Generalization Bounds for Sparse Random Feature Expansions

Abolfazl Hashemi[1], Hayden Schaeffer[2], Robert Shi[1],
Ufuk Topcu[1], Giang Tran[3], and Rachel Ward[*1]

[1]The University of Texas at Austin
[2]Carnegie Mellon University
[3]University of Waterloo

Abstract

Random feature methods have been successful in various machine learning tasks, are easy to compute, and come with theoretical accuracy bounds. They serve as an alternative approach to standard neural networks since they can represent similar function spaces without a costly training phase. However, for accuracy, random feature methods require more measurements than trainable parameters, limiting their use for data-scarce applications or problems in scientific machine learning. This paper introduces the sparse random feature expansion to obtain parsimonious random feature models. Specifically, we leverage ideas from compressive sensing to generate random feature expansions with theoretical guarantees even in the data-scarce setting. In particular, we provide uniform bounds on the approximation error and generalization bounds for functions in a certain class (that is dense in a reproducing kernel Hilbert space) depending on the number of samples and the distribution of features. The error bounds improve with additional structural conditions, such as coordinate sparsity, compact clusters of the spectrum, or rapid spectral decay. In particular, by introducing sparse features, i.e. features with random sparse weights, we provide improved bounds for low order functions. We show that the sparse random feature expansions outperforms shallow networks in several scientific machine learning tasks.

1 Introduction

The *sparsity-of-effects* or *Pareto principle* states that most real-world systems are dominated by a small number of low-complexity interactions. This idea is at the heart of compressive sensing and sparse optimization, which computes a sparse representation for a given dataset using a large set of features. The feature spaces are often constructed using a random matrix, e.g., each element is independent and identically distributed from the normal distribution, or constructed using a bounded orthonormal system, e.g., Fourier or orthonormal polynomials. While completely random matrices are useful for compression, their lack of structure can limit applications to problems that require physical or meaningful constraints. On the other hand, while bounded orthonormal systems provide meaningful structure to the feature space, they often require knowledge of the sampling measure and the target functions themselves, e.g., that the target function is well-represented by polynomials.

In the high-dimensional setting, neural networks can achieve high test accuracy when there are reasonable models for the local interactions between variables. For example, a convolutional

[*]Authors are listed in alphabetical order. This work was supported in part by AFOSR MURI FA9550-19-1-0005, AFOSR MURI FA9550-21-1-0084, NSERC Discovery Grant RGPIN 2018-06135, NSF DMS-1752116, and NSF DMS-1952735. The code is available on our github page `https://github.com/GiangTTran/SparseRandomFeatures`

neural network imposes local spatial dependencies between pixels or nodes. In addition, neural networks can construct data-driven feature spaces that far exceed the limitations of pre-specified bases such as polynomials. However, standard neural networks often rely on back-propagation or greedy algorithms to train the weights, which is a computationally intensive procedure. Furthermore, the trained models do not provide interpretable results, i.e., they remain black-boxes. Randomized networks are a class of neural networks that randomize and fix the weights within the architecture [6,31,34,37,39]. When only the final layer is trained, the training problem becomes linear and can have a much lower cost than the non-convex optimization-based approaches. This method has motivated new algorithms and theory, for example, see [10,30,37–39,46,47,49]. Recently, generalization bounds for over-parameterized random features ridge regression were provided in [33], when the Tikhonov regularization parameter tends to zero. The analysis is asymptotic and is restricted to the ReLU activation function, with data and features drawn on the sphere.

In this work, we introduce a new framework for approximating high-dimensional functions in the case where measurements are *expensive* and *scarce*. We propose the *sparse random feature expansion* (SRFE), which enhances the compressive sensing approach by allowing for more flexible functional relationships between inputs, as well as a more complex feature space. The choice of basis is inspired by the random Fourier feature (RFF) method [37,39], which uses a basis comprised of simple (often trigonometric) functions with randomized parameters. In the RFF method, the model is learned using ridge regression, which leads to dense (or full) representations. By using sparsity, our approach could be viewed as a way to leverage structure in the data-scarce setting while retaining the accuracy and representation capabilities of the randomized feature methods. In addition, the use of sparsity allows for reasonable error bounds even in the very overcomplete setting, which is proving to be a powerful modern tool related to over-parameterized neural networks [3,18,25,29].

In terms of the approximation error, the randomized methods can achieve similar results to those associated with shallow networks. In [4,26], it was shown that if the Fourier transform of the target function f, denoted by $\hat{f}$, has finite integral $\int_{\mathbb{R}^d} |\boldsymbol{\omega}||\hat{f}(\boldsymbol{\omega})|d\boldsymbol{\omega}$ then there is a two-layer neural network with N terms that can approximate f up to an L^2 error of $\mathcal{O}(N^{-\frac{1}{2}})$. These results (and their generalizations) often require specific (greedy) algorithms to achieve. In addition, neural networks often only achieve good performance in the data-rich and over-parameterized regimes. On the other hand, the RFF method achieves uniform errors on the order of $\mathcal{O}(N^{-\frac{1}{2}})$ for functions in a certain class (associated with the choice of the basis functions) without the need for a particular algorithm or construction [37].

One of the most popular techniques in the area of uncertainty quantification is the Polynomial Chaos Expansion (PCE). PCE models are built up from univariate orthonormal polynomial regression; in particular, each basis term is the product of univariate orthonormal polynomials and is characterized by the multi-index of polynomial degrees in each direction. The standard PCE approach solves for the coefficients of the polynomials using the ordinary least squares method. The sparse PCE has recently gained traction, where the coefficient vector is determined through sparse regression. Many sparse regression methods used in PCE were originally developed for compressive sensing [8,16,17,40]. The success of sparse PCE is due in part to the method's ability to incorporate higher degree terms without overfitting. However, the polynomial basis must be orthogonalized with respect to the sampling measure. Moreover, good performance is limited to functions which are well-represented by moderate degree polynomials. This serves as another motivation for the use of randomized features, which may increase the richness of the approximation.

1.1 Contribution

We propose a sparse feature model (the SRFE) which improves on compressive sensing and PCE approaches by utilizing random features from the RFF model. Also, the SRFE outperforms a standard shallow neural network in the limited data regime. We incorporate sparsity in the proposed model in two ways. The first is in our approximation of the target function by using a small number of terms from a large feature space to represent the dominate behavior (this is the sparse expansion component). The second level of sparsity can be considered as side information on the variables and is incorporated by sampling random low-order interactions between variables (the sparse features). Building upon these ideas, as part of our theoretical contributions we derive sample and feature complexity bounds such that the error between the SRFE and the target function is controlled by the richness of the random features, the compressibility of the representation, and the noise on the samples (formalized in Section 3). This also shows the tractability of sparse expansions in the context of randomized feature models.

The SRFE offers additional freedom through redundancy of the basis and does not restrict the model class to low-order interactions in the form of polynomials. While our analysis is mainly for trigonometric functions, extensions and applications with ReLU and other standard activation functions are discussed. In addition, our method and analysis could be extended to include different sampling strategies such as those used in the recovery of dynamical systems [43, 44].

In order to provide generalization bounds, we first characterize the approximation power of the best fit approximator; then, we bound the error between the best fit and the sparse random feature expansion. The best fit results are extensions of [37, 39], but we provide the proof for completeness. The generalization bounds and the sparse approximation results are both novel. While we utilize standard coherence-based results for sparse recovery, we prove new bounds for the coherence and the sample complexity based on the randomized features (for both dense and sparse features). In [50], a sparse random feature algorithm is proposed which iteratively adds random features by using a combination of LASSO and hard thresholding. In our work, we provide sample complexity, sparsity guarantees, and generalization bounds which did not appear in previous works. In addition, we introduce sparse feature weights within our model, which can help with the curse-of-dimensionality for recovering low order functions. It is worth noting that our method extends to any algorithm that uses coherence-based sparsity guarantees, for example, greedy methods such as orthogonal matching pursuit.

2 Approximation via Sparse Random Feature Expansion (SRFE)

Notation. Throughout this paper, we use bold letters and bold capital letters to denote column vectors and matrices, respectively (e.g., $\mathbf{x}$ and $\mathbf{A}$). Let $[N] = \{1, \dots, N\}$ for any positive integer N and $\|\mathbf{c}\|$ denote the Euclidean norm of a vector $\mathbf{c}$. Throughout the paper, f denotes functions of d variables while g denotes functions of $q \ll d$ variables. Furthermore, $\mathbb{B}^d(M)$ denotes the Euclidean ball in $\mathbb{R}^d$ of radius M. A vector $\mathbf{c} \in \mathbb{C}^N$ is said to be s-sparse if the number of nonzero components of $\mathbf{c}$ is at most s. For a vector $\mathbf{c} \in \mathbb{C}^N$, let $\kappa_{s,p}(\mathbf{c})$ denote the error of best s-term approximation to $\mathbf{c}$ in the ℓ_p sense, $\kappa_{s,p}(\mathbf{c}) := \min\{\|\mathbf{c} - \mathbf{z}\|_{\ell_p} : \mathbf{z} \text{ is } s\text{-sparse}\}$ [21]. Note in particular that $\kappa_{s,p}(\mathbf{c}) = 0$ if $\mathbf{c}$ is s-sparse, and $\kappa_{s,p}(\mathbf{c}) \leq \|\mathbf{c}\|_{\ell_p}$ always.

We are interested in identifying an unknown function $f : \mathbb{R}^d \to \mathbb{C}$, belonging to a certain class (defined in Section 3), from a set of samples. We assume that the m sampling points $\mathbf{x}_k$'s are drawn with a probability measure $\mu(\mathbf{x})$ with the corresponding output values

$$y_k = f(\mathbf{x}_k) + e_k, \quad |e_k| \leq E, \quad \forall k \in [m], \tag{1}$$

where e_k is the noise.

A fundamental approach in approximation theory relies on the assumption that f has an approximate linear representation with respect to a suitable collection of N functions $\phi_j(\mathbf{x})$,

Algorithm 1 Sparse Random Feature Expansion (SRFE)

1: **Input:** parametric basis function $\phi(\,;\boldsymbol{\omega}) = \phi(\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$, stability parameter η.
2: Draw m data points $\mathbf{x}_k \sim \mathcal{D}_x$ and observe outputs $y_k = f(\mathbf{x}_k) + e_k$ with $|e_k| \leq E$.
3: Draw N random weights $\boldsymbol{\omega}_j \sim \mathcal{D}_\omega$.
4: Construct the random feature matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ such that $a_{kj} = \phi(\mathbf{x}_k; \boldsymbol{\omega}_j)$.
5: Solve

$$\mathbf{c}^\sharp = \arg\min_{\mathbf{c}} \quad \|\mathbf{c}\|_1 \qquad \text{s.t.} \qquad \|\mathbf{A}\mathbf{c} - \mathbf{y}\| \leq \eta\sqrt{m}.$$

6: **Output:** Form the approximation

$$f^\sharp(\mathbf{x}) = \sum_{j=1}^{N} c_j^\sharp \phi(\mathbf{x}; \boldsymbol{\omega}_j).$$

$j \in [N]$:

$$f(\mathbf{x}) \approx \sum_{j=1}^{N} c_j \phi_j(\mathbf{x}). \tag{2}$$

Important examples of such families of functions include real and complex trigonometric polynomials as well as Legendre polynomials [1,2,11,40,41].

Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be the random feature matrix with entries $a_{k,j} = \phi_j(\mathbf{x}_k)$, then approximating f in Equation (2) is equivalent to

$$\text{find} \quad \mathbf{c} \in \mathbb{C}^N \quad \text{such that} \quad \mathbf{y} \approx \mathbf{A}\mathbf{c}, \tag{3}$$

where $\mathbf{c} = [c_1, \ldots, c_N]^T$ and $\mathbf{y} = [y_1, \ldots, y_m]^T$. In many applications, it is often the case that f is well-approximated by a small subset of the N functions, which implies that $\mathbf{c}$ is sparse. By exploiting the sparsity, the number of samples m required to obtain an accurate approximation of f may be significantly reduced. One effective approach to learn a sparse vector $\mathbf{c}$ is to solve the basis pursuit (BP) problem:

$$\mathbf{c}^\sharp = \arg\min_{\mathbf{c}} \quad \|\mathbf{c}\|_1 \qquad \text{s.t.} \qquad \|\mathbf{A}\mathbf{c} - \mathbf{y}\| \leq \eta\sqrt{m}, \tag{4}$$

where η is a parameter typically related to the measurement noise. The conditions for stable recovery of any sparse vector $\mathbf{c}^\star$ satisfying $\mathbf{y} \approx \mathbf{A}\mathbf{c}^\star$ is extensively studied in compressed sensing and statistics [7,9,21].

In order to construct a sufficiently rich family of functions, we use a randomized approach. Specifically, consider a collection of functions $\phi(\mathbf{x}; \boldsymbol{\omega}) = \phi(\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$ parameterized by a weight vector $\boldsymbol{\omega}$ drawn randomly from a probability distribution $\rho(\boldsymbol{\omega})$. Some popular choices for ϕ are

1. *Random Fourier features:* $\phi(\mathbf{x}; \boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$.
2. *Random trigonometric features:*
   $\phi(\mathbf{x}; \boldsymbol{\omega}) = \cos(\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$ and $\phi(\mathbf{x}; \boldsymbol{\omega}) = \sin(\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$.
3. *Random ReLU features:* $\phi(\mathbf{x}; \boldsymbol{\omega}) = \max(\langle \mathbf{x}, \boldsymbol{\omega} \rangle, 0)$.

Based on [37,39], we call such $\phi(\cdot\,; \boldsymbol{\omega})$ the random features. Altogether, we propose the *Sparse Random Feature Expansion (SRFE)* to approximate f, which is summarized in Algorithm 1.

3 Low Order Functions

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 $q \ll d$ [15,27]. Low order functions also appear in other applications as a way to reduce modeling complexity. For example, in dimension reduction and surrogate modeling, sensitivity analysis is employed to determine the most influential input variables and thus to reduce the approximation onto a subset of the input space [42]. The notion of low order functions are also connected to low-dimensional structures [35,36] and active subspaces [12,13,20]. Low order additive functions and sparsely connected networks are also well-motivated in computational neuroscience for simple brain architectures [22].

With this side information, we can further reduce the number of samples needed (see Theorem 2). We modify Algorithm 1 to incorporate the potential coordinate sparsity into the weights $\boldsymbol{\omega}$. Since we do not know the set of active variables, we draw a number of sparse random feature weights on every subset $\mathcal{S} \subset [d]$ of size $|\mathcal{S}| = q$. That is, for each such $\mathcal{S}$, we draw the on-support feature components randomly from the given distribution, and we set the remaining components to be zero. In particular, we have the following definition for our random features.

Definition 1 (q**-Sparse Feature Weights**). *Let $d, q, n \in \mathbb{N}$ with $q \leq d$ and a multivariate probability density $\zeta : \mathbb{R}^q \to \mathbb{R}$. A collection of $N = n\binom{d}{q}$ weight vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N$ is said to be a* complete set of q-sparse feature weights (drawn from density ζ) *if they are generated as follows: For each subset $\mathcal{S} \subset [d]$ of size $|\mathcal{S}| = q$, draw n random vectors $z_1, \ldots, z_n \in \mathbb{R}^q$ from ζ, independent of each other and of all previous draws. Then, use $z_1, \ldots, z_n$ to form q-sparse feature weights $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n \in \mathbb{R}^d$ by setting $\operatorname{supp}(\boldsymbol{\omega}_k) = \mathcal{S}$ and $\boldsymbol{\omega}_k\big|_{\mathcal{S}} = z_k$.*

This leads to Sparse Random Feature Expansion with Sparse Features (*SRFE-S*) by modifying Step (3) of Algorithm 1 to "Draw a complete set of N q-sparse feature weights $\boldsymbol{\omega}_j \in \mathbb{R}^d$ sampled with density ζ". We summarize the algorithm below.

Algorithm 2 Sparse Random Feature Expansion with Sparse Feature Weights (SRFE-S)

1: **Input:** parametric basis function $\phi(\mathbf{x}; \boldsymbol{\omega}) = \phi(\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$, feature sparsity level q, probability density $\zeta : \mathbb{R}^q \to \mathbb{R}$, stability parameter η.
2: Draw m data points $\mathbf{x}_k \sim \mathcal{D}_x$ and observe outputs $y_k = f(\mathbf{x}_k) + e_k$ with $|e_k| \leq E$.
3: Draw a complete set of N q-sparse feature weights $\boldsymbol{\omega}_j \in \mathbb{R}^d$ sampled from density $\zeta : \mathbb{R}^q \to \mathbb{R}$ as defined in Definition 1.
4: Construct a random feature matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ such that $a_{kj} = \phi(\mathbf{x}_k; \boldsymbol{\omega}_j)$.
5: Solve
$$\mathbf{c}^\sharp = \arg\min_{\mathbf{c}} \quad \|\mathbf{c}\|_1 \qquad \text{s.t.} \qquad \|\mathbf{A}\mathbf{c} - \mathbf{y}\| \leq \eta\sqrt{m}.$$
6: **Output:** Form the approximation
$$f^\sharp(\mathbf{x}) = \sum_{j=1}^{N} c_j^\sharp \phi(\mathbf{x}; \boldsymbol{\omega}_j).$$

Remark 1. *Drawing a complete set of q-sparse feature weights can be slow and cumbersome. In the case where $\zeta(x_1, \ldots, x_q) = \prod_{j=1}^{q} \zeta(x_j)$ is a tensor product of univariate densities, a significantly more practical method for drawing sparse features is as follows: we randomly generate a size q subset of $[d]$ and then define the on-support values using ζ. Alternatively, one can draw sparse feature weights by the procedure: for each coordinate $j \in [d]$, with probability q/d set $w_j = 0$, and with probability $1 - q/d$ draw $\boldsymbol{\omega}_j \sim \zeta$. We further note that any side-information on the feasibility of the low-order support subsets can be incorporated in the procedure outlined in Algorithm 2 to further reduce the required number of sparse features.*

4 Theoretical Analysis

In this section, we provide theoretical performance guarantees on the function approximation given in Algorithm 1. In particular, we derive an explicit bound on the required number of data samples for a stable approximation within a target region. Given the connections to Fourier analysis and its desired characteristics, we mainly focus on the case where $\phi(\mathbf{x};\boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$. Nonetheless, we discuss extensions in Section 5.

Before stating the main results, we recall some useful definitions. The first definition is a complex-valued extension of the class introduced in [37].

Definition 2 (**Bounded ρ-norm functions**). *Fix a probability density function $\rho : \mathbb{R}^d \to \mathbb{R}$ and a function $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$. A function $f : \mathbb{R}^d \to \mathbb{C}$ has finite ρ-norm with respect to $\phi(\mathbf{x};\boldsymbol{\omega})$ if it belongs to the class*

$$\mathcal{F}(\phi,\rho) \;:=\; \left\{ f(\mathbf{x}) \;=\; \int_{\boldsymbol{\omega}\in\mathbb{R}^d} \alpha(\boldsymbol{\omega})\phi(\mathbf{x};\boldsymbol{\omega}) \;\; d\boldsymbol{\omega} \quad : \quad \|f\|_\rho \;:=\; \sup_{\boldsymbol{\omega}} \left|\frac{\alpha(\boldsymbol{\omega})}{\rho(\boldsymbol{\omega})}\right| \;<\; \infty \right\}. \tag{5}$$

Note that in the above definition, if $\phi(\mathbf{x};\boldsymbol{\omega}) = \phi(\langle \mathbf{x}, \boldsymbol{\omega}\rangle) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$, $\alpha : \mathbb{R}^d \to \mathbb{C}$ is the inverse Fourier transform of f. Next, we formalize the notion of low order functions by extending the definition from [27].

Definition 3 (**Order-q functions**). *Fix $d, q, K \in \mathbb{N}$ and $q \le d$. A function $f : \mathbb{R}^d \to \mathbb{C}$ is an order-q function of at most K terms if there exist K functions $g_1, \ldots, g_K : \mathbb{R}^q \to \mathbb{C}$ such that*

$$f(x_1,\ldots,x_d) = \frac{1}{K}\sum_{j=1}^{K} g_j(x_{j_1},\ldots,x_{j_q}) = \frac{1}{K}\sum_{j=1}^{K} g_j(\mathbf{x}|_{\mathcal{S}_j}), \tag{6}$$

where $\mathcal{S}_j = \{j_1,\ldots,j_q\}$ is a subset of the index set $[d]$ and $\mathbf{x}|_{\mathcal{S}_j}$ is the restriction of $\mathbf{x}$ onto $\mathcal{S}_j$. For an order-q function f of at most K terms where every term in the summation belongs to $\mathcal{F}(\phi,\rho)$, we define the complexity measure

$$|||f||| := \binom{d}{q}^{\frac{1}{2}} \left(\frac{1}{K}\sum_{j=1}^{K} \|g_j\|_\rho \right). \tag{7}$$

Note that in general, such a decomposition is not unique and the value $|||f|||$ depends on the chosen decomposition of f. Furthermore, we are interested in the smallest q to refer to the order of a function; trivially, any order-q function $f : \mathbb{R}^d \to \mathbb{C}$ is also order-d.

4.1 Uniform Error

Our first main result establishes the required number of features N as well as the number of data samples m for stable recovery of a bounded ρ-norm function. For this class, we consider $q = d$.

Theorem 1 (**Function Approximation: Bounded ρ-norm Functions**). *Let $f \in \mathcal{F}(\phi,\rho)$, where $\phi(\mathbf{x};\boldsymbol{\omega}) = \phi(\langle \mathbf{x}, \boldsymbol{\omega}\rangle) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$ and $\rho(\boldsymbol{\omega})$ is the density corresponding to a spherical Gaussian with variance σ^2, $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$. For a fixed γ, consider a set of data samples $\mathbf{x}_1,\ldots,\mathbf{x}_m \sim \mathcal{N}(\mathbf{0}, \gamma^2\mathbf{I}_d)$ and frequencies $\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_N \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$. Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ denote the associated random feature matrix where $a_{k,j} = \phi(\mathbf{x}_k;\boldsymbol{\omega}_j)$ and $f^\sharp$ be defined from Algorithm 1 and Equation (4) with $\eta = \sqrt{2(\epsilon^2\|f\|_\rho^2 + E^2)}$.*

For a given s, suppose the feature parameters σ and N, the confidence δ, the accuracy ϵ, and the target radius R are chosen so that the following conditions hold:

1. γ-σ uncertainty principle

$$\gamma^2\sigma^2 \geq \frac{1}{2}\left(\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{2}{d}} - 1\right), \tag{8}$$

2. Number of features

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4R\sigma\sqrt{d} + \sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2, \tag{9}$$

3. Number of measurements

$$m \geq 4(2\gamma^2\sigma^2+1)^d \log\frac{N^2}{\delta}, \tag{10}$$

4. Target radius

$$R \geq \gamma\sqrt{d + \sqrt{12d\log\left(\frac{m}{\delta}\right)}}. \tag{11}$$

Then, with probability at least $1-4\delta$ the following error bound holds

$$\sup_{\|\mathbf{x}\|_2\leq R} |f(\mathbf{x}) - f^\sharp(\mathbf{x})| \leq \epsilon\|f\|_\rho + C'\kappa_{s,1}(\mathbf{c}^\star) + C\eta\sqrt{s}, \tag{12}$$

where $C, C' > 0$ are constants, and $\mathbf{c}^\star$ is the vector

$$\mathbf{c}^\star = \left[\frac{\alpha(\boldsymbol{\omega}_1)}{N\,\rho(\boldsymbol{\omega}_1)}, \ldots, \frac{\alpha(\boldsymbol{\omega}_N)}{N\,\rho(\boldsymbol{\omega}_N)}\right]^T.$$

Note that $\kappa_{s,1}(\mathbf{c}^\star) \leq \|\mathbf{c}^\star\|_1 \leq \|f\|_\rho$.

Remark 2. *Note that, in general, the solution to the BP formulation is not unique. Nonetheless, given that the stability guarantees of the BP problem hold for any solution $\mathbf{c}^\sharp$, the established uniform approximation result holds for any solution found by Algorithm 1.*

Remark 3. *Consider a function $f \in \mathcal{F}(\phi, \rho)$ whose Fourier transform is supported within a compact set $\Omega \subset \mathbb{R}^d$ such that that $\int_\Omega \rho(\boldsymbol{\omega})\,d\boldsymbol{\omega} =: \beta < 1$. Then the vector $\mathbf{c}^\star$ will be sparse with high probability, as its expected sparsity scales like $s = \beta N$. Thus, functions with compactly clustered spectral energy are well-approximated by the SRFE method.*

Remark 4. *An interesting aspect of Theorem 1 is the appearance of a Heisenberg-type uncertainty principle between "frequency-domain" and "space-domain" variances, σ^2 and γ^2 [23]. In Theorem 1, the product of the variances are bounded below by an $\mathcal{O}(s^{\frac{1}{d}})$ term.*

Theorem 1 shows that the error bound consists of three terms. The first term depends on the strength of the random features in representing f. By decreasing ϵ, thereby increasing N, we can increase the power of our representation and thus reduce this error term. The second term depends on the quality of the best s-term approximation of f with respect to the random feature basis. Since $\kappa_{s,1}(\mathbf{c}^\star)$ is bounded by $\|f\|_\rho$, the second error term is related to the complexity of the function class. Lastly, the third term is proportional to the level of noise on the samples and, in general, cannot be reduced arbitrarily.

Next, we adapt the result of Theorem 1 to the case where f is low order. In particular, Theorem 2 shows that one can improve the error bound when using q-sparse feature weights for small q.

Theorem 2 (**Function Approximation: Order-q Functions**). *Let f be an order-q function of at most K terms as defined in Definition 3, such that each term $g_\ell, \ell = 1, 2, \ldots, K$ belongs to $\mathcal{F}(\phi, \rho)$ with $\phi(\mathbf{x}; \boldsymbol{\omega}) = \phi(\langle \mathbf{x}, \boldsymbol{\omega} \rangle) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$, and $\rho : \mathbb{R}^q \to \mathbb{R}$ the density for a spherical Gaussian with variance σ^2, $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_q)$. Let $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N$ be a complete set of q-sparse feature weights drawn from density ρ. Fix γ, and draw i.i.d. sampling points $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathcal{N}(\mathbf{0}, \gamma^2 \mathbf{I}_d)$. Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ denote the associated random feature matrix where $a_{k,j} = \phi(\langle \mathbf{x}_k, \boldsymbol{\omega}_j \rangle)$ and $f^\sharp$ be defined from Algorithm 2 and Equation (4) with $\eta = \sqrt{2(\epsilon^2 |||f|||^2 + E^2)}$.*

Fix a target sparsity $s \in \mathbb{N}$, and suppose the feature parameters σ and N, the confidence δ, the accuracy ϵ, and the target radius R are chosen so that the following conditions hold:

1. *γ-σ uncertainty principle*

$$\gamma^2 \sigma^2 \geq \frac{1}{2} \left(\left(\frac{\sqrt{41}(2s-1)}{2} \right)^{\frac{2}{q}} - 1 \right), \tag{13}$$

2. *Number of features*

$$N \geq \frac{4}{\epsilon^2} \left(1 + 4R\sigma\sqrt{q} + \sqrt{\frac{q}{2} \log\left(\frac{d}{\delta}\right)} \right)^2, \tag{14}$$

3. *Number of measurements*

$$m \geq 4(2\gamma^2\sigma^2 + 1)^{\max\{2q-d,0\}} (\gamma^2\sigma^2 + 1)^{\min\{2q, 2d-2q\}} \log \frac{N^2}{\delta}, \tag{15}$$

4. *Target radius*

$$R \geq \gamma \sqrt{d + \sqrt{12d \log\left(\frac{m}{\delta}\right)}}. \tag{16}$$

Then, with probability at least $1 - 4\delta$ it holds that

$$\sup_{\mathbf{x} \in \mathbb{R}^d :\, \||\mathbf{x}|_{S_j}\|_2 \leq R,\, \forall j \in [K]} |f(\mathbf{x}) - f^\sharp(\mathbf{x})| \leq \epsilon |||f||| + C' \kappa_{s,1}(\tilde{\mathbf{c}}^\star) + C\eta\sqrt{s}, \tag{17}$$

where $C, C' > 0$ are constants and the vector $\tilde{\mathbf{c}} = [\tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_N]^T \in \mathbb{C}^N$ is defined as follows

$$\tilde{\mathbf{c}}_j^\star := \frac{1}{K} \sum_{\ell=1}^{K} \tilde{c}_{\ell,j}^\star, \quad \text{with} \quad \tilde{c}_{\ell,j}^\star = \begin{cases} \frac{\alpha_\ell(\boldsymbol{\omega}_j)}{N\, \rho(\boldsymbol{\omega}_j)}, & \text{if } \operatorname{supp}(\boldsymbol{\omega}_j) = \mathcal{S}_\ell \\ 0, & \text{otherwise.} \end{cases}$$

The function $\alpha_\ell(\boldsymbol{\omega})$ is the transform of g_ℓ using Definition 2 and Definition 3.

Note that $\{x \in \mathbb{R}^d : \|x\|_2 \leq R\} \subset \{x \in \mathbb{R}^d : \|x|_{S_j}\|_2 \leq R,\ \forall j \in [K]\}$, so we immediately get the bound

$$\sup_{\mathbf{x} \in \mathbb{R}^d :\, \|\mathbf{x}\|_2 \leq R} |f(\mathbf{x}) - f^\sharp(\mathbf{x})| \leq \epsilon |||f||| + C' \kappa_{s,1}(\tilde{\mathbf{c}}^\star) + C\eta\sqrt{s}. \tag{18}$$

We further note that, $\kappa_{s,1}(\tilde{\mathbf{c}}^\star) \leq \|\tilde{\mathbf{c}}^\star\|_1 \leq |||f|||$, in the worst case. We highlight the following remarks.

Remark 5. *From the proof, the bound for N is*

$$N \geq \frac{4}{\epsilon^2} \left(1 + 4R\sigma\sqrt{q} + \sqrt{\frac{1}{2} \log\left(\frac{K}{\delta}\right)} \right)^2$$

and we obtain Equation (14) by noting that $K \leq \binom{d}{q} \leq \left(\frac{ed}{q}\right)^q$.

Remark 6. *Note that in the bound for the number of measurements, the term* $(\gamma^2\sigma^2+1)^2$ *is in the range*

$$2\gamma^2\sigma^2+1 \leq (\gamma^2\sigma^2+1)^2 \leq (2\gamma^2\sigma^2+1)^2$$

and thus, if we choose the variances so that uncertainty principle holds with equality, then we see that m *scales between* s^4 *for* $q \leq \frac{d}{2}$ *and* s^2 *for* $q = d$.

We have established that Algorithm 1 uniformly approximates any function with bounded ρ-norm terms up to an error term that is proportional to the quantity $|||f|||$. In addition, the results of Theorem 1 and Theorem 2 suggest that the radius of approximation R depends more on the confidence δ (coming from the bounds on m) than it does on the accuracy parameter ϵ of the best ϕ-based approximation to f.

4.2 Generalization Error

The uniform bounds found in Theorem 1 and Theorem 2 can be extended to generalization error bounds. Recall that $\mu(\mathbf{x})$ denotes the probability measure for sampling $\mathbf{x}$.

Theorem 3 (**Generalization Bound for Bounded ρ-norm Functions**). *Instate the assumptions from Theorem 1. Extend the assumptions by allowing the measurement noise* e_k *to either be bounded by* $E = \sqrt{2}\nu$ *or to be drawn i.i.d. from* $\mathcal{N}(0,\nu^2)$. *Let* $\mathbf{A} \in \mathbb{C}^{m\times N}$ *denote the associated random feature matrix where* $a_{k,j} = \phi(\mathbf{x}_k;\boldsymbol{\omega}_j)$ *and* $f^\sharp$ *be defined from Algorithm 1 and Equation (4) with* $\eta = \sqrt{2(\epsilon^2\|f\|_\rho^2 + 2\nu^2)}$.

For a given s, *if the feature parameters* σ *and* N, *the confidence* δ, *and the accuracy* ϵ *are chosen so that the following conditions hold:*

1. γ-σ *uncertainty principle*

$$\gamma^2\sigma^2 \geq \frac{1}{2}\left(\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{2}{d}} - 1\right), \tag{19}$$

2. *Number of features*

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4\gamma\sigma d\sqrt{1+\sqrt{\frac{12}{d}\log\frac{m}{\delta}}} + \sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2, \tag{20}$$

3. *Number of measurements*

$$m \geq 4(2\gamma^2\sigma^2+1)^d \log\frac{N^2}{\delta}. \tag{21}$$

Then, with probability at least $1-5\delta$ *the following error bound holds*

$$\sqrt{\int_{\mathbb{R}^d} |f^\#(\mathbf{x}) - f^\star(\mathbf{x})|^2 \,\mathrm{d}\mu} \leq \epsilon\|f\|_\rho + C'\kappa_{s,1}(\mathbf{c}^\star) + C\eta\sqrt{s}, \tag{22}$$

where $C, C' > 0$ *are constants and* $\mathbf{c}^\star$ *is the vector*

$$\mathbf{c}^\star = \left[\frac{\alpha(\boldsymbol{\omega}_1)}{N\,\rho(\boldsymbol{\omega}_1)}, \cdots, \frac{\alpha(\boldsymbol{\omega}_N)}{N\,\rho(\boldsymbol{\omega}_N)}\right]^T.$$

The proof of Theorem 3 follows from a variation on the proof of Theorem 1 and can be found in Appendix D. Note that Theorem 1 and Theorem 2 can both be extended to also include normally distributed noise.

Theorem 4 (**Generalization Bounds for Lower Order Functions**). *Instate the assumptions of Theorem 2, extended so as to allow the measurement noise e_k to be either bounded by $E = \sqrt{2}\nu$ or drawn i.i.d. from $\mathcal{N}(0, \nu^2)$. Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ denote the associated random feature matrix where $a_{k,j} = \phi(\mathbf{x}_k; \boldsymbol{\omega}_j)$ and $f^\sharp$ be defined from Algorithm 2 and Equation (4) with $\eta = \sqrt{2\epsilon^2 |||f|||^2 + 4\nu^2}$. For a given s, suppose the feature parameters σ and N, the confidence δ, the accuracy ϵ, and the target radius R are chosen so that the following conditions hold:*

1. *γ-σ uncertainty principle*

$$\gamma^2\sigma^2 \geq \frac{1}{2}\left(\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{2}{q}} - 1\right), \tag{23}$$

2. *Number of features*

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4\gamma\sigma d\sqrt{1 + \sqrt{\frac{12}{d}\log\frac{m}{\delta}}} + \sqrt{\frac{q}{2}\log\left(\frac{d}{\delta}\right)}\right)^2, \tag{24}$$

3. *Number of measurements*

$$m \geq 4(2\gamma^2\sigma^2 + 1)^{\max\{2q-d,0\}}(\gamma^2\sigma^2 + 1)^{\min\{2q,2d-2q\}}\log\frac{N^2}{\delta}. \tag{25}$$

Then, with probability at least $1 - 4\delta$ it holds that

$$\sqrt{\int_{\mathbb{R}^d} |f(\mathbf{x}) - f^\star(\mathbf{x})|^2 \,\mathrm{d}\mu} \leq \epsilon |||f||| + C'\kappa_{s,1}(\tilde{\mathbf{c}}^\star) + C\eta\sqrt{s}, \tag{26}$$

where $C, C' > 0$ are constants and the vector $\tilde{\mathbf{c}}^\star = [\tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_N]^T \in \mathbb{C}^N$ is defined as follows:

$$\tilde{\mathbf{c}}_j^\star := \frac{1}{K}\sum_{\ell=1}^{K} \tilde{c}_{\ell,j}^\star, \quad \text{with} \quad \tilde{c}_{\ell,j}^\star = \begin{cases} \frac{\alpha_\ell(\boldsymbol{\omega}_j)}{N\,\rho(\boldsymbol{\omega}_j)}, & \text{if } \operatorname{supp}(\boldsymbol{\omega}_j) = \mathcal{S}_\ell \\ 0, & \text{otherwise.} \end{cases}$$

The function $\alpha_\ell(\boldsymbol{\omega})$ is the transform of g_ℓ using Definition 2 and Definition 3.

For Theorem 4, observe

$$\begin{aligned} \sqrt{\int_{\mathbb{R}^d} |f(\mathbf{x}) - f^\star(\mathbf{x})|^2 \,\mathrm{d}\mu} &= \left\| \frac{1}{K}\sum_{j=1}^{K} \left(g_j(\mathbf{x}|_{S_j}) - g_j^\star(\mathbf{x}|_{S_j})\right) \right\|_{L^2(d\mu)} \\ &\leq \frac{1}{K}\sum_{j=1}^{K} \left\| g_j(\mathbf{x}|_{S_j}) - g_j^\star(\mathbf{x}|_{S_j}) \right\|_{L^2(d\mu)} \\ &= \frac{1}{K}\sum_{j=1}^{K} \left\| g_j(\mathbf{x}|_{S_j}) - g_j^\star(\mathbf{x}|_{S_j}) \right\|_{L^2(d\tilde{\mu})} \end{aligned} \tag{27}$$

where $\mu(\mathbf{x})$ is the probability measure associated with the d-dimensional spherical Gaussian $\mathcal{N}(\mathbf{0}, \gamma^2\mathbf{I}_d)$ and $\tilde{\mu}(\mathbf{x})$ is the probability measure associated with the q-dimensional spherical Gaussian $\mathcal{N}(\mathbf{0}, \gamma^2\mathbf{I}_q)$. The proof of Theorem 4 follows by applying the lemmata from Appendix D to each error term $\left\| g_j(\mathbf{x}|_{S_j}) - g_j^\star(\mathbf{x}|_{S_j}) \right\|_{L^2(d\tilde{\mu})}$.

4.3 Proof of Theorem 1

In this section, we discuss our main technical arguments, which lead to Theorem 1. Note that the approximation error can be written as

$$\sup_{\mathbf{x}:\|x\|\leq R} |f(\mathbf{x}) - f^\sharp(\mathbf{x})| \leq \sup_{\mathbf{x}:\|x\|\leq R} |f(\mathbf{x}) - f^\star(\mathbf{x})| + \sup_{\mathbf{x}:\|x\|\leq R} |f^\sharp(\mathbf{x}) - f^\star(\mathbf{x})|, \tag{28}$$

where

$$f^\star(\mathbf{x}) = \sum_{j=1}^{N} c_j^\star \exp(i\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle), \quad c_j^\star := \frac{\alpha(\boldsymbol{\omega}_j)}{N\rho(\boldsymbol{\omega}_j)}. \tag{29}$$

We then aim to study these two sources of error in the following lemmata.

4.3.1 Bounding the first error term

We first extend an argument from [37] to derive a bound on how well a function in $\mathcal{F}(\phi, \rho)$ can be approximated by SRFE. That is, we characterize the approximation power of $f^\star$, the best ϕ-based approximation to f.

Lemma 1. *Fix the confidence parameter $\delta > 0$ and the accuracy parameter $\epsilon > 0$. Recall the setting of Algorithm 1 and suppose $f \in \mathcal{F}(\phi, \rho)$ where $\phi(\mathbf{x}; \boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$ and $\rho(\boldsymbol{\omega})$ is the probability density function (with finite second moment) used for sampling the random weights $\boldsymbol{\omega}$. Consider a set $\mathcal{X} \subset \mathbb{R}^d$ with diameter $R = \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|$. Suppose*

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4R\sqrt{\mathbb{E}\|\boldsymbol{\omega}\|^2} + \sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2. \tag{30}$$

Consider the approximation $f^\star$ defined in Equation (29). Then, with probability at least $1 - \delta$ with respect to the draw of the $\boldsymbol{\omega}$'s the following holds

$$\sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x}) - f^\star(\mathbf{x})| \leq \epsilon \|f\|_\rho. \tag{31}$$

The proof of Lemma 1, similar to the result of [37], uses McDiarmid's inequality [45] and bounds on the Rademacher complexity for complex valued functions to establish the lower bound on N. We modified the proof to allow for general ϕ (see Appendix B). The result in Lemma 1 is not constructive given that $\mathbf{c}^\star$ depends on the unknown function $\alpha(\boldsymbol{\omega})$. Nonetheless, Lemma 1 establishes a useful bound on the first source of error in (28).

4.3.2 Bounding the second error term

The next lemma controls the second source of error.

Lemma 2. *Let $f \in \mathcal{F}(\phi, \rho)$, where $\phi(\mathbf{x}; \boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$ and $\rho(\boldsymbol{\omega})$ is the density corresponding to a spherical Gaussian with variance σ^2, $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$. For a fixed γ and q with $q \leq d$, consider a set of data samples $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathcal{N}(\mathbf{0}, \gamma^2\mathbf{I}_d)$ and a complete set of q-sparse feature weights in $\mathbb{R}^d$, $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N$ drawn from $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_q)$. Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ denote the associated random feature matrix where $a_{k,j} = \phi(\mathbf{x}_k; \boldsymbol{\omega}_j)$ and $f^\sharp$ be define from Algorithm 1 and Equation (4) with $\eta = \sqrt{2(\epsilon^2\|f\|_\rho^2 + E^2)}$. Let $f^\star$ be defined as*

$$f^\star(\mathbf{x}) := \sum_{j=1}^{N} c_j^\star \exp(i\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle), \quad \text{where} \quad c_j^\star := \frac{\alpha(\boldsymbol{\omega}_j)}{N\rho(\boldsymbol{\omega}_j)}. \tag{32}$$

For a given s, if the feature parameters σ and N, the confidence δ, the accuracy ϵ, and the target radius R are chosen so that the following conditions hold:

$$\gamma^2\sigma^2 \geq \frac{1}{2}\left(\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{2}{q}} - 1\right),$$

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4R\sigma\sqrt{q} + \sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2,$$

$$m \geq 4(2\gamma^2\sigma^2+1)^{\max\{2q-d,0\}}(\gamma^2\sigma^2+1)^{\min\{2q,2d-2q\}}\log\frac{N^2}{\delta}$$

$$R \geq \gamma\sqrt{d + \sqrt{12d\log\left(\frac{m}{\delta}\right)}}.$$

Then, with probability at least $1-3\delta$ the following error bound holds:

$$\sup_{\mathbf{x}:\|x\|\leq R} |f^\sharp(\mathbf{x}) - f^\star(\mathbf{x})| \leq C'\,\kappa_{s,1}(\mathbf{c}^\star) + C\eta\sqrt{s}, \tag{33}$$

where $C, C' > 0$ are constants.

The proof of this lemma (see Appendix C) relies on demonstrating that given the assumptions on the data samples $\mathbf{x}_k$ and random weights $\boldsymbol{\omega}_j$, the corresponding random feature matrix $\mathbf{A}$ (see Step 4 in Algorithm 1) has a small mutual coherence $\mu_{\mathbf{A}}$, which we recall below.

Definition 4 (**Mutual Coherence** [21])**.** *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ be a matrix with columns $\mathbf{a}_1, \ldots, \mathbf{a}_N$. The mutual coherence of $\mathbf{A}$ is defined as*

$$\mu_{\mathbf{A}} = \sup_{\ell\neq j}\left\{|\mu_{j\ell}|,\ \mu_{j\ell} := \frac{\langle \mathbf{a}_j, \mathbf{a}_\ell\rangle}{\|\mathbf{a}_j\|_2\|\mathbf{a}_\ell\|_2}\right\}. \tag{34}$$

To establish Lemma 2, we argue that a small mutual coherence $\mu_{\mathbf{A}}$ is itself a consequence of the *bounded separation* of the randomly drawn weights. That is, consider a collection of random weights $\{\boldsymbol{\omega}_j\}_{j=1}^N$ in $\mathbb{R}^d$. For $\gamma > 0$ and a function $\psi : \mathbb{R}^d \to \mathbb{R}$, we define the quantities

$$\Gamma_{j\ell} := \psi\left(\gamma(\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell)\right), \quad \Gamma_{min} := \min_{j\neq\ell}\Gamma_{j\ell}, \quad \Gamma_{max} := \max_{j\neq\ell}\Gamma_{j\ell}. \tag{35}$$

We can quantify its separation with respect to ψ by bounding Γ_{max} and Γ_{min} by values depending on N and other dimensional constants. In the setting of Theorem 1 where the sampling points $\mathbf{x}_i$'s are i.i.d. Gaussian, the bounded separations hold for $\psi\left(\gamma(\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell)\right) = \exp\left(-2\gamma^2\pi^2\|\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell\|_2^2\right)$. Consequently, by utilizing the fact that the weights $\boldsymbol{\omega}$'s are normally distributed, we show that the collection $\{\boldsymbol{\omega}_j\}_{j=1}^N$ has bounded separation by establishing suitable bounds on Γ_{max} and Γ_{min} depending on N.

Given the bounds on Γ_{max} and Γ_{min}, by employing the Bernstein's inequality, we then establish that $\mu_{\mathbf{A}} \leq 2\Gamma_{max}$ with high probability, as long as

$$m \geq \frac{4}{\Gamma_{min}^2}\log\frac{N^2}{\delta}. \tag{36}$$

Upon establishing this upper bound on $\mu_{\mathbf{A}}$, we then utilize a result from compressive sensing regarding the stability of the BP formulation (see, e.g., [21]) to complete the proof of Lemma 2.

4.4 Proof of Theorem 2

To establish Theorem 2, we apply variations of Lemma 1 and Lemma 2 to each function g_j involved in the decomposition of f. Since f is an order-q function of at most K terms, it can be written as:

$$f(x_1, \ldots, x_d) = \frac{1}{K} \sum_{j=1}^{K} g_j(x_{j_1}, \ldots, x_{j_q}), \tag{37}$$

for a fixed q and K. For each term, define $g_j^\star$ as

$$g_j^\star(\mathbf{x}|_{\mathcal{S}_j}) = \sum_{\ell=1}^{N} \tilde{\mathbf{c}}_{j,\ell}^\star \ \exp(i\langle \mathbf{x}|_{\mathcal{S}_j}, \boldsymbol{\omega}_\ell|_{\mathcal{S}_j}\rangle), \quad \text{where} \quad \tilde{\mathbf{c}}_{j,\ell}^\star = \begin{cases} \frac{\alpha_j(\boldsymbol{\omega}_\ell)}{N\,\rho(\boldsymbol{\omega}_\ell)}, & \text{if } \operatorname{supp}(\boldsymbol{\omega}_\ell) = \mathcal{S}_j \\ 0, & \text{otherwise.} \end{cases} \tag{38}$$

We define $f^\star$ as

$$\begin{aligned} f^\star(\mathbf{x}) := \frac{1}{K} \sum_{j=1}^{K} g_j^\star(\mathbf{x}|_{\mathcal{S}_j}) &= \frac{1}{K} \sum_{j=1}^{K} \sum_{\ell=1}^{N} \tilde{\mathbf{c}}_{j,\ell}^\star \ \exp(i\langle \mathbf{x}|_{\mathcal{S}_j}, \boldsymbol{\omega}_\ell|_{\mathcal{S}_j}\rangle) \\ &= \frac{1}{K} \sum_{j=1}^{K} \sum_{\ell=1}^{N} \tilde{\mathbf{c}}_{j,\ell}^\star \ \exp(i\langle \mathbf{x}, \boldsymbol{\omega}_\ell\rangle), \\ &= \sum_{\ell=1}^{N} \left(\frac{1}{K} \sum_{j=1}^{K} \tilde{\mathbf{c}}_{j,\ell}^\star \right) \ \exp(i\langle \mathbf{x}, \boldsymbol{\omega}_\ell\rangle) \\ &= \sum_{\ell=1}^{N} \tilde{\mathbf{c}}_\ell^\star \ \exp(i\langle \mathbf{x}, \boldsymbol{\omega}_\ell\rangle), \end{aligned} \tag{39}$$

where in the second line we use Equation (38) and we define $\tilde{\mathbf{c}}_\ell^\star := \frac{1}{K} \sum_{j=1}^{K} \tilde{\mathbf{c}}_{j,\ell}^\star$. For each term g_j, only n out of the N features are active, so $\mathbf{c}^\star$ is nK-sparse. Since there are K such terms, by applying the union bound, if

$$n \geq \frac{4}{\epsilon^2} \left(1 + 4R\sigma\sqrt{q} + \sqrt{\frac{1}{2} \log\left(\frac{K}{\delta}\right)} \right)^2, \tag{40}$$

then

$$\sup_{\|\mathbf{x}|_{\mathcal{S}_j}\| \leq R} |g_j(\mathbf{x}|_{\mathcal{S}_j}) - g_j^\star(\mathbf{x}|_{\mathcal{S}_j})| \leq \epsilon \|g_j\|_\rho, \tag{41}$$

holds for each $j \in [K]$. By the triangle inequality, we have

$$\begin{aligned} \sup_{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}|_{\mathcal{S}_j}\|_2 \leq R,\ \forall j \in [K]} |f(\mathbf{x}) - f^\star(\mathbf{x})| &\leq \frac{1}{K} \sup_{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}|_{\mathcal{S}_j}\|_2 \leq R,\ \forall j \in [K]} \sum_{j=1}^{K} |g_j(\mathbf{x}|_{\mathcal{S}_j}) - g_j^\star(\mathbf{x}|_{\mathcal{S}_j})| \\ &\leq \frac{1}{K} \sum_{j=1}^{K} \left(\sup_{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}|_{\mathcal{S}_j}\|_2 \leq R} |g_j(\mathbf{x}|_{\mathcal{S}_j}) - g_j^\star(\mathbf{x}|_{\mathcal{S}_j})| \right) \\ &\leq \frac{\epsilon}{K} \sum_{j=1}^{K} \|g_j\|_\rho. \end{aligned} \tag{42}$$

Following the proof of Lemma 2 with $\eta = \sqrt{2\left(\epsilon^2 \left(\frac{1}{K} \sum_{j=1}^{K} \|g_j\|_\rho\right)^2 + E^2\right)}$, letting $g_j^\star$ and $f^\star$ be defined as above, and $N = n\binom{d}{q}$, then, with probability at least $1 - 3\delta$ the following error bound

holds:

$$\sup_{\mathbf{x}\in\mathbb{R}^d:\|\mathbf{x}|_{\mathcal{S}_j}\|_2\leq R,\ \forall j\in[K]} |f^\sharp(\mathbf{x}) - f^\star(\mathbf{x})| \leq C'\,\kappa_{s,1}(\mathbf{c}^\star) + C\eta\sqrt{s}, \tag{43}$$

where $C, C' > 0$ are constants. By rescaling the term ϵ to $\epsilon\binom{d}{q}^{\frac{1}{2}}$ and assuming ϵ is sufficiently small, we conclude the proof.

5 Extensions

In this section, we discuss how some of the requirements in Theorem 1 and Theorem 2 can be relaxed to further strengthen our results. For simplicity we state the formal results of this section within the scope of Theorem 1, i.e., without assuming the low order property. However, we note that similar results hold for the case of low order functions by following the steps outlined in the proof of Theorem 2.

Also, we can easily extend the results to the case of $\mathbf{x}_k \sim \mathcal{N}(\bar{\mathbf{x}}, \gamma^2\mathbf{I}_d)$ with the random Fourier features, thus the zero-mean assumption can be relaxed. In practice, we may further consider features involving a random bias term, e.g. $\exp(i(\langle\mathbf{x},\boldsymbol{\omega}\rangle + p))$, where $p \sim \mathcal{U}[0, 2\pi]$. Due to the intrinsic symmetry of the Fourier features, the bias term can be readily accounted for as well without a significant modification to the statements. Note that these extensions do not affect the result of Lemma 1 that bounded the first source of error.

5.1 Real-Valued Functions

If the unknown function is real-valued, we prefer to work with real-valued SRFE such as $\phi(\mathbf{x};\boldsymbol{\omega}) = \cos(\langle\mathbf{x},\boldsymbol{\omega}\rangle)$, $\phi(\mathbf{x};\boldsymbol{\omega}) = \sin(\langle\mathbf{x},\boldsymbol{\omega}\rangle)$ or a combination of the two. Our theoretical results readily extend to this case up to some constants by using a variation of the Euler's formula.

Theorem 5. *Let $f \in \mathcal{F}(\phi,\rho)$, where $\phi_{\cos}(\mathbf{x};\boldsymbol{\omega}) = \cos(\langle\mathbf{x},\boldsymbol{\omega}\rangle)$, $\phi_{\sin}(\mathbf{x};\boldsymbol{\omega}) = \sin(\langle\mathbf{x},\boldsymbol{\omega}\rangle)$, and $\rho(\boldsymbol{\omega})$ is the density corresponding to a spherical Gaussian with variance σ^2, $\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)$, with the additional condition that f is real valued. For a fixed γ, consider a set of data samples $\mathbf{x}_1,\dots,\mathbf{x}_m \sim \mathcal{N}(\mathbf{0},\gamma^2\mathbf{I}_d)$ and a complete set of q-sparse feature weights in $\mathbb{R}^d$, $\boldsymbol{\omega}_1,\dots,\boldsymbol{\omega}_N$ drawn from $\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_q)$. Let $\mathbf{A} \in \mathbb{R}^{m\times 2N}$ denote the associated random feature matrix where $a_{k,j} = \phi_{\cos}(\mathbf{x}_k;\boldsymbol{\omega}_j)$ for $k \leq N$, $a_{k,j} = \phi_{\sin}(\mathbf{x}_k;\boldsymbol{\omega}_j)$ for $N < k \leq 2N$, and $f^\sharp$ be defined from Algorithm 1 and Equation (4) with $\eta = \sqrt{2(\epsilon^2\|f\|_\rho^2 + E^2)}$.*

For a given s, if the feature parameters σ and N, the confidence δ, the accuracy ϵ, and the target radius R are chosen so that the following conditions hold:

1. *γ-σ uncertainty principle*

$$\gamma^2\sigma^2 \geq \frac{1}{2}\left(\left(\sqrt{41}(2s-1)\right)^{\frac{2}{d}} - 1\right), \tag{44}$$

2. *Number of features*

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4R\sigma\sqrt{d} + \sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2 \tag{45}$$

3. *Number of measurements*

$$m \geq 4(2\gamma^2\sigma^2+1)^d \log\frac{N^2}{\delta} \tag{46}$$

4. *Target radius*

$$R \geq \gamma\sqrt{d + \sqrt{12d\log\left(\frac{m}{\delta}\right)}}. \tag{47}$$

Then, with probability at least $1 - 5\delta$ *the following error bound holds:*

$$\sup_{\mathbf{x}:\|x\|\leq R} |f(\mathbf{x}) - f^\sharp(\mathbf{x})| \leq \epsilon \|f\|_\rho + C'\kappa_{s,1}(\mathbf{c}^\star) + C\eta\sqrt{s}, \tag{48}$$

where $C, C' > 0$ *are constants and the vector* $\mathbf{c}^\star$ *is defined by*

$$\mathbf{c}^\star = \left[\frac{\alpha(\boldsymbol{\omega}_1)}{2N\,\rho(\boldsymbol{\omega}_1)}, \ldots, \frac{\alpha(\boldsymbol{\omega}_{2N})}{2N\,\rho(\boldsymbol{\omega}_{2N})}\right]^T.$$

5.2 Uniformly Distributed Data

We can extend our results uniformly distributed data and establish comparable conditions for the uniform approximation of low order functions. Note that this extension does not affect our study of the first source of error in Lemma 1.

Theorem 6. *Let* $f \in \mathcal{F}(\phi, \rho)$*, where* $\phi(\mathbf{x}; \boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$ *and* $\rho(\boldsymbol{\omega})$ *is the density corresponding to a spherical Gaussian with variance* σ^2*,* $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$*. For a fixed* γ*, consider a set of data samples* $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathcal{U}[-\gamma, \gamma]^d$ *and frequencies* $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$*. Let* $\mathbf{A} \in \mathbb{C}^{m\times N}$ *denote the associated random feature matrix where* $a_{k,j} = \phi(\mathbf{x}_k; \boldsymbol{\omega}_j)$ *and* $f^\sharp$ *be defined from Algorithm 1 and Equation (4) with* $\eta = \sqrt{2(\epsilon^2\|f\|_\rho^2 + E^2)}$*.*

For a given s*, if the feature parameters* σ *and* N*, the confidence* δ*, and the accuracy* ϵ *are chosen so that the following conditions hold:*

1. *γ-σ uncertainty principle*

$$\gamma\sigma \geq \min\left\{\sqrt{\frac{\pi}{4}}\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{1}{d}}, 2\right\}, \tag{49}$$

2. *Number of features*

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4\gamma\sigma d + \sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2, \tag{50}$$

3. *Number of measurements*

$$m \geq 4\left(\frac{2\sqrt{2}\gamma\sigma}{\sqrt{\pi}}\right)^{2d}\log\frac{N^2}{\delta}. \tag{51}$$

Then, with probability at least $1 - 4\delta$ *the following error bound holds:*

$$\sup_{\mathbf{x}\in[-\gamma,\gamma]^d} |f(\mathbf{x}) - f^\sharp(\mathbf{x})| \leq \epsilon \|f\|_\rho + C'\kappa_{s,1}(\mathbf{c}^\star) + C\eta\sqrt{s}, \tag{52}$$

where $C, C' > 0$ *are constants and the vector* $\mathbf{c}^\star$ *is defined by*

$$\mathbf{c}^\star = \left[\frac{\alpha(\boldsymbol{\omega}_1)}{N\,\rho(\boldsymbol{\omega}_1)}, \ldots, \frac{\alpha(\boldsymbol{\omega}_N)}{N\,\rho(\boldsymbol{\omega}_N)}\right]^T.$$

The proof of a more general version of Theorem 6 is detailed in Appendix E.3. As a direct result of Theorem 6, the following generalization bound holds.

Corollary 6.1. *Let $f \in \mathcal{F}(\phi, \rho)$, where $\phi(\mathbf{x}; \boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$ and ρ is the probability distribution associated with sampling $\boldsymbol{\omega}$. For a fixed γ, consider a set of data samples $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathcal{U}[-\gamma, \gamma]^d$ and frequencies $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$. Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ denote the associated random feature matrix where $a_{k,j} = \phi(\mathbf{x}_k; \boldsymbol{\omega}_j)$ and $f^\sharp$ be defined from Algorithm 1 and Equation (4) with $\eta = \sqrt{2(\epsilon^2 \|f\|_\rho^2 + E^2)}$.*

For a given s, if the feature parameters σ and N, the confidence δ, and the accuracy ϵ are chosen so that the following conditions hold:

1. *γ-σ uncertainty principle*

$$\gamma\sigma \geq \min\left\{ \sqrt{\frac{\pi}{4}} \left(\frac{\sqrt{41}(2s-1)}{2} \right)^{\frac{1}{d}}, \; 2 \right\}, \tag{53}$$

2. *Number of features*

$$N \geq \frac{4}{\epsilon^2} \left(1 + 4\gamma\sigma d + \sqrt{\frac{1}{2} \log\left(\frac{1}{\delta} \right)} \right)^2, \tag{54}$$

3. *Number of measurements*

$$m \geq 4 \left(\frac{2\sqrt{2}\gamma\sigma}{\sqrt{\pi}} \right)^{2d} \log \frac{N^2}{\delta}. \tag{55}$$

Then, with probability at least $1 - 4\delta$ the following error bound holds:

$$\sqrt{\int_{\mathbb{R}^d} |f(\mathbf{x}) - f^\star(\mathbf{x})|^2 \, \mathrm{d}\mu} \leq \epsilon \|f\|_\rho + C' \kappa_{s,1}(\mathbf{c}^\star) + C\eta\sqrt{s}, \tag{56}$$

where $C, C' > 0$ are constants, and

$$\mathbf{c}^\star = \left[\frac{\alpha(\boldsymbol{\omega}_1)}{N \, \rho(\boldsymbol{\omega}_1)}, \ldots, \frac{\alpha(\boldsymbol{\omega}_N)}{N \, \rho(\boldsymbol{\omega}_N)} \right]^T.$$

5.3 Beyond Gaussian Features

We can extend our main theorems to the case where the nonzero entries of the parameters $\boldsymbol{\omega}$'s are drawn from a *subgaussian* distribution [48]. Although we can directly apply the result of Lemma 1 to bound the first source of error, this extension requires certain modifications to the requirements on the variance parameters σ^2 and γ^2, which stem from the fact that we need to resort to (in some sense) weaker concentration inequalities.

Theorem 7. *Let $f \in \mathcal{F}(\phi, \rho)$, where $\phi(\mathbf{x}; \boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$ and $\rho(\boldsymbol{\omega})$ is the density corresponding to a spherical Gaussian with variance σ^2, $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$. For a fixed γ, consider a set of data samples $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathcal{N}(\mathbf{0}, \gamma^2 \mathbf{I}_d)$ and frequencies $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N$ which are i.i.d. centered σ^2-subgaussian random vectors. Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ denote the associated random feature matrix where $a_{k,j} = \phi(\mathbf{x}_k; \boldsymbol{\omega}_j)$ and $f^\sharp$ be defined from Algorithm 1 and Equation (4) with $\eta = \sqrt{2(\epsilon^2 \|f\|_\rho^2 + E^2)}$.*

For a given s, if the feature parameters σ and N, the confidence δ, the accuracy ϵ, and the target radius R are chosen so that the following conditions hold:

1. *γ-σ uncertainty principle*

$$\begin{aligned} \gamma^2 \mathbb{E}[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2] &\geq \frac{2}{(1-t)} \log\left(\frac{\sqrt{41}(2s-1)}{2} \right), \\ t &:= \frac{2^{9/4} \sigma \sqrt{q}}{\mathbb{E}[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2]} \sqrt{\log \frac{N^2}{\delta}} \end{aligned} \tag{57}$$

where $\{\boldsymbol{\omega}_j\}_{j=1}^N$ *are chosen so that* $t < 1$ *holds,*

2. *Number of features*

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4R\sigma\sqrt{d} + \sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2, \tag{58}$$

3. *Number of measurements*

$$m \geq 4\exp\left(2\gamma^2\mathbb{E}[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2]\right)\log\frac{N^2}{\delta}, \tag{59}$$

4. *Target radius*

$$R \geq \gamma\sqrt{d + \sqrt{12d\log\left(\frac{m}{\delta}\right)}}. \tag{60}$$

Then, with probability at least $1 - 5\delta$ *the following error bound holds:*

$$\sup_{\mathbf{x}:\|x\|\leq R} |f(\mathbf{x}) - f^\sharp(\mathbf{x})| \leq \epsilon\|f\|_\rho + C'\kappa_{s,1}(\mathbf{c}^\star) + C\eta\sqrt{s}, \tag{61}$$

where $C, C' > 0$ *are constants and*

$$\mathbf{c}^\star = \left[\frac{\alpha(\boldsymbol{\omega}_1)}{N\,\rho(\boldsymbol{\omega}_1)}, \ldots, \frac{\alpha(\boldsymbol{\omega}_N)}{N\,\rho(\boldsymbol{\omega}_N)}\right]^T.$$

6 Experimental Results

In the first example, we show that the SRFE in Algorithm 1 outperforms a shallow neural network on the approximation of an order-2 function:

$$f(x_1, \ldots, x_{10}) = \frac{1}{10}\sum_{\ell=1}^{9}\frac{\exp(-x_\ell^2)}{1 + x_{\ell+1}^2}$$

in the data-scarce regime. For Algorithm 1, we set $\eta = 0.01$, $q = 2$ or $q = 10$, $\sigma = 1$, and the bias $p \sim \mathcal{U}[0, 2\pi]$. In all of the examples we set $\phi(\mathbf{x}; \boldsymbol{\omega}) = \sin(\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$, unless otherwise specified. We define the relative testing error to be:

$$\text{Error} = \sqrt{\frac{\sum_{k\in\text{Test}} |f(\mathbf{x}_k) - f^\sharp(\mathbf{x}_k)|^2}{\sum_{k\in\text{Test}} |f(\mathbf{x}_k)|^2}},$$

where $f^\sharp$ denotes the approximation found by Algorithm 1 or the benchmarking schemes.

In Figure 1, we compare the SRFE (with $N = 5000$) to a two-layer ReLU network with 500 and 5000 trainable parameters. The ReLU network with 500 trainable parameters is included so as to match the number of active parameters in the SRFE. The SRFE with $q = d = 10$ is more accurate than the shallow network in this data regime. When $q = 2$, the error associated with the SRFE-S is smaller than that of the SRFE results with $q = d$ and is about one order of magnitude smaller than the neural network.

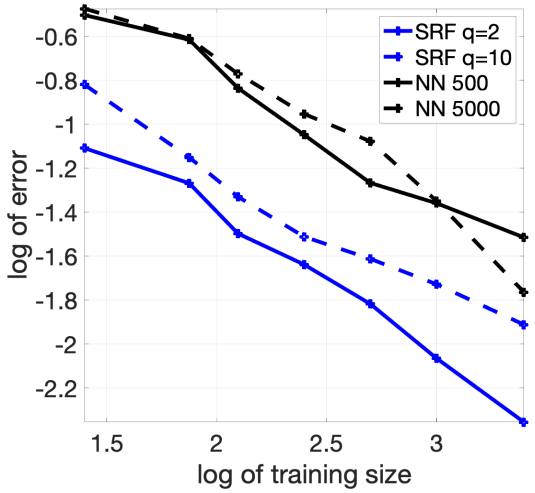

Figure 1: **Function Approximation:** Comparison of relative testing error versus the size of the training set for the sparse random feature model with $q = 2$ and $q = 10$ and for the two-layer ReLU network using 500 and 5000 trainable parameters.

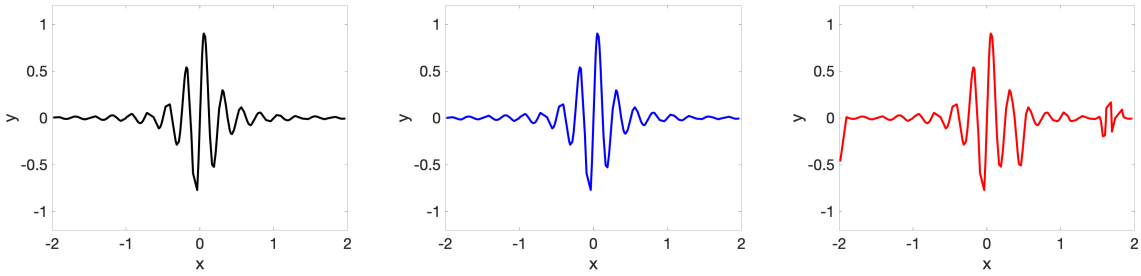

Figure 2: **Comparison, Overfitting**: The first figure is the target function, the second and third figures are the approximations via the SRFE and the OLS methods respectively with the same 200 randomly sampled points.

6.1 Overfitting and Noise

In this example, we provide a visual comparison of the recovery of one-dimensional functions using the SRFE algorithm and the ordinary least squares (OLS) approach. The first plot of Figure 2 is the target function (a sine packet), the second and third plots are the approximations using the SRFE and the OLS methods respectively with the same 200 randomly sampled points. The features are sampled using $\sigma = 2\pi$. Note the appearance of high-frequency aliasing with the OLS approximation.

In Figure 3, noisy one dimensional data is considered. The first column includes the Runge function (top) and a triangle function (bottom) each with 5% relative noise. The second and third columns are the approximations using the SRFE and the OLS methods respectively with the same 200 randomly sampled points. The first row uses $\sigma = \pi$ and the second row uses $\sigma = 2\pi$. The results using the SRFE are more accurate and contain less noise artifacts. Note that since the basis is trigonometric, the approximations are smooth. The OLS results have overfit the data, even when the feature parameter N is varied.

6.2 Low Order Approximations

In Table 1, we test the effect of varying q for different functions using Algorithm 2 and recording the relative errors. The highlighted (purple) values represent the explicit order of the function.

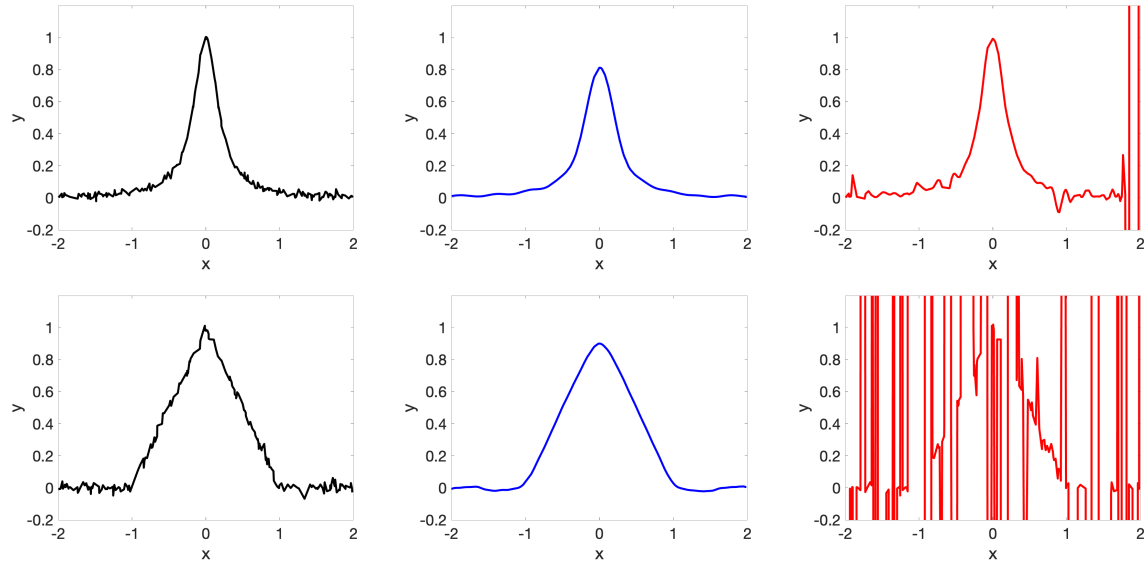

Figure 3: **Comparison, Noise**: The first column includes the Runge function (top) and a triangle function (bottom) each with 5% relative noise. The second and third columns are the approximations via the SRFE and the OLS methods respectively with the same 200 randomly sampled points.

$f(\mathbf{x})$	σ	d	$q=1$	$q=2$	$q=3$	$q=5$
$\left(\sum_{i=1}^{d} x_j\right)^2$	0.1	1	0.82	5.71×10^{-6}	6.92×10^{-5}	8.3×10^{-4}
$(1+\|\mathbf{x}\|_2^2)^{-1/2}$	1	5	3.27	1.60	1.95	1.72
$\sqrt{1+\|\mathbf{x}\|_2^2}$	1	5	1.02	0.73	0.80	1.10
$\mathrm{sinc}(x_1)\mathrm{sinc}(x_3)^3+\mathrm{sinc}(x_2)$	π	5	12.90	1.19	1.13	3.51
$\frac{x_1 x_2}{1+x_3^6}$	1	5	100.30	21.53	4.95	5.06
$\sum_{i=1}^{d} \exp(-\|x_i\|)$	1	100	0.91	1.43	1.57	1.96

Table 1: **Low Order Examples** The table contains the relative test error (as a percentage) for approximating various functions using different q values. The purple values represent the order of the function. We fix $m = 1000$ and $N = 10000$ with random sine features. We draw $\mathbf{x} \sim \mathcal{U}[-1,1]^d$ and the nonzero values of $\boldsymbol{\omega}$ are drawn from $\mathcal{N}(\mathbf{0}, \sigma^2)$.

We fixed $m = 1000$, $N = 10000$ and used the random sine features. The data is sampled from $\mathcal{U}[-1,1]^d$ and the nonzero values of $\boldsymbol{\omega}$ are drawn from $\mathcal{N}(\mathbf{0}, \sigma^2)$, where σ and d are included in the table for each example.

In the second and third examples, while the functions are order $q = d$ functions, they enjoy better accuracy for $q = 2$. This could be due to several phenomena. The first is that, with fixed m and N, the error may increase as q increase (see Theorem 2). However, this should partially be mitigated since we chose $N = 10000$ large enough. Another reason is that, with respect to some expansion (i.e. Fourier or Taylor), the functions can be written as an order $q < d$ function within some level of accuracy. This motivates further investigations in future work. The other examples show a clear transition when the correct range for q is obtained.

6.3 HyShot 30 Data

In Table 2, we apply the SRFE on the HyShot dataset (Hypersonics Flow Data [14]) and measure the relative testing error as a function of N (the number of random features). The input space is $d = 7$ dimensional and the dataset includes 52 total samples (which we split into 26-26). We set

$\eta = 0.01$, $\sigma = 2\pi$, $p \sim \mathcal{U}[0,1]$, and $q = 7$ (no coordinate sparsity is assumed). In this setting, we have $N \gg m$, which causes the RFF model and the two-layer ReLU network to overfit on the data (the training loss is small). When using $\phi(\mathbf{x};\boldsymbol{\omega}) = \sin(\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$, the SRFE produces consistent testing error which decreases as N increases. On the other hand, when $\phi(\mathbf{x};\boldsymbol{\omega}) = \mathrm{ReLU}(\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$, the results using SRFE achieve a smaller overall testing error but do not improve with N. Table 2 shows that unlike the SRFE, there are no gains made from increasing the number of trainable parameters in the shallow NN model.

6.4 NACA Sound Dataset

We comparing the SRFE and the RFF models without coordinate sparsity on the National Advisory Committee for Aeronautics (NACA) sound dataset [19] and measure the relative training and testing error as a function of N. The input space is $d = 5$ dimensional, the total number of samples is 1503, the train-test split $80 - 20$, $\eta = 100$, $\sigma = 1$, and $p \sim \mathcal{U}[0,1]$. The relative testing errors in Table 2 indicate an overall consistent result, in terms of the coefficient sparsity and the errors, when using the SRFE approach. The RFF model overfits as N increases beyond the size of the training set.

6.5 Comparison with Sparse PCE

In Figure 4, we compare the SRFE-S approach with the Sparse PCE approach [32] using various random sampling methods on the Ishigami example:

$$f(x_1, x_2, x_3) = \sin(x_1) + 7\sin^2(x_2) + 0.1x_3^4 \sin(x_1)$$

which is of order 2. The first row of Figure 4 uses i.i.d. samples $\mathbf{x}_k \sim \mathcal{U}[-\pi, \pi]^d$, the third row uses i.i.d. samples $\mathbf{x}_k \sim \mathcal{N}\left(0, \frac{1}{4}\mathbf{I}_d\right)$, and the second row uses a mixed distribution $\mathbf{x}_k = \mathbf{x}_{k,1} + \mathbf{x}_{k,2}$ where $\mathbf{x}_{k,1} \sim \mathcal{N}\left(0, \frac{1}{100}\mathbf{I}_d\right)$ and $\mathbf{x}_{k,1} \sim \mathcal{U}[-\pi, \pi]^d$. Each model uses $N = 3276$ features (which is equivalent to a degree-25 polynomial system in the case of the Sparse PCE approach) and (the same) 200 random samples. The hyper-parameters for the SRFE are set to $q = 2$, $\sigma = \frac{3\pi}{2}$, and $p \sim \mathcal{U}[0, 2\pi]$. When using uniformly random samples, the Sparse PCE approach produces lower testing error (0.24% versus 1.43%), which continues to perform well as N increases. This is due in part to the fact that the orthogonal polynomial basis (in this case, the Legendre basis) has knowledge of the input distribution. When the samples are Gaussian, the SRFE produces a more accurate solution than the Sparse PCE method (0.44% versus 6.24%). For the mixture case, the SRFE outperforms the Sparse PCE method (2.11% versus 15.05%). Note that the Sparse PCE

HyShot 30	$N = 100$	$N = 200$	$N = 400$	$N = 800$
SRFE with Sine	6.95	6.23	5.76	5.64
SRFE with ReLU	1.40	1.45	1.51	1.59
Random Fourier Features	84.23	89.99	95.17	97.84
Two-layer ReLU Network	7.29	11.50	11.19	11.33
NACA Sound	$N = 250$	$N = 1500$	$N = 5000$	$N = 10000$
SRFE (Train)	3.22	2.30	2.30	2.31
SRFE (Test)	3.22	3.04	2.77	2.78
SRFE (Average Sparsity)	250	364.4	185.7	185.7
Random Fourier Features (Train)	3.22	0.25	0.20	0.19
Random Fourier Features (Test)	7.45	2.13×10^8	1.69×10^8	1.48×10^8

Table 2: **HyShot 30 and NACA Sound Datasets**: Average relative train and test errors over 10 random trials (as a percentage). For the shallow NN, we choose the hidden layer so that the total number of parameters match N.

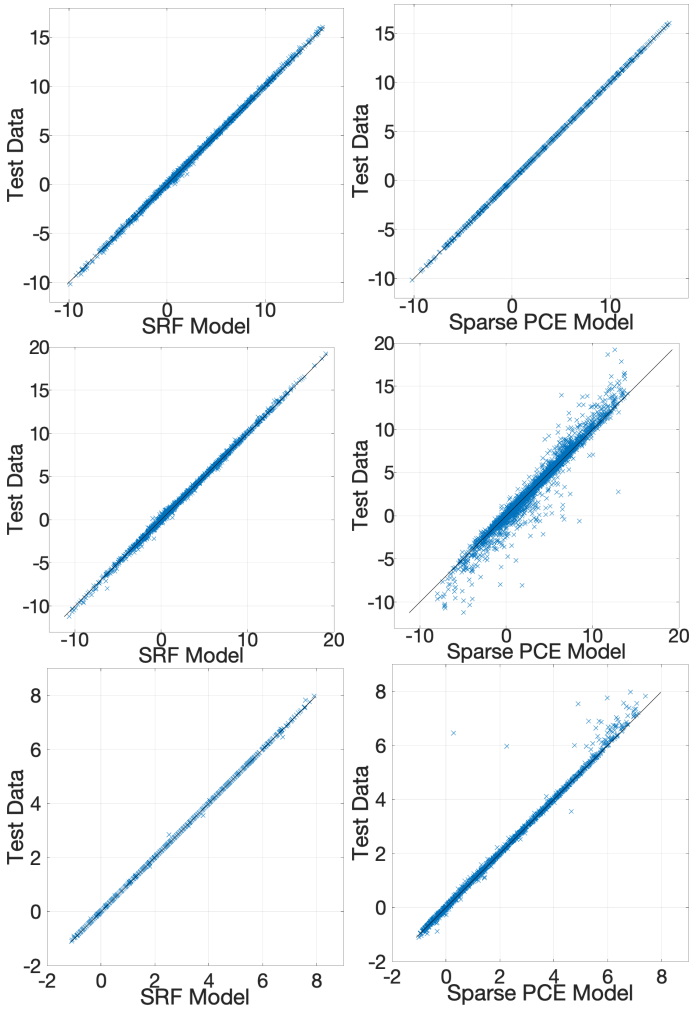

Figure 4: **Comparison with Sparse PCE.** Each scatter plot is the model response versus the data. The first column is the SRFE and the second column is the Sparse PCE model. The first row uses i.i.d. samples from $\mathcal{U}[-\pi,\pi]^d$, the third row uses i.i.d. samples from $\mathcal{N}\left(0,\frac{1}{4}\mathbf{I}_d\right)$, and the second row uses the sum of samples from $\mathcal{N}\left(0,\frac{1}{100}\mathbf{I}_d\right)$ and $\mathcal{U}[-\pi,\pi]^d$. Each model uses $N = 3276$ features (which is equivalent to a degree-25 polynomial system in the case of the Sparse PCE approach) and (the same) $m = 200$ random samples. While the Sparse PCE performs well on the uniform distribution (first row), the SRFE produces accurate approximations in all cases.

must derive the orthogonal basis from the data (or use the Legendre basis as its default), where as, at least experimentally, our approach is applicable to a larger class of input distributions.

7 Conclusion

We proposed the sparse random features method as a new approach in function approximation. For low order functions, i.e. functions that admit a decomposition to terms depending on only a few of the independent variables, we introduce low order random features. By utilizing techniques from compressive sensing and probability, we provided uniform bounds on the approximation error of the proposed scheme and established sample and feature complexities. On several examples, we should improved accuracy over other popular approximation schemes. As part of the future work, we intend to explore the avenues to incorporate additional functional structures into the proposed framework with the hope of further improving the approximation properties of the proposed scheme.

References

[1] Adcock, B., Brugiapaglia, S, and Webster, C. G. Compressed Sensing and its Applications. *Springer* (2017), pp. 93–124.

[2] Adcock, B. Infinite-dimensional compressed sensing and function interpolation. *Foundations of Computational Mathematics* 18(3) (2018), pp. 661–701.

[3] Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning* (2019), PMLR, pp. 322–332.

[4] Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory 39*, 3 (1993), 930–945.

[5] Bartlett, P. L., and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research 3*, Nov (2002), 463–482.

[6] Block, H.-D. The perceptron: A model for brain functioning. i. *Reviews of Modern Physics 34*, 1 (1962), 123.

[7] Cai, T. T., Xu, G., and Zhang, J. On recovery of sparse signals via ℓ_1 minimization. *IEEE Transactions on Information Theory 55*, 7 (2009), 3388–3397.

[8] Candes, E. J., Romberg, J. K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences 59*, 8 (2006), 1207–1223.

[9] Candes, E. J., and Tao, T. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory 52*, 12 (2006), 5406–5425.

[10] Chitta, R., Jin, R., and Jain, A. K. Efficient kernel clustering using random Fourier features. In *2012 IEEE 12th International Conference on Data Mining* (2012), IEEE, pp. 161–170.

[11] Chkifa, A., Dexter, N, Tran, H. and Webster, C. Polynomial approximation via compressed sensing of high-dimensional functions on lower sets. *Mathematics of Computation* 311(87) (2018), pp. 1415–1450.

[12] Constantine, P. G., Dow, E., and Wang, Q. Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM Journal on Scientific Computing 36*, 4 (2014), A1500–A1524.

[13] Constantine, P. G., Eftekhari, A., Hokanson, J., and Ward, R. A. A near-stationary subspace for ridge approximation. *Computer Methods in Applied Mechanics and Engineering 326* (2017), 402–421.

[14] Constantine, P. G., Emory, M., Larsson, J., and Iaccarino, G. Exploiting active subspaces to quantify uncertainty in the numerical simulation of the HyShot II scramjet. *Journal of Computational Physics 302* (2015), 1–20.

[15] DeVore, R., Petrova, G., and Wojtaszczyk, P. Approximation of functions of few variables in high dimensions. *Constructive Approximation 33*, 1 (2011), 125–143.

[16] Donoho, D. L. Compressed sensing. *IEEE Transactions on information theory 52*, 4 (2006), 1289–1306.

[17] Doostan, A., and Owhadi, H. A non-adapted sparse approximation of PDEs with stochastic inputs. *Journal of Computational Physics 230*, 8 (2011), 3015–3034.

[18] Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054* (2018).

[19] Dua, D., and Graff, C. UCI machine learning repository, 2017.

[20] Fornasier, M., Schnass, K., and Vybiral, J. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics 12*, 2 (2012), 229–262.

[21] Foucart, S., and Rauhut, H. *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

[22] Harris, K. D. Additive function approximation in the brain. *arXiv preprint arXiv:1909.02603*, 2019.

[23] Heisenberg, W. About the descriptive content of quantum theoretical kinematics and mechanics. *Zeitschrift für Physik* (1927), 172—-198.

[24] Honorio, J., and Jaakkola, T. Tight bounds for the expected risk of linear classifiers and PAC-Bayes finite-sample guarantees. In *Artificial Intelligence and Statistics* (2014), pp. 384–392.

[25] Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572* (2018).

[26] Jones, L. K. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The annals of Statistics 20*, 1 (1992), 608–613.

[27] Kuo, F., Sloan, I., Wasilkowski, G., and Woźniakowski, H. On decompositions of multivariate functions. *Mathematics of computation 79*, 270 (2010), 953–966.

[28] Ledoux, M., and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[29] Li, Y., and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. *arXiv preprint arXiv:1808.01204* (2018).

[30] Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. Towards a unified analysis of random Fourier features. In *International Conference on Machine Learning* (2019), PMLR, pp. 3905–3914.

[31] Maass, W., and Markram, H. On the computational power of circuits of spiking neurons. *Journal of computer and system sciences 69*, 4 (2004), 593–616.

[32] Marelli, S., and Sudret, B. UQlab: A framework for uncertainty quantification in Matlab. In *Vulnerability, uncertainty, and risk: quantification, mitigation, and management*. 2014, pp. 2554–2563.

[33] Mei, S., and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355* (2020).

[34] Moosmann, F., Triggs, B., and Jurie, F. Randomized clustering forests for building fast and discriminative visual vocabularies. In *NIPS* (2006), NIPS.

[35] Potts, D. and Schmischke, M. Approximation of high-dimensional periodic functions with Fourier-based methods. *arXiv preprint arXiv:1907.11412* (2019).

[36] Potts, D. and Schmischke, M. Learning multivariate functions with low-dimensional structures using polynomial bases. *arXiv preprint arXiv:1912.03195* (2019).

[37] Rahimi, A., and Recht, B. Uniform approximation of functions with random bases. *2008 46th Annual Allerton Conference on Communication, Control, and Computing* (2008), IEEE, pp. 555–561.

[38] Rahimi, A., and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems 21* (2008), 1313–1320.

[39] Rahimi, A., Recht, B., et al. Random features for large-scale kernel machines. In *NIPS* (2007), NIPS, pp. 1–10.

[40] Rauhut, H., and Ward, R. Sparse Legendre expansions via l1-minimization. *Journal of approximation theory 164*, 5 (2012), 517–533.

[41] Rauhut, H., and Ward, R. Interpolation via weighted l1 minimization. *Applied and Computational Harmonic Analysis 40*, 2 (2016), 321–351.

[42] Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

[43] Schaeffer, H., Tran, G., and Ward, R. Extracting sparse high-dimensional dynamics from limited data. *SIAM Journal on Applied Mathematics 78*, 6 (2018), 3279–3295.

[44] Schaeffer, H., Tran, G., Ward, R., and Zhang, L. Extracting structured dynamical systems using sparse optimization with very few samples. *Multiscale Modeling & Simulation 18*, 4 (2020), 1435–1461.

[45] Shalev-Shwartz, S., and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[46] Sriperumbudur, B. K., and Szabó, Z. Optimal rates for random Fourier features. *arXiv preprint arXiv:1506.02155* (2015).

[47] Sutherland, D. J., and Schneider, J. On the error of random Fourier features. *arXiv preprint arXiv:1506.02785* (2015).

[48] Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* (2010).

[49] Yang, T., Li, Y.-F., Mahdavi, M., Jin, R., and Zhou, Z.-H. Nyström method vs random Fourier features: A theoretical and empirical comparison. *Advances in neural information processing systems 25* (2012), 476–484.

[50] Yen, I. E.-H., Lin, T.-W., Lin, S.-D., Ravikumar, P. K. and Dhillon, I. S. Sparse random feature algorithm as coordinate descent in Hilbert space. *Advances in neural information processing systems* (2014), 2456–2464.

A Useful Tools and Definitions

Definition 5 (**Normal Distribution**). *The real random vector $\boldsymbol{\omega} = (w_1, \ldots, w_d) \in \mathbb{R}^d$ is called a centered normal random vector with mean zero and variance σ^2 if it has density function*[1]

$$\rho(\boldsymbol{\omega}) = (2\pi\sigma^2)^{-d/2} \exp\left(-\frac{\|\boldsymbol{\omega}\|^2}{2\sigma^2}\right). \tag{62}$$

Furthermore, it holds that $\mathbb{E}\|\boldsymbol{\omega}\|^2 = d\sigma^2$, and by a standard concentration argument for any $0 < t < 1$

$$\mathbb{P}((1-t)d\sigma^2 \leq \|\boldsymbol{\omega}\|^2 \leq (1+t)d\sigma^2) \geq 1 - 2\exp\left(-\frac{t^2 d}{12}\right) \tag{63}$$

Definition 6 (**Subgaussian Distribution**). *A random variable X is a centered σ^2-subgaussian (denoted by $X \sim \mathrm{SG}(\sigma^2)$) if $\mathbb{E}[X] = 0$ and*

$$\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right), \quad \forall s \in \mathbb{R}. \tag{64}$$

Let $X \sim \mathrm{SG}(\sigma^2)$. Then, for any $t > 0$,

$$\mathbb{P}(|X| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{65}$$

Definition 7 (**Subexponential Distribution**). *A random variable X is a centered subexponential with parameters $\nu, \alpha > 0$ ($X \sim \mathrm{SE}(\nu^2, \alpha)$) if $\mathbb{E}[X] = 0$ and*

$$\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{s^2\nu^2}{2}\right), \qquad \forall |s| < \frac{1}{\alpha}. \tag{66}$$

Next, we overview the basic facts and relationship between subgaussian and subexponential random variables.

Lemma 3 ([24, 48]). *The following statements hold*

1. *If $X \sim \mathrm{SG}(\sigma^2)$, then $X^2 \sim \mathrm{SE}(4\sqrt{2}\sigma^2, 4\sigma^2)$.*
2. *Let $X \sim \mathrm{SE}(\nu^2, \alpha)$. Then, for any $0 < t < \frac{\nu^2}{\alpha}$, it holds that*

$$\mathbb{P}(|X| \geq t) \leq 2\exp\left(-\frac{t^2}{2\nu^2}\right). \tag{67}$$

Lemma 4 (**Khintchine Inequality** [21]). *If $\{\xi_j\}_{j=1}^N$ is a collection of i.i.d. Rademacher random variables, the Khintchine inequality states that for any $q_1, \ldots, q_N \in \mathbb{C}$, and $0 < p \leq 2$,*

$$\left(\mathbb{E}_\xi\left[\left|\sum_{j=1}^N \xi_j q_j\right|^p\right]\right)^{\frac{1}{p}} \leq \sqrt{\sum_{j=1}^N |q_j|^2}. \tag{68}$$

Lemma 5 (**Rademacher Complexity** [21]). *Assume that $\{\mathbf{v}_j\}_{j=1}^M$ is a sequence of independent random vectors in a finite-dimensional vector space V with norm $\|\cdot\|$. Let $F : V \to \mathbb{R}$ be a convex function. Then*

$$\mathbb{E}_{\mathbf{v}} F\left(\sum_{j=1}^M \mathbf{v}_j - \mathbb{E}[\mathbf{v}_j]\right) \leq \mathbb{E}_{\xi,\mathbf{v}} F\left(2\sum_{j=1}^M \xi_j \mathbf{v}_j\right), \tag{69}$$

where $\{\xi_j\}$ is a Rademacher sequence independent of $\mathbf{v}$.

[1] Note that we only consider the homoscedastic case.

Lemma 6 (**Contraction of Rademacher Complexity [5, 28]**). *Let $\{\phi_j\}_{i=1}^N$ be a collection of real-valued functions defined on $\mathbb{R}$ that are L-Lipschitz and satisfy $\phi_j(0) = 0$. Then,*

$$\mathbb{E}_\xi \sup_{x\in\mathcal{X}} \left| \sum_{j=1}^N \xi_j \phi_j(x_j) \right| \leq 2L\mathbb{E}_\xi \sup_{x\in\mathcal{X}} \left| \sum_{j=1}^N \xi_j x_j \right| \tag{70}$$

for any bounded subset $\mathcal{X}$ in $\mathbb{R}^N$. Here, $\{\xi_j\}$ is a Rademacher sequence independent of $\{x_j\}$.

Lemma 7 (**Stability of BP-based Sparse Reconstruction [21]**). *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ be a matrix with coherence $\mu_\mathbf{A}$. If the coherence of A satisfies*

$$\mu_\mathbf{A} \leq \frac{4}{\sqrt{41}(2s-1)}, \tag{71}$$

then, for any vector $\mathbf{c}^\star \in \mathbb{C}^N$ satisfying $\mathbf{y} = \mathbf{A}\mathbf{c}^\star + \mathbf{e}$ with $\|\mathbf{e}\| \leq \eta\sqrt{m}$, a minimizer $\mathbf{c}^\sharp$ of the BP method (4) approximates the vector $\mathbf{c}^\star$ with the error bounds

$$\|\mathbf{c}^\star - \mathbf{c}^\sharp\|_1 \leq C'\kappa_{s,1}(\mathbf{c}^\star) + C\sqrt{s}\eta, \tag{72}$$

where $C, C' > 0$ are constants.

B Proof of Lemma 1

Lemma 1. *Fix confidence parameter $\delta > 0$ and accuracy parameter $\epsilon > 0$. Recall the setting of Algorithm 1 and suppose $f \in \mathcal{F}(\phi, \rho)$ where $\phi(\mathbf{x};\boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$ and ρ is the probability density function (with finite second moment) used for sampling the random weights $\boldsymbol{\omega}$. Consider a set $\mathcal{X} \subset \mathbb{R}^d$ with diameter $R = \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|$. Suppose*

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4R\sqrt{\mathbb{E}\|\boldsymbol{\omega}\|^2} + \sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2. \tag{73}$$

Consider the random feature approximation

$$f^\star(\mathbf{x}) := \sum_{j=1}^N c_j^\star \exp(i\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle), \quad \text{where} \quad c_j^\star := \frac{\alpha(\boldsymbol{\omega}_j)}{N\rho(\boldsymbol{\omega}_j)}. \tag{74}$$

Then, with probability at least $1-\delta$ with respect to the draw of the $\boldsymbol{\omega}$'s, the following holds

$$\sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x}) - f^\star(\mathbf{x})| \leq \epsilon\|f\|_\rho. \tag{75}$$

Proof. First, by construction we have $|c_j^\star| \leq \frac{\|f\|_\rho}{N}$ and, for fixed $\mathbf{x}$, $\mathbb{E}_{\boldsymbol{\omega}}[f^\star(\mathbf{x})] = f(\mathbf{x})$. Define the random variable

$$v(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N) := \|f - f^\star\|_{L^\infty(\mathcal{X})} = \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x}) - f^\star(\mathbf{x})| = \sup_{\mathbf{x}\in\mathcal{X}} \left|\mathbb{E}_\omega\left[f^\star(\mathbf{x})\right] - f^\star(\mathbf{x})\right|.$$

Following [37], we prove the lemma using McDiarmid's inequality.

First, observe that v is stable under perturbation of any one of its coordinates. Specifically, we have that

$$\begin{aligned} |v(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_k, \ldots, \boldsymbol{\omega}_N) - v(\boldsymbol{\omega}_1, \ldots, \tilde{\boldsymbol{\omega}}_k, \ldots, \boldsymbol{\omega}_N)| &\leq \frac{1}{N}\sup_{\mathbf{x}\in X}\left|\frac{\alpha(\boldsymbol{\omega}_k)}{\rho(\boldsymbol{\omega}_k)}\phi(\mathbf{x};\boldsymbol{\omega}_k) - \frac{\alpha(\tilde{\boldsymbol{\omega}}_k)}{\rho(\tilde{\boldsymbol{\omega}}_k)}\phi(\mathbf{x};\tilde{\boldsymbol{\omega}}_k)\right| \\ &\leq \frac{2\|f\|_\rho}{N} =: \Delta_v, \end{aligned} \tag{76}$$

where we used the triangle inequality for the $\|\cdot\|_{L^\infty}$ norm, and uniform bounded on ϕ.

We would like to apply McDiarmid's concentration inequality: $\mathbb{P}(v \geq \mathbb{E}[v]+t) \leq \exp(-\frac{2t^2}{N\Delta_v^2})$, which requires us to estimate the expectation of v. To do this, following [37, 38] we exploit properties of Rademacher random variables [5]. Using the triangle inequality and Lemma 5 yields

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\omega}}[v] &\leq 2\mathbb{E}_{\boldsymbol{\omega},\xi} \sup_{\mathbf{x}\in\mathcal{X}} \left| \sum_{j=1}^{N} \xi_j c_j^\star \phi(\mathbf{x};\boldsymbol{\omega}_j) \right| \\
&= 2\mathbb{E}_{\boldsymbol{\omega},\xi} \sup_{\mathbf{x}\in\mathcal{X}} \left| \sum_{j=1}^{N} \xi_j c_j^\star \left(\phi(\mathbf{x};\boldsymbol{\omega}_j) - \phi(0) + 1\right) \right| \\
&\leq 2\mathbb{E}_{\boldsymbol{\omega},\xi} \sup_{\mathbf{x}\in\mathcal{X}} \left| \sum_{j=1}^{N} \xi_j c_j^\star \left(\phi(\mathbf{x};\boldsymbol{\omega}_j) - \phi(0)\right) \right| + 2\mathbb{E}_{\xi} \left| \sum_{j=1}^{N} \xi_j c_j^\star \right|,
\end{aligned} \tag{77}$$

noting that $\phi(0) = 1$. The second term above can be bounded using Lemma 4 and recalling that $|c_j^\star| \leq \|f\|_\rho/N$:

$$2\mathbb{E}_{\xi} \left| \sum_{j=1}^{N} \xi_j c_j^\star \right| \leq 2\sqrt{\sum_{j=1}^{N} |c_j^\star|^2} \leq \frac{2\|f\|_\rho}{\sqrt{N}}. \tag{78}$$

We now bound the first term. Let $c_j^\star := |c_j^\star| \exp(i\theta_j)$ and note that by Euler's formula

$$c_j^\star \phi(\mathbf{x};\boldsymbol{\omega}_j) = |c_j^\star| \cos(\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle + \theta_j) + i|c_j^\star| \sin(\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle + \theta_j).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\omega},\xi} \sup_{\mathbf{x}\in\mathcal{X}} \left| \sum_{j=1}^{N} \xi_j c_j^\star \left(\phi(\mathbf{x};\boldsymbol{\omega}_j) - \phi(0)\right) \right| \\
&\leq \mathbb{E}_{\boldsymbol{\omega},\xi} \sup_{\mathbf{x}\in\mathcal{X}} \left| \sum_{j=1}^{N} \xi_j |c_j^\star| (\cos(\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle + \theta_j) - \cos(\theta_j) + i\sin(\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle + \theta_j) - i\sin(\theta_j)) \right| \\
&\leq \mathbb{E}_{\boldsymbol{\omega},\xi} \sup_{\mathbf{x}\in\mathcal{X}} \left| \sum_{j=1}^{N} \xi_j |c_j^\star| \left(\cos(\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle + \theta_j) - \cos(\theta_j)\right) \right| \\
&\quad + \mathbb{E}_{\boldsymbol{\omega},\xi} \sup_{\mathbf{x}\in\mathcal{X}} \left| \sum_{j=1}^{N} \xi_j |c_j^\star| \left(\sin(\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle + \theta_j) - \sin(\theta_j)\right) \right|
\end{aligned} \tag{79}$$

The functions $|c_j^\star| \left(\cos(\cdot + \theta_j) - \cos(\theta_j)\right)$ and $|c_j^\star| \left(\sin(\cdot + \theta_j) - \sin(\theta_j)\right)$ are $\|f\|_\rho/N$-Lipschitz and are zero at zero. Thus, using Lemma 6 we may write,

$$\begin{aligned}
2\mathbb{E}_{\boldsymbol{\omega},\xi} \sup_{\mathbf{x}\in\mathcal{X}} \left| \sum_{j=1}^{N} \xi_j c_j^\star (\phi(\mathbf{x};\boldsymbol{\omega}_j) - \phi(0)) \right| &\leq \frac{8\|f\|_\rho}{N} \mathbb{E}_{\boldsymbol{\omega},\xi} \sup_{\mathbf{x}\in\mathcal{X}} \left| \sum_{j=1}^{N} \xi_j \langle \mathbf{x}, \boldsymbol{\omega}_j\rangle \right| \\
&\leq \frac{8\|f\|_\rho}{N} \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\| \; \mathbb{E}_{\boldsymbol{\omega},\xi} \left\| \sum_{j=1}^{N} \xi_j \boldsymbol{\omega}_j \right\| \\
&\leq \frac{8\|f\|_\rho R}{N} \; \mathbb{E}_{\boldsymbol{\omega},\xi} \left\| \sum_{j=1}^{N} \xi_j \boldsymbol{\omega}_j \right\|
\end{aligned} \tag{80}$$

where we used the Cauchy-Schwartz inequality to establish the second inequality. Next, note that by Jensen's inequality and Lemma 4

$$\mathbb{E}_{\boldsymbol{\omega},\xi}\left\|\sum_{j=1}^{N}\xi_j\boldsymbol{\omega}_j\right\| \leq \sqrt{\mathbb{E}_{\boldsymbol{\omega},\xi}\left\|\sum_{j=1}^{N}\xi_j\boldsymbol{\omega}_j\right\|^2} \leq \sqrt{N\mathbb{E}_{\boldsymbol{\omega}}\|\boldsymbol{\omega}\|^2}. \tag{81}$$

Altogether, we have the bound

$$\mathbb{E}[v] \leq \frac{2\|f\|_\rho(1+4R\sqrt{\mathbb{E}\|\boldsymbol{\omega}\|^2})}{\sqrt{N}} =: M_v. \tag{82}$$

We are now in a position to apply McDiarmid's concentration inequality to obtain

$$\mathbb{P}(v \geq M_v + t) \leq \mathbb{P}(v \geq \mathbb{E}[v] + t) \leq \exp\left(-\frac{2t^2}{N\Delta_v^2}\right). \tag{83}$$

By setting $M_v + t \leq \epsilon\|f\|_\rho$ and the probability bound on the right hand side of the equation above to δ, we solve for N and t to obtain

$$t = \|f\|_\rho\sqrt{\frac{2}{N}\log\left(\frac{1}{\delta}\right)}, \tag{84}$$

and

$$N \geq \frac{4}{\epsilon^2}\left(1+4R\sqrt{\mathbb{E}\|\boldsymbol{\omega}\|^2}+\sqrt{\frac{1}{2}\log(1/\delta)}\right)^2 \tag{85}$$

which completes the proof. ■

C Proof of Lemma 2

Lemma 2. *Let $f \in \mathcal{F}(\phi,\rho)$, where $\phi(\mathbf{x};\boldsymbol{\omega}) = \exp(i\langle\mathbf{x},\boldsymbol{\omega}\rangle)$ and $\rho(\boldsymbol{\omega})$ is the density corresponding to a spherical Gaussian with variance σ^2, $\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)$. For a fixed γ and q with $q \leq d$, consider a set of data samples $\mathbf{x}_1,\ldots,\mathbf{x}_m \sim \mathcal{N}(\mathbf{0},\gamma^2\mathbf{I}_d)$ and a complete set of q-sparse feature weights $\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_N$ drawn from $\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_q)$. Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ denote the associated random feature matrix where $a_{k,j} = \phi(\mathbf{x}_k;\boldsymbol{\omega}_j)$ and $f^\sharp$ be define from Algorithm 1 and Equation (4) with $\eta = \sqrt{2(\epsilon^2\|f\|_\rho^2 + E^2)}$. Let $f^\star$ be defined as*

$$f^\star(\mathbf{x}) := \sum_{j=1}^{N} c_j^\star \exp(i\langle\mathbf{x},\boldsymbol{\omega}_j\rangle), \quad \text{where} \quad c_j^\star := \frac{\alpha(\boldsymbol{\omega}_j)}{N\rho(\boldsymbol{\omega}_j)}. \tag{86}$$

For a given s, if the feature parameters σ and N, the confidence δ, the accuracy ϵ, and the target radius R are chosen so that the following conditions hold:

$$\gamma^2\sigma^2 \geq \frac{1}{2}\left(\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{2}{q}} - 1\right),$$

$$N \geq \frac{4}{\epsilon^2}\left(1+4R\sigma\sqrt{q}+\sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2,$$

$$m \geq 4(2\gamma^2\sigma^2+1)^{\max\{2q-d,0\}}(\gamma^2\sigma^2+1)^{\min\{2q,2d-2q\}}\log\frac{N^2}{\delta}$$

$$R \geq \gamma\sqrt{d+\sqrt{12d\log\left(\frac{m}{\delta}\right)}}.$$

Then, with probability at least $1-3\delta$ *the following error bound holds:*

$$\sup_{\mathbf{x}:\|x\|\leq R} |f^\sharp(\mathbf{x}) - f^\star(\mathbf{x})| \leq C' \kappa_{s,1}(\mathbf{c}^\star) + C\eta\sqrt{s}, \tag{87}$$

where $C, C' > 0$ *are constants.*

The proof of this lemma relies on two intermediate results which we outline next.

Lemma 8 (**Samples Lie in the Domain**). *Suppose that* $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathcal{N}(0, \gamma^2 \mathbf{I}_d)$ *are i.i.d. Gaussian points and thus* $\mathbb{E}\|\mathbf{x}\|^2 = d\gamma^2$. *Let* $R > 0$ *be a fixed radius. Then, for any* $0 < \delta < 1$, *the probability of all* m *samples* $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{B}^d(R)$ *is at least* $1-\delta$ *provided that:*

$$R \geq \gamma \sqrt{d + \sqrt{12d \log\left(\frac{m}{\delta}\right)}}. \tag{88}$$

Proof. Straightforward using the concentration result in Definition 5 and the union bound. ■

Lemma 9 (**Coherence Analysis**). *Consider a complete set of* q*-sparse feature weights in* $\mathbb{R}^d$, $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N$ *drawn from* $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_q)$ *and a set of data samples* $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathcal{N}(\mathbf{0}, \gamma^2 \mathbf{I}_d)$. *Define the random features* $\phi(\mathbf{x}; \boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$ *and let* $\mathbf{A} \in \mathbb{C}^{m \times N}$ *denote the associated random feature matrix where* $a_{k,j} = \phi(\mathbf{x}_k; \boldsymbol{\omega}_j)$. *For a fixed* $0 < \delta < 1$ *and for some integer* $s \geq 1$, *suppose*

$$m \geq 4(2\gamma^2\sigma^2 + 1)^{\max\{2q-d,0\}} (\gamma^2\sigma^2 + 1)^{\min\{2q, 2d-2q\}} \log \frac{N^2}{\delta} \tag{89}$$

$$\gamma^2\sigma^2 \geq \frac{1}{2}\left(\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{2}{q}} - 1\right), \tag{90}$$

then we have with probability at least $1-\delta$, *that the coherence of* $\mathbf{A}$ *is bounded by*

$$\mu_{\mathbf{A}} \leq \frac{4}{\sqrt{41}(2s-1)}. \tag{91}$$

Proof. Let $\mathbf{a}_j, \mathbf{a}_\ell$ denote two columns of $\mathbf{A}$. Let $\mathcal{S}_j, \mathcal{S}_\ell$ denote the support sets of $\boldsymbol{\omega}_j, \boldsymbol{\omega}_\ell$, respectively, and let $\mathcal{G} = \mathcal{S}_j \cap \mathcal{S}_\ell$. Then using the characteristic function of the Gaussian distribution

$$\begin{aligned}
\mathbb{E}[\langle \mathbf{a}_j, \mathbf{a}_\ell\rangle \mid \mathcal{S}_j, \mathcal{S}_\ell] &= \mathbb{E}_{\boldsymbol{\omega}_j, \boldsymbol{\omega}_\ell}\left[\mathbb{E}_{\mathbf{x}_k}\left[\sum_{j=1}^m \exp(i\langle \boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell, \mathbf{x}_k\rangle) \mid \boldsymbol{\omega}_j, \boldsymbol{\omega}_\ell\right] \mid \mathcal{S}_j, \mathcal{S}_\ell\right] \\
&= \mathbb{E}_{\boldsymbol{\omega}_j, \boldsymbol{\omega}_\ell}\left[m \exp\left(-\frac{\gamma^2}{2}\|\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell\|^2\right) \mid \mathcal{S}_j, \mathcal{S}_\ell\right] \\
&= m\mathbb{E}_{\boldsymbol{\omega}_j, \boldsymbol{\omega}_\ell}\left[\exp\left(-\frac{\gamma^2}{2}(\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell)_1^2\right) \ldots \exp\left(-\frac{\gamma^2}{2}(\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell)_d^2\right) \mid \mathcal{S}_j, \mathcal{S}_\ell\right] \\
&= m\left(\frac{1}{\sqrt{2\gamma^2\sigma^2 + 1}}\right)^{|\mathcal{G}|}\left(\frac{1}{\sqrt{\gamma^2\sigma^2 + 1}}\right)^{2q-2|\mathcal{G}|} =: m\Gamma_{j\ell},
\end{aligned} \tag{92}$$

where $0 \leq \max\{2q-d, 0\} \leq |\mathcal{G}| \leq q \leq d$. Assuming that $q < \frac{d}{2}$ we have the following bound

$$\left(\frac{1}{\sqrt{\gamma^2\sigma^2 + 1}}\right)^{2q} =: \Gamma_{\min} \leq \Gamma_{j,\ell} \leq \left(\frac{1}{\sqrt{2\gamma^2\sigma^2 + 1}}\right)^{q}, \tag{93}$$

using the inequality $2\gamma^2\sigma^2+1 \leq \left(\gamma^2\sigma^2+1\right)^2$. Given that $\mathbf{x}_k$'s are i.i.d., applying the Bernstein's inequality and recalling that $\mu_{j\ell} = \frac{\langle \mathbf{a}_j, \mathbf{a}_\ell \rangle}{m}$, yields

$$\mathbb{P}(|\mu_{j\ell} - \Gamma_{j\ell}| \geq \Gamma_{j\ell} \mid \mathcal{S}_j, \mathcal{S}_\ell) \leq 2\exp\left(-\frac{\frac{1}{2}m^2{\Gamma_{j\ell}}^2}{m+\frac{2}{3}m\Gamma_{j\ell}}\right). \tag{94}$$

Since $\Gamma_{j\ell} \leq \frac{3}{2}$, the last result simplifies to

$$\mathbb{P}(|\mu_{j\ell} - \Gamma_{j\ell}| \geq \Gamma_{j\ell} \mid \mathcal{S}_j, \mathcal{S}_\ell) \leq 2\exp\left(-\frac{1}{4}m{\Gamma_{j\ell}}^2\right) \leq 2\exp\left(-\frac{1}{4}m\Gamma_{min}^2\right). \tag{95}$$

Taking a union bound over all $\binom{N}{2} \leq \frac{N^2}{2}$ pairs of columns implies that

$$\mathbb{P}(\exists\, k,\, \ell \text{ s.t. } |\mu_{j\ell} - \Gamma_{j\ell}| \geq \Gamma_{j\ell}) \leq N^2\exp\left(-\frac{1}{4}m\Gamma_{min}^2\right) \leq \delta. \tag{96}$$

Therefore, if

$$m \geq \frac{4}{\Gamma_{min}^2}\log\frac{N^2}{\delta} \tag{97}$$

then with probability at least $1-\delta$,

$$\mu_{\mathbf{A}} \leq 2\max_{k,\ell}\Gamma_{j\ell}.$$

For stable recovery, we enforce that

$$\mu_{\mathbf{A}} \leq 2\max_{k,\ell}\Gamma_{j\ell} \leq 2\left(\frac{1}{\sqrt{2\gamma^2\sigma^2+1}}\right)^q \leq \frac{4}{\sqrt{41}(2s-1)}. \tag{98}$$

This implies the following uncertainty principle

$$\gamma^2\sigma^2 \geq \frac{1}{2}\left(\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{2}{q}} - 1\right). \tag{99}$$

To establish a suitable lower bound for m, we impose

$$\frac{4}{\Gamma_{min}^2}\log\frac{N^2}{\delta} = 4(\gamma^2\sigma^2+1)^{2q}\log\frac{N^2}{\delta} \leq m. \tag{100}$$

When $q \geq d/2$, we have that $|\mathcal{G}|$ ranges between $2q-d$ and q. Then analogously

$$\left(\frac{1}{\sqrt{2\gamma^2\sigma^2+1}}\right)^{2q-d}\left(\frac{1}{\sqrt{\gamma^2\sigma^2+1}}\right)^{2d-2q} =: \Gamma_{\min} \leq \Gamma_{j,\ell} \leq \left(\frac{1}{\sqrt{2\gamma^2\sigma^2+1}}\right)^q$$

which implies the uncertainty principle

$$\gamma^2\sigma^2 \geq \frac{1}{2}\left(\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{2}{q}} - 1\right), \tag{101}$$

and the lower bound for the number of samples

$$m \geq 4(2\gamma^2\sigma^2+1)^{2q-d}(\gamma^2\sigma^2+1)^{2d-2q}\log\frac{N^2}{\delta}. \tag{102}$$

This concludes the proof of Lemma 9. ■

Proof of Lemma 2. Fix a radius R. Then, by Hölder's inequality and the fact that $\|\phi(\cdot;\boldsymbol{\omega})\|_\infty \leq 1$, we have

$$\sup_{\|x\|_2 \leq R} |f^\sharp(\mathbf{x}) - f^\star(\mathbf{x})| \leq \|\mathbf{c}^\sharp - \mathbf{c}^\star\|_1. \tag{103}$$

The error between $\mathbf{y}$ and $\mathbf{A}\mathbf{c}^\star$ is bounded by

$$\begin{aligned}
\|\mathbf{y} - \mathbf{A}\mathbf{c}^\star\|^2 &= \sum_{k=1}^m (f(\mathbf{x}_k) - f^\star(\mathbf{x}_k) + e_k)^2 \\
&\leq 2\left(\sum_{k=1}^m (f(\mathbf{x}_k) - f^\star(\mathbf{x}_k))^2 + \sum_{k=1}^m e_k^2\right) \\
&\leq 2m\left(\sup_{\|\mathbf{x}\|_2 \leq R} |f(\mathbf{x}) - f^\star(\mathbf{x})|^2 + E^2\right) \\
&\leq 2m(\epsilon^2 \|f\|_\rho^2 + E^2) =: m\eta^2
\end{aligned} \tag{104}$$

where by Lemma 8 the second to last inequality holds with probability exceeding $1-\delta$ provided that

$$R \geq \gamma \sqrt{d + \sqrt{12d \log\left(\frac{m}{\delta}\right)}},$$

and the final inequality holds with probability exceeding $1-\delta$ if

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4R\sqrt{\mathbb{E}\|\boldsymbol{\omega}\|^2} + \sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2.$$

The final inequality above holds with probability at least $1-2\delta$. Therefore, if the number of measurements m is between

$$\delta \exp\left(\frac{d}{12}\left(\frac{R^2}{\gamma^2 d} - 1\right)^2\right) \geq m \geq 4(2\gamma^2\sigma^2+1)^{\max\{2q-d,0\}}(\gamma^2\sigma^2+1)^{\min\{2q,2d-2q\}} \log\frac{N^2}{\delta} \tag{105}$$

and the other stated conditions hold, then with probability at least $1-3\delta$, the coherence of $\mathbf{A}$ is bounded by

$$\mu_{\mathbf{A}} \leq \frac{4}{\sqrt{41}(2s-1)}. \tag{106}$$

Thus by Lemma 7,

$$\sup_{\|\mathbf{x}\|_2 \leq R} |f^\sharp(\mathbf{x}) - f^\star(\mathbf{x})| \leq \|\mathbf{c}^\sharp - \mathbf{c}^\star\|_1 \leq C'\kappa_{s,1}(\mathbf{c}^\star) + C\sqrt{s}\eta, \tag{107}$$

which concludes the proof of Lemma 2. ∎

D Proofs for the Generalization Bounds

To bound the generalization error, we extend Lemma 1 and Lemma 2. We cannot directly bound the generalization error with the sup-norm results when the probability measure μ does not have compact support. This is circumvented by introducing a $\log^{1/2}\left(\frac{m}{\delta}\right)$ term to the complexity bound for N.

Lemma 10 (Generalization Error, Term 1). *Fix the confidence parameter $\delta > 0$ and accuracy parameter $\epsilon > 0$. Recall the setting of Algorithm 1 and suppose $f \in \mathcal{F}(\phi, \rho)$ where*

$\phi(\mathbf{x};\boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$. *The data samples* $\mathbf{x}_k$ *have probability measure* $\mu(\mathbf{x})$ *and weights* $\boldsymbol{\omega}_j$ *are sampled using the probability density* $\rho(\boldsymbol{\omega})$. *Consider the random feature approximation*

$$f^\star(\mathbf{x}) := \sum_{j=1}^{N} c_j^\star \exp(i\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle), \quad \textit{where} \quad c_j^\star := \frac{\alpha(\boldsymbol{\omega}_j)}{N\rho(\boldsymbol{\omega}_j)}. \tag{108}$$

If the number of features N *satisfies the bound*

$$N \geq \frac{1}{\epsilon^2}\left(1 + \sqrt{2\log\left(\frac{1}{\delta}\right)}\right)^2, \tag{109}$$

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

$$\sqrt{\int_{\mathbb{R}^d} |f(\mathbf{x}) - f^\star(\mathbf{x})|^2 \,\mathrm{d}\mu} \leq \epsilon \|f\|_\rho. \tag{110}$$

Proof. The proof follows similar arguments to those found in [37,38]. The coefficients are bounded by $|c_j^\star| \leq \dfrac{\|f\|_\rho}{N}$ and, for fixed $\mathbf{x}$, $\mathbb{E}_{\boldsymbol{\omega}}[f^\star(\mathbf{x})] = f(\mathbf{x})$. Define the random variable

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

To apply McDiarmid's inequality, 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 $(\boldsymbol{\omega}_1, \ldots, \tilde{\boldsymbol{\omega}}_k, \ldots, \boldsymbol{\omega}_N)$, then

$$\begin{aligned}
|v(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_k, \ldots, \boldsymbol{\omega}_N) - v(\boldsymbol{\omega}_1, \ldots, \tilde{\boldsymbol{\omega}}_k, \ldots, \boldsymbol{\omega}_N)| &\leq \left\|f^\star - \tilde{f}^\star\right\|_{L^2(d\mu)} \\
&= \|c_k^\star \exp(i\langle \cdot, \boldsymbol{\omega}_k\rangle) - \tilde{c}_k^\star \exp(i\langle \cdot, \tilde{\boldsymbol{\omega}}_k\rangle)\|_{L^2(d\mu)} \\
&= \frac{1}{N}\left\|\frac{\alpha(\boldsymbol{\omega}_k)}{\rho(\boldsymbol{\omega}_k)} \exp(i\langle \cdot, \boldsymbol{\omega}_k\rangle) - \frac{\alpha(\tilde{\boldsymbol{\omega}}_k)}{\rho(\tilde{\boldsymbol{\omega}}_k)} \exp(i\langle \cdot, \tilde{\boldsymbol{\omega}}_k\rangle)\right\|_{L^2(d\mu)} \\
&\leq \frac{1}{N} \sup_{\mathbf{x}} \left|\frac{\alpha(\boldsymbol{\omega}_k)}{\rho(\boldsymbol{\omega}_k)} \exp(i\langle \mathbf{x}, \boldsymbol{\omega}_k\rangle) - \frac{\alpha(\tilde{\boldsymbol{\omega}}_k)}{\rho(\tilde{\boldsymbol{\omega}}_k)} \exp(i\langle \mathbf{x}, \tilde{\boldsymbol{\omega}}_k\rangle)\right| \\
&\leq \frac{2\|f\|_\rho}{N} =: \Delta_v,
\end{aligned} \tag{111}$$

where we used the triangle inequality for $\|\cdot\|_{L^2(d\mu)}$ in the first line, Hölder's inequality in the fourth line, and the uniform bound $|\exp(i\langle \mathbf{x}, \boldsymbol{\omega}\rangle)| = 1$ in the fifth line.

To estimate the expectation of v, we bound the expectation of the second moment [38]. 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

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\omega}}[v^2] &= \mathbb{E}_{\boldsymbol{\omega}}\|\mathbb{E}_{\boldsymbol{\omega}}\left[f^\star\right] - f^\star\|^2_{L^2(\mathrm{d}\mu)} \\
&= \frac{1}{N}\left(\mathbb{E}_{\boldsymbol{\omega}}\left\|\frac{\alpha(\boldsymbol{\omega})}{\rho(\boldsymbol{\omega})}\exp(i\langle \cdot, \boldsymbol{\omega}\rangle)\right\|^2_{L^2(\mathrm{d}\mu)} - \left\|\mathbb{E}_{\boldsymbol{\omega}}\left[\frac{\alpha(\boldsymbol{\omega})}{\rho(\boldsymbol{\omega})}\exp(i\langle \cdot, \boldsymbol{\omega}\rangle)\right]\right\|^2_{L^2(\mathrm{d}\mu)}\right) \\
&\leq \frac{1}{N}\,\mathbb{E}_{\boldsymbol{\omega}}\left|\frac{\alpha(\boldsymbol{\omega})}{\rho(\boldsymbol{\omega})}\right|^2 \\
&\leq \frac{\|f\|^2_\rho}{N}.
\end{aligned} \tag{112}$$

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

$$\mathbb{E}_{\boldsymbol{\omega}}[v] \leq \left(\mathbb{E}_{\boldsymbol{\omega}}[v^2]\right)^{\frac{1}{2}} \leq \frac{\|f\|_\rho}{\sqrt{N}}. \tag{113}$$

Applying McDiarmid's concentration inequality yields

$$\mathbb{P}\left(v \geq \frac{\|f\|_\rho}{\sqrt{N}} + t\right) \leq \mathbb{P}(v \geq \mathbb{E}[v] + t) \leq \exp\left(-\frac{2t^2}{N\Delta_v^2}\right). \tag{114}$$

Setting t and N to

$$t = \|f\|_\rho \sqrt{\frac{2}{N}\log\left(\frac{1}{\delta}\right)}, \tag{115}$$

and

$$N \geq \frac{1}{\epsilon^2}\left(1 + \sqrt{2\log\left(\frac{1}{\delta}\right)}\right)^2 \tag{116}$$

enforces that $M_v + t \leq \epsilon\|f\|_\rho$ and that the probability of failure is less than δ. This completes the proof. ■

Lemma 11 (Generalization Error, Term 2). *Let $f \in \mathcal{F}(\phi, \rho)$, where $\phi(\mathbf{x}; \boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$. For a fixed γ and q, consider a set of data samples $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathcal{N}(\mathbf{0}, \gamma^2 \mathbf{I}_d)$ with $\mu(\mathbf{x})$ denoting the associated probability measure and a complete set of N q-sparse feature weights in $\mathbb{R}^d$, $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N$ drawn from $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_q)$. Assume that the noise is bounded by $E = \sqrt{2}\nu$ or that the noise terms e_j are drawn i.i.d. from $\mathcal{N}(0, \nu^2)$. Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ denote the associated random feature matrix where $a_{k,j} = \phi(\mathbf{x}_k; \boldsymbol{\omega}_j)$ and $f^\sharp$ be define from Algorithm 1 and Equation (4) with $\eta = 2\sqrt{\epsilon^2\|f\|_\rho^2 + 2\nu^2}$. Let the random feature approximation $f^\star$ be defined as*

$$f^\star(\mathbf{x}) := \sum_{j=1}^N c_j^\star \exp(i\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle), \quad \text{where} \quad c_j^\star := \frac{\alpha(\boldsymbol{\omega}_j)}{N\rho(\boldsymbol{\omega}_j)}. \tag{117}$$

For a given s, if the feature parameters σ and N, the confidence δ, and the accuracy ϵ are chosen so that the following conditions hold:

$$\gamma^2\sigma^2 \geq \frac{1}{2}\left(\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{2}{q}} - 1\right),$$

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4\gamma\sigma d\sqrt{1 + \sqrt{\frac{12}{d}\log\frac{m}{\delta}}} + \sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2$$

$$m \geq 4(2\gamma^2\sigma^2 + 1)^{\max\{2q-d,0\}}(\gamma^2\sigma^2 + 1)^{\min\{2q,2d-2q\}}\log\frac{N^2}{\delta}.$$

Then, with probability at least $1 - 3\delta$ the following error bound holds:

$$\sqrt{\int_{\mathbb{R}^d} |f^\#(\mathbf{x}) - f^\star(\mathbf{x})|^2 \,\mathrm{d}\mu} \leq C'\,\kappa_{s,1}(\mathbf{c}^\star) + C\eta\sqrt{s}, \tag{118}$$

where $C, C' > 0$ are constants.

Proof of Lemma 2. Using Hölder's twice and $\|\phi(\cdot; \boldsymbol{\omega})\|_{L^\infty(\mathbb{R}^d)} \leq 1$, we have

$$\sqrt{\int_{\mathbb{R}^d} |f^\#(\mathbf{x}) - f^\star(\mathbf{x})|^2 \,\mathrm{d}\mu} \leq \sup_{\mathbf{x}\in\mathbb{R}^d} |f^\sharp(\mathbf{x}) - f^\star(\mathbf{x})| \leq \|\mathbf{c}^\sharp - \mathbf{c}^\star\|_1. \tag{119}$$

To apply the sparse recovery results, we must bound the difference between $\mathbf{y}$ and $\mathbf{A}\mathbf{c}^\star$, in particular,

$$\begin{aligned}\|\mathbf{y}-\mathbf{A}\mathbf{c}^\star\|_2^2 &= \sum_{j=1}^m (f(\mathbf{x}_j)-f^\star(\mathbf{x}_j)+e_j)^2 \\ &\leq 2\left(\sum_{j=1}^m (f(\mathbf{x}_j)-f^\star(\mathbf{x}_j))^2+\sum_{j=1}^m e_j^2\right) \\ &= 2\left(\sum_{j=1}^m (f(\mathbf{x}_j)-f^\star(\mathbf{x}_j))^2+\|\mathbf{e}\|_2^2\right).\end{aligned} \tag{120}$$

where $\mathbf{e} := [e_1,\ldots,e_m]^T \in \mathbb{R}^m$ is the noise vector. The assumption is that either $\|\mathbf{e}\|_2^2 \leq 4\nu^2 m$ or that $\mathbf{e}$ is a random vector with i.i.d. elements drawn from $\mathcal{N}(0,\nu^2)$. In the second case, with probability at least $1-\delta$, the norm is bounded by $\|\mathbf{e}\|_2^2 \leq 4\nu^2 m$, as long as $m \geq 2\log\left(\frac{1}{\delta}\right)$ which always holds by assumption.

Following the proof of Lemma 2, by setting $R(d,m,\delta,\gamma) := \gamma\sqrt{d+d\sqrt{\frac{12}{d}\log\frac{m}{\delta}}}$, we have:

$$\frac{1}{m}\sum_{j=1}^m (f(\mathbf{x}_j)-f^\star(\mathbf{x}_j))^2 \leq \sup_{\|\mathbf{x}\|_2\leq R(d,m,\delta,\gamma)} |f(\mathbf{x})-f^\star(\mathbf{x})|^2 \leq \epsilon^2\|f\|_\rho^2. \tag{121}$$

Therefore, if the number of features satisfies

$$N \geq \frac{4}{\epsilon^2}\left(1+4\gamma\sigma d\sqrt{1+\sqrt{\frac{12}{d}\log\frac{m}{\delta}}}+\sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2$$

then with probability exceeding $1-2\delta$, $\eta^2 = 2(\epsilon^2\|f\|_\rho^2+2\nu^2)$ holds. Note that the bound in Equation (121) is taken over the ball of radius $R(d,m,\delta,\gamma)$ since this is the domain in $\mathbb{R}^d$ which (with high probability) contains the samples $\{\mathbf{x}_j\}_{j=1}^m$. It is important to note that this bound is needed for determining the stability parameter η, which directly controls the errors through the sparse recovery bounds.

We can apply Lemma 9, since the stated conditions hold, thus with probability at least $1-3\delta$, the coherence of $\mathbf{A}$ is bounded by

$$\mu_{\mathbf{A}} \leq \frac{4}{\sqrt{41}(2s-1)}. \tag{122}$$

The upper bound for m that appears in Lemma 9 holds directly by the choice of $R = R(d,m,\delta,\gamma)$. By Lemma 7, the sup-norm is controlled by

$$\sup_{\mathbf{x}} |f^\sharp(\mathbf{x})-f^\star(\mathbf{x})| \leq \|\mathbf{c}^\sharp-\mathbf{c}^\star\|_1 \leq C'\kappa_{s,1}(\mathbf{c}^\star)+C\sqrt{s}\eta. \tag{123}$$

It is worth noting that the supremum in Equation (123) is taken over all $\mathbb{R}^d$ since $\|\phi(\cdot;\boldsymbol{\omega})\|_{L^\infty(\mathbb{R}^d)} \leq 1$, therefore,

$$\sqrt{\int_{\mathbb{R}^d} |f^\#(\mathbf{x})-f^\star(\mathbf{x})|^2\,\mathrm{d}\mu} \leq \sup_{\mathbf{x}} |f^\sharp(\mathbf{x})-f^\star(\mathbf{x})| \leq \|\mathbf{c}^\sharp-\mathbf{c}^\star\|_1 \leq C'\kappa_{s,1}(\mathbf{c}^\star)+C\sqrt{s}\eta, \tag{124}$$

which completes the proof. ■

E Extensions Outlined in Section 5

We provide the proof of the results in Section 5.

E.1 Bias Parameter

Consider the case where we use features of the form $\exp(i\langle \mathbf{x}, \boldsymbol{\omega}_j\rangle + ip_j)$, where $\{p_j\}$ are i.i.d. drawn from a symmetric distribution (e.g., Gaussian or uniform) and independent of $\{\boldsymbol{\omega}_j\}$. Note that these features satisfy the conditions of Lemma 1, hence after redefining

$$c_j^\star := \frac{\alpha(\boldsymbol{\omega}_j)\exp\left(-i\,p_j\right)}{N\rho(\boldsymbol{\omega}_j)},$$

we can directly apply this lemma. However, regarding Lemma 2, we need to account for the inclusion of the bias term p_j. To this end, consider the derivation of the separation quantities in (92). We can write

$$\begin{aligned}
\mathbb{E}[\langle \mathbf{a}_j, \mathbf{a}_\ell\rangle \mid \mathcal{S}_j, \mathcal{S}_\ell] &= \mathbb{E}_{\boldsymbol{\omega},p}\left[\mathbb{E}_{\mathbf{x}_k}\left[\sum_{k=1}^m \exp(i\langle \boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell, \mathbf{x}_k\rangle + i(p_j - p_\ell)) \mid \boldsymbol{\omega}_j, \boldsymbol{\omega}_\ell\right] \mid \mathcal{S}_j, \mathcal{S}_\ell\right] \\
&= \mathbb{E}_{p_j,p_\ell}[\exp(i(p_j - p_\ell))]\ \mathbb{E}_{\boldsymbol{\omega}_j,\boldsymbol{\omega}_\ell}\left[m\exp\left(-\frac{\gamma^2}{2}\|\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell\|^2\right) \mid \mathcal{S}_j, \mathcal{S}_\ell\right] \\
&= m\,\psi^2(p)\left(\frac{1}{\sqrt{2\gamma^2\sigma^2+1}}\right)^{|\mathcal{G}|}\left(\frac{1}{\sqrt{\gamma^2\sigma^2+1}}\right)^{2q-2|\mathcal{G}|} =: m\Gamma_{j\ell},
\end{aligned} \tag{125}$$

where ψ is the characteristic function for the probability distribution for p. Therefore, the new separation quantities are proportional to $\psi^2(p)$. For instance, if $p_k \sim \mathcal{N}(0,1)$, then $\psi^2(p) = 1/e$. Therefore, the conditions derived in Lemma 2 hold up to some constants depending on the characteristic function of p.

Note that we can further consider the case where $\mathbf{x}_k \sim \mathcal{N}(\bar{\mathbf{x}}, \gamma^2\mathbf{I}_d)$. Specifically, $\bar{\mathbf{x}}$ can be thought of as a deterministic bias; hence in (92) we have the additional term $\exp(i\bar{\mathbf{x}})$. Since $|\exp(i\bar{\mathbf{x}})| = 1$, we can directly apply Lemma 2 without any adjustments.

E.2 Proof of Theorem 5

Our proofs naturally extend to the case where the activation functions are real trigonometric functions. The integral representation is now a sum of two terms, one cosine and one sine. In this setting, when we draw N random weights $\boldsymbol{\omega}$, then the matrix $\mathbf{A}$ will have $2N$ columns. Therefore, up to constants, we obtain the same complexity bound N as in the case of complex exponential. To extend the proof for the mutual coherence bound on $\mathbf{A}$, we have to consider four cases for the inner products between columns

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^m \cos(\langle \boldsymbol{\omega}_j, \mathbf{x}_k\rangle)\cos(\langle \boldsymbol{\omega}_\ell, \mathbf{x}_k\rangle) \mid S_j, S_\ell\right] = m\left(\frac{1}{\sqrt{2\gamma^2\sigma^2+1}}\right)^{|\mathcal{G}|}\left(\frac{1}{\sqrt{\gamma^2\sigma^2+1}}\right)^{2q-2|\mathcal{G}|}, \\
&\mathbb{E}\left[\sum_{k=1}^m \sin(\langle \boldsymbol{\omega}_j, \mathbf{x}_k\rangle)\sin(\langle \boldsymbol{\omega}_\ell, \mathbf{x}_k\rangle) \mid S_j, S_\ell\right] = 0, \\
&\mathbb{E}\left[\sum_{k=1}^m \sin(\langle \boldsymbol{\omega}_j, \mathbf{x}_k\rangle)\cos(\langle \boldsymbol{\omega}_\ell, \mathbf{x}_k\rangle) \mid S_j, S_\ell\right] = 0, \\
&\mathbb{E}\left[\sum_{k=1}^m \sin(\langle \boldsymbol{\omega}_j, \mathbf{x}_k\rangle)\cos(\langle \boldsymbol{\omega}_j, \mathbf{x}_k\rangle) \mid S_j, S_\ell\right] = 0.
\end{aligned} \tag{126}$$

The norm of each column is given by

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^{m}\cos^2(\langle\boldsymbol{\omega}_j,\mathbf{x}_k\rangle)\right] &= \frac{m}{2}+\frac{m}{2}\left(\frac{1}{\sqrt{4\gamma^2\sigma^2+1}}\right)^q,\\ \mathbb{E}\left[\sum_{k=1}^{m}\sin^2(\langle\boldsymbol{\omega}_j,\mathbf{x}_k\rangle)\right] &= \frac{m}{2}-\frac{m}{2}\left(\frac{1}{\sqrt{4\gamma^2\sigma^2+1}}\right)^q.\end{aligned} \tag{127}$$

The proof follows from a modification of Lemma 9. Note that we have the following bound:

$$0 \leq \Gamma_{j,\ell} \leq 2\left(\frac{1}{\sqrt{2\gamma^2\sigma^2+1}}\right)^q, \tag{128}$$

where the minimum is obtained when at least one column is sine and the maximum is obtained when both columns are cosine. Consider the case when $q < \frac{d}{2}$ and define the following value

$$\Gamma_{\min^+} := \left(\frac{1}{\sqrt{\gamma^2\sigma^2+1}}\right)^{2q},$$

which is the minimum non-zero value of $\Gamma_{k,\ell}$. Then by Bernstein's inequality we have

$$\mathbb{P}(|\mu_{j\ell}-\Gamma_{j\ell}| \geq \Gamma_{\min^+} \mid \mathcal{S}_j,\mathcal{S}_\ell) \leq 2\exp\left(-\frac{\frac{1}{2}m^2\Gamma^2_{\min^+}}{m+\frac{2}{3}m\Gamma_{\min^+}}\right), \tag{129}$$

since the sum of the second moments are bounded by m. Since $\Gamma_{\min^+} \leq \frac{3}{2}$, we have

$$\mathbb{P}(|\mu_{j\ell}-\Gamma_{j\ell}| \geq \Gamma_{\min^+} \mid \mathcal{S}_j,\mathcal{S}_\ell) \leq 2\exp\left(-\frac{1}{4}m\Gamma^2_{\min^+}\right). \tag{130}$$

As before, taking the union bound leads to

$$\mathbb{P}(\exists\, k,\, \ell \text{ s.t. } |\mu_{j\ell}-\Gamma_{j\ell}| \geq \Gamma_{\min^+}) \leq N^2\exp\left(-\frac{1}{4}m\Gamma^2_{\min^+}\right) \leq \delta \tag{131}$$

and thus

$$m \geq \frac{4}{\Gamma^2_{\min^+}}\log\frac{N^2}{\delta}. \tag{132}$$

Therefore, with probability at least $1-\delta$,

$$\mu_{\mathbf{A}} \leq \Gamma_{\min^+}+\Gamma_{j\ell} \leq 2\max_{j,\ell}\Gamma_{j\ell} \leq 4\left(\frac{1}{\sqrt{2\gamma^2\sigma^2+1}}\right)^q.$$

The rest follows from Lemma 9.

E.3 Proof of Theorem 6

Note that we can directly apply Lemma 1 to bound the first source of error in both settings of bounded ρ-norm and low order functions. To study the second error terms, we can obtain analogous bounds on the separation quantities and hence the coherence constant if $\mathbf{x}_1,\ldots,\mathbf{x}_m \sim \mathcal{U}(-\gamma,\gamma)^d$. To this end, we first we need an analogous of Lemma 8 whose proof is a direct consequence of the Hoeffding's inequality and the union bound.

Lemma 12 (Samples Lie in the Domain). *Suppose that* $\mathbf{x}_1, \dots, \mathbf{x}_m \sim \mathcal{U}(-\gamma, \gamma)^d$ *are i.i.d. uniform points and thus* $\mathbb{E}\|\mathbf{x}\|^2 = d\gamma^2/3$. *Let* $0 < R < \gamma\sqrt{d}$ *be a fixed radius. Then, for any* $0 < \delta < 1$, *the probability of all* m *samples* $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{B}^d(R)$ *is at least* $1 - \delta$ *provided that:*

$$m \leq \delta \exp\left(\frac{2d}{9}\left(\frac{3R^2}{\gamma^2 d} - 1\right)^2\right). \tag{133}$$

If $R \geq \gamma\sqrt{d}$, *then* $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{B}^d(R)$ *for any* $m \geq 1$.

Therefore, we replace (92) by

$$\mathbb{E}[\langle \mathbf{a}_j, \mathbf{a}_\ell \rangle \mid \mathcal{S}_j, \mathcal{S}_\ell] = m\left(\frac{\sqrt{\pi}\,\mathrm{erf}(\gamma\sigma)}{2\gamma\sigma}\right)^{|\mathcal{G}|}\left(\frac{\sqrt{\pi}\,\mathrm{erf}\left(\frac{\gamma\sigma}{\sqrt{2}}\right)}{\sqrt{2}\gamma\sigma}\right)^{2q-2|\mathcal{G}|} =: m\Gamma_{j\ell}. \tag{134}$$

Since

$$\frac{\sqrt{\pi}\,\mathrm{erf}(\gamma\sigma)}{2\gamma\sigma} \geq \left(\frac{\sqrt{\pi}\,\mathrm{erf}\left(\frac{\gamma\sigma}{\sqrt{2}}\right)}{\sqrt{2}\gamma\sigma}\right)^2,$$

we can apply the same steps as the proof for normally distributed data in Lemma 9, as long as

$$\begin{aligned}
&m \leq \delta \exp\left(\frac{2d}{9}\left(\frac{3R^2}{\gamma^2 d} - 1\right)^2\right) \quad \text{only needed if} \quad R < \gamma\sqrt{d}, \\
&m \geq 4\left(\frac{2\gamma\sigma}{\sqrt{\pi}\,\mathrm{erf}(\gamma\sigma)}\right)^{2\max\{2q-d,0\}}\left(\frac{\sqrt{2}\gamma\sigma}{\sqrt{\pi}\,\mathrm{erf}\left(\frac{\gamma\sigma}{\sqrt{2}}\right)}\right)^{2\min\{2q,2d-2q\}}\log\frac{N^2}{\delta}, \\
&\frac{\gamma\sigma}{\mathrm{erf}(\gamma\sigma)} \geq \sqrt{\frac{\pi}{4}}\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{1}{q}},
\end{aligned} \tag{135}$$

with probability at least $1 - \delta$, we have that

$$\mu_{\mathbf{A}} \leq \frac{4}{\sqrt{41}(2s-1)}. \tag{136}$$

To simplify the equations, if we consider $\gamma\sigma \geq 2$, then $\gamma\sigma \leq \frac{\gamma\sigma}{\mathrm{erf}(\gamma\sigma)} \leq \sqrt{2}\gamma\sigma$. Therefore, the bounds can be replaced with

$$\begin{aligned}
&m \leq \delta \exp\left(\frac{2d}{9}\left(\frac{3R^2}{\gamma^2 d} - 1\right)^2\right) \quad \text{only needed if} \quad R < \gamma\sqrt{d}, \\
&m \geq 4\left(\frac{2\sqrt{2}\gamma\sigma}{\sqrt{\pi}}\right)^{2\max\{2q-d,0\}}\left(\frac{2\gamma\sigma}{\sqrt{\pi}}\right)^{2\min\{2q,2d-2q\}}\log\frac{N^2}{\delta}, \\
&\gamma\sigma \geq \min\left\{\sqrt{\frac{\pi}{4}}\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{1}{q}}, 2\right\}.
\end{aligned} \tag{137}$$

E.4 Proof of Theorem 7

We will show that we obtain similar results when the non-zero entries of $\boldsymbol{\omega}$ are drawn from a centered subgaussian distribution (see Definition 5). We restrict ourselves to the case where the data comes from a normal distribution. Let $\mathbf{x}_1, \dots, \mathbf{x}_m \sim \mathcal{N}(\mathbf{0}, \gamma^2\mathbf{I}_d)$ and q-sparse $\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_N$ be

from a centered subgaussian distribution with variance σ^2, and construct the matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ so that $a_{kj} = \exp(i\langle \mathbf{x}_k, \boldsymbol{\omega}_j\rangle)$. Note that we can directly apply the result of Lemma 1 to bound the first source of error. We provide bounds on the mutual coherence of $\mathbf{A}$ by bounding the separation quantities.

Lemma 13. *Construct the random feature matrix $\mathbf{A}$ as above. Fix $0 < \delta < 1$ and define*

$$t := \frac{2^{9/4}\sigma\sqrt{q}}{\mathbb{E}[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2]}\sqrt{\log\frac{N^2}{\delta}}. \tag{138}$$

Assume that $t < 1$ and

$$\begin{aligned} \gamma^2\mathbb{E}[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2] &\geq \frac{2}{(1-t)}\log\left(\frac{\sqrt{41}(2s-1)}{2}\right), \\ m &\geq 4\exp\left(2\gamma^2\mathbb{E}[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2]\right)\log\frac{N^2}{\delta}, \\ R &\geq \gamma\sqrt{d + \sqrt{12d\log\left(\frac{m}{\delta}\right)}}. \end{aligned} \tag{139}$$

Then, with probability at least $1 - 2\delta$, we have

$$\mu_{\mathbf{A}} \leq \frac{4}{\sqrt{41}(2s-1)}. \tag{140}$$

To proceed, we prove the following lemmata.

Lemma 14. *Given $\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_N \in \mathbb{R}^d$, assume that the data points $\mathbf{x}_1, \dots, \mathbf{x}_m$ are i.i.d. sampled from $\mathcal{N}(\mathbf{0}, \gamma^2\mathbf{I}_d)$. Define*

$$\Gamma_{j\ell} := \exp\left(-\frac{\gamma^2}{2}\|\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell\|^2\right)$$

and let Γ_{min} and Γ_{max} be a lower and upper bound respectively with $\Gamma_{max} \leq \frac{3}{2}$. For a given $0 < \delta < 1$, if

$$m \geq \frac{4}{\Gamma_{min}^2}\log\frac{N^2}{\delta} \tag{141}$$

then with probability at least $1 - \delta$ with respect to the draw of the $\mathbf{x}_j's$, the mutual coherence of $\mathbf{A}$ satisfies

$$\mu_{\mathbf{A}} \leq 2\Gamma_{max}. \tag{142}$$

Proof. Since the features are complex exponential, the norm of each is column is $\sqrt{m}$, thus it suffices to only consider the inner products between pairs of columns. For a fixed pair of weights $\boldsymbol{\omega}_j$ and $\boldsymbol{\omega}_\ell$, we have

$$\mathbb{E}_x[a_{kj}a_{k\ell}^\star] = \mathbb{E}_{\mathbf{x}}[\exp(i\langle \mathbf{x}_k, \boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell\rangle)] = \exp\left(-\frac{\gamma^2}{2}\|\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell\|^2\right) = \Gamma_{j\ell}, \tag{143}$$

where we used the characteristic function of the Gaussian distribution. In addition, we have that $\mathrm{Var}_x[a_{kj}a_{k\ell}^\star] \leq 1$ almost surely. Given that $\mathbf{x}_k$'s are i.i.d., we can apply Bernstein inequality to obtain

$$\mathbb{P}\left(|\mu_{j\ell} - \Gamma_{j\ell}| \geq \Gamma_{j\ell}\right) = \mathbb{P}\left(\left|\sum_{j=1}^m a_{jk}a_{j\ell}^\star - m\Gamma_{j\ell}\right| \geq m\Gamma_{j\ell}\right) \leq 2\exp\left(\frac{-\frac{1}{2}m^2{\Gamma_{j\ell}}^2}{m + \frac{2}{3}m\Gamma_{j\ell}}\right) \tag{144}$$

The rest follows from the arguments in the proof of Lemma 9. ∎

This analysis is similar to the Gaussian case; however, since we do not have access to the probability distribution for $\boldsymbol{\omega}$, we use properties of subgaussian distributions to obtain concentration results for the key quantities

$$\Delta_{j\ell} := \|\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell\|, \quad \Delta_{max} := \max_{j\neq\ell} \Delta_{j\ell}, \quad \Delta_{min} := \min_{j\neq\ell} \Delta_{j\ell}. \tag{145}$$

Lemma 15. *Let $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N$ be a collection of q-sparse i.i.d. random vectors in $\mathbb{R}^d$ with arbitrary supports where their nonzero entries are i.i.d. subgaussian. For any $0 < \delta < 1$ define*

$$t := \frac{2^{9/4}\sigma\sqrt{q}}{\mathbb{E}[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2]} \sqrt{\log \frac{N^2}{\delta}}. \tag{146}$$

Then with probability at least $1-\delta$

$$\Delta_{max} < (1+t)\mathbb{E}[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2], \qquad \Delta_{min} > (1-t)\mathbb{E}[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2]. \tag{147}$$

Proof. Define $u = \|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2$. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ denote the support of $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$, respectively, each with cardinality q. Define

$$\mathcal{G} = \mathcal{S}_1 \cap \mathcal{S}_2, \quad \mathcal{G}_1 = \mathcal{S}_1 \backslash \mathcal{G}, \quad \mathcal{G}_2 = \mathcal{S}_2 \backslash \mathcal{G}. \tag{148}$$

Then,

$$u = \|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2 = \sum_{i=1}^{d} (\boldsymbol{\omega}_{1,i} - \boldsymbol{\omega}_{2,i})^2 = \sum_{i\in\mathcal{G}} (\boldsymbol{\omega}_{1,i} - \boldsymbol{\omega}_{2,i})^2 + \sum_{i\in\mathcal{G}_1} \boldsymbol{\omega}_{1,i}^2 + \sum_{i\in\mathcal{G}_2} \boldsymbol{\omega}_{2,i}^2. \tag{149}$$

Let Y, Y_1, and Y_2 be the first, second, and third terms in the above expression, respectively, and note that they are conditionally independent given $\mathcal{S}_1, \mathcal{S}_2$. By the first part of Lemma 3, and the fact that Y, Y_1, and Y_2 are sums of squared of $\mathrm{SG}(2\sigma^2)$ random variables, it holds that [24]

$$Y \sim \mathrm{SE}(8\sqrt{2}|G|\sigma^2, 8\sigma^2), \quad Y_1, Y_2 \sim \mathrm{SE}(4\sqrt{2}(q-|G|)\sigma^2, 4\sigma^2). \tag{150}$$

Thus, conditioned on $\mathcal{S}_1$ and $\mathcal{S}_2$, we conclude that $u \sim \mathrm{SE}(8\sqrt{2}q\sigma^2, 8\sigma^2)$. Therefore, using Lemma 3 yields

$$\mathbb{P}(|u - \mathbb{E}[u]| \geq t\mathbb{E}[u]) \leq 2\exp\left(\frac{t^2\mathbb{E}[u]^2}{16\sqrt{2}q\sigma^2}\right), \tag{151}$$

by noting that we can remove the conditioning given that the parameters of u are uniformly bounded irrespective of the location of the nonzero entries in $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$.

Using the bounded separation properties for any ϵ, ϵ' we may write

$$\begin{aligned} \mathbb{P}\left(\Delta_{min} > \epsilon\right]) &= \mathbb{P}\left(\cap_{j,\ell} \left\{\|\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell\|^2 > \epsilon\right\}\right) \\ &= 1 - \mathbb{P}\left(\cup_{j,\ell} \left\{\|\boldsymbol{\omega}_j - \boldsymbol{\omega}_\ell\|^2 < \epsilon\right\}\right) \\ &\geq 1 - \binom{N}{2} \mathbb{P}\left(\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2 < \epsilon\right). \end{aligned} \tag{152}$$

Similarly,

$$\mathbb{P}\left(\Delta_{max} < \epsilon'\right) \geq 1 - \binom{N}{2} \mathbb{P}\left(\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2 > \epsilon'\right). \tag{153}$$

Setting $\epsilon = (1-t)\mathbb{E}[u]$ and $\epsilon' = (1+t)\mathbb{E}[u]$, yields

$$\mathbb{P}\left(\Delta_{max} < (1+t)\mathbb{E}[u] \text{ and } \Delta_{min} > (1-t)\mathbb{E}[u]\right) \geq 1 - N^2 \exp\left(-\frac{t^2\mathbb{E}[u]^2}{16\sqrt{2}q\sigma^2}\right). \tag{154}$$

Setting the RHS of above to $1-\delta$ establishes the stated result. ∎

Now we prove Lemma 13.

Proof. We use the notation $u = \|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2$ throughout. We need to show that

$$\mu_{\mathbf{A}} \leq 2\Gamma_{\max} \leq \frac{4}{\sqrt{41}(2s-1)} \tag{155}$$

holds with high probability. Conditioned on the event from Lemma 15, which does not occur with probability at most δ,

$$\Delta_{max} < (1+t)\mathbb{E}[u], \quad \Delta_{min} > (1-t)\mathbb{E}[u], \tag{156}$$

we have, which again does not occur with probability at most δ,

$$\Gamma_{max} = \exp\left(-\frac{\gamma^2}{2}\Delta_{min}\right) \implies \Gamma_{max} \leq \exp\left(-\frac{\gamma^2}{2}(1-t)\mathbb{E}[u]\right) \leq \frac{2}{\sqrt{41}(2s-1)}, \tag{157}$$

and so

$$\gamma^2\mathbb{E}[u] \geq \frac{2}{(1-t)}\log\left(\frac{\sqrt{41}(2s-1)}{2}\right). \tag{158}$$

Similarly, we have conditioned on the above event,

$$\Gamma_{min}^2 \geq \exp(-\gamma^2\Delta_{max}) \implies \Gamma_{min}^2 \geq \exp(-\gamma^2(1+t)\mathbb{E}[u]) \geq \exp(-2\gamma^2\mathbb{E}[u]). \tag{159}$$

■

Corollary 7.1. *In the case where we have* $\mathbb{E}[\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|^2] = 2q\sigma^2$*, which occurs for many distributions of interest, if we set*

$$t = \frac{2^{5/4}}{\sigma\sqrt{q}}\sqrt{\log\frac{N^2}{\delta}} \tag{160}$$

then the bounds become

$$\begin{aligned} \gamma^2\sigma^2\left(1 - \frac{2^{5/4}}{\sigma\sqrt{q}}\sqrt{\log\frac{N^2}{\delta}}\right) &\geq \frac{1}{q}\log\left(\frac{\sqrt{41}(2s-1)}{2}\right), \\ m &\geq 4\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{2(1+t)}{1-t}}\log\frac{N^2}{\delta}. \end{aligned} \tag{161}$$