in every case only the terms with non-zero variance in the index-set U from (7.5) are significant. In contrast to the case $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$, for dependent variables ($\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$) the terms f_1 and f_2 are non-zero, which Algorithm 2 finds. Using only 2-sparse random Fourier features and the HARFE algorithm as proposed in [32], leads to results shown in their Fig.4, which clearly blur the importance of the variables x_1 to x_5 in comparison to the other non-necessary variables, see also Example 4.4. Furthermore, they only study the simpler case of independent input variables.

Figure 7.1: The indices $S_{\boldsymbol{u},\text{var}}^{\text{MC}}$ for Gaussian input samples with different covariance $\boldsymbol{\Sigma}$. The pie charts are normalized to the sum of all indices $S_{\boldsymbol{u},\text{var}}^{\text{MC}}$ for $|\boldsymbol{u}| \leq 2$, but the numbers represent the actual indices.

To conclude our numerical section, we want to compare approximation results of random feature algorithms with and without ANOVA boosting from Algorithm 2. We study the same test functions (7.1) to (7.3) as for the case of independent random variables. To define the dependence among random variables, it is usual to use copula functions, [11, 23]. Denote the cumulative distribution function of the samples by

$$R_{\mathcal{X}}(\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} \mu(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t},$$

and $R_1, \ldots, R_d$ are the marginal cumulative distribution functions of $x_1, \ldots, x_d$, i.e.,

$$R_i(x_i) = \int_{-\infty}^{x_i} \mu_i(t)\,\mathrm{d}t.$$

Sklar's theorem [35] is the building block of the theory of copulas. It states, that for continuous functions $R_1, \ldots, R_d$ there exists a d-dimensional Copula C, such that for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$R_{\mathcal{X}}(\boldsymbol{x}) = C(R_1(x_1), \cdots, R_d(x_d)).$$

The copula C contains all information on the dependence structure between components of $(x_1, \ldots, x_d)$, whereas the cumulative distribution functions R_i contain all information on the marginal distribution of x_i. Especially Archimedean copulas are an important class of multivariate dependence models, since it is very easy to generate random numbers from them. Every Archimedean copula has the simple algebraic form

$$C(y_1, \ldots, y_d) = \psi\left(\psi^{-1}(y_1) + \ldots, \psi^{-1}(y_d)\right),$$

where ψ is the generator function of the copula. We test our algorithm for the following well-known copulas with parameter θ:

$$\psi(t) = \frac{1}{\theta}\left(t^{-\theta} - 1\right) \qquad\qquad \theta > 0 \qquad\qquad \text{Clayton copula,}$$

$$\psi(t) = (-\ln t)^{\theta} \qquad\qquad \theta \geq 1 \qquad\qquad \text{Gumbel copula,}$$

$$\psi(t) = -\ln\left(\frac{\exp{-\theta t} - 1}{\exp{-\theta} - 1}\right) \qquad\qquad \theta > 0 \qquad\qquad \text{Frank copula.}$$

We use these copulas on the marginal distributions $\mathcal{N}(0, 1)$ (f_{T1}), uniform on $[-\pi, \pi]$ (f_{T2}) or uniform on $[0, 1]$ (f_{T3}).

We apply Algorithm 2 for different settings of d, M, q with fixed parameter $N = 5M$ and Gaussian random Fourier features drawn from $\mathcal{N}(\boldsymbol{0}, \frac{1}{q}\mathcal{I}_d)$. The used input parameters for our algorithm are summarized in Table 7.2 for the different settings. For a better comparison we use the same regularization parameter $\lambda = 10^{-6}$ for the SHRIMP steps in both cases. The resulting MSE on test data are summarized in Table 7.2, we did the procedure 10 times and show the mean. In any case, the ANOVA boost leads to a clear improvement of the approximation results.

function	d	q	M	C	θ	λ	ϵ	SHRIMP	ANOVA-S
f_{T2}	10	3	500	1	3	100	0.01	$4.47855 \cdot 10^{-5}$	$1.04677 \cdot 10^{-6}$
				2	2	100	0.01	$8.55061 \cdot 10^{-5}$	$1.15639 \cdot 10^{-6}$
				3	4	100	0.01	0.00028	$7.44458 \cdot 10^{-7}$
	20	2	500	1	2	100	0.01	0.00504	$1.80438 \cdot 10^{-6}$
				2	1	100	0.01	0.00634	$1.15239 \cdot 10^{-6}$
				3	4	100	0.01	0.00400	$8.27697 \cdot 10^{-7}$
f_{T2}	5	2	500	1	2	200	0.05	0.02287	0.00034
				2	2	100	0.05	0.00282	0.00050
				3	5	200	0.05	0.01732	0.00341
	10	2	500	1	2	200	0.05	0.31271	0.00075
				2	2	100	0.05	0.40442	0.00075
				3	5	200	0.05	0.47998	0.00591
f_{T3}	10	3	200	1	5	100	0.01	0.13449	0.00952
				2	2	100	0.01	0.41547	0.01573
				3	4	100	0.01	0.21362	0.03427
	20	2	500	1	3	100	0.01	0.02844	0.01603
				2	3	100	0.01	0.02651	0.00494
				3	3	100	0.01	0.02870	0.00071

Table 7.2: Approximation results: MSE on test data for different settings. The column C belongs to the copula: $1, 2, 3$ corresponds to Clayton, Gumbel and Frank copula, respectively. We compare the performance of the SHRIMP and the ANOVA boosted SHRIMP using Algorithm 2 for dependent input variables. The random Fourier features were distributed according to $\rho_{\mathcal{N}} = \mathcal{N}(\boldsymbol{0}, \frac{1}{q}\mathcal{I}_d)$. In every case we did the approximation 10 times and show the mean.

During the numerical experiments we found that the more the input variables are related, the bigger should be the regularization parameter λ in the minimization problem of Algorithm 2. Notice that, depending on the type of dependence, we do not obtain the same variances for the ANOVA terms, but the non-zero terms are in any case a subset

of

$$U_{T1} = \{\varnothing, \{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{2,3\}\}, \tag{7.6}$$
$$U_{T2} = \{\varnothing, \{1\}, \{2\}, \{3\}, \{1,3\}, \{2,3\}\}, \tag{7.7}$$
$$U_{T3} = \{\varnothing, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{1,2\}\}, \tag{7.8}$$

for the three functions respectively. The variances of the terms $f_{\{3\}}$ and $f_{\{5\}}$ are relatively small for function f_{T3}. For that reason, we set $\epsilon = 0.01$ in this case.

The numerical results show that even for a small amount of samples in high dimensions our procedure is able to find the correct non-zero terms, which results in much smaller approximation error, compared to the plain SHRIMP algorithm [39] with fixed effective dimension q.

Conclusion and outlook

We propose a new method, ANOVA boosting, which exploits sparse structure in the ANOVA terms of a function in a learning problem, which often occurs in many domains of interest. This method is a possible extension of random Fourier feature algorithms which finds the ANOVA terms with variance above some threshold before the actual approximation. Our algorithms are able to handle independent as well as dependent input variables.

Maybe it would be also possible to incorporate the ANOVA boosting in every step of the iterative algorithm for sparse random Fourier features. Another possible future direction is the analysis of the impact of noise and the analysis of a good choice of the regularization parameter λ in the boosting step as well as in the random feature algorithm.

Acknowledgement

References

[1] S. Bochner. *Harmonic Analysis and the Theory of Probability*. University of California Press, Berkeley, 1955.

[2] E. Borgonovo, G. Li, J. Barr, E. Plischke, and H. Rabitz. Global sensitivity analysis with mixtures: A generalized functional ANOVA approach. *Risk Analysis*, 42(2):304–333, 2022.

[3] R. Caflisch, W. Morokoff, and A. Owen. Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension. *J. Comput. Finance*, 1(1):27–46, 1997.

[4] G. Chastaing, F. Gamboa, and C. Prieur. Generalized Hoeffding-Sobol decomposition for dependent variables - application to sensitivity analysis. *Electron. J. Stat.*, 6:2420 – 2448, 2012.

[5] R. DeVore, G. Petrova, and P. Wojtaszczyk. Approximation of functions of few variables in high dimensions. *Constr. Approx.*, 33(1):125–143, 2010.

[6] C. Gu. *Smoothing spline ANOVA models*, volume 297 of *Springer Series in Statistics*. Springer, New York, second edition, 2013.

[7] A. Hashemi, H. Schaeffer, R. Shi, U. Topcu, G. Tran, and R. Ward. Generalization bounds for sparse random feature expansions. *Appl. Comput. Harmon. Anal.*, 62:310–330, 2023.

[8] J. Hertrich, F. A. Ba, and G. Steidl. Sparse mixture models inspired by ANOVA. *Electron. Trans. Numer. Anal.*, 55:142–168, 2024.

[9] M. Holtz. *Sparse grid quadrature in high dimensions with applications in finance and insurance*, volume 77 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin, 2011.

[10] G. Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *J. Comput. Graph. Statist.*, 16(3):709–732, 2007.

[11] P. Jaworski, F. Durante, W. K. Härdle, and T. Rychlik. *Copula Theory and Its Applications*. Lecture Notes in Statistics. Springer, Berlin, Heidelberg, 2009.

[12] S. Kucherenko, S. Tarantola, and P. Annoni. Estimation of global sensitivity indices for models with dependent variables. *Comput. Phys. Commun.*, 183(4):937–946, 2012.

[13] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski. On decompositions of multivariate functions. *Math. Comp.*, 79(270):953–966, 2009.

[14] G. Li and H. Rabitz. General formulation of HDMR component functions with independent and correlated variables. *J. Math. Chem.*, 50:99–130, 2012.

[15] Z. Li, J.-F. Ton, D. Oglic, and D. Sejdinovic. Towards a unified analysis of random Fourier features. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3905–3914. PMLR, 09–15 Jun 2019.

[16] L. Lippert and D. Potts. Variable transformations in combination with wavelets and ANOVA for high-dimensional approximation. *ArXiv e-prints*, 2022.

[17] L. Lippert, D. Potts, and T. Ullrich. Fast hyperbolic wavelet regression meets ANOVA. *Numer. Math.*, 154:155–207, 2023.

[18] F. Liu, X. Huang, Y. Chen, and J. A. K. Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):7128–7148, 2022.

[19] R. Liu and A. B. Owen. Estimating mean dimensionality of analysis of variance decompositions. *J. Amer. Statist. Assoc.*, 101(474):712–721, 2006.

[20] K. Märtens and C. Yau. Neural decomposition: Functional ANOVA with variational autoencoders. In S. Chiappa and R. Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2917–2927. PMLR, 2020.

[21] S. Mei, T. Misiakiewicz, and A. Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. Special Issue on Harmonic Analysis and Machine Learning.

[22] C. Molnar. *Interpretable Machine Learning*. 2nd edition, 2022.

[23] R. B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer, New York, NY, 2 edition, 2006.

[24] A. B. Owen. *Practical Quasi-Monte Carlo Integration*. https://artowen.su.domains/mc/practicalqmc.pdf, 2023.

[25] G. Plonka, D. Potts, G. Steidl, and M. Tasche. *Numerical Fourier Analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2nd edition, 2023.

[26] D. Potts and M. Schmischke. Approximation of high-dimensional periodic functions with Fourier-based methods. *SIAM J. Numer. Anal.*, 59(5):2393–2429, 2021.

[27] H. Rabitz and O. F. Alis. General foundations of high dimensional model representations. *J. Math. Chem.*, 25:197–233, 1999.

[28] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[29] S. Rahman. A generalized ANOVA dimensional decomposition for dependent probability measures. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):670–697, 2014.

[30] C. Rieger and H. Wendland. On the approximability and curse of dimensionality of certain classes of high-dimensional functions. *SIAM J. Numer. Anal.*, 62(2):842–871, 2024.

[31] A. Rudi and L. Rosasco. Generalization properties of learning with random features. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3218–3228. Curran Associates Inc., 2017.

[32] E. Saha, H. Schaeffer, and G. Tran. HARFE: hard-ridge random feature expansion. *Sampl. Theory Signal Process. Data Anal.*, 21(27):2730–5724, 2023.

[33] A. Saltelli and I. Sobol'. Sensitivity analysis for nonlinear mathematical models: Numerical experience. *Mat. Model.*, 7, 01 1995.

[34] M. Schmischke. *Interpretable Approximation of High-Dimensional Data based on the ANOVA Decomposition*. PhD thesis, Chemnitz University of Technology, 2022.

[35] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. stat. Univ. Paris*, 8:229–231, 1959.

[36] J. M. Steele. *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge University Press, 2004.

[37] C. J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Appl. Stat.*, 22(1):118–171, 1994.

[38] C. F. J. Wu and M. S. Hamada. *Experiments - Planning, Analysis, and Optimization*. John Wiley & Sons, New York, 2011.

[39] Y. Xie, R. Shi, H. Schaeffer, and R. Ward. SHRIMP: sparser random feature models via iterative magnitude pruning. In B. Dong, Q. Li, L. Wang, and Z.-Q. J. Xu, editors, *Proceedings of Mathematical and Scientific Machine Learning*, volume 190 of *Proceedings of Machine Learning Research*, pages 303–318. PMLR, 2022.

A Appendix

Proofs of Section 2

First, we need an auxiliary result.

Lemma A.1. *Let $g \in L_2(\mathbb{R}^d)$ and K be a symmetric kernel function. Then*

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} g(\boldsymbol{\omega})\overline{g(\boldsymbol{v})} K(\boldsymbol{\omega}, \boldsymbol{v}) \, \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{v} \le \int_{\mathbb{R}^d} |g(\boldsymbol{\omega})|^2 k(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega},$$

where $k(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} |K(\boldsymbol{\omega}, \boldsymbol{v})| \, \mathrm{d}\boldsymbol{v}$.

Proof. According to [36, Chapter 1], the generalization of the Cauchy-Schwarz inequality for double integrals yields

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} g(\boldsymbol{\omega})\overline{g(\boldsymbol{v})} K(\boldsymbol{\omega}, \boldsymbol{v}) \, \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{v} \le \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left| g(\boldsymbol{\omega})\overline{g(\boldsymbol{v})} K(\boldsymbol{\omega}, \boldsymbol{v}) \right| \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{v}$$

$$\le \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |g(\boldsymbol{\omega})|^2 |K(\boldsymbol{\omega}, \boldsymbol{v})| \, \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{v} \right)^{1/2} \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\overline{g(\boldsymbol{v})}|^2 |K(\boldsymbol{\omega}, \boldsymbol{v})| \, \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{v} \right)^{1/2}$$

$$= \left(\int_{\mathbb{R}^d} |g(\boldsymbol{\omega})|^2 \int_{\mathbb{R}^d} |K(\boldsymbol{\omega}, \boldsymbol{v})| \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{\omega} \right)^{1/2} \left(\int_{\mathbb{R}^d} |g(\boldsymbol{v})|^2 \int_{\mathbb{R}^d} |K(\boldsymbol{\omega}, \boldsymbol{v})| \, \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{v} \right)^{1/2}$$

$$= \int_{\mathbb{R}^d} |g(\boldsymbol{\omega})|^2 k(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega}.$$

This finishes the proof. ∎

Proof of Lemma 2.5

Proof. Lemma 2.2 describes the ANOVA terms of the function f. In order to calculate the variance of the ANOVA terms we have

$$\sigma^2(f_{\boldsymbol{u}}) = \int_{\mathbb{R}^{|\boldsymbol{u}|}} \left| \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\omega}) E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \boldsymbol{u}) \, \mathrm{d}\boldsymbol{\omega} \right|^2 \mu_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) \, \mathrm{d}\boldsymbol{x}$$

$$= \frac{1}{(2\pi)^{2d}} \int_{\mathbb{R}^{|\boldsymbol{u}|}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\omega})\overline{\hat{f}(\boldsymbol{v})} E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \boldsymbol{u})\overline{E(\boldsymbol{x}, \boldsymbol{v}, \mu, \boldsymbol{u})} \, \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{v}\mu_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) \, \mathrm{d}\boldsymbol{x}$$

$$= \frac{1}{(2\pi)^{2d}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^{|\boldsymbol{u}|}} \hat{f}(\boldsymbol{\omega})\hat{f}(-\boldsymbol{v}) E(\boldsymbol{x}, \boldsymbol{\omega}, \mu, \boldsymbol{u}) E(\boldsymbol{x}, -\boldsymbol{v}, \mu, \boldsymbol{u}) \, \mu_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) \, \mathrm{d}\boldsymbol{x}_{\boldsymbol{u}} \, \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{v}$$

$$= \frac{1}{(2\pi)^{2d}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\omega})\hat{f}(-\boldsymbol{v}) \prod_{i \in \boldsymbol{u}} (\hat{\mu}_i(-\omega_i + v_i) - \hat{\mu}_i(-\omega_i)\hat{\mu}_i(v_i)) \prod_{i \in \boldsymbol{u}^c} \hat{\mu}_i(-\omega_i)\hat{\mu}_i(v_i) \, \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\boldsymbol{v}$$

To apply Lemma A.1 we choose $K(\boldsymbol{\omega}, \boldsymbol{v}) = \prod_{i \in \boldsymbol{u}} (\hat{\mu}_i(-\omega_i + v_i) - \hat{\mu}_i(-\omega_i)\hat{\mu}_i(v_i)) \prod_{i \in \boldsymbol{u}^c} \hat{\mu}_i(-\omega_i)\hat{\mu}_i(v_i)$ with

$$k(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} \left| \prod_{i \in \boldsymbol{u}} (\hat{\mu}_i(-\omega_i + v_i) - \hat{\mu}_i(-\omega_i)\hat{\mu}_i(v_i)) \prod_{i \in \boldsymbol{u}^c} \hat{\mu}_i(-\omega_i)\hat{\mu}_i(v_i) \right| \mathrm{d}\boldsymbol{v}$$

$$= \prod_{i \in [d]} \|\hat{\mu}_i\|_{L_1(\mathbb{R})} \prod_{i \in \boldsymbol{u}} |1 - \hat{\mu}_i(-\omega_i)| \prod_{i \in \boldsymbol{u}^c} |\hat{\mu}_i(-\omega_i)|$$

$$= \|\hat{\mu}\|_{L_1(\mathbb{R}^d)} |E(\mathbf{0}, \boldsymbol{\omega}, \mu, \boldsymbol{u})|. \qquad \blacksquare$$

Proof of Theorem 2.7

Proof. For this proof we introduce the notation

$$A(\boldsymbol{\omega}, d, q) := \left(\prod_{i \in [d]} (1 + |\omega_i|^2)^{-s} \sum_{|\boldsymbol{u}| > q} |E(\boldsymbol{0}, \boldsymbol{\omega}, \mu, \boldsymbol{u})| \right). \tag{A.1}$$

First, note that since every measure μ_i is symmetric, $\|\hat{\mu}_i\|_{L_\infty(\mathbb{R})} \le \|\mu_i\|_{L_1(\mathbb{R})} = 1$ and $-1 \le \hat{\mu}_i(-\omega_i) \le 1$. We start with applying Lemma 2.5.

$$\|f - \mathcal{T}_q f\|^2_{L_2(\mathbb{R}^d, \mu)} = \sum_{|\boldsymbol{u}| > q} \sigma^2(f_{\boldsymbol{u}}) \le \frac{\|\hat{\mu}\|_{L_1(\mathbb{R}^d)}}{(2\pi)^{2d}} \sum_{|\boldsymbol{u}| \ge q} \int_{\mathbb{R}^d} |\hat{f}(\boldsymbol{\omega})|^2 \prod_{i \in \boldsymbol{u}} |1 - \hat{\mu}_i(-\omega_i)| \prod_{i \in \boldsymbol{u}^c} |\hat{\mu}_i(-\omega_i)| \, \mathrm{d}\boldsymbol{\omega}$$

$$= \frac{\|\hat{\mu}\|_{L_1(\mathbb{R}^d)}}{(2\pi)^{2d}} \int_{\mathbb{R}^d} |\hat{f}(\boldsymbol{\omega})|^2 \prod_{i \in [d]} \frac{(1 + |\omega_i|^2)^s}{(1 + |\omega_i|^2)^s} \sum_{|\boldsymbol{u}| > q} \prod_{i \in \boldsymbol{u}} |1 - \hat{\mu}_i(-\omega_i)| \prod_{i \in \boldsymbol{u}^c} |\hat{\mu}_i(-\omega_i)| \, \mathrm{d}\boldsymbol{\omega}$$

$$= \frac{\|\hat{\mu}\|_{L_1(\mathbb{R}^d)}}{(2\pi)^{2d}} \max_{\boldsymbol{\omega} \in \mathbb{R}^d} \left(\prod_{i \in [d]} (1 + |\omega_i|^2)^{-s} \sum_{|\boldsymbol{u}| > q} \prod_{i \in \boldsymbol{u}} |1 - \hat{\mu}_i(-\omega_i)| \prod_{i \in \boldsymbol{u}^c} |\hat{\mu}_i(-\omega_i)| \right) \int_{\mathbb{R}^d} |\hat{f}(\boldsymbol{\omega})|^2 \prod_{i \in [d]} (1 + |\omega_i|^2)^s \, \mathrm{d}\boldsymbol{\omega}$$

$$= \frac{\|\hat{\mu}\|_{L_1(\mathbb{R}^d)}}{(2\pi)^{2d}} \|f\|^2_{H^s_{\mathrm{mix}}(\mathbb{R}^d)} \max_{\boldsymbol{\omega} \in \mathbb{R}^d} \left(\prod_{i \in [d]} (1 + |\omega_i|^2)^{-s} \sum_{|\boldsymbol{u}| > q} \prod_{i \in \boldsymbol{u}} |1 - \hat{\mu}_i(-\omega_i)| \prod_{i \in \boldsymbol{u}^c} |\hat{\mu}_i(-\omega_i)| \right)$$

$$= \frac{\|\hat{\mu}\|_{L_1(\mathbb{R}^d)}}{(2\pi)^{2d}} \|f\|^2_{H^s_{\mathrm{mix}}(\mathbb{R}^d)} \max_{\boldsymbol{\omega} \in \mathbb{R}^d} A(\boldsymbol{\omega}, d, q). \tag{A.2}$$

Let us have a closer look at the involved term $A(\boldsymbol{\omega}, d, q)$. Let $\boldsymbol{v}$ be the support of the $\boldsymbol{\omega}$, which attains the maximum. Since $\hat{\mu}_i(-\omega_i) = 0$ for every $i \in \boldsymbol{v}^c$, which means that $|\boldsymbol{v}^c| > q$ and $|\boldsymbol{v}| < d - q$, we have

$$\max_{\boldsymbol{\omega} \in \mathbb{R}^d} A(\boldsymbol{\omega}, d, q) = \prod_{i \in [d]} (1 + |\omega_i|^2)^{-s} \sum_{|\boldsymbol{u}| > q, \boldsymbol{v}^c \subseteq \boldsymbol{u}} \prod_{i \in \boldsymbol{u}} |1 - \hat{\mu}_i(-\omega_i)| \prod_{i \in \boldsymbol{u}^c} |\hat{\mu}_i(-\omega_i)|$$

$$\le \sum_{\boldsymbol{u}' \supseteq \boldsymbol{v}^c} c_{\mu,s}^{|\boldsymbol{v}^c|} \left(\prod_{i \in \boldsymbol{u}' \setminus \boldsymbol{v}^c} \frac{|1 - \hat{\mu}_i(-\omega_i)|}{(1 + |\omega_i|^2)^s} \prod_{i \in \boldsymbol{u}'^c} \frac{|\hat{\mu}_i(-\omega_i)|}{(1 + |\omega_i|^2)^s} \right)$$

$$= c_{\mu,s}^{|\boldsymbol{v}^c|} \sum_{\boldsymbol{u}' \subseteq \boldsymbol{v}} \left(\prod_{i \in \boldsymbol{u}'} \frac{|1 - \hat{\mu}_i(-\omega_i)|}{(1 + |\omega_i|^2)^s} \prod_{i \in \boldsymbol{u}'^c} \frac{|\hat{\mu}_i(-\omega_i)|}{(1 + |\omega_i|^2)^s} \right)$$

$$= c_{\mu,s}^{|\boldsymbol{v}^c|} \prod_{i \in \boldsymbol{v}} \frac{|1 - \hat{\mu}_i(-\omega_i)| + |\hat{\mu}_i(-\omega_i)|}{(1 + |\omega_i|^2)^s}$$

$$\le c_{\mu,s}^{q+1}.$$

The last inequality follows by either demanding a symmetric measure μ with positive Fourier transform or the condition (2.10). The equality (2.11) follows by the fact that the maximum of $g(\omega_i) := (1 + |\omega_i|^2)^{-s} (1 - \hat{\mu}_i(-\omega_i))$ is attained where $g'(\omega_i) = 0$, i.e.

$$0 \overset{!}{=} \frac{-2\omega_i s}{(1 + |\omega_i|^2)^{s+1}} (1 - \hat{\mu}_i(\omega_i)) - \frac{\mu_i'(\omega_i)}{(1 + |\omega_i|^2)^s},$$

$$\hat{\mu}_i(\omega_i) = 1 + \hat{\mu}_i'(\omega_i) \frac{1 + \omega_i^2}{2\omega_i s}.$$

Inserting this into $g(\omega_i)$ yields

$$c_{\mu,s} = \sup_{\omega_i \in \mathbb{R}} (1 + |\omega_i|^2)^{-s} \left(-\hat{\mu}_i'(\omega_i) \frac{1 + \omega_i^2}{2\omega_i s} \right)$$

$$= \sup_{\omega_i \in \mathbb{R}} \frac{1}{2\omega_i s (1 + |\omega_i|^2)^{s-1}}.$$

∎

29

A.1 McDiarmids inequality

Theorem A.2 (Mc Diarmids inequality). *Let a function $v\colon \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N \to \mathbb{R}$ satisfy the bounded differences property, i.e. for all $k \in [N]$, and all $x_1 \in \mathcal{X}, \ldots, x_N \in \mathcal{X}_N$,*

$$\sup_{x_k' \in \mathcal{X}_k} |v(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_N) - v(x_1, \ldots, x_{k-1}, x_k', x_{k+1}, \ldots, x_N)| \leq \Delta_k.$$

Consider independent random variables $X_1, X_2, \ldots, X_N$ where $X_k \in \mathcal{X}_k$ for all k. Then, for any $\varepsilon > 0$

$$\mathbb{P}\left(v(X_1, X_2, \ldots, X_N) - \mathbb{E}[v(X_1, X_2, \ldots, X_N)] > \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{k=1}^{N} \Delta_k^2}\right).$$