# 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 $\varphi$ 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 $\Delta$ 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 $\sigma$.

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 $\lambda_n$, noting that in our notation, $\lambda$ is scaled by an $n$ factor as compared to those works. As the ratio between $n$ and $\lambda$ 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 $\Delta$-*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 $\Delta$-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 $\Delta$-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 $\lambda$ 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 $\sigma_\nu^2$. 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* $\Delta$*-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 $\Delta$ 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 $\Delta$-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 $\Delta$-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 $n/\lambda$, 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 $\lambda$ 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 $\Delta$-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 $\Delta$ 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 $\Delta$-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 $\Delta$ 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 $n_\lambda$ 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, $\lambda$ 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 $n_\lambda$, 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 $\eta$. 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 $\Delta$-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 $\eta$ 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 $\ell_2$ 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 $\lambda$ on the statistical dimension!

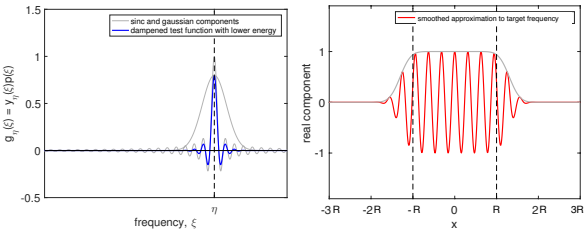


*Figure 1.* 'Soft spike' function $y$ and its Fourier transform $\mathbf{\Phi}y$, which is approximately a pure cosine wave on $[-R, R]$.

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

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

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 $\eta$ 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.