Random Fourier Features for Kernel Ridge Regression: Approximation Bounds and Statistical Guarantees

Haim Avron [1] **Michael Kapralov** [2] **Cameron Musco** [3]
Christopher Musco [3] **Ameya Velingker** [2] **Amir Zandieh** [2]

Abstract

Random Fourier features is one of the most popular techniques for scaling up kernel methods, such as kernel ridge regression. However, despite impressive empirical results, the statistical properties of random Fourier features are still not well understood. In this paper we take steps toward filling this gap. Specifically, we approach random Fourier features from a spectral matrix approximation point of view, give tight bounds on the number of Fourier features required to achieve a spectral approximation, and show how spectral matrix approximation bounds imply statistical guarantees for kernel ridge regression.

1. Introduction

Kernel methods constitute a powerful paradigm for devising non-parametric modeling techniques for a wide range of problems in machine learning. One of the most elementary is *Kernel Ridge Regression (KRR)*. Given training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is an input domain and $\mathcal{Y} \subseteq \mathbb{R}$ is an output domain, a positive definite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and a regularization parameter $\lambda > 0$, the response for a given input $\mathbf{x}$ is estimated as:

$$\bar{f}(\mathbf{x}) \equiv \sum_{j=1}^{n} k(\mathbf{x}_j, \mathbf{x})\alpha_j$$

where $\boldsymbol{\alpha} = (\alpha_1 \cdots \alpha_n)^{\mathrm{T}}$ is the solution of the equation

$$(\mathbf{K} + \lambda \mathbf{I}_n)\boldsymbol{\alpha} = \mathbf{y}. \tag{1}$$

In the above, $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the *kernel matrix* or *Gram matrix* defined by $\mathbf{K}_{ij} \equiv k(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{y} \equiv [y_1 \cdots y_n]^{\mathrm{T}}$ is the vector of responses. The KRR estimator can be derived by minimizing a regularized square loss objective function over a hypothesis space defined by the reproducing kernel Hilbert space associated with $k(\cdot, \cdot)$; however, the details are not important for this paper.

While simple, KRR is a powerful technique that is well understood statistically and capable of achieving impressive empirical results. Nevertheless, the method has a key weakness: computing the KRR estimator can be prohibitively expensive for large datasets. Solving (1) generally requires $\Theta(n^3)$ time and $\Theta(n^2)$ memory. Thus, the design of scalable methods for KRR (and other kernel based methods) has been the focus of intensive research in recent years (Zhang et al., 2015; Alaoui & Mahoney, 2015; Musco & Musco, 2016; Avron et al., 2016).

One of the most popular approaches to scaling up kernel based methods is random Fourier features sampling, originally proposed by Rahimi & Recht (2007). For shift-invariant kernels (e.g. the Gaussian kernel), Rahimi & Recht (2007) presented a distribution D on functions from $\mathcal{X}$ to $\mathbb{C}^s$ (s is a parameter) such that for every $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$

$$k(\mathbf{x}, \mathbf{z}) = \mathbb{E}_{\varphi \sim D}\left[\varphi(\mathbf{x})^* \varphi(\mathbf{z})\right] .$$

The idea is to sample φ from D and use $\tilde{k}(\mathbf{x}, \mathbf{z}) \equiv \varphi(\mathbf{x})^* \varphi(\mathbf{y})$ as a surrogate kernel. The resulting approximate KRR estimator can be computed in $O(ns^2)$ time and $O(ns)$ memory (see §2.2 for details), giving substantial computational savings if $s \ll n$.

This approach naturally raises the question: how large should s be to ensure a high quality estimator? Or, using the exact KRR estimator as a natural baseline: how large should s be for the random Fourier features estimator to be almost as good as the exact KRR estimator? Answering this question can help us determine when random Fourier features can be useful, whether the method needs to be improved, and how to go about improving it.

The original random Fourier features analysis (Rahimi & Recht, 2007) bounds the point-wise distance between

*Equal contribution [1]School of Mathematical Sciences, Tel Aviv University, Israel [2]School of Computer and Communication Sciences, EPFL, Switzerland [3]Computer Science and Artificial Intelligence Laboratory, MIT, USA. Correspondence to: Haim Avron <haimav@post.tau.ac.il>, Michael Kapralov <michael.kapralov@epfl.ch>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

$k(\cdot,\cdot)$ and $\tilde{k}(\cdot,\cdot)$ (for other approaches for analyzing random Fourier features, see §2.3). However, the bounds do not naturally lead to an answer to the aforementioned question. In contrast, spectral approximation bounds on the entire kernel matrix, i.e. of the form

$$(1-\Delta)(\mathbf{K}+\lambda\mathbf{I}_n) \preceq \tilde{\mathbf{K}}+\lambda\mathbf{I}_n \preceq (1+\Delta)(\mathbf{K}+\lambda\mathbf{I}_n)\,, \quad (2)$$

naturally have statistical and algorithmic implications. Indeed, in §3 we show that when (2) holds we can bound the excess risk introduced by the random Fourier features estimator when compared to the KRR estimator. We also show that $\tilde{\mathbf{K}}+\lambda\mathbf{I}_n$ can be used as an effective preconditioner for the solution of (1). This motivates the study of how large s should be as a function of Δ for (2) to hold.

In this paper we rigorously analyze the relation between the number of random Fourier features and the spectral approximation bound (2). Our main results are the following:

- We give an upper bound on the number of random features needed to achieve (2) (Theorem 7). This bound, in conjunction with the results in §3, positively shows that random Fourier features can give guarantees for KRR under reasonable assumptions.
- We give a lower bound showing that our upper bound is tight for the Gaussian kernel (Theorem 8).
- We show that the upper bound can be improved dramatically by modifying the sampling distribution used in classical random Fourier features (§4). Our sampling distribution is based on an appropriately defined *leverage function* of the kernel, closely related to so-called leverage scores frequently encountered in the analysis of sampling based methods for linear regression. Unfortunately, it is unclear how to efficiently sample using the leverage function.
- To address the lack of an efficient way to sample using the leverage function, we propose a novel, easy-to-sample distribution for the Gaussian kernel which approximates the true leverage function distribution and allows random Fourier features to achieve a significantly improved upper bound (Theorem 10). The bound has an exponential dependence on the data dimension, so it is only applicable to low dimensional datasets. Nevertheless, it demonstrate that classic random Fourier features can be improved for spectral approximation and motivates further study. As an application, our improved understanding of the leverage function yields a novel asymptotic bound on the statistical dimension of Gaussian kernel matrices over bounded datasets, which may be of independent interest (Corollary 15).

2. Preliminaries

2.1. Setup and Notation

The complex conjugate of $x \in \mathbb{C}$ is denoted by x^*. For a vector $\mathbf{x}$ or a matrix $\mathbf{A}$, $\mathbf{x}^*$ or $\mathbf{A}^*$ denotes the Hermitian transpose. The $l \times l$ identity matrix is denoted $\mathbf{I}_l$. We use the convention that vectors are column-vectors.

A Hermitian matrix $\mathbf{A}$ is positive semidefinite (PSD) if $\mathbf{x}^*\mathbf{A}\mathbf{x} \geq 0$ for every vector $\mathbf{x}$. It is positive definite (PD) if $\mathbf{x}^*\mathbf{A}\mathbf{x} > 0$ for every vector $\mathbf{x} \neq 0$. For any two Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$ of the same size, $\mathbf{A} \preceq \mathbf{B}$ means that $\mathbf{B}-\mathbf{A}$ is PSD.

We use $L_2(d\rho) = L_2(\mathbb{R}^d, d\rho)$ to denote the space of complex-valued square-integrable functions with respect to some measure $\rho(\cdot)$. $L_2(d\rho)$ is a Hilbert space equipped with the inner product

$$\begin{aligned}\langle f, g\rangle_{L_2(d\rho)} &= \int_{\mathbb{R}^d} f(\boldsymbol{\eta})g(\boldsymbol{\eta})^* d\rho(\boldsymbol{\eta}) \\ &= \int_{\mathbb{R}^d} f(\boldsymbol{\eta})g(\boldsymbol{\eta})^* p_\rho(\boldsymbol{\eta}) d\boldsymbol{\eta}\,.\end{aligned}$$

In the above, $p_\rho(\cdot)$ is the density associated with $\rho(\cdot)$.

We denote the training set by $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$. Note that n denotes the number of training examples, and d their dimension. We denote the kernel, which is a function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$, by k. We denote the kernel matrix by $\mathbf{K}$, with $\mathbf{K}_{ij} \equiv k(\mathbf{x}_i, \mathbf{x}_j)$. The associated reproducing kernel Hilbert space (RKHS) is denoted by $\mathcal{H}_k$, and the associated inner product by $(\cdot,\cdot)_{\mathcal{H}_k}$. Some results are stated for the Gaussian kernel $k(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x}-\mathbf{z}\|_2^2/2\sigma^2)$ for some bandwidth parameter σ.

We use $\lambda = \lambda_n$ to denote the ridge regularization parameter. While for brevity we omit the n subscript, the choice of regularization parameter generally depends on n. Typically, $\lambda_n = \omega(1)$ and $\lambda_n = o(n)$. See Caponnetto & De Vito (2007) and Bach (2013) for discussion on the asymptotic behavior of λ_n, noting that in our notation, λ is scaled by an n factor as compared to those works. As the ratio between n and λ will be an important quantity in our bounds, we denote it as $n_\lambda \equiv n/\lambda$.

The *statistical dimension* or *effective degrees of freedom* is denoted by $s_\lambda(\mathbf{K}) \equiv \mathbf{Tr}\left((\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{K}\right)$.

2.2. Random Fourier Features

2.2.1. Classical Random Fourier Features

Random Fourier features (Rahimi & Recht, 2007) is an approach to scaling up kernel methods for shift-invariant kernels. A shift-invariant kernel is a kernel of the form $k(\mathbf{x}, \mathbf{z}) = k(\mathbf{x}-\mathbf{z})$ where $k(\cdot)$ is a positive definite func-

tion (we abuse notation by using k to denote both the kernel and the defining positive definite function).

The underlying observation behind random Fourier features is a simple consequence of Bochner's Theorem: for every shift-invariant kernel for which $k(0) = 1$ there is a probability measure $\mu_k(\cdot)$ and a corresponding probability density function $p_k(\cdot)$, both on $\mathbb{R}^d$, such that

$$\begin{aligned} k(\mathbf{x}, \mathbf{z}) &= \int_{\mathbb{R}^d} e^{-2\pi i \boldsymbol{\eta}^{\mathrm{T}}(\mathbf{x}-\mathbf{z})} d\mu_k(\boldsymbol{\eta}) \\ &= \int_{\mathbb{R}^d} e^{-2\pi i \boldsymbol{\eta}^{\mathrm{T}}(\mathbf{x}-\mathbf{z})} p_k(\boldsymbol{\eta}) d\boldsymbol{\eta} \,. \end{aligned} \tag{3}$$

In other words, the inverse Fourier transform of the kernel $k(\cdot)$ is a probability density function, $p_k(\cdot)$. For simplicity we typically drop the k subscript, writing $\mu(\cdot) = \mu_k(\cdot)$ and $p(\cdot) = p_k(\cdot)$, with the associated kernel function clear from context.

If $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s$ are drawn according to $p(\cdot)$, and we define $\varphi(\mathbf{x}) \equiv \frac{1}{\sqrt{s}} \left(e^{-2\pi i \boldsymbol{\eta}_1^T \mathbf{x}}, \cdots, e^{-2\pi i \boldsymbol{\eta}_s^T \mathbf{x}} \right)^*$, then it is not hard to see that

$$k(\mathbf{x}, \mathbf{z}) = \mathbb{E}_\varphi \left[\varphi(\mathbf{x})^* \varphi(\mathbf{z}) \right] .$$

The idea of the Random Fourier features method is then to define

$$\tilde{k}(\mathbf{x}, \mathbf{z}) \equiv \varphi(\mathbf{x})^* \varphi(\mathbf{z}) = \frac{1}{s} \sum_{l=1}^{s} e^{-2\pi i \boldsymbol{\eta}_l^{\mathrm{T}}(\mathbf{x}-\mathbf{z})} \tag{4}$$

as a substitute kernel.

Now suppose that $\mathbf{Z} \in \mathbb{C}^{n \times s}$ is the matrix whose j^{th} row is $\varphi(\mathbf{x}_j)^*$, and let $\tilde{\mathbf{K}} = \mathbf{Z}\mathbf{Z}^*$. $\tilde{\mathbf{K}}$ is the kernel matrix corresponding to $\tilde{k}(\cdot, \cdot)$. The resulting random Fourier features KRR estimator is $\tilde{f}(\mathbf{x}) \equiv \sum_{j=1}^{n} \tilde{k}(\mathbf{x}_j, \mathbf{x}) \tilde{\alpha}_j$ where $\tilde{\boldsymbol{\alpha}}$ is the solution of $(\tilde{\mathbf{K}} + \lambda \mathbf{I}_n)\tilde{\boldsymbol{\alpha}} = \mathbf{y}$. Typically, $s < n$ and we can represent $\tilde{f}(\cdot)$ more efficiently as:

$$\tilde{f}(\mathbf{x}) = \varphi(\mathbf{x})^* \mathbf{w}$$

where

$$\mathbf{w} = (\mathbf{Z}^*\mathbf{Z} + \lambda \mathbf{I}_s)^{-1} \mathbf{Z}^* \mathbf{y}$$

We can compute $\mathbf{w}$ in $O(ns^2)$ time, making random Fourier features computationally attractive if $s < n$.

2.2.2. Modified Random Fourier Features

While it seems to be a natural choice, there is no fundamental reason that we must sample the frequencies $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s$ using the Fourier transform density function $p(\cdot)$. In fact, our results show that it is advantageous to use a different sampling distribution based on the kernel leverage function (defined later).

Let $q(\cdot)$ be any probability density function whose support includes that of $p(\cdot)$. If we sample $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s$ using $q(\cdot)$, and define

$$\varphi(\mathbf{x}) \equiv \frac{1}{\sqrt{s}} \left(\sqrt{\frac{p(\boldsymbol{\eta}_1)}{q(\boldsymbol{\eta}_1)}} e^{-2\pi i \boldsymbol{\eta}_1^T \mathbf{x}}, \cdots, \sqrt{\frac{p(\boldsymbol{\eta}_s)}{q(\boldsymbol{\eta}_s)}} e^{-2\pi i \boldsymbol{\eta}_s^T \mathbf{x}} \right)^*$$

we still have $k(\mathbf{x}, \mathbf{z}) = \mathbb{E}_\varphi \left[\varphi(\mathbf{x})^* \varphi(\mathbf{z}) \right]$. We refer to this method as *modified random Fourier features* and remark that it can be viewed as a form of importance sampling.

2.2.3. Additional Notations and Identities

Now that we have defined (modified) random Fourier features, we can introduce some additional notation and identities that shall prove useful in the rest of the paper.

The (j, l) entry of $\mathbf{Z}$ is given by

$$\mathbf{Z}_{jl} = \frac{1}{\sqrt{s}} e^{-2\pi i \mathbf{x}_j^{\mathrm{T}} \boldsymbol{\eta}_l} \sqrt{p(\boldsymbol{\eta}_l)/q(\boldsymbol{\eta}_l)}. \tag{5}$$

Let $\mathbf{z} : \mathbb{R}^d \to \mathbb{C}^n$ be defined by

$$\mathbf{z}(\boldsymbol{\eta})_j = e^{-2\pi i \mathbf{x}_j^{\mathrm{T}} \boldsymbol{\eta}} \,.$$

Note that column l of $\mathbf{Z}$ from the previous section is exactly $\mathbf{z}(\boldsymbol{\eta}_l)\sqrt{p(\boldsymbol{\eta}_l)/[s \cdot q(\boldsymbol{\eta}_l)]}$. So we have:

$$\mathbf{Z}\mathbf{Z}^* = \frac{1}{s} \sum_{l=1}^{s} \frac{p(\boldsymbol{\eta}_l)}{q(\boldsymbol{\eta}_l)} \mathbf{z}(\boldsymbol{\eta}_l)\mathbf{z}(\boldsymbol{\eta}_l)^*.$$

Finally, by (3) we have $\mathbb{E}\left[\mathbf{Z}\mathbf{Z}^*\right] = \mathbf{K}$ since

$$\mathbf{K} = \int_{\mathbb{R}^d} \mathbf{z}(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^* d\mu(\boldsymbol{\eta}) = \int_{\mathbb{R}^d} \mathbf{z}(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^* p(\boldsymbol{\eta}) d\boldsymbol{\eta} \,.$$

2.3. Related Work

Rahimi & Recht (2007)'s original analysis of random Fourier features bounded the point-wise distance between $k(\cdot, \cdot)$ and $\tilde{k}(\cdot, \cdot)$. In follow-up work, they give learning rate bounds for a broad class of estimators using random Fourier features. However, their results do not apply to classic KRR (Rahimi & Recht, 2008). Their main bound becomes relevant only when the number of sampled features is on order of the training set size.

Rudi et al. (2016) prove generalization properties for KRR with random features, under somewhat difficult to verify technical assumptions, some of which can be seen as constraining the leverage function distribution that we study. They leave open improving their bounds via a more refined sampling approach. Bach (2017) analyzes random Fourier features from a function approximation point of view. He defines a similar leverage function distribution to the one that we consider, but leaves open establishing

bounds on and effectively sampling from this distribution, both of which we address in this work. Finally, Tropp (2015) analyzes the distance between the kernel matrix and its approximation in terms of the spectral norm, $\|\mathbf{K}-\tilde{\mathbf{K}}\|_2$, which can be a significantly weaker error metric than (2).

Outside of work on random Fourier features, risk inflation bounds for approximate KRR and leverage score sampling have been used to analyze and improve the Nyström method for kernel approximation (Bach, 2013; Alaoui & Mahoney, 2015; Rudi et al., 2015; Musco & Musco, 2016). We apply a number of techniques from this line of work.

Spectral approximation bounds, such as (2), are quite popular in the sketching literature; see Woodruff (2014). Most closely related to our work is analysis of spectral approximation bounds without regularization (i.e. $\lambda = 0$) for the polynomial kernel (Avron et al., 2014). Improved bounds with regularization (still for the polynomial kernel) were recently proved by Avron et al. (2016).

3. Spectral Bounds and Statistical Guarantees

Given a feature transformation, like random Fourier features, how do we analyze it and relate its use to non-approximate methods? A common approach, taken for example in the original paper on random Fourier features (Rahimi & Recht, 2007), is to bound the difference between the true kernel $k(\cdot,\cdot)$ and the approximate kernel $\tilde{k}(\cdot,\cdot)$. However, it is unclear how such bounds translate to downstream guarantees on statistical learning methods, such as KRR. In this paper we advocate and focus on spectral approximation bounds on the regularized kernel matrix, specifically, bounds of the form

$$(1-\Delta)(\mathbf{K}+\lambda\mathbf{I}_n) \preceq \mathbf{Z}\mathbf{Z}^* + \lambda\mathbf{I}_n \preceq (1+\Delta)(\mathbf{K}+\lambda\mathbf{I}_n) \quad (6)$$

for some $\Delta < 1$.

Definition 1. We say that a matrix $\mathbf{A}$ is a Δ-*spectral approximation* of another matrix $\mathbf{B}$, if $(1-\Delta)\mathbf{B} \preceq \mathbf{A} \preceq (1+\Delta)\mathbf{B}$.

Remark 1. When $\lambda = 0$, bounds of the form of (6) can be viewed as a low-distortion subspace embedding bounds. Indeed, when $\lambda = 0$ it follows from (6) that $\mathbf{Sp}\left(k(\mathbf{x}_1,\cdot),\ldots,k(\mathbf{x}_n,\cdot)\right) \subseteq \mathcal{H}_k$ can be embedded with Δ-distortion in $\mathbf{Sp}\left(\varphi(\mathbf{x}_1),\ldots,\varphi(\mathbf{x}_n)\right) \subseteq \mathbb{R}^s$.

The main mathematical question we seek to address in this paper is: when using random Fourier features, how large should s be in order to guarantee that $\mathbf{Z}\mathbf{Z}^* + \lambda\mathbf{I}_n$ is a Δ-spectral approximation of $\mathbf{K}+\lambda\mathbf{I}_n$? To motivate this question, in the following two subsections we show that such bounds can be used to derive risk inflation bounds for approximate kernel ridge regression. We also show that such bounds can be used to analyze the use of $\mathbf{Z}\mathbf{Z}^* + \lambda\mathbf{I}_n$ as a preconditioner for $\mathbf{K}+\lambda\mathbf{I}_n$.

While this paper focuses on KRR for conciseness, we remark that in the sketching literature, spectral approximation bounds also form the basis for analyzing sketching based methods for tasks like low-rank approximation, k-means and more. In the kernel setting, such bounds where analyzed, without regularization, for the polynomial kernel (Avron et al., 2014). Cohen et al. (2017) recently showed that (6) along with a trace condition on $\mathbf{Z}\mathbf{Z}^*$ (which holds for all sampling approaches we consider) yields a so called "projection-cost preservation" condition for the kernel approximation. With λ chosen appropriately, this condition ensures that $\mathbf{Z}\mathbf{Z}^*$ can be used in place of $\mathbf{K}$ for approximately solving kernel k-means clustering and for certain versions of kernel PCA and kernel CCA. See Musco & Musco (2016) for details, where this analysis is carried out for the Nyström method.

3.1. Risk Bounds

One way to analyze estimators is via risk bounds; several recent papers on approximate KRR employ such an analysis (Bach, 2013; Alaoui & Mahoney, 2015; Musco & Musco, 2016). In particular, these papers consider the fixed design setting and seek to bound the expected in-sample predication error of the KRR estimator $\bar{f}$, viewing it as an empirical estimate of the statistical risk. More specifically, the underlying assumption is that y_i satisfies

$$y_i = f^\star(\mathbf{x}_i) + \nu_i \quad (7)$$

for some $f^\star : \mathcal{X} \to \mathbb{R}$. The $\{\nu_i\}$'s are i.i.d noise terms, distributed as normal variables with variance σ_ν^2. The empirical risk of an estimator f, which can be viewed as a measure of the quality of the estimator, is

$$\mathcal{R}(f) \equiv \mathbb{E}_{\{\nu_i\}}\left[\frac{1}{n}\sum_{j=1}^{n}(f(\mathbf{x}_i) - f^\star(\mathbf{x}_i))^2\right]$$

(note that f itself might be a function of $\{\nu_i\}$).

Let $\mathbf{f} \in \mathbb{R}^n$ be the vector whose j^{th} entry is $f^\star(\mathbf{x}_j)$. It is quite straightforward to show that for the KRR estimator $\bar{f}$ we have (Bach, 2013; Alaoui & Mahoney, 2015):

$$\mathcal{R}(\bar{f}) = n^{-1}\lambda^2\mathbf{f}^{\mathrm{T}}(\mathbf{K}+\lambda\mathbf{I}_n)^{-2}\mathbf{f} + n^{-1}\sigma_\nu^2\mathbf{Tr}\left(\mathbf{K}^2(\mathbf{K}+\lambda\mathbf{I}_n)^{-2}\right).$$

Since $\lambda^2\mathbf{f}^{\mathrm{T}}(\mathbf{K}+\lambda\mathbf{I}_n)^{-2}\mathbf{f} \leq \lambda\mathbf{f}^{\mathrm{T}}(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{f}$ and $\mathbf{Tr}\left(\mathbf{K}^2(\mathbf{K}+\lambda\mathbf{I}_n)^{-2}\right) \leq \mathbf{Tr}\left(\mathbf{K}(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\right) = s_\lambda(\mathbf{K})$, we define

$$\widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f}) \equiv n^{-1}\lambda\mathbf{f}^{\mathrm{T}}(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{f} + n^{-1}\sigma_\nu^2 s_\lambda(\mathbf{K})$$

and note that $\mathcal{R}(\bar{f}) \leq \widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f})$. The first term in the above expressions for $\mathcal{R}(\bar{f})$ and $\widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f})$ is frequently referred to as the bias term, while the second is the variance term.

Lemma 2. *Suppose that (7) holds, and let* $\mathbf{f} \in \mathbb{R}^n$ *be the vector whose* j^{th} *entry is* $f^\star(\mathbf{x}_j)$. *Let* $\bar{f}$ *be the KRR estimator, and let* $\tilde{f}$ *be KRR estimator obtained using some other kernel* $\tilde{k}(\cdot,\cdot)$ *whose kernel matrix is* $\tilde{\mathbf{K}}$. *Suppose that* $\tilde{\mathbf{K}} + \lambda\mathbf{I}_n$ *is a* Δ*-spectral approximation to* $\mathbf{K} + \lambda\mathbf{I}_n$ *for some* $\Delta < 1$, *and that* $\|\mathbf{K}\|_2 \geq 1$. *The following bound holds:*

$$\mathcal{R}(\tilde{f}) \leq (1-\Delta)^{-1}\widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f}) + \frac{\Delta}{(1+\Delta)} \cdot \frac{\mathbf{rank}(\tilde{\mathbf{K}})}{n} \cdot \sigma_\nu^2 \tag{8}$$

The proof appears in the supplementary material (Appendix B).

In short, Lemma 2 bounds the risk of the approximate KRR estimator as a function of both the risk upper bound $\widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f})$ (8) and an additive term which is small if the rank of $\mathbf{rank}(\tilde{\mathbf{K}})$ and/or Δ is small. In particular, it is instructive to compare the additive term $(\Delta/(1+\Delta))n^{-1}\sigma_\nu^2 \cdot \mathbf{rank}(\tilde{\mathbf{K}})$ to the variance term $n^{-1}\sigma_\nu^2 \cdot s_\lambda(\mathbf{K})$. Since approximation $\tilde{\mathbf{K}}$ is only useful computationally if $\mathbf{rank}(\tilde{\mathbf{K}}) \ll n$ we should expect the additive term in (8) to also approach 0 an generally be small when n is large.

Remark 2. An approximation $\tilde{\mathbf{K}}$ is only useful computationally if $\mathbf{rank}(\tilde{\mathbf{K}}) \ll n$ so $\tilde{\mathbf{K}}$ gives a significantly compressed approximation to the original kernel matrix. Ideally we should have $\mathbf{rank}(\tilde{\mathbf{K}})/n \to 0$ as $n \to \infty$ and so the additive term in (8) will also approach 0 and generally be small when n is large.

3.2. Random Features Preconditioning

Suppose we choose to solve $(\mathbf{K} + \lambda\mathbf{I}_n)\boldsymbol{\alpha} = \mathbf{y}$ using an iterative method (e.g. CG). In this case, we can apply $\mathbf{Z}\mathbf{Z}^* + \lambda\mathbf{I}_n$ as a preconditioner. Using standard analysis of Krylov-subspace iterative methods it is immediate that if $\mathbf{Z}\mathbf{Z}^* + \lambda\mathbf{I}_n$ is a Δ-spectral approximation of $\mathbf{K} + \lambda\mathbf{I}_n$ then the number of iterations until convergence is $O(\sqrt{(1+\Delta)/(1-\Delta)}))$. Thus, if $\mathbf{Z}\mathbf{Z}^* + \lambda\mathbf{I}_n$ is, say, a $1/2$-spectral approximation of $\mathbf{K}+\lambda\mathbf{I}_n$, then the number of iterations is bounded by a constant. The preconditioner can be efficiently applied (after preprocessing) via the Woodbury formula, giving cost per iteration (if $s \leq n$) of $O(n^2)$. The overall cost of computing the KRR estimator is therefore $O(ns^2 + n^2)$. Thus, as long as $s = o(n)$ this approach gives an advantage over direct methods which cost $O(n^3)$. For small s it also beats non-preconditioned iterative methods cost $O(n^2\sqrt{\kappa(\mathbf{K})})$. We reach again the question that was poised earlier: how big should s be so that $\mathbf{Z}\mathbf{Z}^* + \lambda\mathbf{I}_n$ is a $1/2$-spectral approximation of $\mathbf{K} + \lambda\mathbf{I}_n$?

See Cutajar et al. (2016) and Avron et al. (2016) for more details and discussion on random features preconditioning.

4. Ridge Leverage Function Sampling and Random Fourier Features

In this section we present upper bounds on the number of random Fourier features needed to guarantee that $\mathbf{Z}\mathbf{Z}^* + \lambda\mathbf{I}_n$ is a Δ-spectral approximation to $\mathbf{K} + \lambda\mathbf{I}_n$. Our bounds are applicable to *any* shift-invariant kernel, and a wide range of feature sampling distributions (and, in particular, for classical random Fourier features).

Our analysis is based on relating the sampling density to an appropriately defined *ridge leverage function*. This function is a continuous generalization of the popular leverage scores (Mahoney & Drineas, 2009) and ridge leverage scores (Alaoui & Mahoney, 2015; Cohen et al., 2017) used in the analysis of linear methods. Bach (2017) defined the leverage function of the integral operator given by the kernel function and the data distribution. For our purposes, a more appropriate definition is with respect to a fixed input dataset:

Definition 3. For given $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and shift-invariant kernel $k(\cdot,\cdot)$, define the *ridge leverage function* as

$$\tau_\lambda(\boldsymbol{\eta}) \equiv p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{z}(\boldsymbol{\eta})\,.$$

In the above, $\mathbf{K}$ is the kernel matrix and $p(\cdot)$ is the distribution associated with $k(\cdot,\cdot)$.

Proposition 4.

$$p(\boldsymbol{\eta})n/(n+\lambda) \leq \tau_\lambda(\boldsymbol{\eta}) \leq p(\boldsymbol{\eta})n/\lambda$$

$$\int_{\mathbb{R}^d} \tau_\lambda(\boldsymbol{\eta})d\eta = s_\lambda(\mathbf{K})$$

The (simple) proof of the proposition is given in the supplementary material (Appendix C).

Recall that we denote the ratio n/λ, which appears frequently in our analysis, by $n_\lambda = n/\lambda$. As discussed, theoretical bounds generally set $\lambda = \omega(1)$ (as a function of n) so $n_\lambda = o(n)$. However we remark that in practice, it may frequently be the case that λ is very small and $n_\lambda \gg n$.

Corollary 5. *For any* $\mathbf{K}$, $\mathbf{s}_\lambda(\mathbf{K}) \leq n_\lambda$.

For any shift-invariant kernel with $k(\mathbf{x},\mathbf{x}) = 1$ and $k(\mathbf{x},\mathbf{z}) \to 0$ as $\|\mathbf{x}-\mathbf{z}\|_2 \to \infty$ (e.g., the Gaussian kernel) if we allow points to be arbitrarily spread out, the kernel matrix converges to the identity matrix, and $s_\lambda(\mathbf{I}_n) = n/(1+\lambda) = \Omega(n_\lambda)$ if $\lambda = \Omega(1)$ so the above bound is tight. However, this requires datasets of increasingly large diameter (as n grows). In contrast, the usual assumption in statistical learning is that the data is sampled from a bounded domain $\mathcal{X}$. In §7.2 we show via a leverage function upper bound that for the important Gaussian kernel, for bounded datasets we have $s_\lambda(\mathbf{K}) = o(n_\lambda)$.

In the matrix sketching literature it is well known that spectral approximation bounds similar to (6) can be constructed by sampling columns relative to upper bounds on the leverage scores. In the following, we generalize this for the case of sampling Fourier features from a continuous domain.

Lemma 6. *Let $\tilde{\tau} : \mathbb{R}^d \to \mathbb{R}$ be a measurable function such that $\tilde{\tau}(\boldsymbol{\eta}) \geq \tau_\lambda(\boldsymbol{\eta})$ for all $\boldsymbol{\eta} \in \mathbb{R}^d$, and furthermore assume that*

$$s_{\tilde{\tau}} \equiv \int_{\mathbb{R}^d} \tilde{\tau}(\boldsymbol{\eta}) d\boldsymbol{\eta}$$

is finite. Denote $p_{\tilde{\tau}}(\boldsymbol{\eta}) = \tilde{\tau}(\boldsymbol{\eta})/s_{\tilde{\tau}}$. Let $\Delta \leq 1/2$ and $\rho \in (0,1)$. Assume that $\|\mathbf{K}\|_2 \geq \lambda$. Suppose we take $s \geq \frac{8}{3}\Delta^{-2} s_{\tilde{\tau}} \ln(16 s_\lambda(\mathbf{K})/\rho)$ samples $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s$ from the distribution associated with the density $p_{\tilde{\tau}}(\cdot)$ and the construct the matrix $\mathbf{Z}$ according to (5) with $q = p_{\tilde{\tau}}$. Then $\mathbf{Z}\mathbf{Z}^* + \lambda \mathbf{I}_n$ is Δ-spectral approximation of $\mathbf{K} + \lambda \mathbf{I}_n$ with probability of at least $1 - \rho$.

The proof is based on matrix concentration inequalities, and appears in the supplementary material (Appendix D).

Lemma 6 shows that if we could sample using the ridge leverage function, then $O(s_\lambda(\mathbf{K}) \log(s_\lambda(\mathbf{K})))$ samples suffice for spectral approximation of $\mathbf{K}$ (for a fixed Δ and failure probability). While there is no straightforward way to perform this sampling, we can consider how well the classic random Fourier features sampling distribution approximates the leverage function, obtaining a bound on its performance (the proof is in Appendix D as well):

Theorem 7. *Let $\Delta \leq 1/2$ and $\delta \in (0,1)$. Assume that $\|\mathbf{K}\|_2 \geq \lambda$. If we use $s \geq \frac{8}{3}\Delta^{-2} n_\lambda \ln(16 s_\lambda(\mathbf{K})/\rho)$ random Fourier features (i.e., sampled according to $p(\cdot)$), then $\mathbf{Z}\mathbf{Z}^* + \lambda \mathbf{I}_n$ is Δ-spectral approximation of $\mathbf{K} + \lambda \mathbf{I}_n$ with probability of at least $1 - \rho$.*

Theorem 7 establishes that if $\lambda = \omega(\log(n))$ and Δ is fixed, $o(n)$ random Fourier features suffice for spectral approximation, and so the method can provably speed up KRR. Nevertheless, the bound depends on n_λ instead of $s_\lambda(\mathbf{K})$, as is possible with true leverage function sampling (see Lemma 6). This gap arises from our use of the simple, often loose, ridge leverage function upper bound given by Proposition 4.

Unfortunately, as the next section shows, the bound in Theorem 7 cannot be improved since the classic random Fourier features sampling distribution can be far enough from the ridge leverage distribution that $\Omega(n_\lambda)$ features may be needed even when $s_\lambda(\mathbf{K}) = o(n_\lambda)$.

5. Lower Bound

Our lower bound shows that the upper bound of Theorem 7 on the number of samples required by classic random Fourier features to obtain a spectral approximation to $\mathbf{K} + \lambda \mathbf{I}_n$ is essentially best possible. The full proof is given in the supplementary material (Appendix I).

Theorem 8. *Consider the Gaussian kernel with $\sigma = (2\pi)^{-1}$ (so $p(\eta) = \frac{1}{\sqrt{2\pi}} e^{-\eta^2/2}$). Suppose $n \geq 17$ is an odd integer, λ satisfies $\frac{10}{n} < \lambda \leq \frac{n}{2}$, and R satisfies $3000 \log^{1.5}(n_\lambda) \leq R \leq \frac{n}{500\sqrt{\log(n_\lambda)}}$. Then, there exists a dataset of n points $\{x_j\}_{j=1}^n \subseteq [-R, R]$ such that if s random Fourier features (i.e., sampled according to $p(\cdot)$) are used for some $s \leq \frac{n_\lambda}{400}$, then with probability at least $1/2$, there exists a vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that*

$$\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{K} + \lambda \mathbf{I}_n)\boldsymbol{\alpha} < \frac{2}{3}\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{Z}\mathbf{Z}^* + \lambda \mathbf{I}_n)\boldsymbol{\alpha}. \tag{9}$$

Furthermore, for the said dataset we have $s_\lambda(\mathbf{K}) = O(R \cdot \mathbf{poly}(\log n_\lambda))$.

Thus, the number of samples s required for $\mathbf{Z}\mathbf{Z}^* + \lambda \mathbf{I}_n$ to be a $1/2$-spectral approximation to $\mathbf{K} + \lambda \mathbf{I}_n$ for a bounded dataset of points must either depend exponentially on the radius of the point set, or at least linearly on n_λ, and there is an asymptotic gap between what is achieved with classical random Fourier features and what is achieved by modified random Fourier features using leverage function sampling.

We note that the above lower bound is proven for a one-dimensional point set, which makes it only stronger: even at low dimensions, and for the common Gaussian kernel, there is a large gap between the performance of classic random Fourier features and leverage function sampling.

The bound applies for datasets bounded on the range $[-R, R]$ for $R = \Omega\left(\log^{1.5} n_\lambda\right)$. As we will see in §7, the key idea behind the proof is to show that for such a dataset, the ridge leverage function is large on a range of low frequencies. In contrast, the classic random Fourier features distribution is very small at the edges of this frequency range, and so significantly undersamples some frequencies and does not achieve spectral approximation.

We remark that it would be preferable if Theorem 8 applied to bounded datasets (i.e. with R fixed), as the usual assumption in statistical learning theory is that data is sampled from a bounded domain. However, our current techniques are unable to address this scenario. Nevertheless, our analysis allows R to grow very slowly with n and we conjecture that the upper bound is tight even for bounded domains.

6. Improved Sampling (Gaussian Kernel)

Contrasting with the lower bound of Theorem 8, we now give a modified Fourier feature sampling distribution that does perform well for the Gaussian kernel on bounded input sets. Furthermore, unlike the true ridge leverage function, this distribution is simple and efficient to sample from.

To reduce clutter, we state the result for a fixed bandwidth $\sigma = (2\pi)^{-1}$. This is without loss of generality since we can rescale the points and adjust the bounding interval.

Our modified distribution essentially corrects the classic distribution by "capping" the probability of sampling low frequencies near the origin. This allows it to allocate more samples to higher frequencies, which are undersampled by classical random Fourier features. For simplicity, we focus on the one-dimensional setting. Our results extend to higher dimensions, albeit with an exponential in the dimension loss.

Definition 9 (Improved Fourier Feature Distribution for the Gaussian Kernel)**.** Define the function

$$\bar{\tau}_R(\eta) \equiv \begin{cases} 25\max(R, 3000\log^{1.5} n_\lambda) & |\eta| \le 10\sqrt{\log(n_\lambda)} \\ p(\eta) n_\lambda & \text{o.w.} \end{cases}$$

Let $s_{\bar{\tau}_R} = \int_{\mathbb{R}} \bar{\tau}_R(\eta) d\eta$ and define the probability density function $\bar{p}_R(\eta) = \bar{\tau}_R(\eta)/s_{\bar{\tau}_R}$.

Note that $\bar{p}_R(\eta)$ is just the uniform distribution for low frequencies with $|\eta| \le 10\sqrt{\log(n_\lambda)}$, and the classic Fourier features distribution, appropriately scaled, outside this range. As we show in §7, $\bar{\tau}_R(\eta)$ upper bounds the true ridge leverage function $\tau_\lambda(\eta)$ for all η. Hence, simply applying Lemma 6:

Theorem 10. *For any integer n and parameter $0 < \lambda \le \frac{n}{2}$, consider the one dimensional Gaussian kernel with $\sigma = (2\pi)^{-1}$ (so $p(\eta) = \frac{1}{\sqrt{2\pi}} e^{-\eta^2/2}$) and any dataset of n points $\{x_j\}_{j=1}^n \subseteq [-R, R]$ with any radius $R > 0$. If we sample $s \ge \frac{8}{3}\Delta^{-2} s_{\bar{\tau}_R} \ln(16 s_{\bar{\tau}_R}/\rho)$ random Fourier features according to $\bar{p}_R(\cdot)$ and construct $\mathbf{Z}$ according to (5), then with probability at least $1-\rho$, $\mathbf{Z}\mathbf{Z}^* + \lambda \mathbf{I}_n$ is Δ-spectral approximation of $\mathbf{K} + \lambda \mathbf{I}_n$ for any $\Delta \le 1/2$ and $\rho \in (0,1)$. Furthermore, $s_{\bar{\tau}_R} = O(R\sqrt{\log(n_\lambda)} + \log^2 n_\lambda)$ and $\bar{p}_R(\cdot)$ can be sampled from in $O(1)$ time.*

Theorem 10 represents a possibly exponential improvement over the bound obtainable by classic random Fourier features. For $R \ge \log^{1.5}(n_\lambda)$ our modified distribution requires $O(R\sqrt{\log(n_\lambda)})$ samples, as compared to the lower bound of $\frac{n_\lambda}{400}$ given by Theorem 8.

7. Bounding the Ridge Leverage Function

We conclude by discussing our approach to bounding the ridge leverage function of the Gaussian kernel, which leads to Theorems 8 and 10. The key idea is to reformulate the leverage function as the solution of two dual optimization problems. By exhibiting suitable test functions for these optimization problems, we are able to give both upper and lower bounds on the ridge leverage function, and correspondingly on the sampling performance of classic and modified Fourier feature sampling.

7.1. Primal-Dual Characterization

In this section we prove two alternative characterizations of the ridge leverage function: one as a minimization, and the other as a maximization. These characterization are useful for bounding the leverage function, as we exhibit in the next subsection for the Gaussian kernel.

Define the operator $\mathbf{\Phi} : L_2(d\mu) \to \mathbb{C}^n$ by

$$\mathbf{\Phi} y \equiv \int_{\mathbb{R}^d} \mathbf{z}(\boldsymbol{\xi}) y(\boldsymbol{\xi}) d\mu(\boldsymbol{\xi}). \tag{10}$$

The following two lemmas constitute the main result of this subsection. The proofs can be found in the supplementary material (Appendix E).

Lemma 11. *The ridge leverage function can alternatively be defined as follows:*

$$\tau_\lambda(\boldsymbol{\eta}) = \min_{y \in L_2(d\mu)} \lambda^{-1} \|\mathbf{\Phi} y - \sqrt{p(\boldsymbol{\eta})}\mathbf{z}(\boldsymbol{\eta})\|_2^2 + \|y\|_{L_2(d\mu)}^2 \tag{11}$$

Lemma 12. *The ridge leverage function can alternatively be defined as follows:*

$$\tau_\lambda(\boldsymbol{\eta}) = \max_{\boldsymbol{\alpha} \in \mathbb{C}^n} \frac{p(\boldsymbol{\eta}) \cdot |\boldsymbol{\alpha}^* \mathbf{z}(\boldsymbol{\eta})|^2}{\|\mathbf{\Phi}^* \boldsymbol{\alpha}\|_{L_2(d\mu)}^2 + \lambda \|\boldsymbol{\alpha}\|_2^2} \tag{12}$$

Similar results are well known for the finite dimensional case. Here we extend these results to an infinite dimensional case. Lemma 11 allows us to upper bound the leverage function at any point $\boldsymbol{\eta} \in \mathbb{R}^d$ by exhibiting a carefully constructed function $y(\cdot)$ and upper bounding the ratio in (11), while Lemma 12 allows us to lower bound it in a similar fashion.

7.2. Leverage Function: the Gaussian Case

In this section we prove nearly matching bounds on the leverage score function for the one-dimensional Gaussian kernel on bounded datasets. For simplicity of presentation we focus on the one-dimensional setting. Our results extend to higher dimensions, albeit with an exponential in the dimension loss in the gap between upper and lower bounds.

Our bounds are parameterized by the width of the point set, which we denote by R. To reduce clutter, we present all results for fixed $\sigma = (2\pi)^{-1}$. This is without loss of generality since we can rescale the points. All the proofs appear in the supplementary material (Appendices F–H).

Theorem 13. *Consider the one dimensional Gaussian kernel with $\sigma = (2\pi)^{-1}$. For any integer n and parameter $0 < \lambda \le \frac{n}{2}$, and any radius $R > 0$, if $x_1, ..., x_n \in [-R, R]$, for every $|\eta| \le 10\sqrt{\log n_\lambda}$:*

$$\tau_\lambda(\eta) \le 25\max(R, 3000\log^{1.5} n_\lambda)\,.$$

Theorem 14. *Consider the one dimensional Gaussian kernel with $\sigma = (2\pi)^{-1}$. For any integer $n \geq 17$, any parameter $\frac{10}{n} \leq \lambda \leq \frac{n}{16}$, and every radius $1000 \log^{1.5} n_\lambda \leq R \leq \frac{n}{500\sqrt{\log(n_\lambda)}}$, there exist $x_1, ..., x_n \in [-R, R]$ such that for every $\eta \in [-100\sqrt{\log n_\lambda}, +100\sqrt{\log n_\lambda}]$ we have:*

$$\tau_\lambda(\eta) \geq \frac{R}{150}\left(\frac{p(\eta)}{p(\eta) + 2Rn_\lambda^{-1}}\right).$$

The last two theorems lead to a tight bound on the statistical dimension matrices corresponding to bounded points sets:

Corollary 15. *Consider the Gaussian kernel with $\sigma = (2\pi)^{-1}$. For any integer n and parameter $0 < \lambda \leq \frac{n}{2}$, and any $R > 0$, if $x_1, ..., x_n \in [-R, R]$ then we have:*

$$\begin{aligned} s_\lambda(\mathbf{K}) &\leq 500 \cdot \max(R, 3000 \log^{1.5} n_\lambda)\sqrt{\log n_\lambda} + 1 \\ &= O(R\sqrt{\log n_\lambda} + \log^2 n_\lambda) \end{aligned}$$

Furthermore, if $1000 \log^{1.5} n_\lambda \leq R \leq \frac{n}{500\sqrt{\log(n_\lambda)}}$ there exists a set of points $x_1, \ldots, x_n \subseteq [-R, R]$ such that:

$$s_\lambda(\mathbf{K}) = \Omega\Big(R\sqrt{\log(n_\lambda/R)}\Big).$$

The bounds above match up to constant factors if $1000 \log^{1.5} n_\lambda \leq R \leq n_\lambda^{0.99}$. For any $1000 \log^{1.5} n_\lambda \leq R \leq \frac{n}{500\sqrt{\log(n_\lambda)}}$ they match up to a $\sqrt{\log n_\lambda}$ factor.

7.3. Theorems 13 and 14: Proof Outline

Lemma 11 allows us to bound $\tau_\lambda(\eta)$ simply by exhibiting any $y(\cdot)$ which makes the cost function small. One simple attempt might be $y_\eta^{(s)}(\xi) = \delta(\eta - \xi)$ where $\delta(\cdot)$ is the Dirac delta function. This choice zeros out the first term. However the delta function is not square integrable, $y_\eta^{(s)} \notin L_2(d\mu)$, so the lemma cannot be used. Another trivial attempt is $y^{(0)}(\xi) = 0$, which zeros out the second term and recovers the trivial bound $\tau_\lambda(\eta) \leq p(\eta)n_\lambda$. Nevertheless, a smarter test functions $y(\cdot)$ can yield improved bounds, yielding results on the leverage score function that are parameterized by the diameter of the point set.

At a high level, our approach is to replace the spike function at η with a 'soft spike' whose Fourier transform still looks approximately like a cosine wave on $[-R, R]$, yet is still square integrable. The smaller R is, the more spread out this function will be able to be, and hence the smaller its ℓ_2 norm, and the better the leverage score bound. A natural candidate for a 'soft spike' is a Gaussian of appropriate variance, but this choice does not suffice to obtain tight bounds, due to two difficulties. First, for the **upper bound** a simple Gaussian does not result in a function that is close enough to a pure frequency in time domain (first term of the objective function in Lemma 11) unless we settle for an upper bound of $O(R \cdot \mathbf{poly}(n_\lambda))$ as opposed to the tight $O(R)$ on the leverage score density function. Second, the **lower bound** on the leverage score function resulting from using a Gaussian pulse would only be of the form $\Omega(R/\sqrt{\log n_\lambda})$, leading to a weak lower bound on the statistical dimension, namely $\Omega(R)$ as opposed to $\Omega(R \cdot \sqrt{\log n_\lambda})$, thereby missing entirely the effect of the regularization parameter λ on the statistical dimension!

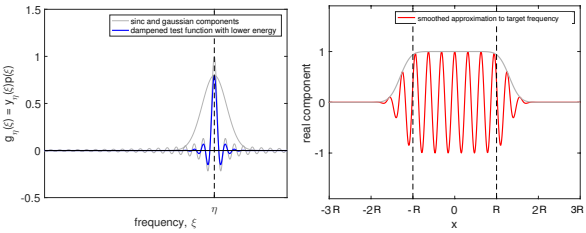

Figure 1. 'Soft spike' function y and its Fourier transform $\mathbf{\Phi}y$, which is approximately a pure cosine wave on $[-R, R]$.

The remedy to the issues above turns out to be the convolution of a (modulated) Gaussian with a rectangular pulse in time domain (product of a shifted Gaussian with the sinc function in frequency domain). Specifically, our bounds are based on variants of a flattened Gaussian spike function

$$y_{\eta,b,v}(\xi) \equiv e^{-(\xi-\eta)^2 b^2/4} \cdot v \cdot \mathbf{sinc}\left(v(\xi - \eta)\right). \qquad (13)$$

for some $b > 0$, $v > 0$ and $\eta \in \mathbb{R}$.

It turns out that with a proper setting of parameters (where one should think of b as large, i.e. the spike y is rather narrow) the function $\mathbf{\Phi}y_{\eta,b,v}$ satisfies

$$(\mathbf{\Phi}y_{\eta,b,v})(x) \quad \approx p(\eta) \cdot \exp(2\pi i \eta x) \int_{x-\frac{v}{2}}^{x+\frac{v}{2}} \frac{1}{\sqrt{2\pi}b} e^{-t^2/2b^2} dt.$$

An illustration of this function in y is given in Fig. 1, (left) and the function $\mathbf{\Phi}y$ in Fig. 1, (right). Note that if the parameter v is chosen to be large, then for x not too large we have $\int_{x-\frac{v}{2}}^{x+\frac{v}{2}} \frac{1}{\sqrt{2\pi}b} e^{-t^2/2b^2} dt \approx \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}b} e^{-t^2/2b^2} dt$, i.e. the second multiplier is essentially constant, i.e. flat as a function of x (hence the term 'flattened Gaussian spike'). This means that $\mathbf{\Phi}y_{\eta,b,v}$ is essentially the kernel density evaluated at η times a pure harmonic term $\exp(2\pi i \eta x)$, which is exactly what one needs to minimize the first term on the rhs of (11) in Lemma 11, up to a factor of $\sqrt{p(\eta)}$ – see Appendix F. One can also see that setting v to be not too large results in a good function to use in the maximization problem in (12) in Lemma 12 – see Appendix G. Obtaining tight bounds and in particular achieving the right dependence on $\sqrt{\log n_\lambda}$ requires several modifications to the function y above, but the intuition we just described works!

Acknowledgements

The authors thank Arturs Backurs helpful discussions at early stages of this project. Haim Avron acknowledges the support from the XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323 and an IBM Faculty Award. Cameron Musco acknowledges the support by NSF Graduate Research Fellowship, AFOSR grant FA9550-13-1-0042 and the NSF Center for Science of Information.

References

Alaoui, Ahmed El and Mahoney, Michael W. Fast randomized kernel ridge regression with statistical guarantees. In *Neural Information Processing Systems (NIPS)*, 2015.

Avron, Haim, Nguyen, Huy, and Woodruff, David. Subspace embeddings for the polynomial kernel. In *Neural Information Processing Systems (NIPS)*, 2014.

Avron, Haim, Clarkson, Kenneth L., and Woodruff, David P. Faster kernel ridge regression using sketching and preconditioning. *CoRR*, abs/1611.03220, 2016. URL http://arxiv.org/abs/1611.03220.

Bach, Francis. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017. URL http://jmlr.org/papers/v18/15-178.html.

Bach, Francis R. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory (COLT)*, 2013. URL http://jmlr.org/proceedings/papers/v30/Bach13.html.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. ISSN 1615-3383. doi: 10.1007/s10208-006-0196-8. URL http://dx.doi.org/10.1007/s10208-006-0196-8.

Cohen, Michael B., Musco, Cameron, and Musco, Christopher. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pp. 1758–1777, Philadelphia, PA, USA, 2017. Society for Industrial and Applied Mathematics. URL http://dl.acm.org/citation.cfm?id=3039686.3039801.

Cutajar, Kurt, Osborne, Michael, Cunningham, John, and Filippone, Maurizio. Preconditioning kernel matrices. In *International Conference on Machine Learning (ICML)*, 2016. URL http://jmlr.org/proceedings/papers/v48/cutajar16.html.

Feller, William. *An introduction to probability theory and its applications. Volume 1*. Wiley series in probability and mathematical statistics. John Wiley & sons, New York, Chichester, Brisbane, 1968. ISBN 0-471-25711-7. URL http://opac.inria.fr/record=b1122219.

Mahoney, Michael W. and Drineas, Petros. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009. doi: 10.1073/pnas.0803205106. URL http://www.pnas.org/content/106/3/697.abstract.

Musco, Cameron and Musco, Christopher. Recursive sampling for the Nyström method. *CoRR*, abs/1605.07583, 2016. URL http://arxiv.org/abs/1605.07583.

Ogawa, Hidemitsu. An operator pseudo-inversion lemma. *SIAM Journal on Applied Mathematics*, 48(6):1527–1531, 1988. doi: 10.1137/0148095. URL http://dx.doi.org/10.1137/0148095.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Neural Information Processing Systems (NIPS)*, 2007.

Rahimi, Ali and Recht, Benjamin. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Neural Information Processing Systems (NIPS)*, 2008.

Rudi, Alessandro, Camoriano, Raffaello, and Rosasco, Lorenzo. Less is more: Nyström computational regularization. In *Neural Information Processing Systems (NIPS)*, 2015.

Rudi, Alessandro, Camoriano, Raffaello, and Rosasco, Lorenzo. Generalization properties of learning with random features. *ArXiv e-prints*, feb 2016.

Tropp, Joel A. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015. ISSN 1935-8237. doi: 10.1561/2200000048. URL http://dx.doi.org/10.1561/2200000048.

Woodruff, David P. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157, October 2014. URL http://dx.doi.org/10.1561/0400000060.

Zhang, Yuchen, Duchi, John, and Wainwright, Martin. Divide and conquer kernel ridge regression: A distributed

algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, 16(1):3299–3340, January 2015. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2789272.2912104.

Random Fourier Features for Kernel Ridge Regression: Approximation Bounds and Statistical Guarantees Appendix: Proofs

A. Preliminaries

Our upper and lower bound analysis relies predominantly on Fourier analysis, so we now introduce some additional notation and state some useful facts about these.

A.1. Properties of Fourier Transforms

Definition 16 (Fourier Transform)**.** The *Fourier transform* of a continuous function $f : \mathbb{R}^d \to \mathbb{C}$ in $L_1(\mathbb{R}^n)$ is defined to be the function $\mathcal{F}f : \mathbb{R}^d \to \mathbb{C}$ as follows:

$$(\mathcal{F}f)(\boldsymbol{\xi}) = \int_{\mathbb{R}^d} f(\mathbf{t}) e^{-2\pi i \mathbf{t}^T \boldsymbol{\xi}} \, d\mathbf{t}.$$

We also sometimes use the notation $\hat{f}$ for the Fourier transform of f. We often informally refer to f as representing the function in *time domain* and $\hat{f}$ as representing the function in *frequency domain*.

The original function f can also be obtained from $\hat{f}$ by the *inverse Fourier transform*:

$$f(\mathbf{t}) = \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\xi}) e^{2\pi i \boldsymbol{\xi}^T \mathbf{t}} \, d\boldsymbol{\xi}$$

Definition 17 (Convolution)**.** The *convolution* of two functions $f : \mathbb{R}^d \to \mathbb{C}$ and $g : \mathbb{R}^d \to \mathbb{C}$ is defined to be the function $(f * g) : \mathbb{R}^d \to \mathbb{C}$ given by

$$(f * g)(\boldsymbol{\eta}) = \int_{\mathbb{R}^d} f(\mathbf{t}) g(\boldsymbol{\eta} - \mathbf{t}) \, d\mathbf{t}.$$

The convolution theorem shows that the Fourier transform of the convolution of two functions is simply the product of the individual Fourier transforms:

Claim 18 (Convolution Theorem)**.** *Given functions $f : \mathbb{R}^d \to \mathbb{C}$ and $g : \mathbb{R}^d \to \mathbb{C}$ whose convolution is $h = f * g$, we have*

$$\hat{h}(\boldsymbol{\xi}) = \hat{f}(\boldsymbol{\xi}) \cdot \hat{g}(\boldsymbol{\xi})$$

for all $\boldsymbol{\xi} \in \mathbb{R}^d$.

Now, suppose $d = 1$, i.e., the functions we consider take inputs in $\mathbb{R}$. We define the *rectangle function* and *normalized sinc function*, which we use extensively in our analysis.

Definition 19 (Rectangle Function)**.** We define the *rectangle function* $\mathrm{rect}_a : \mathbb{R} \to \mathbb{C}$ as

$$\mathrm{rect}_a(x) = \begin{cases} 0 & \text{if } |x| > a/2 \\ \frac{1}{2} & \text{if } |x| = a/2 \\ 1 & \text{if } |x| < a/2 \end{cases}.$$

If $a = 1$, then we often omit the subscript and simply write rect.

Definition 20 (Normalized Sinc Function)**.** We define the *normalized sinc function* $\mathrm{sinc} : \mathbb{R} \to \mathbb{C}$ as

$$\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}.$$

It is well known that the Fourier transform of the rectangle function (with $a = 1$) is the normalized sinc function:

$$\mathcal{F}(\text{rect}) = \text{sinc}$$

We use δ to denote the *Dirac delta function*. Recall that the Dirac delta function satisfies the following useful property for any function f:

$$\int_{-\infty}^{\infty} f(x)\delta(x-a)\,dx = f(a),$$

i.e. the integral of a function multiplied by a shifted Dirac delta functions picks out the value of the function at a particular point. Thus, it is not hard to see that the Fourier transform of a δ is the constant function which is 1 everywhere:

$$(\mathcal{F}\delta)(\xi) = \int_{-\infty}^{\infty} e^{-2\pi i t\xi} \cdot \delta(t)\,dt = e^{-2\pi i \cdot 0 \cdot \xi} = 1$$

for all ξ. Similarly, the Fourier transform of a shifted delta function is as follows:

$$(\mathcal{F}\delta(\cdot - a))(\xi) = \int_{-\infty}^{\infty} e^{-2\pi i t\xi} \cdot \delta(t-a)\,dt = e^{-2\pi i a\xi}.$$

Moreover, it is not hard to see that convolving a function by a shifted delta funciton results in a shift of the original function:

$$(f * \delta(\cdot - a))(x) = f(x-a).$$

Thus, by the convolution theorem, we obtain the following identity:

Claim 21. *Given a function $f : \mathbb{R} \to \mathbb{C}$, we have*

$$(\mathcal{F}f(\cdot - a))(\xi) = (\mathcal{F}(f * \delta(\cdot - a)))(\xi) = \hat{f}(\xi) \cdot e^{-2\pi i a\xi}.$$

Similarly,

Claim 22. *Given a function $f : \mathbb{R} \to \mathbb{C}$, we have*

$$(\mathcal{F}(f(x) \cdot e^{2\pi i a x}))(\xi) = \hat{f}(\xi - a).$$

Finally, we introduce a useful function known as the *Dirac comb function*:

Definition 23. The *Dirac comb function* with period T is defined as f satisfying

$$f(x) = \sum_{j=-\infty}^{\infty} \delta(x - jT).$$

It is a standard fact that the Fourier transform of a Dirac comb function is another Dirac comb function which is scaled and has the inverse period:

Claim 24. *Let*

$$f(x) = \sum_{j=-\infty}^{\infty} \delta(x - jT)$$

be the Dirac comb function with period T. Then,

$$(\mathcal{F}f)(\xi) = \frac{1}{T} \sum_{j=-\infty}^{\infty} \delta\left(\xi - \frac{j}{T}\right).$$

We use the Dirac comb function in our lower bound constructions.

A.2. Properties of Gaussian Distributions

We also need several useful facts about Gaussian distributions. The following is a standard fact about the cumulative distribution function of the standard Gaussian distribution:

Claim 25 ((Feller, 1968))**.** *For any $x > 0$, we have*

$$\frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\, dt \leq \frac{e^{-x^2/2}}{x\sqrt{2\pi}}.$$

Moreover, as a direct consequence, for any $\sigma, x > 0$, we have that

$$\frac{1}{\sqrt{2\pi}\sigma} \int_x^\infty e^{-t^2/2\sigma^2}\, dt \leq \frac{\sigma e^{-x^2/2\sigma^2}}{x\sqrt{2\pi}}.$$

Also, if $x \geq 1$, then

$$\left(\frac{1}{x} - \frac{1}{x^3}\right) \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \leq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2}\, dt.$$

We also need the following property about Gaussian samples.

Claim 26. *Let $t \geq 10$, and $a_1, a_2, \ldots, a_t$ be sampled according to the Gaussian distribution given by probability density function $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Also, let $a^* = \max_{1 \leq j \leq t} |a_j|$. Then,*

$$\Pr\left[\frac{1}{\sqrt{2\pi}} e^{-a^{*2}/2} \leq \frac{8\sqrt{\log t}}{t}\right] \geq \frac{1}{2}.$$

Proof. Choose q_1 such that

$$\int_{q_1}^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \frac{1}{t}. \tag{14}$$

Note that by Claim 25, we have

$$\int_{2\sqrt{\log t}}^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \leq \frac{1}{2\sqrt{2\pi} t^2 \sqrt{\log t}} \leq \frac{1}{t}.$$

Thus, $q_1 \leq 2\sqrt{\log t}$.

Also, since $\frac{1}{t} \leq \frac{1}{4}$, we have that $q_1 \geq \frac{6}{5}$. Thus, by another application of Claim 25,

$$\frac{1}{t} = \int_{q_1}^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \geq \left(\frac{1}{q_1} - \frac{1}{q_1^3}\right) \frac{1}{\sqrt{2\pi}} e^{-q_1^2/2} \geq \frac{1}{4q_1} \cdot \frac{1}{\sqrt{2\pi}} e^{-q_1^2/2},$$

and so,

$$\frac{1}{\sqrt{2\pi}} e^{-q_1^2/2} \leq \frac{4q_1}{t} \leq \frac{8\sqrt{\log t}}{t}.$$

Therefore,

$$\begin{aligned}
\Pr\left[\frac{1}{\sqrt{2\pi}} e^{-a^{*2}/2} \leq \frac{8\sqrt{\log t}}{t}\right] &\geq \Pr[a^* \geq q_1] \\
&= 1 - \left(1 - \frac{1}{t}\right)^t \\
&\geq 1 - \frac{1}{e} \\
&\geq \frac{1}{2},
\end{aligned}$$

as desired. □

B. Proof of Lemma 2

Note that $\mathbf{A} \preceq \mathbf{B}$ implies that $\mathbf{B}^{-1} \preceq \mathbf{A}^{-1}$ so for the bias term we have:

$$\mathbf{f}^{\mathrm{T}}(\tilde{\mathbf{K}} + \lambda \mathbf{I}_n)^{-1}\mathbf{f} \leq (1-\Delta)^{-1}\mathbf{f}^{\mathrm{T}}(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{f}. \tag{15}$$

We now consider the variance term. Denote $s = \mathbf{rank}(\tilde{\mathbf{K}})$, and let $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ denote the eigenvalues of a matrix $\mathbf{A}$. We have:

$$\begin{aligned}
s_\lambda(\tilde{\mathbf{K}}) = \mathbf{Tr}\left((\tilde{\mathbf{K}} + \lambda \mathbf{I}_n)^{-1}\tilde{\mathbf{K}}\right) &= \sum_{i=1}^{s} \frac{\lambda_i(\tilde{\mathbf{K}})}{\lambda_i(\tilde{\mathbf{K}}) + \lambda} \\
&= s - \sum_{i=1}^{s} \frac{\lambda}{\lambda_i(\tilde{\mathbf{K}}) + \lambda} \\
&\leq s - (1+\Delta)^{-1} \sum_{i=1}^{s} \frac{\lambda}{\lambda_i(\mathbf{K}) + \lambda} \\
&= s - \sum_{i=1}^{s} \frac{\lambda}{\lambda_i(\mathbf{K}) + \lambda} + \frac{\Delta}{1+\Delta} \sum_{i=1}^{s} \frac{\lambda}{\lambda_i(\mathbf{K}) + \lambda} \\
&\leq n - \sum_{i=1}^{n} \frac{\lambda}{\lambda_i(\mathbf{K}) + \lambda} + \frac{\Delta \cdot s}{1+\Delta} \\
&= s_\lambda(\mathbf{K}) + \frac{\Delta \cdot s}{1+\Delta} \\
&\leq (1-\Delta)^{-1} s_\lambda(\mathbf{K}) + \frac{\Delta \cdot s}{1+\Delta}
\end{aligned}$$

where we use the fact that $\mathbf{A} \preceq \mathbf{B}$ implies that $\lambda_i(\mathbf{A}) \leq \lambda_i(\mathbf{B})$ (this is a simple consequence of the Courant-Fischer minimax theorem).

Combining the above variance bound with the bias bound in (15) yields:

$$\widehat{\mathcal{R}}_{\tilde{\mathbf{K}}}(\mathbf{f}) \leq (1-\Delta)^{-1}\widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f}) + \frac{\Delta}{(1+\Delta)} \cdot \frac{\mathbf{rank}(\tilde{\mathbf{K}})}{n} \cdot \sigma_\nu^2$$

and the bound $\mathcal{R}(\tilde{f}) \leq \widehat{\mathcal{R}}_{\tilde{\mathbf{K}}}(\mathbf{f})$ completes the proof.

C. Proof of Proposition 4

Since k is positive definite and $k(0) = 1$, $|k(\mathbf{x}, \mathbf{z})| \leq 1$ for all $\mathbf{x}$ and $\mathbf{z}$. This implies that the maximum eigenvalue of $\mathbf{K}$ is bounded by n, and the lower bound follows immediately. The upper bound on $\tau_\lambda(\boldsymbol{\eta})$ follows from the fact that $\|\mathbf{z}(\boldsymbol{\eta})\|_2^2 = n$ and all eigenvalues of $\mathbf{K} + \lambda \mathbf{I}_n$ are bounded from below by λ. The bound also establishes that the integral converges. We now have,

$$\begin{aligned}
\int_{\mathbb{R}^d} \tau_\lambda(\boldsymbol{\eta}) d\boldsymbol{\eta} &= \int_{\mathbb{R}^d} p(\boldsymbol{\eta}) \mathbf{z}(\boldsymbol{\eta})^* (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{z}(\boldsymbol{\eta}) d\boldsymbol{\eta} \\
&= \int_{\mathbb{R}^d} \mathbf{Tr}\left(p(\boldsymbol{\eta})(\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{z}(\boldsymbol{\eta}) \mathbf{z}(\boldsymbol{\eta})^*\right) d\boldsymbol{\eta} \\
&= \mathbf{Tr}\left(\int_{\mathbb{R}^d} p(\boldsymbol{\eta})(\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{z}(\boldsymbol{\eta}) \mathbf{z}(\boldsymbol{\eta})^* d\boldsymbol{\eta}\right) \\
&= \mathbf{Tr}\left((\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \int_{\mathbb{R}^d} p(\boldsymbol{\eta}) \mathbf{z}(\boldsymbol{\eta}) \mathbf{z}(\boldsymbol{\eta})^* d\boldsymbol{\eta}\right) \\
&= \mathbf{Tr}\left((\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{K}\right) = s_\lambda(\mathbf{K}).
\end{aligned}$$

The second equality is due to the fact that $\mathbf{z}(\eta)$ is a rank one matrix, and third equality is due to linearity of the trace operation and the fact that all diagonal entries are positive.

D. Proof of Lemma 6 and Theorem 7

To prove Lemma 6 we need the following lemma which is essentially a restatement of Corollary 7.3.3 from (Tropp, 2015). However, the minimum t in the following statement is much lower than the bound that appears in (Tropp, 2015) which is unnecessarily loose (possibly, a typo in (Tropp, 2015)). For completeness, we include a proof.

Lemma 27. *Let $\mathbf{B}$ be a fixed $d_1 \times d_2$ matrix. Construct a $d_1 \times d_2$ random matrix $\mathbf{R}$ that satisfies*

$$\mathbb{E}[\mathbf{R}] = \mathbf{B} \quad \textit{and} \quad \|\mathbf{R}\|_2 \leq L.$$

Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be semidefinite upper bounds for the expected squares:

$$\mathbb{E}[\mathbf{R}\mathbf{R}^*] \preceq \mathbf{M}_1 \quad \textit{and} \quad \mathbb{E}[\mathbf{R}^*\mathbf{R}] \preceq \mathbf{M}_2.$$

Define the quantities

$$m = \max(\|\mathbf{M}_1\|_2, \|\mathbf{M}_2\|_2) \quad \textit{and} \quad d = (\mathbf{Tr}(\mathbf{M}_1) + \mathbf{Tr}(\mathbf{M}_2))/m.$$

Form the matrix sampling estimator

$$\bar{\mathbf{R}}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{R}_k$$

where each $\mathbf{R}_k$ is an independent copy of $\mathbf{R}$. Then, for all $t \geq \sqrt{m/n} + 2L/3n$,

$$\Pr(\|\bar{\mathbf{R}}_n - \mathbf{B}\|_2 \geq t) \leq 4d \exp\left(\frac{-nt^2/2}{m + 2Lt/3}\right). \tag{16}$$

Proof. The proof mirrors the proof of Corollary 6.2.1 in (Tropp, 2015), using Theorem 7.3.1 instead of Theorem 6.1.1 (both from (Tropp, 2015)).

Since $\mathbb{E}[\mathbf{R}] = \mathbf{B}$, we can write

$$\mathbf{Z} \equiv \bar{\mathbf{R}}_n - \mathbf{B} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{R}_k - \mathbb{E}[\mathbf{R}]) = \sum_{k=1}^{n}\mathbf{S}_k,$$

where we have define $\mathbf{S}_k \equiv n^{-1}(\mathbf{R}_k - \mathbb{E}[\mathbf{R}])$. These random matrices are i.i.d and each has zero mean.

Now, we can bound each of the summands:

$$\|\mathbf{S}_k\|_2 \leq \frac{1}{n}(\|\mathbf{R}_k\|_2 + \|\mathbb{E}[\mathbf{R}]\|_2) \leq \frac{1}{n}(\|\mathbf{R}_k\|_2 + \mathbb{E}[\|\mathbf{R}\|_2]) \leq \frac{2L}{n},$$

where the first inequality is the triangle inequality and the second is Jensen's inequality.

To find semidefinite upper bounds $\mathbf{V}_1$ and $\mathbf{V}_2$ on the matrix-valued variances we note that

$$\begin{aligned}
\mathbb{E}[\mathbf{S}_1\mathbf{S}_1^*] &= n^{-2}\mathbb{E}[(\mathbf{R} - \mathbb{E}[\mathbf{R}])(\mathbf{R} - \mathbb{E}[\mathbf{R}])^*] \\
&= n^{-2}\left(\mathbb{E}[\mathbf{R}\mathbf{R}^*] - \mathbb{E}[\mathbf{R}]\,\mathbb{E}[\mathbf{R}]^*\right) \\
&\preceq n^{-2}\mathbb{E}[\mathbf{R}\mathbf{R}^*].
\end{aligned}$$

Likewise, $\mathbb{E}[\mathbf{S}_1^*\mathbf{S}_1] \preceq n^{-2}\mathbb{E}[\mathbf{R}^*\mathbf{R}]$. Since the summands are i.i.d, if we define $\mathbf{V}_1 \equiv n^{-1}\mathbf{M}_1$ and $\mathbf{V}_2 \equiv n^{-1}\mathbf{M}_2$, we have $\mathbb{E}[\mathbf{Z}\mathbf{Z}^*] \preceq \mathbf{V}_1$ and $\mathbb{E}[\mathbf{Z}^*\mathbf{Z}] \preceq \mathbf{V}_2$.

We now calculate,

$$\nu \equiv \max(\|\mathbf{V}_1\|_2, \|\mathbf{V}_2\|_2) = \frac{m}{n}$$

and

$$\frac{\mathbf{Tr}(\mathbf{V}_1) + \mathbf{Tr}(\mathbf{V}_2)}{\max(\|\mathbf{V}_1\|_2, \|\mathbf{V}_2\|_2)} = d.$$

Noticing, that the condition $t \geq \sqrt{m/n} + 2L/3n$ meets the required lower bound in Theorem 7.3.1 in (Tropp, 2015) we can now apply this theorem, which along with the above calculations translates to (16). □

We can now prove Lemma 6.

Proof of Lemma 6. Let $\mathbf{K}+\lambda\mathbf{I}_n = \mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^2\mathbf{V}$ be an eigendecomposition of $\mathbf{K}+\lambda\mathbf{I}_n$. Note that the Δ-spectral approximation guarantee (2) is equivalent to

$$\mathbf{K} - \Delta(\mathbf{K}+\lambda\mathbf{I}_n) \preceq \mathbf{Z}\mathbf{Z}^* \preceq \mathbf{K} + \Delta(\mathbf{K}+\lambda\mathbf{I}_n)\,,$$

so by multiplying by $\mathbf{\Sigma}^{-1}\mathbf{V}$ on the left and $\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}$ on the right we find that it suffices to show that

$$\|\mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{Z}\mathbf{Z}^*\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{K}\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\|_2 \leq \Delta \tag{17}$$

holds with probability of at least $1-\rho$. Let

$$\mathbf{Y}_l = \frac{p(\boldsymbol{\eta}_l)}{p_{\tilde{\tau}}(\boldsymbol{\eta}_l)}\mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{z}(\boldsymbol{\eta}_l)\mathbf{z}(\boldsymbol{\eta}_l)^*\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\,.$$

Note that $\mathbb{E}\left[\mathbf{Y}_l\right] = \mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{K}\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}$ and $\frac{1}{s}\sum_{l=1}^{s}\mathbf{Y}_l = \mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{Z}\mathbf{Z}^*\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}$. Thus, we can use the matrix concentration result above to prove (17).

To apply this bound we need to bound the norm of $\mathbf{Y}_l$ and the stable rank $\mathbb{E}\left[\mathbf{Y}_l^2\right]$. Since $\mathbf{Y}_l$ is always a rank one matrix we have

$$\begin{aligned}
\|\mathbf{Y}_l\|_2 &= \frac{p(\boldsymbol{\eta}_l)}{p_{\tilde{\tau}}(\boldsymbol{\eta}_l)}\mathbf{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{z}(\boldsymbol{\eta}_l)\mathbf{z}(\boldsymbol{\eta}_l)^*\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\right)\\
&= \frac{p(\boldsymbol{\eta}_l)}{p_{\tilde{\tau}}(\boldsymbol{\eta}_l)}\mathbf{z}(\boldsymbol{\eta}_l)^*\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{z}(\boldsymbol{\eta}_l)\\
&= \frac{p(\boldsymbol{\eta}_l)}{p_{\tilde{\tau}}(\boldsymbol{\eta}_l)}\mathbf{z}(\boldsymbol{\eta}_l)^*(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta}_l)\\
&= \frac{s_{\tilde{\tau}}\cdot\tau(\boldsymbol{\eta}_l)}{\tilde{\tau}(\boldsymbol{\eta}_l)} \leq s_{\tilde{\tau}}
\end{aligned}$$

since $\tilde{\tau}(\boldsymbol{\eta}_l) \geq \tau(\boldsymbol{\eta}_l)$. We also have

$$\begin{aligned}
\mathbf{Y}_l^2 &= \frac{p(\boldsymbol{\eta}_l)^2}{p_{\tilde{\tau}}(\boldsymbol{\eta}_l)^2}\mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{z}(\boldsymbol{\eta}_l)\mathbf{z}(\boldsymbol{\eta}_l)^*\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{z}(\boldsymbol{\eta}_l)\mathbf{z}(\boldsymbol{\eta}_l)^*\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\\
&= \frac{p(\boldsymbol{\eta}_l)^2}{p_{\tilde{\tau}}(\boldsymbol{\eta}_l)^2}\mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{z}(\boldsymbol{\eta}_l)\mathbf{z}(\boldsymbol{\eta}_l)^*(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta}_l)^*\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\\
&= \frac{p(\boldsymbol{\eta}_l)\tau(\boldsymbol{\eta}_l)}{p_{\tilde{\tau}}(\boldsymbol{\eta}_l)^2}\mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{z}(\boldsymbol{\eta}_l)\mathbf{z}(\boldsymbol{\eta}_l)^*\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\\
&= \frac{\tau(\boldsymbol{\eta}_l)}{p_{\tilde{\tau}}(\boldsymbol{\eta}_l)}\mathbf{Y}_l\\
&= \frac{s_{\tilde{\tau}}\tau(\boldsymbol{\eta}_l)}{\tilde{\tau}(\boldsymbol{\eta}_l)}\mathbf{Y}_l \preceq s_{\tilde{\tau}}\mathbf{Y}_l
\end{aligned}$$

Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of $\mathbf{K}$. We have

$$\begin{aligned}
\mathbb{E}\left[s_{\tilde{\tau}}\mathbf{Y}_l\right] &= s_{\tilde{\tau}}\mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{K}\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\\
&= s_{\tilde{\tau}}\left(\mathbf{I}_n - \lambda\mathbf{\Sigma}^{-2}\right)\\
&= s_{\tilde{\tau}}\cdot\mathbf{diag}\left(\lambda_1/(\lambda_1+\lambda),\ldots,\lambda_n/(\lambda_n+\lambda)\right) := \mathbf{D}\,.
\end{aligned}$$

So,

$$\begin{aligned}\Pr\left(\left\|\frac{1}{s}\sum_{l=1}^{s}\mathbf{Y}_l-\boldsymbol{\Sigma}^{-1}\mathbf{V}\mathbf{K}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\right\|_2\geq\Delta\right) &\leq \frac{8\mathbf{Tr}\left(\mathbf{D}\right)}{\|\mathbf{D}\|_2}\exp\left(\frac{-s\Delta^2/2}{\|\mathbf{D}\|_2+2s_{\tilde{\tau}}\Delta/3}\right)\\ &\leq 8\frac{s_{\tilde{\tau}}\cdot s_\lambda(\mathbf{K})}{s_{\tilde{\tau}}\cdot\lambda_1/(\lambda_1+\lambda)}\exp\left(\frac{-s\Delta^2}{2s_{\tilde{\tau}}(1+2\Delta/3)}\right)\\ &\leq 16s_\lambda(\mathbf{K})\exp\left(\frac{-s\Delta^2}{2s_{\tilde{\tau}}(1+2\Delta/3)}\right)\\ &\leq 16s_\lambda(\mathbf{K})\exp\left(\frac{-3s\Delta^2}{8s_{\tilde{\tau}}}\right)\leq\rho\end{aligned}$$

where the third inequality is due to the assumption that $\lambda_1=\|\mathbf{K}\|_2\geq\lambda$ and the last inequality is due to the bound on s. □

Proof of Theorem 7. Define $\tilde{\tau}(\boldsymbol{\eta})=p(\boldsymbol{\eta})\cdot n_\lambda$ and note that $\tilde{\tau}(\boldsymbol{\eta})\geq\tau_\lambda(\boldsymbol{\eta})$ by Proposition 4 and that $s_{\tilde{\tau}}=n_\lambda$. Finally, note that $p_{\tilde{\tau}}(\boldsymbol{\eta})=p(\boldsymbol{\eta})$, the classic Fourier features sampling probability. □

E. Proof of Lemmas 11 and 12

Let $R(\boldsymbol{\Phi})\subseteq\mathbb{C}^n$ denote the range of $\boldsymbol{\Phi}$. Here we first prove that the operator $\boldsymbol{\Phi}$ is defined on all $L_2(d\mu)$ and is a bounded linear operator. Indeed, for $y\in L_2(d\mu)$ we have:

$$\begin{aligned}\|\boldsymbol{\Phi}y\|_2^2 &= \left\|\int_{\mathbb{R}^d}\mathbf{z}(\boldsymbol{\xi})y(\boldsymbol{\xi})d\mu(\boldsymbol{\xi})\right\|_2^2\\ &\text{(by Jensen's inequality)}\\ &\leq \int_{\mathbb{R}^d}\|\mathbf{z}(\boldsymbol{\xi})y(\boldsymbol{\xi})\|_2^2 d\mu(\boldsymbol{\xi})\\ &= \int_{\mathbb{R}^d}\|y(\xi)\|_2^2\cdot\|\mathbf{z}(\boldsymbol{\xi})\|_2^2 d\mu(\boldsymbol{\xi})\\ &= n\cdot\|y\|_{L_2(d\mu)}^2.\end{aligned}$$

Thus, $R(\boldsymbol{\Phi})$ is a linear subspace of $\mathbb{C}^n$. Therefore, there is a unique adjoint operator $\boldsymbol{\Phi}^*:R(\boldsymbol{\Phi})\to L_2(d\mu)$, such that $\langle\boldsymbol{\Phi}y,\mathbf{x}\rangle_{\mathbb{C}^n}=\langle y,\boldsymbol{\Phi}^*\mathbf{x}\rangle_{L_2(d\mu)}$ for every $y\in L_2(d\mu)$ and $\mathbf{x}\in R(\boldsymbol{\Phi})$. It is easy to verify that $(\boldsymbol{\Phi}^*\mathbf{x})(\boldsymbol{\eta})=\mathbf{z}(\boldsymbol{\eta})^*\mathbf{x}$. We now have the following:

Proposition 28.

$$\boldsymbol{\Phi}\boldsymbol{\Phi}^*=\mathbf{K}$$

Proof. We have that for every $\mathbf{x}\in\mathbb{C}^n$,

$$\begin{aligned}\boldsymbol{\Phi}\boldsymbol{\Phi}^*\mathbf{x} &= \int_{\mathbb{R}^d}\mathbf{z}(\boldsymbol{\xi})(\boldsymbol{\Phi}^*\mathbf{x})(\boldsymbol{\xi})d\mu(\boldsymbol{\xi})\\ &= \int_{\mathbb{R}^d}\mathbf{z}(\boldsymbol{\xi})\mathbf{z}(\boldsymbol{\xi})^*\mathbf{x}d\mu(\boldsymbol{\xi})\\ &= \left(\int_{\mathbb{R}^d}\mathbf{z}(\boldsymbol{\xi})\mathbf{z}(\boldsymbol{\xi})^*d\mu(\boldsymbol{\xi})\right)\mathbf{x}=\mathbf{K}\mathbf{x}\end{aligned}$$

so $\boldsymbol{\Phi}\boldsymbol{\Phi}^*=\mathbf{K}$. □

We are now ready to prove the two lemmas.

Proof of Lemma 11. The minimizer of the right-hand side of (11) can be obtained from the usual normal equations, and simplified using the matrix inversion lemma for operators (Ogawa, 1988):

$$\begin{aligned} y^\star &= \sqrt{p(\boldsymbol{\eta})}(\boldsymbol{\Phi}^*\boldsymbol{\Phi} + \lambda \mathbf{I}_{L_2(d\mu)})^{-1}\boldsymbol{\Phi}^*\mathbf{z}(\boldsymbol{\eta}) \\ &= \sqrt{p(\boldsymbol{\eta})}\boldsymbol{\Phi}^*(\boldsymbol{\Phi}\boldsymbol{\Phi}^* + \lambda \mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta}) \\ &= \sqrt{p(\boldsymbol{\eta})}\boldsymbol{\Phi}^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta})\,. \end{aligned}$$

So, $y^\star(\xi) = \sqrt{p(\boldsymbol{\eta})}\mathbf{z}(\xi)^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta})$. We now have

$$\begin{aligned} \|y^\star\|^2_{L_2(d\mu)} &= p(\boldsymbol{\eta}) \int_{\mathbb{R}^d} |\mathbf{z}(\boldsymbol{\xi})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta})|^2 d\mu(\boldsymbol{\xi}) \\ &= p(\boldsymbol{\eta}) \int_{\mathbb{R}^d} \mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{z}(\xi)\mathbf{z}(\boldsymbol{\xi})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta}) d\mu(\boldsymbol{\xi}) \\ &= p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \left(\int_{\mathbb{R}^d} \mathbf{z}(\boldsymbol{\xi})\mathbf{z}(\boldsymbol{\xi})^* d\mu(\boldsymbol{\xi}) \right) (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta}) \\ &= p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{K}(\mathbf{K} + \lambda I_n)^{-1}\mathbf{z}(\boldsymbol{\eta}) \\ &= p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}(\mathbf{K} + \lambda I_n - \lambda I_n)(\mathbf{K} + \lambda I_n)^{-1}\mathbf{z}(\boldsymbol{\eta}) \\ &= p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta}) - \lambda p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-2}\mathbf{z}(\boldsymbol{\eta}) \end{aligned}$$

and

$$\begin{aligned} \|\boldsymbol{\Phi}y^\star - \sqrt{p(\boldsymbol{\eta})}\mathbf{z}(\boldsymbol{\eta})\|_2^2 &= p(\boldsymbol{\eta})\|\boldsymbol{\Phi}\boldsymbol{\Phi}^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta}) - \mathbf{z}(\boldsymbol{\eta})\|_2^2 \\ &= p(\boldsymbol{\eta})\|(\mathbf{K}(\mathbf{K} + \lambda \mathbf{I}_n)^{-1} - \mathbf{I}_n)\mathbf{z}(\boldsymbol{\eta})\|_2^2 \\ &= p(\boldsymbol{\eta})\|\big((\mathbf{K} + \lambda I_n - \lambda I_n)(\mathbf{K} + \lambda \mathbf{I}_n)^{-1} - \mathbf{I}_n\big)\mathbf{z}(\boldsymbol{\eta})\|_2^2 \\ &= p(\boldsymbol{\eta})\|\big(\lambda(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\big)\mathbf{z}(\boldsymbol{\eta})\|_2^2 \\ &= \lambda^2 p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-2}\mathbf{z}(\boldsymbol{\eta})\,. \end{aligned}$$

Now plugging these into (11) gives:

$$\begin{aligned} &\|y^\star\|^2_{L_2(d\mu)} + \lambda^{-1}\|\boldsymbol{\Phi}y^\star - \sqrt{p(\boldsymbol{\eta})}\mathbf{z}(\boldsymbol{\eta})\|_2^2 \\ &\quad = p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta}) - \lambda p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-2}\mathbf{z}(\boldsymbol{\eta}) \\ &\quad + \lambda p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-2}\mathbf{z}(\boldsymbol{\eta}) \\ &\quad = p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta}) \\ &\quad = \tau_\lambda(\boldsymbol{\eta}) \end{aligned}$$

□

Proof of Lemma 12. The optimization problem (11) can equivalently be reformulated as the following problem:

$$\begin{aligned} \tau(\boldsymbol{\eta}) = \text{minimum} \quad & \|y\|^2_{L_2(d\mu)} + \|\mathbf{u}\|_2^2 \\ y \in L_2(d\mu); \quad & \mathbf{u} \in \mathbb{C}^n \\ \text{subject to:} \quad & \boldsymbol{\Phi}y + \sqrt{\lambda}\mathbf{u} = \sqrt{p(\boldsymbol{\eta})}\mathbf{z}(\boldsymbol{\eta}) \end{aligned}$$

First we show that for any $\boldsymbol{\alpha} \in \mathbb{C}^n$, the argument of the minimization problem in (12) is no bigger than $\tau_\lambda(\boldsymbol{\eta})$. That is because for the optimal solution to above optimization, namely $\bar{\mathbf{u}}$ and $\bar{y}$, we have:

$$\boldsymbol{\Phi}\bar{y} + \sqrt{\lambda}\bar{\mathbf{u}} = \sqrt{p(\boldsymbol{\eta})}\mathbf{z}(\boldsymbol{\eta})$$

hence,

$$\begin{aligned}
|\sqrt{p(\boldsymbol{\eta})}\boldsymbol{\alpha}^*\mathbf{z}(\boldsymbol{\eta})| &= |\boldsymbol{\alpha}^*(\boldsymbol{\Phi}\bar{y} + \sqrt{\lambda}\bar{\mathbf{u}})| \\
&= |\boldsymbol{\alpha}^*\boldsymbol{\Phi}\bar{y} + \boldsymbol{\alpha}^*\sqrt{\lambda}\bar{\mathbf{u}}| \\
&\leq |\boldsymbol{\alpha}^*\boldsymbol{\Phi}\bar{y}| + |\boldsymbol{\alpha}^*\sqrt{\lambda}\bar{\mathbf{u}}| \\
&= |\langle \boldsymbol{\alpha}, \boldsymbol{\Phi}\bar{y}\rangle_{\mathbb{C}^n}| + |\boldsymbol{\alpha}^*\sqrt{\lambda}\bar{\mathbf{u}}| \\
&= |\langle \boldsymbol{\Phi}^*\boldsymbol{\alpha}, \bar{y}\rangle_{L_2(d\mu)}| + |\boldsymbol{\alpha}^*\sqrt{\lambda}\bar{\mathbf{u}}| \\
&\leq \|\boldsymbol{\Phi}^*\boldsymbol{\alpha}\|_{L_2(d\mu)} \cdot \|\bar{y}\|_{L_2(d\mu)} + \sqrt{\lambda}\|\boldsymbol{\alpha}^*\|_2 \cdot \|\bar{\mathbf{u}}\|_2
\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz inequality ($|\boldsymbol{\alpha}^*\boldsymbol{\Phi}\bar{y}| = |(\boldsymbol{\alpha}^*\boldsymbol{\Phi}\bar{y})^*| = |(\boldsymbol{\Phi}\bar{y})^*\boldsymbol{\alpha}| = |\langle \bar{y}, \boldsymbol{\Phi}^*\boldsymbol{\alpha}\rangle_{L_2(d\mu)}| \leq \|\boldsymbol{\Phi}^*\boldsymbol{\alpha}\|_{L_2(d\mu)} \cdot \|\bar{y}\|_{L_2(d\mu)}$). By another use of Cauchy-Schwarz we have:

$$\begin{aligned}
p(\boldsymbol{\eta})|\boldsymbol{\alpha}^*\mathbf{z}(\boldsymbol{\eta})|^2 &\leq \left(\|\boldsymbol{\Phi}^*\boldsymbol{\alpha}\|_{L_2(d\mu)}\|\bar{y}\|_{L_2(d\mu)} + \sqrt{\lambda}\|\boldsymbol{\alpha}^*\|_2 \cdot \|\bar{\mathbf{u}}\|_2\right)^2 \\
&\leq \left(\|\boldsymbol{\Phi}^*\boldsymbol{\alpha}\|^2_{L_2(d\mu)} + \lambda\|\boldsymbol{\alpha}^*\|_2^2\right) \cdot \left(\|\bar{y}\|^2_{L_2(d\mu)} + \|\bar{\mathbf{u}}\|_2^2\right)
\end{aligned}$$

therefore, for every $\boldsymbol{\alpha} \in \mathbb{C}^n$,

$$\frac{p(\boldsymbol{\eta})|\boldsymbol{\alpha}^*\mathbf{z}(\boldsymbol{\eta})|^2}{\|\boldsymbol{\Phi}^*\boldsymbol{\alpha}\|^2_{L_2(d\mu)} + \lambda\|\boldsymbol{\alpha}\|_2^2} \leq \|\bar{y}\|^2_{L_2(d\mu)} + \|\bar{\mathbf{u}}\|_2^2 = \tau_\lambda(\boldsymbol{\eta}) \tag{18}$$

Now it is enough to show that at the optimal $\boldsymbol{\alpha}$ the dual problem gives the leverage scores. We show that $\bar{\boldsymbol{\alpha}} = \sqrt{p(\boldsymbol{\eta})}(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta})$ matches the leverage scores. First note that for any $\boldsymbol{\alpha} \in \mathbb{C}^n$ we have

$$\begin{aligned}
\|\boldsymbol{\Phi}^*\boldsymbol{\alpha}\|^2_{L_2(d\mu)} + \lambda\|\boldsymbol{\alpha}\|_2^2 &= \langle \boldsymbol{\Phi}^*\boldsymbol{\alpha}, \boldsymbol{\Phi}^*\boldsymbol{\alpha}\rangle_{L_2(d\mu)} + \lambda\boldsymbol{\alpha}^*\boldsymbol{\alpha} \\
&= \langle \boldsymbol{\Phi}\boldsymbol{\Phi}^*\boldsymbol{\alpha}, \boldsymbol{\alpha}\rangle_{\mathbb{C}^n} + \lambda\boldsymbol{\alpha}^*\boldsymbol{\alpha} \\
&= \langle \mathbf{K}\boldsymbol{\alpha}, \boldsymbol{\alpha}\rangle_{\mathbb{C}^n} + \lambda\boldsymbol{\alpha}^*\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^*(\mathbf{K} + \lambda\mathbf{I}_n)\boldsymbol{\alpha}
\end{aligned}$$

Now by substituting $\bar{\boldsymbol{\alpha}} = \sqrt{p(\boldsymbol{\eta})}(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta})$ we have:

$$\begin{aligned}
\frac{p(\boldsymbol{\eta})|\bar{\boldsymbol{\alpha}}^*\mathbf{z}(\boldsymbol{\eta})|^2}{\|\boldsymbol{\Phi}^*\bar{\boldsymbol{\alpha}}\|^2_{L_2(d\mu)} + \lambda\|\bar{\boldsymbol{\alpha}}\|_2^2} &= \frac{p(\boldsymbol{\eta})^2|\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta})|^2}{p(\boldsymbol{\eta})\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}(\mathbf{K} + \lambda\mathbf{I}_n)(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta})} \\
&= p(\boldsymbol{\eta})|\mathbf{z}(\boldsymbol{\eta})^*(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{z}(\boldsymbol{\eta})| \\
&= \tau_\lambda(\boldsymbol{\eta})
\end{aligned} \tag{19}$$

□

F. Proof of Theorem 13

Recall from Lemma 11 that

$$\tau_\lambda(\eta) = \min_{y \in L_2(d\mu)} \lambda^{-1}\|\boldsymbol{\Phi}y - \sqrt{p(\eta)}\mathbf{z}(\eta)\|_2^2 + \|y\|^2_{L_2(d\mu)} \tag{20}$$

To upper bound $\tau_\lambda(\eta)$ for any η, we will exhibit some test function, $y_\eta(\cdot)$, and compute the quantity under the minimum. $y_\eta(\cdot)$ will be a 'softened spike function' given by:

Definition 29 (Softened spike function)**.** For any η, and any u define:

$$y_{\eta,u}(t) = \frac{\sqrt{p(\eta)}}{p(t)} \cdot e^{-(t-\eta)^2u^2/4} \cdot v \cdot \mathbf{sinc}\left(v(t-\eta)\right) \tag{21}$$

where $v = 2(R + u\sqrt{2\log n_\lambda})$.

The reweighted function $g_{\eta,u}(t) = p(t) \cdot y_{\eta,u}(t)$ is just a Gaussian with standard deviation $\Theta(1/u)$ multiplied by a sinc function with width $\tilde{O}(1/(u+R))$, both centered at η. Taking the Fourier transform of this function yields a Gaussian with standard deviation $\Theta(u)$ convolved with a box of width $\tilde{O}(u) + R$. This width is wide enough such that when centered between $[-R, R]$ the box covers nearly all the mass of the Gaussian, and so the Fourier transform is nearly identically 1 on the range $[-R, R]$. Shifting by η, means that it is very close to a pure cosine wave with frequency η on this range, and hence makes the first term of (20) small. We make this argument formal below.

F.1. Bounding $\lambda^{-1}\|\mathbf{\Phi} y_{\eta,u} - \sqrt{p(\eta)}\mathbf{z}(\eta)\|_2^2$

Lemma 30 (Test Function Fourier Transform Bound)**.** *For any integer n, every parameter $0 < \lambda \leq n$ and every η, u, and any kernel density function $p(\eta)$ if $x_j \in [-R, +R]$ for all $j \in [n]$ for any radius $R > 0$, then:*

$$\lambda^{-1}\|\mathbf{\Phi} y - \sqrt{p(\eta)}\mathbf{z}(\eta)\|_2^2 = \frac{1}{\lambda}\sum_{j=1}^{n}\left|\hat{g}_{\eta,u}(x_j) - \sqrt{p(\eta)} \cdot z(\eta)_j\right|^2 \leq p(\eta). \tag{22}$$

where $g_{\eta,u}(t) \equiv p(t)y_{\eta,u}(t)$.

Proof. We have $g_{\eta,u}(t) = p(t)y_{\eta,u}(t) = p(\eta)e^{-(t-\eta)^2u^2/4} \cdot v \cdot \mathbf{sinc}\left(v(t-\eta)\right)$ and $\hat{g}_{\eta,u}(x_j) = (\mathbf{\Phi} y)_j$. We thus have:

$$\begin{aligned}
\hat{g}_{\eta,u}(x_j) &= \sqrt{p(\eta)}\int_{\mathbb{R}} e^{-2\pi i t x_j} e^{-(t-\eta)^2u^2/4} \cdot v \cdot \mathbf{sinc}\left(v(t-\eta)\right) dt \\
&= \sqrt{p(\eta)}e^{-2\pi i x_j \eta}\int_{\mathbb{R}} e^{-2\pi i t x_j} e^{-t^2u^2/4} \cdot v \cdot \mathbf{sinc}\left(vt\right) dt \\
&= \sqrt{p(\eta)} \cdot z(\eta)_j \cdot h(x_j)
\end{aligned} \tag{23}$$

where $h(x) = \frac{2\sqrt{\pi}}{u}e^{-4\pi^2x^2/u^2} * \mathrm{rect}_v(x)$ by the fact that multiplication in time domain becomes convolution in the Fourier domain (Claim 18), $\mathcal{F}(e^{-t^2u^2/4}) = \frac{2\sqrt{\pi}}{u}e^{-4\pi^2x^2/u^2}$, and $\mathcal{F}(v \cdot \mathbf{sinc}\left(vt\right)) = \mathrm{rect}_v(x)$.

For any x, we have $h(x) \leq \int_{\mathbb{R}} \frac{2\sqrt{\pi}}{u}e^{-4\pi^2x^2/u^2} = 1$. Additionally, for any $x \in [-R, R]$ we have by Claim 25 and the fact that $v = 2R + 2u\sqrt{2\log n_\lambda}$:

$$\begin{aligned}
h(x) &= \int_{x-\frac{v}{2}}^{x+\frac{v}{2}} \frac{2\sqrt{\pi}}{u}e^{-4\pi^2x^2/u^2}\,dx \\
&\geq 1 - 2\int_{v/2-R}^{\infty} \frac{2\sqrt{\pi}}{u}e^{-4\pi^2x^2/u^2}\,dx \\
&\geq 1 - \frac{1}{4\sqrt{\pi}} \cdot \frac{u}{v/2-R}e^{-4\pi^2(v/2-R)^2/u^2} && \text{(by second part of Claim 25)} \\
&\geq 1 - \frac{1}{4\sqrt{\pi}\sqrt{2\log n_\lambda}} \cdot \frac{1}{\sqrt{n_\lambda}} && \text{(since } v = 2R + 2u\sqrt{2\log n_\lambda}\text{)} \\
&\geq 1 - \frac{1}{\sqrt{n_\lambda}} && \text{(by assumption } n_\lambda \geq 2\text{).}
\end{aligned}$$

Plugging into (23) gives

$$\begin{aligned}
\left|\hat{g}_{\eta,u}(x_j) - \sqrt{p(\eta)} \cdot z(\eta)_j\right|^2 &= p(\eta)\left|h(x_j) - 1\right|^2 \\
&\leq \frac{p(\eta)}{n_\lambda},
\end{aligned}$$

and so,

$$\frac{1}{\lambda}\sum_{j=1}^{n}\left[\hat{g}(x_j) - \sqrt{p(\eta)} \cdot z(\eta)_j\right]^2 \leq n_\lambda \cdot p(\eta) \cdot \frac{\lambda}{n} < p(\eta)$$

proving the claim. □

F.2. Bounding $\|y_{\eta,u}\|^2_{L_2(d\mu)}$

Having established Lemma 30, we note that showing that the weighted Fourier transform of $y_{\eta,u}$ is close to $\sqrt{p(\eta)}\mathbf{z}(\eta)$ reduces to bounding the norm of the test function. To that effect, we show the following:

Lemma 31 (Test Function ℓ_2 Norm Bound)**.** *For any integer n, any parameter $0 < \lambda \leq \frac{n}{2}$, every $|\eta| \leq 10\sqrt{\log n_\lambda}$, and every $2000 \log n_\lambda \leq u \leq 500 \log^{1.5} n_\lambda$, if $y_{\eta,u}(t)$ is defined as in (20), as per Definition 29, then we have*

$$\|y\|^2_{L_2(d\mu)} \leq 12\Big(R + u\sqrt{2\log n_\lambda}\Big) \tag{24}$$

Before proving Lemma 31, we first prove a claim:

Claim 32. *Suppose $|\eta| \leq 100\sqrt{\log n_\lambda}$, and*

$$\eta - \frac{c\sqrt{\log n_\lambda}}{b} \leq t \leq \eta + \frac{c\sqrt{\log n_\lambda}}{b}$$

for some absolute constant $c > 0$. If $b \geq 100c \cdot \log n_\lambda$ then,

$$e^{-\frac{t^2}{2}+\frac{\eta^2}{2}} \leq 3.$$

Proof. Let $\Delta = t - \eta$. Then, note that $|\Delta| \leq c\sqrt{\log n_\lambda}/b$, and so,

$$\begin{aligned} e^{-\frac{t^2}{2}+\frac{\eta^2}{2}} &= e^{-\frac{(\Delta+\eta)^2}{2}+\frac{\eta^2}{2}} \\ &= e^{-\Delta\eta-\frac{\Delta^2}{2}} \\ &\leq e^{|\Delta\eta|-\frac{\Delta^2}{2}} \\ &\leq e^{|\Delta|\cdot|\eta|} \\ &\leq e^{(c\sqrt{\log n_\lambda}/b)(100\sqrt{\log n_\lambda})} \\ &\leq e \leq 3, \end{aligned}$$

since $b \geq 100c \cdot \log n_\lambda$. □

Now, we are ready to prove Lemma 31:

Proof of Lemma 31. Recall that for the Gaussian kernel, we have $p(\eta) = \frac{1}{\sqrt{2\pi}}e^{-\eta^2/2}$. We calculate:

$$\begin{aligned} \int_{\mathbb{R}} |y_{\eta,u}(t)|^2 d\mu(t) &= p(\eta)\int_{\mathbb{R}} \sqrt{2\pi}e^{t^2/2}\cdot e^{-(t-\eta)^2u^2/2}\cdot v^2\left(\mathbf{sinc}\left(v(t-\eta)\right)\right)^2 dt \\ &= \sqrt{2\pi}p(\eta)\cdot v^2 \int_{\eta-\frac{20\sqrt{\log n_\lambda}}{u}}^{\eta+\frac{20\sqrt{\log n_\lambda}}{u}} e^{t^2/2}\cdot e^{-(t-\eta)^2u^2/2}\left(\mathbf{sinc}\left(v(t-\eta)\right)\right)^2 dt \\ &\quad + \sqrt{2\pi}p(\eta)\cdot v^2 \int_{|t-\eta|\geq\frac{20\sqrt{\log n_\lambda}}{u}} e^{t^2/2}\cdot e^{-(t-\eta)^2u^2/2}\left(\mathbf{sinc}\left(v(t-\eta)\right)\right)^2 dt \end{aligned} \tag{25}$$

For the integral over $|t-\eta| \geq 20\frac{\sqrt{\log n_\lambda}}{u}$ we have:

$$\begin{aligned} \int_{|t-\eta|\geq 20\frac{\sqrt{\log n_\lambda}}{u}} e^{t^2/2}\cdot e^{-(t-\eta)^2u^2/2}\left(\mathbf{sinc}\left(v(t-\eta)\right)\right)^2 dt &\leq \frac{1}{\left(v\cdot 20\frac{\sqrt{\log n_\lambda}}{u}\right)^2}\int_{|t-\eta|\geq 20\frac{\sqrt{\log n_\lambda}}{u}} e^{t^2/2}\cdot e^{-(t-\eta)^2u^2/2}dt \\ &\leq \frac{1}{v}\int_{|t-\eta|\geq 20\frac{\sqrt{\log n_\lambda}}{u}} e^{t^2/2}\cdot e^{-(t-\eta)^2u^2/2}dt \end{aligned} \tag{26}$$

The first inequality above is because by definition of $\mathbf{sinc}(\cdot)$ we have the following for all $|t-\eta| \geq \frac{20\sqrt{\log n_\lambda}}{u}$:

$$|\mathbf{sinc}(v(t-\eta))|^2 = \frac{\sin^2(\pi v(t-\eta))}{(\pi v(t-\eta))^2} \leq \frac{1}{(v(t-\eta))^2} \leq \frac{1}{\left(v \cdot \frac{20\sqrt{\log n_\lambda}}{u}\right)^2}$$

The last inequality in (26) is because of the following reason:

$$\begin{aligned}
\frac{1}{\left(v \cdot \frac{20\sqrt{\log n_\lambda}}{u}\right)^2} &= \frac{1}{v} \cdot \frac{1}{v \cdot \left(\frac{20\sqrt{\log n_\lambda}}{u}\right)^2} \\
&\leq \frac{1}{v} \cdot \frac{1}{800\left(\frac{\log^{1.5} n_\lambda}{u}\right)} \quad \text{(since } v = 2(R + u\sqrt{2\log n_\lambda}) \geq 2u\sqrt{2\log n_\lambda}\text{, see Definition 29)} \\
&\leq \frac{1}{v} \qquad\qquad\qquad\qquad \text{(since } u \leq 500\log^{1.5} n_\lambda\text{)}
\end{aligned}$$

Now note that $t^2 \leq 2(t-\eta)^2 + 2\eta^2$. We have the following for all $|t-\eta| \geq \frac{20\sqrt{\log n_\lambda}}{u}$:

$$\begin{aligned}
t^2 &\leq 2(t-\eta)^2 + 2\eta^2 \\
&\leq 2(t-\eta)^2 + 200\log n_\lambda \qquad \text{(by the assumption } |\eta| \leq 10\sqrt{\log n_\lambda}\text{)} \\
&\leq 2(t-\eta)^2 + (t-\eta)^2u^2/2 \quad \text{(by the assumption } |t-\eta| \geq \frac{20\sqrt{\log n_\lambda}}{u}\text{)} \\
&\leq \frac{2}{3}(t-\eta)^2u^2
\end{aligned}$$

where the last inequality follows from $u \geq 2000\log n_\lambda \geq 600$ (because $n_\lambda \geq 1/2$). Hence,

$$\begin{aligned}
\frac{1}{v}\int_{|t-\eta|\geq 20\frac{\sqrt{\log n_\lambda}}{u}} e^{t^2/2} \cdot e^{-(t-\eta)^2u^2/2} dt &\leq \frac{1}{v}\int_{|t-\eta|\geq 20\frac{\sqrt{\log n_\lambda}}{u}} e^{-(t-\eta)^2u^2/3} dt \\
&\leq \frac{1}{v} \cdot n_\lambda^{100} \qquad (27)
\end{aligned}$$

Now, the first integral in (25):

$$\begin{aligned}
\int_{\eta-20\frac{\sqrt{\log n_\lambda}}{u}}^{\eta+20\frac{\sqrt{\log n_\lambda}}{u}} e^{t^2/2} \cdot e^{-(t-\eta)^2u^2/2} \left(\mathbf{sinc}(v(t-\eta))\right)^2 dt &\leq 3e^{\frac{\eta^2}{2}} \int_{\mathbb{R}} \left(\mathbf{sinc}(v(t-\eta))\right)^2 dt \\
&= \frac{3e^{\frac{\eta^2}{2}}}{v}. \qquad (28)
\end{aligned}$$

where the inequality follows from Claim 32 with $c = 20$ and $b = u$, since, by assumption, $u \geq 2000\log n_\lambda$ and $|t| \leq |\eta| + |t-\eta| \leq 10\sqrt{\log n_\lambda} + \frac{20}{u}\sqrt{\log n_\lambda} \leq 100\sqrt{\log n_\lambda}$ whenever $t \in \left[\eta - \frac{20\log n_\lambda}{u}, \eta + \frac{20\log n_\lambda}{u}\right]$.

By incorporating (27) and (28) into (25) we have:

$$\int_{\mathbb{R}} |y_{\eta,u}(t)|^2 dt \leq \sqrt{2\pi}p(\eta) \cdot v^2\left(\frac{1}{v} \cdot n_\lambda^{-100} + \frac{3e^{\frac{\eta^2}{2}}}{v}\right) \leq 6v \qquad (29)$$

where the last inequality uses that $\sqrt{2\pi}p(\eta) = \frac{\sqrt{2\pi}}{\sqrt{2\pi}}e^{-\eta^2/2} \leq 1$. □

Proof of Theorem 13. By the assumptions of the theorem n is an integer, parameter $0 < \lambda \leq n/2$, and $R > 0$, and all $x_1, ..., x_n \in [-R, R]$ and $p(\eta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\eta^2}{2}}$, therefore all the preconditions of Lemmas 31, and 30 are satisfied and hence the lemmas go through and the upper bounds in (22) and (24) hold true. The theorem follows by setting $u = 2000 \log n_\lambda$ and then plugging upper bounds (22) and (24) into (20). □

G. Proof of Theorem 14

With the choice of the Gaussian kernel with $\sigma = (2\pi)^{-1}$ we have $p(\eta) = (2\pi)^{-1/2} \exp(-\eta^2/2)$. Recall from Lemma 12 that

$$\tau_\lambda(\eta) = \max_{\boldsymbol{\alpha} \in \mathbb{C}^n} \frac{p(\eta) \cdot |\boldsymbol{\alpha}^* \mathbf{z}(\eta)|^2}{\|\boldsymbol{\Phi}^* \boldsymbol{\alpha}\|^2_{L_2(d\mu)} + \lambda \|\boldsymbol{\alpha}\|_2^2}. \tag{30}$$

In particular, this gives us a method of bounding the leverage scores from below, namely, by exhibiting some $\boldsymbol{\alpha}$ and computing the quantity under the maximum.

The rest of this section is organized as follows. In Section G.1, we construct our candidate set of data points $x_1, x_2, \ldots, x_n$ along with the vector $\boldsymbol{\alpha}$. In particular, $\boldsymbol{\alpha}$ will be chosen to be a vector of samples of a function $f_{\Delta,b,v}$ at each of the data points. Section G.2 then describes basic Fourier properties of the function $f_{\Delta,b,v}$ and $\boldsymbol{\alpha}$ that we will require later. The remaining sections then bound each of the relevant quantities that appear in (30) for our specific choice of $x_1, x_2, \ldots, x_n$ and $\boldsymbol{\alpha}$. In particular, Section G.3 shows a lower bound for $\boldsymbol{\alpha}^* \mathbf{z}(\eta)$, while Section G.4 shows an upper bound for $\|\boldsymbol{\alpha}\|_2^2$ and Section G.5 shows an upper bound for $\|\boldsymbol{\Phi}^* \boldsymbol{\alpha}\|^2_{L_2(d\mu)}$.

G.1. Construction of Data Point Set and the Vector of Coefficients $\boldsymbol{\alpha}$

Definition 33. For parameters Δ, $b > 0$ and $v > 0$, let the function $f_{\Delta,b,v}$ be defined as follows:

$$\begin{aligned} f_{\Delta,b,v}(x) &= 2\cos(2\pi\Delta x) \left(\frac{1}{\sqrt{2\pi}b} e^{-(\cdot)^2/2b^2} * \mathrm{rect}_v \right)(x) \\ &= 2\cos(2\pi\Delta x) \int_{x-\frac{v}{2}}^{x+\frac{v}{2}} \frac{1}{\sqrt{2\pi}b} e^{-t^2/2b^2}\, dt \end{aligned}$$

Lemma 34. *For any $\Delta > 0$, $v > 0$, and $b > 0$, if we define the function $f_{\Delta,b,v}$ as in Definition 33, then*

$$\mathcal{F}\left(f_{\Delta,b,v}\right)(z) = e^{-2\pi^2 b^2 (z-\Delta)^2} (v \cdot \mathbf{sinc}\left(v(z-\Delta)\right) + e^{-2\pi^2 b^2 (z+\Delta)^2} (v \cdot \mathbf{sinc}\left(v(z+\Delta)\right)).$$

Proof. Note that

$$\mathcal{F}\left(\frac{1}{\sqrt{2\pi}b} e^{-(\cdot)^2/2b^2} \right)(z) = e^{-2\pi^2 b^2 z^2}.$$

Thus, by the convolution theorem (see Claim 18),

$$\mathcal{F}\left(\frac{1}{\sqrt{2\pi}b} e^{-(\cdot)^2/2b^2} * \mathrm{rect}_v \right)(z) = e^{-2\pi^2 b^2 z^2} \cdot v \cdot \mathbf{sinc}\left(v(z)\right).$$

Now by the duality of phase shift in time domain and frequency shift in the Fourier domain,

$$\begin{aligned} \mathcal{F}(f_{\Delta,b,v})(z) &= \mathcal{F}\left((e^{2\pi i \Delta(\cdot)} + e^{-2\pi i \Delta(\cdot)}) \cdot \left(\frac{1}{\sqrt{2\pi}b} e^{-(\cdot)^2/2b^2} * \mathrm{rect}_v \right) \right)(z) \\ &= \mathcal{F}\left(\frac{1}{\sqrt{2\pi}b} e^{-(\cdot)^2/2b^2} * \mathrm{rect}_v \right)(z-\Delta) + \mathcal{F}\left(\frac{1}{\sqrt{2\pi}b} e^{-(\cdot)^2/2b^2} * \mathrm{rect}_v \right)(z+\Delta) \\ &= e^{-2\pi^2 b^2 (z-\Delta)^2} \cdot v \cdot \mathbf{sinc}\left(v(z-\Delta)\right) + e^{-2\pi^2 b^2 (z+\Delta)^2} \cdot v \cdot \mathbf{sinc}\left(v(z+\Delta)\right). \end{aligned}$$

□

Intuition for Theorem 14 If, instead of a discrete set of data points, we had a continuum of points, $\boldsymbol{\alpha}$ would be a function (or, alternatively, an infinite-dimensional vector coresponding to the evaluation of the function on the continuum of points). The intuition is that in this case, we would essentially like to choose $\boldsymbol{\alpha}$ to be the function $f_{\Delta,b,v}$ for some suitable choice of parameters Δ, b, v. In this case, the computation of bounds for the various quantities appearing in (30) would be relatively simple and involve bounding integrals. However, since our data points are actually discrete and $\boldsymbol{\alpha}$ is finite-dimensional, we must instead choose $\boldsymbol{\alpha}$ to be the vector of samples of $f_{\Delta,b,v}$ on the data points, and the bounds we deduce require computing Fourier transforms of $f_{\Delta,b,v}$ multiplied by suitable Dirac combs (see Lemma 36). Computation of the necessary bounds is further complicated by the fact that the data points are bounded in $[-R, R]$, which requires us to truncate the aforementioned Dirac combs and have appropriate Fourier tail bounds (see Lemma 37).

Let us provide some intuition about the quantities $|\boldsymbol{\alpha}^* \mathbf{z}(\eta)|^2$, $\|\boldsymbol{\Phi}^* \boldsymbol{\alpha}\|_{L_2(d\mu)}^2$ and $\|\boldsymbol{\alpha}\|_2^2$ that arise in (30) along these lines. If we have $\approx 2R$ equally spaced data points between $-R$ and R, then note that the points are separated by distance ≈ 1. This approximately corresponds to dealing with the continuous case in which $\boldsymbol{\alpha}$ is a function $f_{\Delta,b,v}$ and, therefore, sums in the discrete case can be approximated by corresponding integrals over continuous functions. Suppose $\Delta = \eta$ and $v = R$.

Note that the quantity $\boldsymbol{\alpha}^* \mathbf{z}(\xi)$ corresponds to

$$\begin{aligned}
\boldsymbol{\alpha}^* \mathbf{z}(\xi) &\approx \int_{-\infty}^{\infty} f_{\eta,b,R}(x) e^{-2\pi i \xi x}\, dx \\
&\approx \mathcal{F}(f_{\eta,b,R})(\xi) \\
&= e^{-2\pi^2 b^2 (\xi-\eta)^2} \cdot R \cdot \mathbf{sinc}\left(R(\xi-\eta)\right) + e^{-2\pi^2 b^2 (\xi+\eta)^2} \cdot R \cdot \mathbf{sinc}\left(R(\xi+\eta)\right). \qquad (31)
\end{aligned}$$

Thus, $\boldsymbol{\alpha}^* \mathbf{z}(\xi)$ (which we bound rigorously in Section G.3) can be approximated as follows:

$$\boldsymbol{\alpha}^* \mathbf{z}(\xi) \approx R(1 + e^{-8\pi^2 b^2 \eta^2} \mathbf{sinc}\left(2R\eta\right)) \approx \Omega(R), \qquad (32)$$

where the last transition uses the fact that $\mathbf{sinc}\left(\cdot\right) \geq -1/4$. Next, note that the quantity $\|\boldsymbol{\alpha}\|_2^2$ (which we bound rigorously in Section G.4) is roughly

$$\begin{aligned}
\|\boldsymbol{\alpha}\|_2^2 \approx \int_{-\infty}^{\infty} f_{\eta,b,R}(x)^2\, dx &= \int_{-\infty}^{\infty} 4\cos^2(2\pi\eta x) \left(\int_{x-\frac{R}{2}}^{x+\frac{R}{2}} \frac{1}{\sqrt{2\pi} b} e^{-t^2/2b^2}\, dt \right)^2 dx \\
&\approx 4 \int_{-\frac{3R}{2}}^{\frac{3R}{2}} \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} b} e^{-t^2/2b^2}\, dt \right)^2 dx \\
&\approx O(R). \qquad (33)
\end{aligned}$$

Finally, note that $\|\boldsymbol{\Phi}^* \boldsymbol{\alpha}\|_{L_2(d\mu)}^2$ (which we bound rigorously in Section G.5) is roughly

$$\begin{aligned}
\|\boldsymbol{\Phi}^* \boldsymbol{\alpha}\|_{L_2(d\mu)}^2 &\approx \int_{-\infty}^{\infty} |\boldsymbol{\alpha}^* \mathbf{z}(\xi)|^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}\, d\xi \\
&\approx \int_{-\infty}^{\infty} \left(e^{-2\pi^2 b^2 (\xi-\eta)^2} \cdot R \cdot \mathbf{sinc}\left(R(\xi-\eta)\right) + e^{-2\pi^2 b^2 (\xi+\eta)^2} \cdot R \cdot \mathbf{sinc}\left(R(\xi+\eta)\right) \right)^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}\, d\xi \\
&\approx \frac{1}{\sqrt{2\pi}} e^{-\eta^2/2} R^2 \int_{-\infty}^{\infty} \mathbf{sinc}\left(R(\xi-\eta)\right)^2\, d\xi \\
&\approx O(p(\eta) R), \qquad (34)
\end{aligned}$$

using (31).

Now, going back to the discrete case, consider what happens if we scale up the number of points from $2R$ to n, keeping the points evenly spaced in the interval $[-R, R]$. In this case, the spacing between points decreases by a factor of $\gamma \approx n/2R$. Thus, this corresponds to the measure of integration over $\mathbb{R}$ scaling up by a factor of γ. Hence, $|\boldsymbol{\alpha}^* \mathbf{z}(\eta)|$ and $\|\boldsymbol{\alpha}\|_2^2$ can be expected to scale up by a factor of γ, while $\|\boldsymbol{\Phi}^* \boldsymbol{\alpha}\|_{L_2(d\mu)}^2$ would scale up by a factor of γ^2. Thus, along with (32), (33), and (34), we get that

$$\frac{p(\eta) \cdot |\boldsymbol{\alpha}^* \mathbf{z}(\eta)|^2}{\|\boldsymbol{\Phi}^* \boldsymbol{\alpha}\|_{L_2(d\mu)}^2 + \lambda \|\boldsymbol{\alpha}\|_2^2} \approx \frac{(\gamma R)^2 p(\eta)}{\gamma^2 p(\eta) R + \lambda \gamma R} \approx R \cdot \frac{p(\eta)}{p(\eta) + \lambda/\gamma} \approx R \cdot \frac{p(\eta)}{p(\eta) + 2R n_\lambda^{-1}},$$

which is within a constant factor of the expression in Theorem 14.

Definition 35 (Construction of data points and $\boldsymbol{\alpha}$)**.** We first define the set of data points x_j for $j = 1, 2, \ldots, n$ for odd n as follows:

$$x_j = \left(j - \frac{n+1}{2}\right) \cdot \frac{2R}{n}$$

Thus, the data points are on a grid of width $\frac{2R}{n}$ extending from $-R$ to R.

The vector $\boldsymbol{\alpha}$ is chosen to be the tuple of evaluations of $f_{\eta,b,v}$ at the individual x_j, for some parameters b, v, and η. More specifically, for $1 \le j \le n$, we define

$$\begin{aligned}\boldsymbol{\alpha}_j &= f_{\eta,b,v}(x_j) \\ &= 2\cos(2\pi\eta x_j) \int_{x_j - \frac{v}{2}}^{x_j + \frac{v}{2}} \frac{1}{\sqrt{2\pi}b} e^{-t^2/2b^2} dt. \end{aligned} \tag{35}$$

G.2. Basic Properties of $f_{\Delta,b,v}$ and $\boldsymbol{\alpha}$

By the Nyquist-Shannon sampling theorem, we have the following lemma.

Lemma 36. *For any parameters $\Delta > 0$, $v > 0$, and $b > 0$, if we define the function $f_{\Delta,b,v}$ as in Definition 33, then for any $w > 0$,*

$$\begin{aligned}\mathcal{F}\left(f_{\Delta,b,v}(\cdot) \cdot \sum_{j=-\infty}^{\infty} \delta(\cdot - jw)\right)(z) = w^{-1}v \sum_{j=-\infty}^{\infty} e^{-2\pi^2 b^2 (z - jw^{-1} - \Delta)^2} \cdot \mathbf{sinc}\left(v(z - jw^{-1} - \Delta)\right) \\ + w^{-1}v \sum_{j=-\infty}^{\infty} e^{-2\pi^2 b^2 (z - jw^{-1} + \Delta)^2} \cdot \mathbf{sinc}\left(v(z - jw^{-1} + \Delta)\right).\end{aligned}$$

Proof. By the Nyquist-Shannon sampling theorem, we have

$$\begin{aligned}\mathcal{F}\left(f_{\Delta,b,v}(\cdot) \sum_{j=-\infty}^{\infty} \delta(\cdot - jw)\right)(z) &= \left(w^{-1} \sum_{j=-\infty}^{\infty} \delta(\cdot - jw^{-1}) * \mathcal{F}(f_{\Delta,b,v})(\cdot)\right)(z) \\ &= \sum_{j=-\infty}^{\infty} w^{-1} \mathcal{F}(f_{\Delta,b,v})(z - jw^{-1}). \end{aligned} \tag{36}$$

Thus, by Lemma 34, we find that (36) can be written as

$$\begin{aligned}\sum_{j=-\infty}^{\infty} w^{-1}\mathcal{F}(f_{\Delta,b,v})(z - jw^{-1}) = w^{-1} \sum_{j=-\infty}^{\infty} e^{-2\pi^2 b^2 (z - jw^{-1} - \Delta)^2} \cdot v \cdot \mathbf{sinc}\left(v(z - jw^{-1} - \Delta)\right) \\ + w^{-1} \sum_{j=-\infty}^{\infty} e^{-2\pi^2 b^2 (z - jw^{-1} + \Delta)^2} \cdot v \cdot \mathbf{sinc}\left(v(z - jw^{-1} + \Delta)\right),\end{aligned}$$

which completes the proof. □

Lemma 37. *For every odd integer $n \ge 3$ and parameters $0 < \lambda \le \frac{n}{2}$, $\eta > 0$, $v \le R$, and $b \le \frac{R}{4\sqrt{\log n_\lambda}}$, if we define the function $f_{\Delta,b,v}$ as in Definition 33, then*

$$\left|\mathcal{F}\left(\sum_{|j| > \frac{n}{2}} f_{\eta,b,v}\left(j \cdot \frac{2R}{n}\right) \cdot \delta\left(\cdot - j \cdot \frac{2R}{n}\right)\right)(z)\right| \le \sqrt{\lambda n}$$

for all z.

Proof. By definition of $f_{\eta,b,v}$, we have the following for all x:

$$|f_{\eta,b,v}(x)| \le \int_{x-\frac{v}{2}}^{x+\frac{v}{2}} \frac{2}{\sqrt{2\pi}b} e^{-t^2/2b^2}\, dt.$$

Therefore, if $|j| > \frac{n}{2}$, then

$$\begin{aligned}
\left| f_{\eta,b,v}\left(j \cdot \frac{2R}{n}\right)\right| &\le \frac{2}{\sqrt{2\pi}b} \int_{j\cdot\frac{2R}{n}-\frac{v}{2}}^{\infty} e^{-t^2/2b^2}\, dt \\
&\le \frac{2}{\sqrt{2\pi}b} \int_{\frac{jR}{n}}^{\infty} e^{-t^2/2b^2}\, dt \\
&\le \frac{2}{\sqrt{2\pi}} \cdot \frac{nb}{jR} \cdot e^{-\frac{1}{2}\cdot\left(\frac{jR}{nb}\right)^2} \\
&\le \frac{2b}{R} \cdot e^{-\frac{1}{2}\cdot\left(\frac{jR}{nb}\right)^2}, \qquad (37)
\end{aligned}$$

where we have used the fact that $j \cdot \frac{2R}{n} - \frac{v}{2} \ge j \cdot \frac{2R}{n} - \frac{R}{2} \ge j \cdot \frac{R}{n}$, along with Claim 25. Therefore, again using Claim 25, we have

$$\begin{aligned}
\left| \mathcal{F}\left(\sum_{|j|>\frac{n}{2}} f_{\eta,b,v}\left(j\cdot\frac{2R}{n}\right) \cdot \delta\left(\cdot - j\cdot\frac{2R}{n}\right)\right)(z)\right| &\le \sum_{|j|>\frac{n}{2}} \left| f_{\eta,b,v}\left(j\cdot\frac{2R}{n}\right)\right| \\
&\le \sum_{|j|>\frac{n}{2}} \frac{2b}{R} \cdot e^{-\frac{1}{2}\cdot\left(\frac{jR}{nb}\right)^2} \\
&\le \frac{2b}{R} \cdot \left(\frac{nb}{R} \int_{\frac{(n-1)R}{2nb}}^{\infty} e^{-t^2/2}\, dt \right) \\
&\le \frac{n}{4\log n_\lambda} \cdot \int_{\sqrt{\log n_\lambda}}^{\infty} e^{-t^2/2}\, dt \\
&\le \frac{n}{4\log n_\lambda} \cdot \frac{1}{\sqrt{\log n_\lambda}} \cdot e^{-\frac{1}{2}\cdot\left(\sqrt{\log n_\lambda}\right)^2} \\
&\le \frac{1}{4\log^{3/2}(n_\lambda)} \cdot \sqrt{\lambda n} \\
&\le \sqrt{\lambda n},
\end{aligned}$$

since $n \ge 3$, $R \ge 4b\sqrt{\log n_\lambda}$, and $\lambda \le n/2$. □

Lemma 38. *For every odd integer $n \ge 3$, any parameter $0 < \lambda \le \frac{n}{2}$, every frequency η and ξ, and any parameter $v \le R$ and $b \le \frac{R}{4\sqrt{\log n_\lambda}}$, if $\boldsymbol{\alpha}$ is defined as in (35) of Definition 35, then we have,*

$$\begin{aligned}
&\left| \boldsymbol{\alpha}^* \mathbf{z}(\xi) - \frac{nv}{2R} \sum_{j=-\infty}^{\infty} \left(e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} - \eta)^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} - \eta)\right) + e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} + \eta)^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} + \eta)\right)\right)\right| \\
&\qquad \le \sqrt{\lambda n}. \qquad (38)
\end{aligned}$$

Proof. Note that

$$\begin{aligned}\boldsymbol{\alpha}^* \mathbf{z}(\xi) &= \sum_{j=1}^{n} \boldsymbol{\alpha}_j e^{-2\pi i x_j \xi} \\ &= \sum_{|j| \le \frac{n}{2}} f_{\eta,b,v}(2Rj/n) \cdot e^{-2\pi i \left(\frac{2Rj}{n}\right)\xi} \\ &= \mathcal{F}\left(\sum_{|j| \le \frac{n}{2}} f_{\eta,b,v}(2Rj/n) \cdot \delta\left(\cdot - \frac{2Rj}{n}\right)\right)(\xi) \\ &= \mathcal{F}\left(\sum_{j=-\infty}^{\infty} f_{\eta,b,v}(\cdot) \cdot \delta\left(\cdot - \frac{2Rj}{n}\right)\right)(\xi) - \mathcal{F}\left(\sum_{|j| > \frac{n}{2}} f_{\eta,b,v}\left(\frac{2Rj}{n}\right) \cdot \delta\left(\cdot - \frac{2Rj}{n}\right)\right)(\xi). \end{aligned} \tag{39}$$

By Lemma 36 (applied with $w = 2R/n$), we have the following expression for the first term in (39):

$$\begin{aligned}\mathcal{F}\left(\sum_{j=-\infty}^{\infty} f_{\eta,b,v}(\cdot) \cdot \delta\left(\cdot - \frac{2Rj}{n}\right)\right)(\xi) = \frac{nv}{2R} \sum_{j=-\infty}^{\infty} e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} - \eta)^2} \cdot \mathbf{sinc}\left(v\left(\xi - \frac{jn}{2R} - \eta\right)\right) \\ + \frac{nv}{2R} \sum_{j=-\infty}^{\infty} e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} + \eta)^2} \cdot \mathbf{sinc}\left(v\left(\xi - \frac{jn}{2R} + \eta\right)\right). \end{aligned} \tag{40}$$

Now, by the assumption that $R \ge 4b\sqrt{\log n_\lambda}$ and $v \le R$, it follows from Lemma 37 that the second term in (39) can be bounded as

$$\left| \mathcal{F}\left(\sum_{|j| > \frac{n}{2}} f_{\eta,b,v}(2Rj/n) \cdot \delta\left(\cdot - \frac{2Rj}{n}\right)\right)(\xi) \right| \le \sqrt{\lambda n}. \tag{41}$$

Thus, the desired result follows by combining (39), (40), and (41). □

G.3. Bounding $\boldsymbol{\alpha}^* \mathbf{z}(\eta)$

Lemma 39. *For every odd integer $n \ge 17$, any parameter $0 < \lambda \le (\frac{v}{R})^2 \cdot n/16$, every frequency $|\eta| \le \frac{n}{10R}$, and any parameter $v \le R$ and $\frac{R}{2\sqrt{n}} \le b \le \frac{R}{4\sqrt{\log(n_\lambda)}}$, if $\boldsymbol{\alpha}$ is defined as in (35) of Definition 35, then we have*

$$|\boldsymbol{\alpha}^* \mathbf{z}(\eta)| \ge \frac{nv}{5R}.$$

Proof. Since $v \le R$ and $b \le \frac{R}{4\sqrt{\log(n_\lambda)}}$ and $\lambda \le n/2$, Lemma 38 implies that

$$\begin{aligned}\left| \boldsymbol{\alpha}^* \mathbf{z}(\eta) - \frac{nv}{2R} \sum_{j=-\infty}^{\infty} \left(e^{-2\pi^2 b^2 (-\frac{jn}{2R})^2} \mathbf{sinc}\left(v(-jn/2R)\right) + e^{-2\pi^2 b^2 (2\eta - \frac{jn}{2R})^2} \mathbf{sinc}\left(v(2\eta - jn/2R)\right)\right) \right| \\ \le \sqrt{\lambda n}. \end{aligned} \tag{42}$$

Hence,

$$\begin{aligned}
|\boldsymbol{\alpha}^* \mathbf{z}(\eta)| &\geq \frac{nv}{2R} \left| \sum_{j=-\infty}^{\infty} \Big(e^{-2\pi^2 b^2 (-\frac{jn}{2R})^2} \mathbf{sinc}\left(v(-jn/2R)\right) \right. \\
&\qquad \left. + e^{-2\pi^2 b^2 (2\eta - \frac{jn}{2R})^2} \mathbf{sinc}\left(v(2\eta - jn/2R)\right) \Big) \right| - \sqrt{\lambda n} \\
&\geq \frac{nv}{2R} e^{-2\pi^2 b^2 (0)^2} \mathbf{sinc}\left(v(0)\right) + \frac{nv}{2R} e^{-2\pi^2 b^2 (2\eta)^2} \mathbf{sinc}\left(v(2\eta)\right) \\
&\qquad - \frac{nv}{2R} \sum_{|j|\geq 1} \left(e^{-2\pi^2 b^2 (-\frac{jn}{2R})^2} + e^{-2\pi^2 b^2 (2\eta - \frac{jn}{2R})^2} \right) - \sqrt{\lambda n} \\
&\geq \frac{3}{4} \left(\frac{nv}{2R} \right) - \frac{nv}{2R} \sum_{|j|\geq 1} \left(e^{-2\pi^2 b^2 (-\frac{jn}{2R})^2} + e^{-2\pi^2 b^2 (2\eta - \frac{jn}{2R})^2} \right) - \sqrt{\lambda n}, \qquad (43)
\end{aligned}$$

since $|\mathbf{sinc}\left(\cdot\right)| \leq 1$ and $\mathbf{sinc}\left(\cdot\right) \geq -\frac{1}{4}$.

Now we show that $\sum_{|j|\geq 1} \left(e^{-2\pi^2 b^2 (-\frac{jn}{2R})^2} + e^{-2\pi^2 b^2 (2\eta - \frac{jn}{2R})^2} \right)$ is small. Note that by the assumption of $b \geq \frac{R}{2\sqrt{n}}$, we have $e^{-2\pi^2 b^2 (-\frac{jn}{2R})^2} \leq e^{-jn}$ for all $|j| \geq 1$. Also recall that $|\eta| \leq \frac{n}{10R}$, and so, $(2\eta - \frac{jn}{2R})^2 \geq (\frac{jn}{4R})^2$ for all $|j| \geq 1$. Hence, we have

$$\begin{aligned}
\sum_{|j|\geq 1} \left(e^{-2\pi^2 b^2 (-\frac{jn}{2R})^2} + e^{-2\pi^2 b^2 (2\eta - \frac{jn}{2R})^2} \right) &\leq \sum_{|j|\geq 1} \left(e^{-|j|n} + e^{-\frac{|j|n}{4}} \right) \\
&\leq 5 e^{-\frac{n}{4}} \qquad (44)
\end{aligned}$$

by assumption $n \geq 17$. The lemma follows by combining (43) and (44). □

G.4. Bounding $\|\boldsymbol{\alpha}\|_2^2$

Lemma 40. *For every odd integer n and parameters $b, \eta, v > 0$, if $\boldsymbol{\alpha}$ is defined as in (35) of Definition 35, then we have*

$$\|\boldsymbol{\alpha}\|_2^2 \leq 4n.$$

Now we are ready for the proof of Lemma 40.

Proof of Lemma 40. Let $w = 2R/n$. Then, we observe that

$$\begin{aligned}
\|\boldsymbol{\alpha}\|_2^2 &= \sum_{j=1}^{n} \alpha_j^2 \\
&\leq \sum_{|j|\leq \frac{n-1}{2}} \left(\frac{2}{\sqrt{2\pi} b} \cos(2\pi j w \eta) \int_{jw - \frac{v}{2}}^{jw + \frac{v}{2}} e^{-x^2/2b^2} \right)^2 \\
&\leq \sum_{|j|\leq \frac{n-1}{2}} \left(\frac{2}{\sqrt{2\pi} b} \cos(2\pi j w \eta) \int_{-\infty}^{\infty} e^{-x^2/2b^2} \right)^2 \\
&\leq \sum_{|j|\leq \frac{n-1}{2}} \left(\frac{2}{\sqrt{2\pi} b} \int_{-\infty}^{\infty} e^{-x^2/2b^2} \right)^2
\end{aligned}$$

because $|\cos(\cdot)| \leq 1$. Hence,

$$\begin{aligned}
\|\boldsymbol{\alpha}\|_2^2 &\leq \sum_{|j|\leq \frac{n-1}{2}} 4 \\
&= 4n \qquad (45)
\end{aligned}$$

as desired. □

G.5. Bounding $\|\mathbf{\Phi}^* \boldsymbol{\alpha}\|^2_{L_2(d\mu)}$

Note that all the results so far hold for any kernel $p(\eta)$ and are independent of the kernel function. Now, we upper bound $\|\mathbf{\Phi}^* \boldsymbol{\alpha}\|_{L_2(d\mu)}$. This quantity depends on the particular choice of kernel, which is assumed to be Gaussian.

Lemma 41. *For every odd integer $n \geq 17$, any parameter $\frac{10}{n} < \lambda \leq \frac{n}{2}$, every $|\eta| \leq 100\sqrt{\log n_\lambda}$, and any $1000 \log^{1.5} n_\lambda \leq R \leq \frac{n}{500\sqrt{\log n_\lambda}}$, and $\frac{R}{2\sqrt{n}} \leq b \leq \frac{R}{4\sqrt{\log n_\lambda}}$, if $\boldsymbol{\alpha}$ is defined as in (35) of Definition 35 with parameter $v = R$, then for the Gaussian kernel with $p(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}$, we have:*

$$\|\mathbf{\Phi}^* \boldsymbol{\alpha}\|^2_{L_2(d\mu)} \leq 6\frac{n^2}{R} \cdot p(\eta) + 3\lambda n. \tag{46}$$

Proof. Recall from Lemma 38 that:

$$\begin{aligned}
|\boldsymbol{\alpha}^* \mathbf{z}(\xi)|^2 &\leq \left| \frac{nv}{2R} \sum_{j=-\infty}^{\infty} \left(e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} - \eta)^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} - \eta) \right) \right.\right. \\
&\qquad \left.\left. + e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} + \eta)^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} + \eta) \right) + \sqrt{\lambda n} \right) \right|^2 \\
&\leq \frac{n^2}{2} \left| \sum_{j=-\infty}^{\infty} \left(e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} - \eta)^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} - \eta) \right) \right.\right. \\
&\qquad \left.\left. + e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} + \eta)^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} + \eta) \right) \right) \right|^2 + 2\left(\sqrt{\lambda n}\right)^2.
\end{aligned}$$

Now, by the definition of the $L_2(d\mu)$ norm, $\|\mathbf{\Phi}^* \boldsymbol{\alpha}\|^2_{L_2(d\mu)} = \int_{-\infty}^{\infty} |\boldsymbol{\alpha}^* z(\xi)|^2 p(\xi)\, d\xi$, and so, we have

$$\begin{aligned}
\|\mathbf{\Phi}^* \boldsymbol{\alpha}\|^2_{L_2(d\mu)} &\leq \int_{-\infty}^{+\infty} \frac{n^2}{2} \left(\sum_{j=-\infty}^{\infty} e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} - \eta)^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} - \eta) \right) \right. \\
&\qquad \left. + e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} + \eta)^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} + \eta) \right) \right)^2 p(\xi)\, d\xi + \int_{-\infty}^{\infty} 2\left(\sqrt{\lambda n}\right)^2 p(\xi)\, d\xi \\
&\leq \int_{-\infty}^{+\infty} n^2 \left(\sum_{j=-\infty}^{\infty} e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} - \eta)^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} - \eta) \right) \right)^2 p(\xi)\, d\xi \\
&\qquad + \int_{-\infty}^{+\infty} n^2 \left(\sum_{j=-\infty}^{\infty} e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} + \eta)^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} + \eta) \right) \right)^2 p(\xi)\, d\xi + 2\lambda n \\
&= 2n^2 \int_{-\infty}^{\infty} \left(\sum_{j=-\infty}^{\infty} e^{-2\pi^2 b^2 (\xi - \frac{jn}{2R} - \eta)^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} - \eta) \right) \right)^2 p(\xi)\, d\xi + 2\lambda n,
\end{aligned} \tag{47}$$

where we have used the inequality $(a_1 + a_2)^2 \leq 2a_1^2 + 2a_2^2$ in the second step, and the last equality occurs because the kernel probability distribution function $p(\xi)$ is symmetric in our case, along with the fact that the underlying sum is over

all j. Now, the integral in (47) can be split into two integrals as follows:

$$\begin{aligned}
&\int_{-\infty}^{\infty} p(\xi) \left(\sum_{j=-\infty}^{\infty} e^{-2\pi^2(\xi - \frac{jn}{2R} - \eta)^2 b^2} \cdot \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} - \eta) \right) \right)^2 d\xi \\
&\qquad = \int_{-10\sqrt{\log n_\lambda}}^{10\sqrt{\log n_\lambda}} p(\xi) \left(\sum_{j=-\infty}^{\infty} e^{-2\pi^2(\xi - \frac{jn}{2R} - \eta)^2 b^2} \cdot \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} - \eta) \right) \right)^2 d\xi \\
&\qquad\quad + \int_{|\xi| \geq 10\sqrt{\log n_\lambda}} p(\xi) \left(\sum_{j=-\infty}^{\infty} e^{-2\pi^2(\xi - \frac{jn}{2R} - \eta)^2 b^2} \cdot \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} - \eta) \right) \right)^2 d\xi. \qquad (48)
\end{aligned}$$

First, we consider the case in which $|\xi| \leq 10\sqrt{\log n_\lambda}$. By the assumption of the lemma, $|\eta| \leq 100\sqrt{\log n_\lambda}$, and hence, $|\xi - \eta| \leq 110\sqrt{\log n_\lambda}$. This implies that $|\xi - \eta| \leq \frac{1}{2}(\frac{n}{2R})$, since we are assuming that $R \leq \frac{n}{500\sqrt{\log(n/\lambda)}}$. Therefore, for any integer $j \neq 0$,

$$e^{-2\pi^2(\xi - \frac{jn}{2R} - \eta)^2 b^2} \leq e^{-(\frac{jn}{R})^2 b^2}.$$

Hence, we have

$$\begin{aligned}
\sum_{|j|\geq 1} e^{-2\pi^2(\xi - \frac{jn}{2R} - \eta)^2 b^2} &\leq \sum_{|j|\geq 1} e^{-(\frac{jn}{R})^2 b^2} \\
&\leq \sum_{|j|\geq 1} e^{-j(\frac{n}{R})^2 b^2} \\
&\leq 3e^{-n/4}, \qquad (49)
\end{aligned}$$

where we used assumptions $b \geq \frac{R}{2\sqrt{n}}$ and $n \geq 17$.

Now, using (49), we see that the first integral in (48) can be bounded as follows:

$$\begin{aligned}
&\int_{-10\sqrt{\log n_\lambda}}^{10\sqrt{\log n_\lambda}} p(\xi) \left(\sum_{j=-\infty}^{\infty} e^{-2\pi^2(\xi - \frac{jn}{2R} - \eta)^2 b^2} \cdot \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} - \eta) \right) \right)^2 d\xi \\
&\leq 2 \int_{-10\sqrt{\log n_\lambda}}^{10\sqrt{\log n_\lambda}} p(\xi) \left(e^{-2\pi^2 b^2 (\xi - \eta)^2} \mathbf{sinc}\left(v(\xi - \eta) \right)^2 \right)^2 d\xi \\
&\qquad + 2 \int_{-10\sqrt{\log n_\lambda}}^{10\sqrt{\log n_\lambda}} p(\xi) \left(\sum_{|j|\geq 1} e^{-2\pi^2(\xi - \frac{jn}{2R} - \eta)^2 b^2} \mathbf{sinc}\left(v(\xi - \frac{jn}{2R} - \eta) \right) \right)^2 d\xi \\
&\leq 2 \int_{-10\sqrt{\log n_\lambda}}^{10\sqrt{\log n_\lambda}} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} \left(e^{-2\pi^2 b^2 (\xi - \eta)^2} \mathbf{sinc}\left(v(\xi - \eta) \right)^2 + 9e^{-n/2} \right) d\xi \\
&= 2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} e^{-b^2(\xi - \eta)^2} \cdot \mathbf{sinc}\left(v(\xi - \eta) \right)^2 d\xi + 18 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} e^{-n/2} d\xi \\
&\leq \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} e^{-b^2(\xi - \eta)^2} \cdot \mathbf{sinc}\left(v(\xi - \eta) \right)^2 d\xi + 18e^{-n/2}. \qquad (50)
\end{aligned}$$

Next, by Claim 32, we have $e^{-\xi^2/2} \leq 3e^{-\eta^2/2}$ for $|\xi - \eta| \leq \frac{10\sqrt{\log n_\lambda}}{b}$. Hence,

$$\begin{aligned}
&\int_{\eta-\frac{10\sqrt{\log n_\lambda}}{b}}^{\eta+\frac{10\sqrt{\log n_\lambda}}{b}} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} e^{-b^2(\xi-\eta)^2} \cdot \mathbf{sinc}\left(v(\xi-\eta)\right)^2 \, d\xi \\
&\quad \le 3 \cdot \frac{1}{\sqrt{2\pi}} e^{-\eta^2/2} \int_{-\infty}^{+\infty} e^{-b^2(\xi-\eta)^2} \cdot \mathbf{sinc}\left(v(\xi-\eta)\right)^2 \, d\xi \\
&\quad \le 3 \cdot \frac{1}{\sqrt{2\pi}} e^{-\eta^2/2} \int_{-\infty}^{+\infty} \mathbf{sinc}\left(v(\xi-\eta)\right)^2 \, d\xi \\
&\quad = \frac{3p(\eta)}{v}
\end{aligned} \tag{51}$$

Note that the last line follows from the fact that $v \cdot \mathbf{sinc}\left(v\eta\right)$ is the Fourier transform of $\mathrm{rect}_v(x)$, and so, by the convolution theorem,

$$\begin{aligned}
\int_{-\infty}^{\infty} \left(v \cdot \mathbf{sinc}\left(vx\right)\right)^2 dx &= \left(\mathrm{rect}_v(x) * \mathrm{rect}_v(x)\right)(0) \\
&= v.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\int_{|\xi-\eta| \ge \frac{10\sqrt{\log n_\lambda}}{b}} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} e^{-b^2(\xi-\eta)^2} \cdot \mathbf{sinc}\left(v(\xi-\eta)\right)^2 \, d\xi &\le \left(\frac{\lambda}{n}\right)^{50} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} \, d\xi \\
&= \left(\frac{\lambda}{n}\right)^{50},
\end{aligned} \tag{52}$$

since the $\mathbf{sinc}\left(\cdot\right)$ function is bounded by 1 in absolute value. Thus, (50), (51), and (52) imply that

$$\int_{-10\sqrt{\log n_\lambda}}^{10\sqrt{\log n_\lambda}} p(\xi) \Big(\sum_{j=-\infty}^{\infty} e^{-2\pi^2(\xi-\frac{jn}{2R}-\eta)^2 b^2} \cdot \mathbf{sinc}\left(v(\xi-\frac{jn}{2R}-\eta)\right) \Big)^2 d\xi \;\le\; \frac{3p(\eta)}{v} + \left(\frac{\lambda}{n}\right)^{50} + 18e^{-n/2}. \tag{53}$$

Next, we bound the second integral in (48). Consider ξ satisfying $|\xi| \ge 10\sqrt{\log n_\lambda}$. Note that the following upper bound holds for any ξ and η:

$$\begin{aligned}
\sum_{j=-\infty}^{\infty} e^{-2\pi^2(\xi-\frac{jn}{2R}-\eta)^2 b^2} \cdot \mathbf{sinc}\left(v(\xi-\frac{jn}{2R}-\eta)\right) &\le \sum_{j=-\infty}^{\infty} e^{-2\pi^2(\xi-\frac{jn}{2R}-\eta)^2 b^2} \\
&= 1 + \sum_{|j|\ge 1} e^{-2\pi^2(\xi-\frac{jn}{2R}-\eta)^2 b^2} \\
&\le 1 + \frac{2R}{n} \int_{-\infty}^{\infty} e^{-2\pi^2(\xi-x-\eta)^2 b^2} \, dx \\
&\le 2,
\end{aligned} \tag{54}$$

where we have used the fact that $\frac{b}{R} \le \frac{1}{4\sqrt{\log n_\lambda}} \le 1/\sqrt{2}$. Thus,

$$\begin{aligned}
&\int_{|\xi|\ge 10\sqrt{\log n_\lambda}} p(\xi) \left(\sum_{j=-\infty}^{\infty} e^{-2\pi^2(\xi-\frac{jn}{2R}-\eta)^2 b^2} \mathbf{sinc}\left(v(\xi-\frac{jn}{2R}-\eta)\right) \right)^2 d\xi \\
&\quad \le 2 \int_{|\xi|\ge 10\sqrt{\log n_\lambda}} p(\xi) \, d\xi \\
&\quad \le \left(\frac{\lambda}{n}\right)^{50},
\end{aligned} \tag{55}$$

by Claim 25. Combining (47), (48), (53), and (55) now yields the desired result. □

Proof of Theorem 14. Note that we can choose data points $x_1, x_2, \ldots, x_n$ and the vector $\boldsymbol{\alpha}$ according to the construction in Definition 35 with $v = R$ and $b = \frac{R}{4\sqrt{\log n_\lambda}}$. Thus, Lemmas 39, 40, and 41, as well as (30), imply that

$$\begin{aligned}\tau_\lambda(\boldsymbol{\eta}) &\geq \frac{p(\eta)\cdot|\boldsymbol{\alpha}^*\mathbf{z}(\eta)|^2}{\|\boldsymbol{\Phi}^*\boldsymbol{\alpha}\|^2_{L_2(d\mu)}+\lambda\|\boldsymbol{\alpha}\|_2^2}\\ &\geq \frac{p(\eta)\cdot\left(\frac{n}{5}\right)^2}{\left(6\frac{n^2}{R}\cdot p(\eta)+3\lambda n\right)+\lambda(4n)}\\ &\geq \frac{R}{150}\left(\frac{p(\eta)}{p(\eta)+2Rn_\lambda^{-1}}\right),\end{aligned}$$

as desired. □

H. Proof of Corollary 15

First claim of the corollary (upper bound on statistical dimension): Let $t = 10\sqrt{\log n_\lambda}$. We have:

$$s_\lambda(\mathbf{K}) = \int_{\mathbb{R}}\tau(\eta)d\eta = \int_{[-t,t]}\tau(\eta)d\eta + \int_{[-\infty,-t]\cup[t,\infty]}\tau(\eta)d\eta$$

By the naive bound in Proposition 4 and Claim 25 we have:

$$\begin{aligned}\int_{[-\infty,-t]\cup[t,\infty]}\tau(\eta)d\eta &\leq n_\lambda\int_{[-\infty,-t]\cup[t,\infty]}\frac{e^{-\eta^2/2}}{\sqrt{2\pi}}d\eta\\ &\leq n_\lambda\cdot\left(\frac{e^{-t^2/2}}{t}\right)\\ &\leq 1\end{aligned}\tag{56}$$

Further, by the more refined bound of Theorem 13, for any $\eta \leq 10\sqrt{\log n_\lambda} = t$ we have

$$\begin{aligned}\int_{[-t,t]}\tau(\eta)d\eta &\leq \int_{[-t,t]}25(R+3000\log^{1.5}n_\lambda)\,d\eta\\ &\leq 50t\cdot(R+3000\log^{1.5}n_\lambda)\\ &= O\Big(\sqrt{\log n_\lambda}\cdot R+\log^2 n_\lambda\Big).\end{aligned}\tag{57}$$

Combining (56) and (57) gives the lemma.

Second claim of the corollary: Note that for all $|\eta| \leq \sqrt{2\log\frac{n_\lambda}{R}}$ we have $p(\eta) \geq \frac{R}{\sqrt{2\pi}n_\lambda}$, hence we have:

$$p(\eta) + 2R/n_\lambda \leq 7p(\eta)$$

hence, by Theorem 14, we have:

$$\tau(\eta) \geq \frac{R}{150}\left(\frac{1}{7}\right)$$

And for $|\eta| > \sqrt{2\log\frac{n_\lambda}{R}}$ we have:

$$p(\eta) + 2R/n_\lambda \leq 3R/n_\lambda$$

therefore,

$$\begin{aligned}s_\lambda(\mathbf{K}) &= \int_{-\infty}^{\infty}\tau(\eta)d\eta\\ &\geq \int_{-\sqrt{2\log\frac{n_\lambda}{R}}}^{\sqrt{2\log\frac{n_\lambda}{R}}}\frac{R}{1050}d\eta + \int_{|\boldsymbol{\eta}|>\sqrt{2\log\frac{n_\lambda}{R}}}\frac{R}{150}\left(\frac{p(\eta)}{3R/n_\lambda}\right)d\eta\\ &= \Omega(R\sqrt{\log\frac{n_\lambda}{R}})\end{aligned}\tag{58}$$

I. Proof of Theorem 8 and 10

Proof of Theorem 8. We show a lower bound on the number of samples required under the random feature map of Rahimi and Recht by exhibiting a set of data points for which the appropriate number of samples does not suffice.

Our goal is to show that if we take s samples $\eta_1, \eta_2, \ldots, \eta_s$ from the distribution defined by p, for s too small, then there is an $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{R}^n$ such that with at least constant probability,

$$\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{K} + \lambda \mathbf{I}_n)\boldsymbol{\alpha} < \frac{2}{3}\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{Z}\mathbf{Z}^* + \lambda \mathbf{I}_n)\boldsymbol{\alpha}. \tag{59}$$

By (3), we have

$$\begin{aligned}
\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} &= \sum_{j,k} \alpha_j \alpha_k \cdot k(x_j, x_k) \\
&= \sum_{j,k} \int_{-\infty}^{\infty} e^{-2\pi i \eta (x_j - x_k)} \alpha_j \alpha_k p(\eta)\, d\eta \\
&= \int_{-\infty}^{\infty} \left(\sum_{j=1}^{n} \alpha_j e^{-2\pi i \eta x_j} \right) \left(\sum_{k=1}^{n} \alpha_k e^{2\pi i \eta x_k} \right) p(\eta)\, d\eta \\
&= \int_{-\infty}^{\infty} p(\eta) \left| \sum_{j=1}^{n} \alpha_j e^{2\pi i \eta x_j} \right|^2 d\eta.
\end{aligned}$$

Also, by the definition of $\mathbf{Z}$ and φ (see Section 2.2), we have

$$\begin{aligned}
\boldsymbol{\alpha}^T \mathbf{Z}\mathbf{Z}^* \boldsymbol{\alpha} &= \left\| \sum_{j=1}^{n} \alpha_j \varphi(x_j) \right\|_2^2 \\
&= \sum_{k=1}^{s} \left| \sum_{j=1}^{n} \alpha_j \cdot \frac{1}{\sqrt{s}} e^{2\pi i \eta_k x_j} \right|^2 \\
&= \frac{1}{s} \sum_{k=1}^{s} \left| \sum_{j=1}^{n} \alpha_j e^{2\pi i \eta_k x_j} \right|^2,
\end{aligned}$$

where $\eta_1, \eta_2, \ldots, \eta_s$ are the s samples from the distribution given by p. Hence, (59) is equivalent to

$$\int_{-\infty}^{\infty} p(\eta) \left| \sum_{j=1}^{n} \alpha_j e^{2\pi i \eta x_j} \right|^2 d\eta + \frac{1}{3}\lambda \|\boldsymbol{\alpha}\|_2^2 < \frac{2}{3} \cdot \frac{1}{s} \sum_{k=1}^{s} \left| \sum_{j=1}^{n} \alpha_j e^{2\pi i \eta_k x_j} \right|^2. \tag{60}$$

We again use the same construction of n data points $x_1, x_2, \ldots, x_n \in \mathbb{R}$, according to the construction in Definition 33. Moreover, we define η^* to be

$$\eta^* = \max_{1 \le j \le s} |\eta_j|$$

and let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ be given by

$$\alpha_j = f_{\eta^*, b, v}(x_j),$$

where $b = R/4\sqrt{\log n_\lambda}$ and $v = \delta$. We will show that this choice of data points and $\boldsymbol{\alpha}$ satisfy (60).

First, we upper bound the first term on the left side of (60). Note that by Claim 26, with probability at least $1/2$ over the

samples $z_1, z_2, \ldots, z_s$, we have

$$\begin{aligned}\int_{-\infty}^{\infty} p(\eta)\left|\sum_{j=1}^{n}\alpha_j e^{2\pi i\eta x_j}\right|^2 d\eta &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\eta^2/2}\left|\sum_{j=-m}^{m}\alpha_j e^{2\pi i\eta x_j}\right|^2 d\eta \\ &= \|\boldsymbol{\Phi}^*\boldsymbol{\alpha}\|^2_{L_2(d\mu)} \\ &\le \frac{6n^2}{R}\cdot p(\eta^*) + 3\lambda n \\ &\le \frac{48n^2}{R}\cdot\frac{\sqrt{\log s}}{s} + 3\lambda n. \end{aligned} \tag{61}$$

where we have let $\eta = \eta^*$ and applied Lemma 41.

Next, we bound the right side of (60) from below. Note that

$$\begin{aligned}\frac{1}{s}\sum_{k=1}^{s}\left|\sum_{j=1}^{n}\alpha_j e^{2\pi i\eta_k x_j}\right|^2 &\ge \frac{1}{s}\left|\sum_{j=1}^{n}\alpha_j e^{2\pi i\eta^* x_j}\right|^2 \\ &= \frac{1}{s}(\boldsymbol{\alpha}^*\mathbf{z}(\eta^*))^2 \\ &\ge \frac{1}{s}\left(\frac{n}{5}\right)^2 = \frac{n^2}{25s}, \end{aligned} \tag{62}$$

by Lemma 39 applied with $\eta = \eta^*$.

We also require the following estimate of $\|\boldsymbol{\alpha}\|_2^2$, which is provided by Lemma 40:

$$\|\boldsymbol{\alpha}\|_2^2 \le 4n. \tag{63}$$

Note that by combining (61), (62), and (63), we have that with probability at least 1/2,

$$\begin{aligned}\int_{-\infty}^{\infty} p(\eta)\left|\sum_{j=1}^{n}\alpha_j e^{2\pi i\eta x_j}\right|^2 d\eta + \frac{1}{3}\lambda\|\boldsymbol{\alpha}\|_2^2 &\le \frac{48n^2}{R}\cdot\frac{\sqrt{\log s}}{s} + 3\lambda n + \frac{4}{3}\lambda n \\ &\le \frac{2n^2}{75s} \\ &\le \frac{2}{3}\cdot\frac{1}{s}\sum_{k=1}^{s}\left|\sum_{j=1}^{n}\alpha_j e^{2\pi i\eta_k x_j}\right|^2, \end{aligned}$$

since $s \le n_\lambda/400$ and also because $R \ge 3000\log^{1.5}(n_\lambda)$. This completes the proof. □

Proof of Theorem 10. By the assumptions of the theorem n is an integer, parameter $0 < \lambda \le n/2$, and $R > 0$, and all $x_1, ..., x_n \in [-R, R]$ and $p(\eta) = \frac{1}{\sqrt{2\pi}}e^{-\frac{\eta^2}{2}}$, therefore all the preconditions of Proposition 4, and Theorem 13 are satisfied and hence the theorem and proposition go through and for every η we have:

$$\tau_\lambda(\eta) \le \bar{\tau}_R(\eta)$$

Hence applying Lemma 6 with $\tilde{\tau}(\eta) = \bar{\tau}_R(\eta)$ gives the desired spectral approximation with $\frac{8}{3}\Delta^{-2}s_{\bar{\tau}_R}\ln(16s_{\bar{\tau}_R}/\rho)$ samples where $s_{\bar{\tau}_R} = \int_{\mathbb{R}}\bar{\tau}_R(\eta)d\eta$. Now we show that $s_{\bar{\tau}_R} = O(R\sqrt{\log(n_\lambda)} + \log^2 n_\lambda)$.

Let $t = 10\sqrt{\log n_\lambda}$. We have:

$$s_{\bar{\tau}_R} = \int_{\mathbb{R}}\bar{\tau}_R(\eta)d\eta = \int_{[-t,t]}\bar{\tau}_R(\eta)d\eta + \int_{[-\infty,-t]\cup[t,\infty]}\bar{\tau}_R(\eta)d\eta$$

By Definition 9 and Claim 25 we have:

$$\begin{aligned}\int_{[-\infty,-t]\cup[t,\infty]} \bar{\tau}_R(\eta)d\eta &= n_\lambda \int_{[-\infty,-t]\cup[t,\infty]} \frac{e^{-\eta^2/2}}{\sqrt{2\pi}} d\eta \\ &\leq n_\lambda \cdot \left(2\frac{e^{-t^2/2}}{\sqrt{2\pi}t}\right) \\ &\leq n_\lambda \cdot \left(\frac{e^{-t^2/2}}{t}\right) \\ &\leq 1\end{aligned}$$

Furthermore, for any $\eta \leq 10\sqrt{\log n_\lambda} = t$ we have

$$\begin{aligned}\int_{[-t,t]} \tau(\eta)d\eta &\leq \int_{[-t,t]} 25(R + 3000\log^{1.5} n_\lambda)\, d\eta \\ &\leq 50t \cdot (R + 3000\log^{1.5} n_\lambda) \\ &= O\left(\sqrt{\log n_\lambda} \cdot R + \log^2 n_\lambda\right).\end{aligned}$$

Combining the bounds above gives the result.

Sampling from $\bar{p}_R(\cdot)$: Sampling from $\bar{p}_R(\cdot)$ amounts to sampling from a mixture of the uniform distribution on $[-10\sqrt{\log(n_\lambda)}, +10\sqrt{\log(n_\lambda)}]$ and from the tail of the Gaussian distribution: with probability $\frac{25\max(R,3000\log^{1.5} n_\lambda)}{s_{\bar{\tau}_R}}$. $20\sqrt{\log(n_\lambda)}$ sample from the uniform distribution and with remaining probability sample from the tail of the Gaussian. Sampling from the tail of the Gaussian can be easily accomplished via rejection sampling at unit expected cost. Indeed, we only need to generate a sample from the tail with probability proportional to the mass of the tail. On the other hand, once we do, the expected cost of obtaining a sample via rejection sampling is inversely proportional to the amount of mass in the tail, leading to unit cost in expectation.

□