Journal of Machine Learning Research 22 (2021) 1-51 Submitted 12/20; Published 5/21

Towards a Unified Analysis of Random Fourier Features

Zhu Li ZHU.LI@UCL.AC.UK
Gatsby Computational Neuroscience Unit, University College London, London, UK
Department of Statistics, University of Oxford, Oxford, UK *

Jean-Francois Ton JEAN-FRANCOIS.TON@SPC.OX.AC.UK
Department of Statistics, University of Oxford, Oxford, UK

Dino Oglic DINO.OGLIC@ASTRAZENECA.COM
AstraZeneca PLC, Cambridge, UK
Department of Engineering, King's College London, London, UK *

Dino Sejdinovic DINO.SEJDINOVIC@STATS.OX.AC.UK
Department of Statistics, University of Oxford, Oxford, UK

Editor: Kilian Weinberger

Abstract

Random Fourier features is a widely used, simple, and effective technique for scaling up kernel methods. The existing theoretical analysis of the approach, however, remains focused on specific learning tasks and typically gives pessimistic bounds which are at odds with the empirical results. We tackle these problems and provide the first unified risk analysis of learning with random Fourier features using the squared error and Lipschitz continuous loss functions. In our bounds, the trade-off between the computational cost and the learning risk convergence rate is problem specific and expressed in terms of the regularization parameter and the *number of effective degrees of freedom*. We study both the standard random Fourier features method for which we improve the existing bounds on the number of features required to guarantee the corresponding minimax risk convergence rate of kernel ridge regression, as well as a data-dependent modification which samples features proportional to *ridge leverage scores* and further reduces the required number of features. As ridge leverage scores are expensive to compute, we devise a simple approximation scheme which provably reduces the computational cost without loss of statistical efficiency. Our empirical results illustrate the effectiveness of the proposed scheme relative to the standard random Fourier features method.

Keywords: Kernel methods, random Fourier features, stationary kernels, kernel ridge regression, Lipschitz continuous loss, support vector machines, logistic regression, ridge leverage scores.

1. Introduction

Kernel methods are one of the pillars of machine learning (Schölkopf and Smola, 2001; Schölkopf et al., 2004), as they give us a flexible framework to model complex functional relationships in a principled way and also come with well-established statistical properties and theoretical guarantees (Caponnetto and De Vito, 2007; Steinwart and Christmann, 2008). The key ingredient, known as *kernel trick*, allows implicit computation of an inner product between rich feature representations of data through the kernel evaluation $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$, while the actual feature mapping $\varphi : \mathcal{X} \to \mathcal{H}$ between a data domain $\mathcal{X}$ and some high and often infinite dimensional Hilbert space $\mathcal{H}$

$*$. Previous affiliations where significant portion of this work was completed.

©2021 Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic.

License: CC-BY 4.0, see `https://creativecommons.org/licenses/by/4.0/`. Attribution requirements are provided at `http://jmlr.org/papers/v22/20-1369.html`.

is never computed. However, such convenience comes at a price: due to operating on all pairs of observations, kernel methods inherently require computation and storage which is at least quadratic in the number of observations, and hence often prohibitive for large datasets. In particular, the kernel matrix has to be computed, stored, and often inverted. As a result, a flurry of research into scalable kernel methods and the analysis of their performance emerged (Rahimi and Recht, 2007; Mahoney and Drineas, 2009; Bach, 2013; Alaoui and Mahoney, 2015; Rudi et al., 2015; Rudi and Rosasco, 2017; Rudi et al., 2017; Zhang et al., 2015). Among the most popular frameworks for fast approximations to kernel methods are random Fourier features (RFF) due to Rahimi and Recht (2007). The idea of random Fourier features is to construct an explicit feature map which is of a dimension much lower than the number of observations, but with the resulting inner product which approximates the desired kernel function $k(x, y)$. In particular, random Fourier features rely on Bochner's theorem (Bochner, 1932; Rudin, 2017) which tells us that any bounded, continuous and shift-invariant kernel is the Fourier transform of a bounded positive measure, called the spectral measure. The feature map is then constructed using samples drawn from the spectral measure. Essentially, any kernel method can then be adjusted to operate on these explicit feature maps (i.e., primal representations), greatly reducing the computational and storage costs, while in practice mimicking performance of the original method.

Despite their empirical success, the theoretical understanding of statistical properties of random Fourier features is incomplete, and the question of how many features are needed, in order to obtain a method with performance provably comparable to the original one, remains without a definitive answer. Currently, there are two main lines of research addressing this question. The first line considers the approximation error of the kernel matrix itself (e.g., see Rahimi and Recht, 2007; Sriperumbudur and Szabó, 2015; Sutherland and Schneider, 2015, and references therein) and bases performance guarantees on the accuracy of this approximation. However, all of these works require $\Omega(n)$ features (n being the number of observations), which translates to no computational savings at all and is at odds with empirical findings. Realizing that the approximation of kernel matrices is just a means to an end, the second line of research aims at directly studying the risk and generalization properties of random Fourier features in various supervised learning scenarios. Arguably, first such result is already in Rahimi and Recht (2009), where supervised learning with Lipschitz continuous loss functions is studied. However, the bounds therein still require a pessimistic $\Omega(n)$ number of features and cannot demonstrate the efficiency of random Fourier features theoretically. In Bach (2017b), the generalization properties are studied from a function approximation perspective, showing for the first time that fewer features could preserve the statistical properties of the original method, but in the case where a certain data-dependent sampling distribution is used instead of the spectral measure. These results also do not apply to kernel ridge regression and the mentioned sampling distribution is typically itself intractable. Avron et al. (2017) study the random Fourier features for kernel ridge regression in the fixed design setting. They show that it is possible to use $o(n)$ features and have the risk of the linear ridge regression estimator based on random Fourier features close to the risk of the original kernel estimator, also relying on a modification to the sampling distribution. However, their result restricts the data distribution to have finite support, and a tractable method to sample from a modified distribution is proposed for the Gaussian kernel only. A highly refined analysis of kernel ridge regression is given by Rudi and Rosasco (2017), where it is shown that $\Omega(\sqrt{n} \log n)$ features suffices for an optimal $\mathcal{O}(1/\sqrt{n})$ learning rate in a minimax sense (Caponnetto and De Vito, 2007). Moreover, the number of features can be reduced even further if a data-dependent sampling distribution is employed. While these are groundbreaking results, guaranteeing computational savings without any loss of statistical efficiency, they require some technical assumptions that are difficult to verify. Moreover, to what extent the bounds can be improved by utilizing data-dependent distributions still remains unclear. Finally, it does not seem straightforward to generalize the approach of Rudi and Rosasco (2017) to kernel support vector machines (SVM) and/or kernel logistic regression (KLR). Recently, Sun et al. (2018) have provided novel bounds for random Fourier features in the SVM setting, assuming the Massart's low noise condition and that the target hypothesis lies in the corresponding reproducing kernel Hilbert space. The bounds, however, require the sample complexity and the number of features to be exponential in the dimension of the instance space and this can be problematic for high dimensional instance spaces. The theoretical results are also restricted to the hinge loss (without means to generalize to other loss functions) and require optimized features.

In this paper, we address the gaps mentioned above by making the following contributions:

- We devise a simple framework for the unified analysis of generalization properties of random Fourier features, which applies to kernel ridge regression, as well as to kernel support vector machines and logistic regression.

- For the plain random Fourier features sampling scheme (Section 3.1.1), we provide, to the best of our knowledge, the sharpest results on the number of features required. In particular, we show that already with $\Omega(\sqrt{n}\log d_{\mathbf{K}}^{\lambda})$ random features one can obtain the minimax learning rate of kernel ridge regression (Caponnetto and De Vito, 2007), where $d_{\mathbf{K}}^{\lambda}$ corresponds to the notion of *the number of effective degrees of freedom* (Bach, 2013) with $d_{\mathbf{K}}^{\lambda} \ll n$ and $\lambda := \lambda(n)$ is the regularization parameter.

- In the case of a modified data-dependent sampling distribution (Section 3.1.2), the so called *empirical ridge leverage score distribution*, we demonstrate that $\Omega(d_{\mathbf{K}}^{\lambda})$ features suffice for the learning risk to converge at $\mathcal{O}(\lambda)$ rate in kernel ridge regression. In addition, we show that the excess risk convergence rate of the estimator based on random Fourier features can (depending on the decay rate of the spectrum of the kernel function) be upper bounded by $\mathcal{O}(\log n/n)$ or even $\mathcal{O}(1/n)$, which implies much faster convergence than the standard $\mathcal{O}(1/\sqrt{n})$ rate featuring in the majority of previous bounds.

- For plain random Fourier features in the Lipschitz continuous loss setting (Section 3.2.1), we show that $\Omega(1/\lambda)$ features are sufficient to ensure $\mathcal{O}(\sqrt{\lambda})$ learning risk rate in kernel support vector machines and kernel logistic regression. Moreover, using the empirical ridge leverage score distribution, we show that $\Omega(d_{\mathbf{K}}^{\lambda})$ features are sufficient to guarantee $\mathcal{O}(\sqrt{\lambda})$ risk convergence rate in these two learning settings.

- Similarly, under the low noise assumption (Section 3.2.2), our refined analysis for the Lipschitz continuous loss function demonstrates that it is possible to achieve $\mathcal{O}(1/n)$ excess risk convergence rate. The required number of features can be $\Omega(\log n \log\log n)$ when learning using the empirical leverage score distribution, or even constant in some benign cases. To the best of our knowledge, this is the first result offering non-trivial computational savings for approximations in problems with Lipschitz loss functions.

- Finally, as the empirical ridge leverage scores distribution is typically costly to compute, we give a fast algorithm to generate samples from the approximated empirical leverage distribution

(Section 4). Utilizing these samples one can significantly reduce the computation time during the in-sample prediction and testing stages, i.e., $\mathcal{O}(n \log n \log \log n)$ and $\mathcal{O}(\log n \log \log n)$, respectively. We also include a proof that characterizes the trade-off between the computational cost and the learning risk of the algorithm, showing that the statistical efficiency can be preserved while provably reducing the required computational cost.

We remark that a shorter version of this paper has appeared before in Li et al. (2019). In this extended version, we have included a refined analysis of the trade-offs between the computational cost and statistical efficiency for the Lipschitz continuous loss (Theorem 19). Utilizing the notion of local Rademacher complexity, we show that the random Fourier features estimator can obtain a faster learning rate than the traditional minimax optimal rate of $\mathcal{O}(1/\sqrt{n})$. We also provide a theoretical analysis of the trade-offs between the computational cost and accuracy for the proposed approximate leverage score sampling algorithm in the Lipschitz loss setting (Theorem 21).

2. Background

In this section, we provide some notation and preliminary results that will be used throughout the paper. Henceforth, we denote the Euclidean norm of a vector $\mathbf{a} \in \mathbb{R}^n$ with $\|\mathbf{a}\|_2$ and the operator norm of a matrix $A \in \mathbb{R}^{n_1 \times n_2}$ with $\|A\|_2$. Let $\mathcal{H}$ be a Hilbert space with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ as its inner product and $\| \cdot \|_{\mathcal{H}}$ as its norm. We use $\mathrm{Tr}(\cdot)$ to denote the trace of an operator or a matrix. Given a measure $d\rho$, we use $L_2(d\rho)$ to denote the space of square-integrable functions with respect to $d\rho$.

2.1 Supervised Learning with Kernels

We first briefly review the standard problem setting for supervised learning with kernel methods. Let $\mathcal{X}$ be an instance space, $\mathcal{Y}$ a label space, and $P(x,y) = P_x P(y \mid x)$ a joint probability density function on $\mathcal{X} \times \mathcal{Y}$ defining the relationship between an instance $x \in \mathcal{X}$ and a label $y \in \mathcal{Y}$. A training sample is a set of examples $\{(x_i, y_i)\}_{i=1}^n$ sampled independently from $P(x,y)$. The value P_x is called the marginal distribution of an instance $x \in \mathcal{X}$. The goal of a supervised learning task defined with a kernel function k (and the associated reproducing kernel Hilbert space $\mathcal{H}$) is to find a hypothesis $f \colon \mathcal{X} \to \mathcal{Y}$ such that $f \in \mathcal{H}$ and $f(x)$ is a good estimate of the label $y \in \mathcal{Y}$ corresponding to a previously unseen instance $x \in \mathcal{X}$. While in regression tasks $\mathcal{Y} \subset \mathbb{R}$, in classification tasks it is typically the case that $\mathcal{Y} = \{-1, 1\}$. As a result of the representer theorem, an empirical risk minimization problem in this setting can be expressed as (Schölkopf and Smola, 2001)

$$\begin{aligned} \hat{f}^{\lambda} &:= \operatorname*{arg\,min}_{f \in \mathcal{H}} \ \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \\ &= \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \ \frac{1}{n} \sum_{i=1}^{n} l(y_i, (\mathbf{K}\alpha)_i) + \lambda \alpha^T \mathbf{K} \alpha \ , \end{aligned} \tag{1}$$

where $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$ with $\alpha \in \mathbb{R}^n$, $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function, $\mathbf{K}$ is the kernel matrix, and λ is the regularization parameter. The hypothesis $\hat{f}^{\lambda}$ is an empirical estimator and its ability to capture the relationship between instances and labels given by P is measured by the learning risk (Caponnetto and De Vito, 2007)

$$\mathbb{E}_P[l_{\hat{f}^{\lambda}}] = \int_{\mathcal{X} \times \mathcal{Y}} l(y, \hat{f}^{\lambda}(x)) dP(x,y) \ ,$$

where we use l_f to denote $l(y, f(x))$. When it is clear from the context, we will omit P from the expectation, i.e., writing $\mathbb{E}_P[l_{\hat{f}^\lambda}]$ as $\mathbb{E}[l_{\hat{f}^\lambda}]$. The empirical distribution $P_n(x, y)$ is given by a sample of n examples drawn independently from $P(x, y)$. The empirical risk is used to estimate the learning risk $\mathbb{E}[l_{\hat{f}^\lambda}]$ and it is given by

$$\mathbb{E}_n[l_{\hat{f}^\lambda}] = \frac{1}{n}\sum_{i=1}^{n} l(y_i, \hat{f}^\lambda(x_i)) \ .$$

Similar to Rudi and Rosasco (2017) and Caponnetto and De Vito (2007), we will assume [1] the existence of $f_\mathcal{H} \in \mathcal{H}$ such that $f_\mathcal{H} = \arg\inf_{f\in\mathcal{H}} \ \mathbb{E}[l_f]$. The assumption implies that there exists some ball of radius $R > 0$ containing $f_\mathcal{H}$ in its interior. Our theoretical results do not require prior knowledge of this constant and hold uniformly over all finite radii. Furthermore, for all the estimators returned by the empirical risk minimization, we assume that they have bounded reproducing kernel Hilbert space norms. As a result, to simplify our derivations and constant terms in our bounds, unless specifically point out, we have (without loss of generality) assumed that all the estimators appearing in the remainder of the manuscript are within the *unit ball* of our reproducing kernel Hilbert space.

Note that $\mathbb{E}[l_{f_\mathcal{H}}]$ is the lowest learning risk one can achieve in the reproducing kernel Hilbert space $\mathcal{H}$. Hence, the theoretical studies of the estimator $\hat{f}^\lambda$ often concern how fast its learning risk $\mathbb{E}[l_{\hat{f}^\lambda}]$ converges to $\mathbb{E}[l_{f_\mathcal{H}}]$, in other words, how fast the excess risk $\mathbb{E}[l_{\hat{f}^\lambda}] - \mathbb{E}[l_{f_\mathcal{H}}]$ converges to zero. In the remainder of the manuscript, we will refer to the rate at which the excess risk converges to zero as the *learning rate*.

2.2 Random Fourier Features

Random Fourier features is a widely used, simple, and effective technique for scaling up kernel methods. The underlying principle of the approach is a consequence of Bochner's theorem (Bochner, 1932), which states that any bounded, continuous, and shift-invariant kernel is the Fourier transform of a bounded positive measure. This measure can be transformed/normalized into a probability measure which is typically called the spectral measure of the kernel. Assuming the spectral measure $d\tau$ has a density function $p(\cdot)$, the corresponding shift-invariant kernel can be written as

$$k(x, y) = \int_\mathcal{V} e^{-2\pi i v^T(x-y)} d\tau(v) = \int_\mathcal{V} \left(e^{-2\pi i v^T x}\right)\left(e^{-2\pi i v^T y}\right)^* p(v) dv \ , \tag{2}$$

where c^* denotes the complex conjugate of $c \in \mathbb{C}$. Typically, the kernel is real valued and we can ignore the imaginary part in this equation (e.g., see Rahimi and Recht, 2007). The principle can be further generalized by considering the class of kernel functions which can be decomposed as

$$k(x, y) = \int_\mathcal{V} z(v, x) z(v, y) p(v) dv \ , \tag{3}$$

where $z\colon \mathcal{V} \times \mathcal{X} \to \mathbb{R}$ is a continuous and bounded function with respect to v and x. The main idea behind random Fourier features is to approximate the kernel function by its Monte-Carlo estimate

$$\tilde{k}(x, y) = \frac{1}{s}\sum_{i=1}^{s} z(v_i, x) z(v_i, y) \ , \tag{4}$$

[1] The existence of $f_\mathcal{H}$ depends on the complexity of $\mathcal{H}$ which is related to the data distribution $P(y|x)$. For more details, please see Caponnetto and De Vito (2007) and Rudi and Rosasco (2017).

with the reproducing kernel Hilbert space $\tilde{\mathcal{H}}$ (note that in general $\tilde{\mathcal{H}} \not\subseteq \mathcal{H}$) and $\{v_i\}_{i=1}^s$ sampled independently from the spectral measure. In Bach (2017a, Appendix A), it has been established that a function $f \in \mathcal{H}$ can be expressed as: [2]

$$f(x) = \int_{\mathcal{V}} g(v)z(v,x)p(v)dv \qquad (\forall x \in \mathcal{X}) \tag{5}$$

where $g \in L_2(d\tau)$ is a real-valued function such that $\|g\|^2_{L_2(d\tau)} < \infty$ and $\|f\|_{\mathcal{H}} = \min_g \|g\|_{L_2(d\tau)}$, with the minimum taken over all possible decompositions of f. Thus, one can take an independent sample $\{v_i\}_{i=1}^s \sim p(v)$ (we refer to this sampling scheme as *plain RFF*) and approximate a function $f \in \mathcal{H}$ at a point $x_j \in \mathcal{X}$ by

$$\tilde{f}(x_j) = \sum_{i=1}^{s} \alpha_i z(v_i, x_j) \coloneqq \mathbf{z}_{x_j}(\mathbf{v})^T \alpha \quad \text{with} \quad \alpha \in \mathbb{R}^s \ .$$

In standard estimation problems, it is typically the case that for a given set of instances $\{x_i\}_{i=1}^n$ one approximates $\mathbf{f}_x = [f(x_1), \cdots, f(x_n)]^T$ by

$$\tilde{\mathbf{f}}_x = [\mathbf{z}_{x_1}(\mathbf{v})^T\alpha, \cdots, \mathbf{z}_{x_n}(\mathbf{v})^T\alpha]^T \coloneqq \mathbf{Z}\alpha \ ,$$

where $\mathbf{Z} \in \mathbb{R}^{n \times s}$ with $\mathbf{z}_{x_j}(\mathbf{v})^T$ as its j-th row.

As the latter approximation is simply a Monte Carlo estimate, one could also select an importance weighted probability density function $q(\cdot)$ and sample features $\{v_i\}_{i=1}^s$ from q (we refer to this sampling scheme as *weighted RFF*). Then, the function value $f(x_j)$ can be approximated by

$$\tilde{f}_q(x_j) = \sum_{i=1}^{s} \beta_i z_q(v_i, x_j) \coloneqq \mathbf{z}_{q,x_j}(\mathbf{v})^T \beta \ ,$$

with $z_q(v_i, x_j) = \sqrt{p(v_i)/q(v_i)}z(v_i, x_j)$ and $\mathbf{z}_{q,x_j}(\mathbf{v}) = [z_q(v_1, x_j), \cdots, z_q(v_s, x_j)]^T$. Hence, a Monte-Carlo estimate of $\mathbf{f}_x$ can be written in the matrix form as $\tilde{\mathbf{f}}_{q,x} = \mathbf{Z}_q\beta$, where $\mathbf{Z}_q \in \mathbb{R}^{n \times s}$ with $\mathbf{z}_{q,x_j}(\mathbf{v})^T$ as its j-th row.

Let $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{K}}_q$ be Gram-matrices with entries $\tilde{\mathbf{K}}_{ij} = \tilde{k}(x_i, x_j)$ and $\tilde{\mathbf{K}}_{q,ij} = \tilde{k}_q(x_i, x_j)$ such that

$$\tilde{\mathbf{K}} = \frac{1}{s}\,\mathbf{Z}\mathbf{Z}^T \qquad \wedge \qquad \tilde{\mathbf{K}}_q = \frac{1}{s}\,\mathbf{Z}_q\mathbf{Z}_q^T \ .$$

If we now denote the j-th column of $\mathbf{Z}$ by $\mathbf{z}_{v_j}(\mathbf{x})$ and the j-th column of $\mathbf{Z}_q$ by $\mathbf{z}_{q,v_j}(\mathbf{x})$, then the following equalities can be derived easily from Eq. (4):

$$\mathbb{E}_{v\sim p}[\tilde{\mathbf{K}}] = \mathbf{K} = \mathbb{E}_{v\sim q}[\tilde{\mathbf{K}}_q] \quad \wedge \quad \mathbb{E}_{v\sim p}\left[\mathbf{z}_v(\mathbf{x})\mathbf{z}_v(\mathbf{x})^T\right] = \mathbf{K} = \mathbb{E}_{v\sim q}\left[\mathbf{z}_{q,v}(\mathbf{x})\mathbf{z}_{q,v}(\mathbf{x})^T\right] \ .$$

Sampling features from the importance weighted probability density function $q(\cdot)$ has led to much interest in the literature (Bach, 2017b; Alaoui and Mahoney, 2015; Avron et al., 2017; Rudi and Rosasco, 2017) as it can lead to huge computational savings. The reason for this is that when sampling according to $p(v)$, the focus is typically on approximating the leading/top eigenvalues of the corresponding kernel matrix $\mathbf{K}$. In contrast, sampling according to a re-weighted distribution

[2] It is not necessarily true that for any $g \in L_2(d\tau)$, there exists a corresponding $f \in \mathcal{H}$.

is likely to yield the Fourier features that span the whole eigenspectrum of $\mathbf{K}$. To that end, an importance weighted density function based on the notion of *ridge leverage scores* is introduced in Alaoui and Mahoney (2015) for landmark selection in the Nyström method (Nyström, 1930; Smola and Schölkopf, 2000; Williams and Seeger, 2001). For landmarks selected using that sampling strategy, Alaoui and Mahoney (2015) established a sharp convergence rate of the low-rank estimator based on the Nyström method. This result has motivated the pursuit of a similar notion for random Fourier features. Indeed, Bach (2017b) proposed a leverage score function based on an integral operator defined using the kernel function and the marginal distribution of a data-generating process. Building on this work, Avron et al. (2017) proposed the ridge leverage score function with respect to a fixed input dataset, i.e.,

$$l_\lambda(v) = p(v)\mathbf{z}_v(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{z}_v(\mathbf{x}) \ . \tag{6}$$

From our assumption on the decomposition of a kernel function, it follows that there exists a constant z_0 such that $|z(v,x)| \leq z_0$ (for all v and x) and $\mathbf{z}_v(\mathbf{x})^T\mathbf{z}_v(\mathbf{x}) \leq nz_0^2$. We can now deduce the following inequality using a result from Avron et al. (2017, Proposition 4):

$$l_\lambda(v) \leq p(v)\frac{z_0^2}{\lambda} \ .$$

An important property of function $l_\lambda(v)$ is its relation to the effective number of parameters:

$$\int_{\mathcal{V}} l_\lambda(v)dv = \mathrm{Tr}\big[\mathbf{K}(\mathbf{K} + n\lambda\mathbf{I})^{-1}\big] := d_{\mathbf{K}}^\lambda \ ,$$

where $d_{\mathbf{K}}^\lambda$ is known for implicitly determining the number of independent parameters in a learning problem and, thus, it is called the *effective dimension of the problem* (Caponnetto and De Vito, 2007) or the *number of effective degrees of freedom* (Bach, 2013; Hastie, 2017).

We can now observe that $q^*(v) = l_\lambda(v)/d_{\mathbf{K}}^\lambda$ is a probability density function. In Avron et al. (2017), it has been established that sampling according to $q^*(v)$ requires fewer Fourier features in the fixed design setting compared to the standard spectral measure sampling. We refer to $q^*(v)$ as the *empirical ridge leverage score distribution* and, in the remainder of the manuscript, refer to this sampling strategy as *leverage weighted RFF*.

2.3 Rademacher Complexity

To characterize the performance of a learning algorithm, we need to take into account the complexity of its hypothesis space. Below, we first introduce a particular measure of the complexity over function spaces known as *Rademacher complexity* (Bartlett and Mendelson, 2002). Then, we give two lemmas that demonstrate how Rademacher complexity of a reproducing kernel Hilbert space can be linked to the corresponding kernel and how the excess risk can be computed via Rademacher complexity.

Definition 1 *Suppose that $\{x_1 \cdots, x_n\}$ are independent samples selected according to P_x. Let $\mathcal{H}$ be a class of functions mapping $\mathcal{X}$ to $\mathbb{R}$. Then, the random variable known as the empirical Rademacher complexity is defined as*

$$\hat{R}_n(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\left|\frac{2}{n}\sum_{i=1}^n \sigma_i f(x_i)\right| \mid x_1, \cdots, x_n\right] \ ,$$

where $\sigma_1, \cdots, \sigma_n$ are independent samples from the uniform distribution over the two element set $\{\pm 1\}$. The corresponding Rademacher complexity is then defined as the expectation of the empirical Rademacher complexity

$$R_n(\mathcal{H}) = \mathbb{E}\left[\hat{R}_n(\mathcal{H})\right] ,$$

where the expectation is taken with respect to n-element sets of indepedent samples from P_x.

The following lemma provides an upper bound on the Rademacher complexity of a hypothesis space that is a subspace of the reproducing kernel Hilbert space with a kernel k.

Lemma 2 *(Bartlett and Mendelson, 2002) Let $\mathcal{H}_0$ be the unit ball that is centered at the origin of the reproducing kernel Hilbert space $\mathcal{H}$ associated with a kernel k. Then, we have that $R_n(\mathcal{H}_0) \leq (1/n)\mathbb{E}_X[\sqrt{Tr(\mathbf{K})}]$, where $\mathbf{K}$ is the Gram matrix for kernel k over an independent and identically distributed sample $X = \{x_1, \cdots, x_n\}$.*

Lemma 3 states that the expected excess risk convergence rate of a particular estimator in $\mathcal{H}$ not only depends on the number of data points, but also on the complexity of $\mathcal{H}$ and how it interacts with the loss function.

Lemma 3 *(Bartlett and Mendelson, 2002, Theorem 8) Let $\{x_i, y_i\}_{i=1}^n$ be i.i.d samples from P and let $\mathcal{H}$ be the space of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Denote a loss function with $l : \mathcal{Y} \times \mathbb{R} \to [0, 1]$ and recall that, for all $f \in \mathcal{H}$, the expected and corresponding empirical learning risk functions are denoted with $\mathbb{E}[l_f]$ and $\mathbb{E}_n[l_f] = (1/n)\sum_{i=1}^n l(y_i, f(x_i))$, respectively. Then, for a sample of size n, for all $f \in \mathcal{H}$ and $\delta \in (0, 1)$, with probability $1 - \delta$, we have that*

$$\mathbb{E}[l_f] \leq \mathbb{E}_n[l_f] + R_n(l \circ \mathcal{H}) + \sqrt{\frac{8\log(2/\delta)}{n}} ,$$

where $l \circ \mathcal{H} = \{(x, y) \mapsto l(y, f(x)) - l(y, 0) \mid f \in \mathcal{H}\}$.

Note that the risk bound is given by the Rademacher complexity term $R_n(l \circ \mathcal{H})$ defined on the transformed space $l \circ \mathcal{H}$, which is obtained via composition of $f \in \mathcal{H}$ and the loss function l. This term is, in general, different from $R_n(\mathcal{H})$. However, in the case when l is Lipschitz continuous with constant L_l, then $R_n(l \circ \mathcal{H}) \leq 2L_l R_n(\mathcal{H})$ by Theorem 12 in Bartlett and Mendelson (2002).

2.4 Local Rademacher Complexity

When characterizing the finite sample behaviour of learning risk, the notion of Rademacher complexity introduced in the previous section does not typically give the optimal convergence rates. This is because Rademacher complexity considers the behaviour of the empirical learning risk over the whole hypothesis space, while the estimator returned by the regression is typically in a neighbourhood around the optimal estimator. Hence, in our refined analysis we rely on the so called *local Rademacher complexity*. Before illustrating this concept, we first recall that given a hypothesis $f \in \mathcal{H}$, we denote its expectation and finite sample average with $\mathbb{E}[f]$ and $\mathbb{E}_n[f]$, respectively. The notion of local Rademacher complexity is typically introduced via the so called *sub-root function*. This sub-root function is used to obtain a fixed point of the local Rademacher complexity, which gives a sharper convergence rate than the notion introduced in previous section. Below, we first give the definition and a useful property of the sub-root function. We then review a theorem that relates the notion of local Rademacher complexity and learning risk.

Definition 4 *Let $\psi : [0, \infty) \to [0, \infty)$ be a function. Then, $\psi(r)$ is called a sub-root function if, for all $r > 0$, $\psi(r)$ is non-decreasing and $\psi(r)/r$ is non-increasing.*

A sub-root function has the following property.

Lemma 5 *(Bartlett et al., 2005, Lemma 3.2) If $\psi(r)$ is a sub-root function, then $\psi(r) = r$ has a unique positive solution r^*. In addition, we have that $r \geq \psi(r)$ if and only if $r \geq r^*$.*

In Lemma 3, we can see that the difference between the expected and empirical learning risks, $\mathbb{E}[l_f]$ and $\mathbb{E}_n[l_f]$, is upper bounded by $\mathcal{O}(1/\sqrt{n})$. This rate can be further improved with local Rademacher complexity. The reason for the slow learning rate is because the bound accounts for the difference between $\mathbb{E}[l_f]$ and $\mathbb{E}_n[l_f]$ using the global Rademacher complexity. Inspecting the definition of $R_n(\mathcal{H})$ (Definition 1), we can see that $R_n(\mathcal{H})$ is defined by considering the whole hypothesis space, as the supremum operator is applied over all functions in $\mathcal{H}$. However, as discussed before, learning algorithms typically return functions that are in the neighbourhood around the optimal estimator. Hence, using $R_n(\mathcal{H})$ unnecessarily enlarges the space that we are interested in.

As empirical estimators returned by learning algorithms typically have low learning risk as well as low variance, we could instead consider the alternative space $\mathcal{H}_r := \{f \in \mathcal{H} : \mathbb{E}[f^2] \leq r\}$ for some given value $r \in \mathbb{R}$. In this way, we greatly reduce the complexity of the function space at hand and can provide a sharper convergence rate. The following results from Bartlett et al. (2005) details how this idea can be used to describe the learning risk behaviour.

Lemma 6 *(Bartlett et al., 2005, Theorem 4.1) Let $\mathcal{H}$ be a class of functions with bounded ranges and assume that there is some constant $B > 0$ such that, for all $f \in \mathcal{H}$, $\mathbb{E}[f^2] \leq B\mathbb{E}[f]$. Let $\hat{\psi}_n$ be a sub-root function and let $\hat{r}^*$ be the fixed point of $\hat{\psi}_n$, i.e., $\hat{\psi}_n(\hat{r}^*) = \hat{r}^*$. Fix any $\delta \in (0, 1)$, and assume that for any $r \geq \hat{r}^*$,*

$$\hat{\psi}_n(r) \geq c_1 \hat{R}_n\{f \in \operatorname{star}(\mathcal{H}, 0) \mid \mathbb{E}_n[f^2] \leq r\} + \frac{c_2}{n} \log \frac{1}{\delta} ,$$

where

$$\operatorname{star}(\mathcal{H}, f_0) = \{f_0 + \alpha(f - f_0) \mid f \in \mathcal{H} \wedge \alpha \in [0, 1]\} .$$

Then for all $D > 1$ and $f \in \mathcal{H}$, with probability greater than $1 - \delta$,

$$\mathbb{E}[f] \leq \frac{D}{D-1}\mathbb{E}_n[f] + \frac{6D}{B}\hat{r}^* + \frac{c_3}{n} \log \frac{1}{\delta} ,$$

where c_1, c_2 and c_3 are some constants.

Note that this theorem bounds the difference between $\mathbb{E}[f]$ and $\mathbb{E}_n[f]$. We will show later (Section 6.3), with a simple transformation, that this result can be used to bound the difference between the learning and empirical risks for estimators based on random Fourier features.

We have seen that in the above theorem, we can use the fixed point of the sub-root function to upper bound the learning rate. It is, however, not clear how to obtain the explicit formula for the fixed point using this result. Fortunately, in the setting of learning with kernel k and the corresponding reproducing kernel Hilbert space, we can derive such results. The following lemma provides us with an upper bound on local Rademacher complexity through the eigenvalues of the Gram matrix.

Lemma 7 *(Bartlett et al., 2005, Lemma 6.6) Let k be a positive definite kernel function with reproducing kernel Hilbert space $\mathcal{H}$ and let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$ be the eigenvalues of the normalized Gram-matrix $(1/n)\mathbf{K}$. Then, for all $r > 0$ and $f \in \mathcal{H}$,*

$$\hat{R}_n\{f \in \mathcal{H} \mid \mathbb{E}_n[f^2] \leq r\} \leq \left(\frac{2}{n}\sum_{i=1}^{n} \min\{r, \hat{\lambda}_i\}\right)^{1/2}.$$

3. Theoretical Analysis

In this section, we provide a unified analysis for the generalization properties of learning with random Fourier features. Our analysis is split into two cases/settings: *i*) we start with a bound for learning with the squared error loss function (Section 3.1) and *ii*) then extend these results to learning problems with Lipschitz continuous loss functions (Section 3.2). In addition, in each of the cases, we will present two different analyses. In the worst case analysis, we provide the conditions for the estimator to achieve the minimax learning rate of the corresponding kernel-based estimator. After that, we present a refined analysis and show that the estimators based on random Fourier features are able to achieve faster learning rates if the learning problem exhibits certain benign properties. Before proceeding with our theoretical contributions, we first enumerate the assumptions that will be used throughout our analysis:

1. For a learning problem with kernel k (and corresponding reproducing kernel Hilbert space $\mathcal{H}$) defined as in Eq. (1), we assume that $f_{\mathcal{H}} = \arg\inf_{f \in \mathcal{H}} \mathbb{E}[l_f]$ always exists and has a bounded $\mathcal{H}$-norm. Moreover, we (without further loss of generality) restrict our analysis to the unit ball of $\mathcal{H}$, i.e., the hypothesis space is given by $\|f\|_{\mathcal{H}} \leq 1$;

2. We assume that the kernel k has the decomposition as in Eq. (3) with $|z(w, x)| < z_0 \in (0, \infty)$;

3. For kernel k, denote with $\lambda_1 \geq \cdots \geq \lambda_n$ the eigenvalues of the kernel matrix $\mathbf{K}$. We assume that the regularization parameter satisfies $0 \leq n\lambda \leq \lambda_1$.

Intuitively, Assumption 3 requires that the signal λ_1 is stronger than the added regularization term $n\lambda$. More specifically, the in-sample prediction of a kernel ridge regression problem is $\mathbf{K}(\mathbf{K} + n\lambda I)^{-1}Y$. The largest eigenvalue of $\mathbf{K}(\mathbf{K} + n\lambda I)^{-1}$ is $\lambda_1/(\lambda_1+n\lambda)$. If $n\lambda > \lambda_1$, then the in-sample prediction is essentially dominated by $n\lambda$, which could lead to under-fitting.

Throughout our analysis, we will use the assumptions listed above and will not be restating them unless problem-specific clarifications are required.

3.1 Learning with the Squared Error Loss

In this section, we consider learning with the squared error loss, i.e., $l(y, f(x)) = (y - f(x))^2$. For this particular loss function, the optimization problem from Eq. (1) is known as *kernel ridge regression* (KRR). We make the following assumption specific for the KRR problem.

A.1 $y = f^*(x) + \epsilon$ with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}[\epsilon] = \sigma^2$. Furthermore, we assume that y is a bounded random variable, i.e., $|y| \leq y_0$;

For regression problem, we have the target regression function $f^* = \mathbb{E}[y \mid x]$. Note that f^* may be different from $f_{\mathcal{H}}$ as it is not necessarily contained in our hypothesis space $\mathcal{H}$.

In the random Fourier feature setting, the KRR problem can be reduced to solving a linear system $(\tilde{\mathbf{K}}+n\lambda\mathbf{I})\alpha = Y$, with $Y = [y_1, \cdots, y_n]^T$. Typically, an approximation of the kernel function based on random Fourier features is employed in order to effectively reduce the computational cost and scale kernel ridge regression to problems with a large number of examples. More specifically, for a vector of observed labels Y the goal is to find a hypothesis $\tilde{\mathbf{f}}_x = \mathbf{Z}\beta$ that minimizes $\|Y - \tilde{\mathbf{f}}_x\|_2^2$ while having good generalization properties. In order to achieve this, one needs to control the complexity of hypotheses defined by random Fourier features and avoid over-fitting. Hence, we would like to estimate the norm of a function $\tilde{f} \in \tilde{\mathcal{H}}$ for the purpose of regularization. The following proposition (originally from Bach, 2017b) gives an upper bound on that norm (a proof is given in Section 6).

Proposition 8 *Assume that the reproducing kernel Hilbert space $\mathcal{H}$ with kernel k admits a decomposition as in Eq. (3) and denote by $\tilde{\mathcal{H}} := \{\tilde{f} \mid \tilde{f} = \sum_{i=1}^{s} \alpha_i z(v_i, \cdot), \alpha_i \in \mathbb{R}\}$ the reproducing kernel Hilbert space with kernel $\tilde{k}$ (see Eq. 4). Then, for all $\tilde{f} \in \tilde{\mathcal{H}}$ it holds that $\|\tilde{f}\|_{\tilde{\mathcal{H}}}^2 \leq s\|\alpha\|_2^2$.*

According to Proposition 8, the learning problem with random Fourier features and the squared error loss can be cast as

$$\beta_\lambda := \underset{\beta \in \mathbb{R}^s}{\arg\min} \; \frac{1}{n}\|Y - \mathbf{Z}_q\beta\|_2^2 + \lambda s\|\beta\|_2^2 \, . \tag{7}$$

This is simply a linear ridge regression problem in the space of Fourier features. We denote the optimal hypothesis function returned by Eq. (7) with $\tilde{f}_\beta^\lambda$. The function can be parameterized by β_λ and its in-sample evaluation is given by $\tilde{\mathbf{f}}_\beta^\lambda = \mathbf{Z}_q\beta_\lambda$, where $\beta_\lambda = (\mathbf{Z}_q^T\mathbf{Z}_q + ns\lambda\mathbf{I})^{-1}\mathbf{Z}_q^T Y$. Since $\mathbf{Z}_q \in \mathbb{R}^{n\times s}$, the computational and space complexities are $\mathcal{O}(s^3 + ns^2)$ and $\mathcal{O}(ns)$. Thus, significant savings can be achieved using estimators with $s \ll n$. To assess the effectiveness of such estimators, it is important to understand the relationship between the excess learning risk and the choice of s.

3.1.1 Worst Case Analysis

In this section, we provide a bound on the required number of random Fourier features with respect to the worst case (in the minimax sense) of the corresponding kernel ridge regression problem, i.e., learning rate $\mathcal{O}(1/\sqrt{n})$. The following theorem gives a general result while taking into account both the number of features s and a sampling strategy for selecting them.

Theorem 9 *Suppose that Assumption A.1 holds and let $\tilde{l} : \mathcal{V} \to \mathbb{R}$ be a measurable function such that $\tilde{l}(v) \geq l_\lambda(v)$ ($\forall v \in \mathcal{V}$) with $d_{\tilde{l}} = \int_{\mathcal{V}} \tilde{l}(v)dv < \infty$. Suppose also that $\{v_i\}_{i=1}^s$ are sampled independently from the probability density function $q(v) = \tilde{l}(v)/d_{\tilde{l}}$. If*

$$s \;\geq\; 5d_{\tilde{l}} \log \frac{16 d_{\mathbf{K}}^\lambda}{\delta} \; ,$$

then for all $\delta \in (0,1)$, with probability $1-\delta$, the excess risk of $\tilde{f}_\beta^\lambda$ can be upper bounded by

$$\mathbb{E}[l_{\tilde{f}_\beta^\lambda}] - \mathbb{E}[l_{f_\mathcal{H}}] \;\leq\; 4\lambda + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) + \mathbb{E}[l_{\hat{f}^\lambda}] - \mathbb{E}[l_{f_\mathcal{H}}] \, . \tag{8}$$

Theorem 9 expresses the trade-off between the computational and statistical efficiency through the regularization parameter λ, the effective dimension of the problem $d_{\mathbf{K}}^\lambda$, and the normalization constant $d_{\tilde{l}}$ of the sampling distribution. The decay rate of the regularization parameter is used as a key quantity (Caponnetto and De Vito, 2007; Rudi and Rosasco, 2017) and its choice can be linked to the complexity of the target regression function $f^*(x) = \int y d\rho(y \mid x)$. In particular, Caponnetto and De Vito (2007) have shown that the minimax risk convergence rate for kernel ridge regression is $\mathcal{O}(1/\sqrt{n})$. Setting $\lambda \propto 1/\sqrt{n}$, we observe that the estimator $\tilde{f}_{\beta}^{\lambda}$ attains the worst case minimax rate of kernel ridge regression.

As a consequence of Theorem 9, we have the following bounds on the number of required features for the two strategies: *leverage weighted* RFF (Corollary 1) and *plain* RFF (Corollary 2).

Corollary 10 *If the probability density function from Theorem 9 is the empirical ridge leverage score distribution $q^*(v)$, then the upper bound on the risk from Eq. (8) holds for all $s \geq 5d_{\mathbf{K}}^{\lambda} \log \frac{16d_{\mathbf{K}}^{\lambda}}{\delta}$.*

Proof For this corollary, we set $\tilde{l}(v) = l_{\lambda}(v)$ and deduce $d_{\tilde{l}} = \int_{\mathcal{V}} l_{\lambda}(v) dv = d_{\mathbf{K}}^{\lambda}$. ■

Theorem 9 and Corollary 10 have several implications on the choice of λ and s. First, we could pick $\lambda \in \mathcal{O}(n^{-1/2})$ that implies the worst case minimax rate for kernel ridge regression (Caponnetto and De Vito, 2007; Rudi and Rosasco, 2017; Bartlett et al., 2005) and observe that in this case s is proportional to $d_{\mathbf{K}}^{\lambda} \log d_{\mathbf{K}}^{\lambda}$. As $d_{\mathbf{K}}^{\lambda}$ is determined by the learning problem (i.e., the marginal distribution P_x), we can consider several different cases. In the best case, where the number of positive eigenvalues is finite, implying that $d_{\mathbf{K}}^{\lambda}$ does not grow with n, we then have that even with a constant number of features, we are able to achieve the $\mathcal{O}(1/\sqrt{n})$ learning rate. Next, if the eigenvalues of $\mathbf{K}$ exhibit a geometric/exponential decay, i.e., $\lambda_i \propto R_0 r^i$ with a constant $R_0 > 0$ (this can happen in scenario where we have a Gaussian kernel and a sub-Gaussian marginal distribution P_x), we then know that $d_{\mathbf{K}}^{\lambda} \leq \log(R_0/\lambda)$ (Bach, 2017b), implying $s \geq \log n \log \log n$. Hence, significant savings can be obtained with $\mathcal{O}(n \log^4 n + \log^6 n)$ computational and $\mathcal{O}(n \log^2 n)$ storage complexities of linear ridge regression over random Fourier features, as opposed to $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$ costs (respectively) in the kernel ridge regression setting.

In the case of a slower decay (e.g., $\mathcal{H}$ is a Sobolev space of order $t \geq 1$) with $\lambda_i \propto R_0 i^{-2t}$, we have $d_{\mathbf{K}}^{\lambda} \leq (R_0/\lambda)^{1/(2t)}$ and $s \geq n^{1/(4t)} \log n$. Hence, substantial computational savings can be achieved even in this case. Furthermore, in the worst case with λ_i close to $R_0 i^{-1}$, our bound implies that $s \geq \sqrt{n} \log n$ features are sufficient, recovering the result from Rudi and Rosasco (2017).

Corollary 11 *If the probability density function from Theorem 9 is the spectral measure $p(v)$ from Eq. (3), then the upper bound on the learning risk from Eq. (8) holds for all $s \geq 5\frac{z_0^2}{\lambda} \log \frac{16d_{\mathbf{K}}^{\lambda}}{\delta}$.*

Proof We set $\tilde{l}(v) = p(v)\frac{z_0^2}{\lambda}$ and obtain $d_{\tilde{l}} = \int_{\mathcal{V}} p(v) \frac{z_0^2}{\lambda} dv = \frac{z_0^2}{\lambda}$. ■

Corollary 11 addresses plain random Fourier features and states that if s is chosen to be greater than $\sqrt{n} \log d_{\mathbf{K}}^{\lambda}$ and $\lambda \propto 1/\sqrt{n}$ then the minimax risk convergence rate is guaranteed. In the case of finitely many positive eigenvalues, $s \geq \sqrt{n}$ features are needed to obtain $\mathcal{O}(1/\sqrt{n})$ convergence rate. When the eigenvalues have an exponential decay, we obtain the same convergence rate with only $s \geq \sqrt{n} \log \log n$ features, which is an improvement compared to a result by Rudi and Rosasco (2017) where $s \geq \sqrt{n} \log n$ is needed. For the other two cases, we derive $s \geq \sqrt{n} \log n$ and recover the results from Rudi and Rosasco (2017). Table 1 provides a summary of the trade-offs between computational complexity and statistical efficiency for the worst case scenario.

SAMPLING SCHEME	SPECTRUM	NUMBER OF FEATURES	LEARNING RATE
WEIGHTED RFF	finite rank	$s \in \Omega(1)$	$\mathcal{O}(1/\sqrt{n})$
	$\lambda_i \propto A^i$	$s \in \Omega(\log n \cdot \log\log n)$	
	$\lambda_i \propto i^{-2t}$ $(t \geq 1)$	$s \in \Omega(n^{1/2t} \cdot \log n)$	
	$\lambda_i \propto i^{-1}$	$s \in \Omega(\sqrt{n} \cdot \log n)$	
PLAIN RFF	finite rank	$s \in \Omega(\sqrt{n})$	$\mathcal{O}(1/\sqrt{n})$
	$\lambda_i \propto A^i$	$s \in \Omega(\sqrt{n} \cdot \log\log n)$	
	$\lambda_i \propto i^{-2t}$ $(t \geq 1)$	$s \in \Omega(\sqrt{n} \cdot \log n)$	
	$\lambda_i \propto i^{-1}$	$s \in \Omega(\sqrt{n} \cdot \log n)$	

Table 1: The worst case trade-offs between computational complexity and statistical efficiency for the squared error loss.

3.1.2 Refined Analysis

In this section, we provide a more refined analysis with risk convergence rates faster than $\mathcal{O}(1/\sqrt{n})$, depending on the spectrum decay of the kernel and/or the complexity of the target regression function. The main reason for obtaining a faster rate compared to the previous section is the reliance on local Rademacher complexity, instead of the global one (detailed proofs can be found in Section 6.3).

Theorem 12 *Suppose that Assumption A.1 holds and that the conditions on sampling measure $\tilde{l}$ from Theorem 9 apply to this setting. If*

$$s \;\geq\; 5 d_{\tilde{l}} \log \frac{16 d_{\mathbf{K}}^{\lambda}}{\delta}$$

then for all $D > 1$ and $\delta \in (0,1)$, with probability $1-\delta$, the excess risk of $\tilde{f}_{\beta}^{\lambda}$ can be bounded by

$$\mathbb{E}[l_{\tilde{f}_{\beta}^{\lambda}}] - \mathbb{E}[l_{f_{\mathcal{H}}}] \;\leq\; \frac{12D}{B} \hat{r}_{\mathcal{H}}^{*} + 4\frac{D}{D-1}\lambda + \mathcal{O}\left(\frac{1}{n}\right) + \mathbb{E}[l_{\hat{f}^{\lambda}}] - \mathbb{E}[l_{f_{\mathcal{H}}}] \,. \tag{9}$$

Furthermore, denoting the eigenvalues of the normalized kernel matrix $(1/n)\mathbf{K}$ with $\{\hat{\lambda}_i\}_{i=1}^{n}$, we have that

$$\hat{r}_{\mathcal{H}}^{*} \leq \min_{0 \leq h \leq n} \left(e_0 \frac{h}{n} + \sqrt{\frac{1}{n} \sum_{i > h} \hat{\lambda}_i} \right) \,, \tag{10}$$

where $B, e_0 > 0$ are some constant and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$.

Theorem 12 covers a wide range of cases and can provide tight/sharp risk convergence rates. In particular, note that $\hat{r}_{\mathcal{H}}^{*}$ has an upper bound of $\mathcal{O}(1/\sqrt{n})$ in all cases, which happens when we let $h = 0$ and the spectrum decays polynomially as $\mathcal{O}(1/n^t)$ with $t > 1$. On the other hand, if the eigenvalues decay exponentially, then setting $h = \lceil \log n \rceil$ implies that $\hat{r}_{\mathcal{H}}^{*} \leq \mathcal{O}(\log n/n)$. In the best case, when the kernel function has only finitely many positive eigenvalues, we have that

SAMPLING SCHEME	SPECTRUM	NUMBER OF FEATURES	LEARNING RATE
WEIGHTED RFF	finite rank	$s \in \Omega(1)$	$\mathcal{O}(1/n)$
	$\lambda_i \propto A^i$	$s \in \Omega(\log n \cdot \log\log n)$	$\mathcal{O}(\log n/n)$
	$\lambda_i \propto i^{-t}$ $(t > 1)$	$s \in \Omega(n^{1/2t} \cdot \log n)$	$\mathcal{O}(1/\sqrt{n})$
PLAIN RFF	finite rank	$s \in \Omega(n)$	$\mathcal{O}(1/n)$
	$\lambda_i \propto A^i$	$s \in \Omega(n)$	$\mathcal{O}(\log n/n)$
	$\lambda_i \propto i^{-t}$ $(t > 1)$	$s \in \Omega(\sqrt{n} \cdot \log n)$	$\mathcal{O}(1/\sqrt{n})$

Table 2: The refined case trade-offs between computational complexity and statistical efficiency for the squared error loss.

$\hat{r}^*_{\mathcal{H}} \leq \mathcal{O}(1/n)$ by letting h be any fixed value larger than the number of positive eigenvalues. These different upper bounds provide insights into various trade-offs between computational complexity and statistical efficiency. We now split the discussion into two cases: weighted sampling with empirical leverage score and plain sampling.

Under the weighted sampling scheme, if the eigenvalues decay polynomially, i.e., $\lambda_i \propto i^{-t}$ with $t > 1$, then the learning rate is upper bounded by $\mathcal{O}(1/\sqrt{n})$. In this case, we have $d^{\lambda}_{\mathbf{K}} \leq (R_0/\lambda)^{1/t} \leq n^{1/2t}$ and consequently $s \geq n^{1/2t} \log n$. On the other hand, if the eigenvalues decay exponentially, we have $d^{\lambda}_{\mathbf{K}} \leq \log(R_0/\lambda)^{1/t} \leq \log n$. Hence, if $s \geq \log n \log\log n$ we achieve $\mathcal{O}(\log n/n)$ learning rate. In the best case, where we have finitely many positive eigenvalues, then with a constant number of features we achieve $\mathcal{O}(1/n)$ learning rate.

As for the plain sampling strategy, the learning rates and required numbers of features for the three above cases are: *i*) $\mathcal{O}(1/\sqrt{n})$ and $s \geq \sqrt{n} \log n$ (polynomial decay), *ii*) $\mathcal{O}(\log n/n)$ and $s \geq n$ (exponential decay), and *iii*) $\mathcal{O}(1/n)$ and $s \geq n$ (finite many positive eigenvalues). Table 2 summarizes our results for the refined case.

Remark 13 *In Caponnetto and De Vito (2007), the convergence rate of the excess risk has been linked to two constants* (b, c)*, where* $b \in (1, \infty)$ *represents the eigenvalue decay and* $c \in [1, 2]$ *measures the complexity of the target function* $f_{\mathcal{H}}$*. Essentially,* c *determines how fast the coefficients* α_i *of* $f_{\mathcal{H}}$ *decay, where* α_i *represents the coefficient of the expansion of* $f_{\mathcal{H}}$ *along the eigenfunctions of the integral operator defined by the kernel* k *and the data generating distribution* $P(x, y)$*. While* $c = 1$ *is equivalent to assuming* $f_{\mathcal{H}}$ *exists, in literature, it is typical that we have to assume benign cases (i.e.,* $c > 1$*) to obtain fast learning rates.*

Our analysis is different from that in Caponnetto and De Vito (2007) in the sense that we only consider the worst case $c = 1$*. Under this assumption, we compute excess learning risk of the random Fourier features estimator under various eigenvalue decays (the values of constant* b*). Our results demonstrate that even if we only consider case* $c = 1$*, we are still able to obtain the rate* $\mathcal{O}(1/\sqrt{n})$ *in Theorem 1. This is aligned with the worst case rate in Caponnetto and De Vito (2007). In the refined analysis, the local Rademacher complexity technique allows us to obtain sharper/better convergence rates without further assumptions on the constant such as* $c > 1$*, i.e., resulting in an improvement from* $\mathcal{O}(1/\sqrt{n})$ *to* $\mathcal{O}(1/n)$ *learning rate. Moreover, our fast rate range matches that in Caponnetto and De Vito (2007).*

RESULTS	SPECTRUM	NUMBER OF FEATURES	LEARNING RATE
THIS WORK	finite rank	$s \in \Omega(\sqrt{n})$	$\mathcal{O}(1/\sqrt{n})$
	$\lambda_i \propto A^i$	$s \in \Omega(\sqrt{n} \cdot \log\log n)$	
	$\lambda_i \propto i^{-2t}$ $(t \geq 1)$	$s \in \Omega(\sqrt{n} \cdot \log n)$	
	$\lambda_i \propto i^{-1}$		
RUDI & ROSASCO (2017)	finite rank	$s \in \Omega(\sqrt{n} \cdot \log n)$	$\mathcal{O}(1/\sqrt{n})$
	$\lambda_i \propto A^i$		
	$\lambda_i \propto i^{-2t}$ $(t \geq 1)$		
	$\lambda_i \propto i^{-1}$		

Table 3: The comparison of our results to the sharpest learning rates from prior work (Rudi and Rosasco, 2017) in the worst case setting for plain random Fourier features and the squared error loss.

3.1.3 COMPARISON WITH THE SHARPEST BOUNDS FROM PRIOR WORK

The trade-offs between the computational cost and prediction accuracy for random Fourier features in the squared error loss setting have been thoroughly studied in the literature (see, e.g., Sutherland and Schneider, 2015; Sriperumbudur and Szabó, 2015; Avron et al., 2017; Rudi and Rosasco, 2017). In the remainder of this section, we discuss our results relative to Rudi and Rosasco (2017), which provide the sharpest learning rates so far and demonstrate that it is possible to achieve computational savings without sacrificing the prediction accuracy when learning with random Fourier features.

Worst Case Analysis. We start with the worst case setting where the learning rate is *minimax optimal* $\mathcal{O}(n^{-1/2})$. Table 3 provides a summary of our bounds for the plain random Fourier features sampling scheme relative to the results from Rudi and Rosasco (2017). Note that we omit the weighted sampling because Rudi and Rosasco (2017) do not consider the worst case setting for weighted random Fourier features. When the eigenspectrum has a finite rank or geometric decay, we can see that our results achieve sharper bound on the number of features, $s \in \Omega(\sqrt{n})$ versus $s \in \Omega(\sqrt{n} \cdot \log n)$ and $s \in \Omega(\sqrt{n} \cdot \log\log n)$ versus $s \in \Omega(\sqrt{n} \cdot \log n)$, respectively. When the eigenspectrum has a polynomial decay, we can see that the number of required features is the same.

Refined Analysis. We discuss here our results for the refined analysis, where learning algorithms with random Fourier features can achieve rates faster than $\mathcal{O}(1/\sqrt{n})$. We split the comparison into two cases: plain and weighted random Fourier features. Note that to obtain sharp learning rates, Rudi and Rosasco (2017) adopt the *source condition* assumption discussed in Remark 13. In particular, they assume that $c \in [1, 2]$. The assumption is convenient in that it allows one to derive a fast learning rate while at the same time providing a more flexible trade-off between the computational cost and prediction accuracy (please refer to Theorem 2 in Rudi and Rosasco, 2017, for more details). To facilitate the comparison between our results and Rudi and Rosasco (2017), we set $c = 1$ in both cases (plain and weighted random features). Table 4 provides a detailed comparison between the two works. One can see that when the eigenspectrum has a finite rank or exponential decay our results are the same as those in Rudi and Rosasco (2017). On the other hand, when eigenvalues have a fast polynomial decay i^{-t} with $t > 1$, it can be seen that the learning rates are different, i.e., $\mathcal{O}(n^{-1/2})$

RESULTS	SPECTRUM	NUMBER OF FEATURES	LEARNING RATE
THIS WORK	finite rank	$s \in \Omega(n)$	$\mathcal{O}(1/n)$
	$\lambda_i \propto A^i$	$s \in \Omega(n)$	$\mathcal{O}(\log n/n)$
	$\lambda_i \propto i^{-t}$ $(t > 1)$	$s \in \Omega(n^{1/2t} \cdot \log n)$	$\mathcal{O}(1/\sqrt{n})$
RUDI & ROSASCO (2017)	finite rank	$s \in \Omega(n)$	$\mathcal{O}(1/n)$
	$\lambda_i \propto A^i$	$s \in \Omega(n)$	$\mathcal{O}(\log n/n)$
	$\lambda_i \propto i^{-t}$ $(t > 1)$	$s \in \Omega(n^{\frac{2t}{1+2t}} \cdot \log n)$	$\mathcal{O}(n^{-\frac{2t}{2t+1}})$

Table 4: The comparison of our results to the sharpest learning rates from prior work (Rudi and Rosasco, 2017) in the refined case for plain random Fourier features and the squared error loss.

RESULTS	SPECTRUM	NUMBER OF FEATURES	LEARNING RATE
THIS WORK	finite rank	$s \in \Omega(1)$	$\mathcal{O}(1/n)$
	$\lambda_i \propto A^i$	$s \in \Omega(\log n \cdot \log\log n)$	$\mathcal{O}(\log n/n)$
	$\lambda_i \propto i^{-t}$ $(t > 1)$	$s \in \Omega(n^{1/2t} \cdot \log n)$	$\mathcal{O}(1/\sqrt{n})$
RUDI & ROSASCO (2017)	finite rank	$s \in \Omega(1)$	$\mathcal{O}(1/n)$
	$\lambda_i \propto A^i$	$s \in \Omega((\frac{n}{\log n})^{\alpha} \cdot \log n), \alpha \in (0,1)$	$\mathcal{O}(\log n/n)$
	$\lambda_i \propto i^{-t}$ $(t > 1)$	$s \in \Omega(n^{\frac{\alpha}{2t+1}} \cdot \log n), \alpha \in (0,1)$	$\mathcal{O}(n^{-\frac{2t}{2t+1}})$

Table 5: The comparison of our results to the sharpest learning rates from prior work (Rudi and Rosasco, 2017) in the refined case for weighted random Fourier features and the squared error loss.

versus $\mathcal{O}(n^{-2t/(2t+1)})$ obtained by Rudi and Rosasco (2017). However, the latter comes at the cost of requiring more features, i.e., $s \in \Omega(n^{2t/(2t+1)} \log n)$, whereas we only require $s \in \Omega(n^{1/2t} \log n)$.

We conclude this discussion with a comparison for the weighted sampling scheme. To obtain fast learning rates in this setting, Rudi and Rosasco (2017) introduce a further *compatibility condition* which relates random features to the data generating distribution $P(x, y)$ through a parameter α. Table 5 below illustrates the trade-offs between the computational cost and the statistical learning accuracy in this case. We can see that when the eigenspectrum has a finite rank, our results are the same as those in Rudi and Rosasco (2017). However, when the eigenspectrum displays a slower decay the results are quite different. In particular, the computational cost and prediction accuracy trade-offs from Rudi and Rosasco (2017) depend heavily on the compatibility parameter α. Specifically, when the eigenspectrum exhibits an exponential decay, our results show that $s \in \Omega(\log n \log\log n)$ features is guaranteed to achieve the fast learning rate $\mathcal{O}(\log n/n)$. In contrast to this, Rudi and Rosasco (2017) require $s \in \Omega((n/\log n)^{\alpha} \cdot \log n)$ features to achieve the same learning rate. When α is close to zero, the number of features required can be $\Omega(\log n)$, which is better than our results. However, when α is close to one, then the required number of features is $s \in \Omega(n)$, which is significantly worse than our results. Additionally, when the eigenspectrum has a polynomial decay i^{-t} with $t > 1$, our results show that $s \in \Omega(n^{1/2t})$ features can guarantee the minimax optimal learning rate $\mathcal{O}(1/\sqrt{n})$. For the same eigenspectrum decay, Rudi and Rosasco (2017) provide a more flexible trade-off that depends on α. For example, when $\alpha = 0$, then constant number of features is sufficient to obtain a fast learning rate $\mathcal{O}(n^{-2t/(1+2t)})$. However, when $\alpha = 1$ then learning algorithms require $s \in \Omega(n^{1/(2t+1)})$ features to guarantee the same learning rate.

3.2 Learning with a Lipschitz Continuous Loss

We next consider kernel methods with Lipschitz continuous loss functions, examples of which include kernel support vector machines and kernel logistic regression. Similar to the squared error loss case, we approximate y_i with $g_\beta(x_i) = \mathbf{z}_{q,x_i}(\mathbf{v})^T\beta$ and formulate the following learning problem

$$\beta^\lambda := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^s} \; \frac{1}{n}\sum_{i=1}^{n} l(y_i, \mathbf{z}_{q,x_i}(\mathbf{v})^T\beta) + \lambda s\|\beta\|_2^2 \;\; .$$

We let g_β^λ to be the prediction function defined based on β^λ and state an additional assumption that is specific to the Lipschitz continuous loss:

B.1 We assume that l is Lipschitz continuous with constant L:

$$(\forall g, g' \in \mathcal{H})(\forall x \in \mathcal{X})\colon |l_g - l_{g'}| \leq L|g(x) - g'(x)| \; .$$

Remark 14 *While the focus of our analysis in this section is on classification problems with a Lipschitz continuous loss function (e.g., hinge or logistic loss), the upper bounds on the excess risk and the resulting learning rates apply to the setting with* 0-1 *loss. The latter holds because the excess risk under* 0-1 *loss can be upper bounded by the excess risk under the hinge loss (e.g., see Sun et al., 2018; Steinwart and Christmann, 2008, for more details).*

3.2.1 WORST CASE ANALYSIS

The following theorem describes the trade-off between the selected number of features s and the learning risk of the estimator, providing an insight into the choice of s for Lipschitz continuous loss.

Theorem 15 *Suppose that Assumption B.1 holds and that the conditions on sampling measure $\tilde{l}$ from Theorem 9 apply to the setting with a Lipschitz continuous loss. If*

$$s \;\geq\; 5d_{\tilde{l}} \log \frac{16 d_{\mathbf{K}}^\lambda}{\delta}$$

then for all $\delta \in (0,1)$, with probability $1-\delta$, the learning risk of g_β^λ can be upper bounded by

$$\mathbb{E}[l_{g_\beta^\lambda}] \;\leq\; \mathbb{E}[l_{g_\mathcal{H}}] + \sqrt{2\lambda} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right). \tag{11}$$

Remark 16 *Just as Theorem 9 captures the trade-offs between computational complexity and statistical efficiency for the squared error loss, this theorem describes the relationship between s and $\mathbb{E}[l_{g_\beta^\lambda}]$ for the Lipschitz continuous setting. There is, however, an important difference between the two settings. More specifically, a property of the squared error loss function allows us to relate the random Fourier feature estimator $\tilde{f}_\beta^\lambda$ to the kernel ridge regression estimator $\hat{f}^\lambda$ by first computing the excess risk between $\tilde{f}_\beta^\beta$ and $\hat{f}^\lambda$ and then computing the same quantity for $\hat{f}^\lambda$ and $f_\mathcal{H}$. In contrast*

SAMPLING SCHEME	SPECTRUM	NUMBER OF FEATURES	LEARNING RATE
WEIGHTED RFF	finite rank	$s \in \Omega(1)$	$\mathcal{O}(1/\sqrt{n})$
	$\lambda_i \propto A^i$	$s \in \Omega(\log n \cdot \log\log n)$	
	$\lambda_i \propto i^{-2t}, t \geq 1$	$s \in \Omega(n^{1/2t} \cdot \log n)$	
	$\lambda_i \propto i^{-1}$	$s \in \Omega(n \cdot \log n)$	
PLAIN RFF	finite rank	$s \in \Omega(n)$	$\mathcal{O}(1/\sqrt{n})$
	$\lambda_i \propto A^i$	$s \in \Omega(n \cdot \log\log n)$	
	$\lambda_i \propto i^{-2t}, t \geq 1$	$s \in \Omega(n \cdot \log n)$	
	$\lambda_i \propto i^{-1}$		

Table 6: The worst case trade-offs between computational and statistical efficiency for Lipschitz continuous loss.

to this, the Lipschitz continuity property characteristic to the latter setting allows us to compute the excess learning risk directly by computing the risk difference between g_{β}^{λ} and $g_{\mathcal{H}}$. While the derivation in Lipschitz continuous loss is greatly simplified, the upper bound for the learning risk drops from $O(\lambda)$ in the regression case to $O(\sqrt{\lambda})$ in the classification case.

Corollaries 17 and 18 provide bounds for the cases of leverage weighted and plain RFF, respectively. The proofs are similar to the proofs of Corollaries 10 and 11.

Corollary 17 *If the probability density function from Theorem 15 is the empirical ridge leverage score distribution $q^*(v)$, then the upper bound on the risk from Eq. (11) holds for all $s \geq 5d_{\mathbf{K}}^{\lambda} \log \frac{16d_{\mathbf{K}}^{\lambda}}{\delta}$.*

Similar to Theorem 9, we consider four different cases for the effective dimension of the problem $d_{\mathbf{K}}^{\lambda}$. Corollary 17 states that the statistical efficiency is preserved if the leverage weighted RFF strategy is used with $s = \Omega(1)$, $s \geq \log n \log\log n$, $s \geq n^{1/(2t)} \log n$, and $s \geq n \log n$, respectively. Again, significant computational savings can be achieved if the kernel matrix $\mathbf{K}$ has a finite rank, as well as geometrically/exponentially or polynomially decaying eigenvalues.

Corollary 18 *If the probability density function from Theorem 15 is the spectral measure $p(v)$ from Eq. (3), then the upper bound on the risk from Eq. (11) holds for all $s \geq 5\frac{z_0^2}{\lambda} \log \frac{(16d_{\mathbf{K}}^{\lambda})}{\delta}$.*

Corollary 18 states that $n \log n$ features are required to attain $\mathcal{O}(n^{-1/2})$ convergence rate of the learning risk with plain RFF, recovering results from Rahimi and Recht (2009). Similar to the analysis in the squared error loss case, Theorem 15 together with Corollaries 17 and 18 allows theoretically motivated trade-offs between the statistical and computational efficiency of the estimator g_{β}^{λ}. Table 6 summarizes the worst case trade-offs between the computational and statistical efficiency.

3.2.2 Refined Analysis

In general, it is hard for classification problems to achieve learning rates sharper/faster than $\mathcal{O}\left(1/\sqrt{n}\right)$. However, as pointed out by Bartlett et al. (2006) and Steinwart and Christmann (2008), under some

SAMPLING SCHEME	SPECTRUM	NUMBER OF FEATURES	LEARNING RATE
WEIGHTED RFF	finite rank	$s \in \Omega(1)$	$\mathcal{O}(1/n)$
	$\lambda_i \propto A^i$	$s \in \Omega(\log n \cdot \log\log n)$	$\mathcal{O}(\log n/n)$
	$\lambda_i \propto i^{-t}, t > 1$	$s \in \Omega(n^{1/2t} \cdot \log n)$	$\mathcal{O}(1/\sqrt{n})$
PLAIN RFF	finite rank	$s \in \Omega(n^2)$	$\mathcal{O}(1/n)$
	$\lambda_i \propto A^i$	$s \in \Omega(n^2)$	$\mathcal{O}(\log n/n)$
	$\lambda_i \propto i^{-t}, t > 1$	$s \in \Omega(n \cdot \log n)$	$\mathcal{O}(1/\sqrt{n})$

Table 7: The refined case trade-offs between computational and statistical efficiency for Lipschitz continuous loss.

benign conditions, it is possible to obtain $\mathcal{O}(1/n)$ convergence rate. Hence, in this section, by adding an extra assumption, we derive a sharp learning rate for classification problems under random Fourier features setting. Similar to the squared error loss case, we rely on the notion of local Rademacher complexity to derive such a learning rate (details of the proof are presented in Section 6.5). Before we formally specify our result, we state the required additional assumption:

B.2 Recall that $g_{\mathcal{H}}$ is the estimator such that $g_{\mathcal{H}} = \arg\inf_{g\in\mathcal{H}} \mathbb{E}[l_g]$, where P is a probability distribution over $\mathcal{X} \times \mathcal{Y}$. We assume that there exists a constant B such that for all $g \in \mathcal{H}$

$$\mathbb{E}[(g - g_{\mathcal{H}})^2] \leq B\mathbb{E}[l_g - l_{g_{\mathcal{H}}}] .$$

Assumption B.2 is a condition for classification problems to obtain faster learning rates. It typically requires that the function space $\mathcal{H}$ is convex and uniformly bounded, as well as an additional uniform convexity condition on the loss function l. It can be shown that many loss functions satisfy this assumption, including squared loss (Bartlett et al., 2005) and hinge loss (Steinwart and Christmann, 2008, Chapter 8.5). Additional examples of such loss functions are discussed in Bartlett et al. (2006) and Mendelson (2002). As l is Lipschitz continuous, we have

$$\mathbb{E}[(l_g - l_{g_{\mathcal{H}}})^2] \leq L^2\mathbb{E}[(g - g_{\mathcal{H}})^2] \leq BL^2\mathbb{E}[l_g - l_{g_{\mathcal{H}}}] .$$

This is the variance condition described in Steinwart and Christmann (2008, Chapter 7.3), required to achieve faster convergence rates. The variance condition is also linked to the Massart's low noise condition or more generally to the Tsybakov condition (Sun et al., 2018), which intuitively speaking, requires that $P(Y = 1 \mid X = x)$ is not close to $1/2$. For more details, please refer to Tsybakov et al. (2004) and Koltchinskii (2011).

Theorem 19 *Suppose that Assumptions B.1-2 hold and that the conditions on sampling measure $\tilde{l}$ from Theorem 9 apply to the setting with a Lipschitz continuous loss. If*

$$s \;\geq\; 5d_{\tilde{l}} \log \frac{16d_{\mathbf{K}}^{\lambda}}{\delta}$$

then for all $D > 1$ and $\delta \in (0, 1)$ with probability greater than $1 - \delta$, we have

$$\mathbb{E}[l_{g_{\beta}^{\lambda}}] \leq \frac{12D}{B}\hat{r}_{\mathcal{H}}^{*} + \frac{D}{D-1}\sqrt{2\lambda} + \mathcal{O}(1/n) + \mathbb{E}[l_{g_{\mathcal{H}}}] . \quad (12)$$

Furthermore, denoting the eigenvalues of the normalized kernel matrix $(1/n)\mathbf{K}$ with $\{\hat{\lambda}_i\}_{i=1}^n$, we have that $\hat{r}_{\mathcal{H}}^*$ can be upper bounded by

$$\hat{r}_{\mathcal{H}}^* \leq \min_{0 \leq h \leq n} \left(b_0 \frac{h}{n} + \sqrt{\frac{1}{n} \sum_{i>h} \hat{\lambda}_i} \right) , \tag{13}$$

where B and b_0 are some constants.

Theorem 19 provides a sharper learning rate compared to Theorem 15. Similar to Theorem 12, $\hat{r}_{\mathcal{H}}^*$ can be upper bounded by $\mathcal{O}(1/n)$ (Gram-matrix is of finite rank), $\mathcal{O}(\log n/n)$ (eigenvalues decay exponentially), and $\mathcal{O}(1/\sqrt{n})$ (eigenvalues decay proportional to $1/n$). This has various implications on the trade-offs between computational cost and statistical efficiency. Just as in previous sections, we split the discussion into two parts according to the two sampling strategies.

We first discuss the scenario with empirical leverage score sampling. In a finite rank setting, if we choose $\lambda \in \mathcal{O}(1/n^2)$, we can see that the learning rate is of the order $\mathcal{O}(1/n)$. In addition, since we use the weighted sampling strategy and the Gram-matrix has finitely many eigenvalues, random Fourier features learning only requires a constant number of features to achieve $\mathcal{O}(1/n)$ learning rate. To the best of our knowledge, this is the first result that achieves this. In the case of exponential spectrum decay, the learning rate can be bounded with $\hat{r}_{\mathcal{H}}^* \leq {}^{\log n}\!/\!_n$ by setting $\lambda \in \mathcal{O}({}^{\log^2 n}\!/\!_{n^2})$. The number of required features is $s \geq \log n \log\log n$ because $d_{\mathbf{K}}^{\lambda} \leq \log(R^2/\lambda) \leq \log n$. If the eigenvalues decay at the rate $\lambda_i \propto \mathcal{O}(i^{-t})$ with $t > 1$, then the learning rate is $\mathcal{O}(1/\sqrt{n})$ by setting $\lambda \in \mathcal{O}(1/n)$, with the requirement on the number of features given by $d_{\mathbf{K}}^{\lambda} \leq (R^2/\lambda)^{1/t} \leq n^{1/t}$. Since $t > 1$, one can see that with fewer than n features, we could obtain fast $\mathcal{O}(1/n)$ learning rate.

On the other hand, in the plain sampling scheme, if we would like to achieve the fast $\mathcal{O}(1/n)$ learning rate, we need to set $\lambda \in \mathcal{O}(1/n^2)$, implying that the required number of features has to be $s \geq n^2$. This is undesirable as it does not provide any computation savings. The bottleneck here is that in the Lipschitz continuous case, learning rate is upper bounded by $\mathcal{O}(\sqrt{\lambda})$.

3.2.3 COMPARISON WITH THE SHARPEST BOUNDS FROM PRIOR WORK

In the setting with Lipshcitz continuous loss, several previous results provide similar trade-offs between the computational cost and prediction accuracy (see, e.g., Rahimi and Recht, 2009; Bach, 2017b; Sun et al., 2018). We cover below the sharpest bounds from prior work and discuss how our results advance the understanding of learning with random features in this setting.

Worst Case Analysis. Rahimi and Recht (2009) were the first to consider the trade-offs between the computational cost and prediction accuracy for the plain random Fourier features sampling scheme. Building on that work, Bach (2017b) has provided a characterization for the setting with weighted random Fourier features. We start with a comparison to Rahimi and Recht (2009) in the setting with plain random Fourier features. The first block of rows in Table 8 illustrates the difference between our results and that work. We can see that under plain sampling, when the eigenspetrum has a finite rank or exhibits an exponential decay, the required number of features in our bounds is smaller than that required by Rahimi and Recht (2009), i.e., $s \in \Omega(n)$ versus $s \in \Omega(n \cdot \log n)$ and $s \in \Omega(n \cdot \log\log n)$ versus $s \in \Omega(n \cdot \log n)$, respectively. When the eigenspectum follows a polynomial decay, our results match those provided by Rahimi and Recht (2009).

For the worst case setting under the weighted sampling scheme, Bach (2017b) provides a detailed theoretical analysis of learning with random features by establishing the equivalence between random

SAMPLING SCHEME	RESULTS	SPECTRUM	NUMBER OF FEATURES	LEARNING RATE
PLAIN RFF	THIS WORK	finite rank	$s \in \Omega(n)$	$\mathcal{O}(1/\sqrt{n})$
		$\lambda_i \propto A^i$	$s \in \Omega(n \cdot \log\log n)$	
		$\lambda_i \propto i^{-2t}, t \geq 1$	$s \in \Omega(n \cdot \log n)$	
		$\lambda_i \propto i^{-1}$		
	RAHIMI & RECHT (2009)	finite rank	$s \in \Omega(n \cdot \log n)$	
		$\lambda_i \propto A^i$		
		$\lambda_i \propto i^{-2t}, t \geq 1$		
		$\lambda_i \propto i^{-1}$		
WEIGHTED RFF	THIS WORK	finite rank	$s \in \Omega(1)$	$\mathcal{O}(1/\sqrt{n})$
		$\lambda_i \propto A^i$	$s \in \Omega(\log n \cdot \log\log n)$	
		$\lambda_i \propto i^{-2t}, t \geq 1$	$s \in \Omega(n^{1/2t} \cdot \log n)$	
		$\lambda_i \propto i^{-1}$	$s \in \Omega(n \cdot \log n)$	
	BACH (2017B)	finite rank	$-$	
		$\lambda_i \propto A^i$	$s \in \Omega(\log n \cdot \log\log n)$	
		$\lambda_i \propto i^{-2t}, t \geq 1$	$s \in \Omega(n^{1/2t} \cdot \log n)$	
		$\lambda_i \propto i^{-1}$	$s \in \Omega(n \cdot \log n)$	

Table 8: The comparison of our results to the sharpest learning rates from prior work in the worst case setting with a Lipschitz continuous loss and plain/weighted random Fourier features.

features and the kernel quadrature rules. That work was also the first to demonstrate that it is possible to achieve computational savings while preserving the prediction accuracy. The second block of rows in Table 8 illustrates the difference between our results and Bach (2017b). We can see that apart from the finite rank case that was not covered by Bach (2017b), our results match the worst case bounds.

Refined Case Analysis. Having covered the worst case setting for Lipschitz continuous loss, we now proceed to the refined case where learning rates can be sharper than $\mathcal{O}(1/\sqrt{n})$. We remark that while Rahimi and Recht (2009) and Bach (2017b) do not cover the refined case with fast learning rates, Sun et al. (2018) provide a refined analysis for the hinge loss. Similar to our work (i.e., Assumption B.2), Sun et al. (2018) assume a low noise condition and demonstrate that learning algorithms operating on features selected via an optimized sampling distribution can obtain sharper learning rates in classification problems (0-1 loss function) than the standard $\mathcal{O}(1/\sqrt{n})$ rate.

Table 9 illustrates the differences between our results and bounds from Sun et al. (2018). From the table, we can see that the guarantees match when the eigenspectrum has a finite rank. When the eigenspectrum exhibits a slower decay, however, the results differ significantly. More specifically, the results from Sun et al. (2018) suffer from the curse of dimension when the decay is exponential.

SAMPLING SCHEME	SPECTRUM	NUMBER OF FEATURES	LEARNING RATE
THIS WORK	finite rank	$s \in \Omega(1)$	$\mathcal{O}(1/n)$
	$\lambda_i \propto A^i$	$s \in \Omega(\log n \cdot \log\log n)$	$\mathcal{O}(\log n/n)$
	$\lambda_i \propto i^{-t}, t > 1$	$s \in \Omega(n^{1/2t} \cdot \log n)$	$\mathcal{O}(1/\sqrt{n})$
SUN ET AL. (2018)	finite rank	$s \in \Omega(1)$	$\mathcal{O}(1/n)$
	$\lambda_i \propto A^i$	$s \in \Omega(\log^d n \cdot \log\log^d n)$	$\mathcal{O}(\log^{d+2} n/n)$
	$\lambda_i \propto i^{-t}, t > 1$	$s \in \Omega(n^{\frac{2}{2+t}} \cdot \log n)$	$\mathcal{O}(n^{\frac{t}{2+t}})$

Table 9: The refined case trade-offs between computational and statistical efficiency relative to Sun et al. (2018).

This is because both the number of features required and the learning rate obtained depend on the data dimension d, whereas our results do not have this dependency.

When the eigenspectrum exhibits a polynomial decay, we can see that our results achieve the minimax learning rate of $\mathcal{O}(1/\sqrt{n})$ while Sun et al. (2018) obtain a more flexible rate that depends on the magnitude of the decay given by parameter t. When $t < 2$, our results have a better trade-off as both the number of features and learning rate are sharper than those from Sun et al. (2018). When $t > 2$, the learning rate obtained by Sun et al. (2018) is faster than ours. However, that comes at the cost of increasing the required number of features, i.e., $s \in \Omega(n^{\frac{2}{2+t}} \cdot \log n)$ versus $s \in \Omega(n^{\frac{1}{2t}} \cdot \log n)$. In particular, setting $t = 4$ we can see that Sun et al. (2018) obtain a fast $\mathcal{O}(n^{2/3})$ learning rate, but at the cost of requiring $s \in \Omega(n^{1/3} \cdot \log n)$ random features. For the same setting, on the other hand, we obtain the minimax optimal $\mathcal{O}(1/\sqrt{n})$ learning rate with only $s \in \Omega(n^{1/8} \cdot \log n)$ random features.

4. A Fast Approximation of Leverage Weighted RFF

As discussed in Sections 3, sampling according to the empirical ridge leverage score distribution (i.e., leverage weighted RFF) could speed up kernel methods. However, computing ridge leverage scores is as costly as inverting the Gram matrix. To address this computational shortcoming, we propose a simple algorithm to approximate the empirical ridge leverage score distribution and the leverage weights. In particular, we propose to first sample a pool of s features from the spectral measure $p(\cdot)$ and form the feature matrix $\mathbf{Z}_s \in \mathbb{R}^{n \times s}$ (Algorithm 1, lines 1-2). Then, the algorithm associates an approximate empirical ridge leverage score to each feature (Algorithm 1, lines 3-4) and samples a set of $m \ll s$ features from the pool proportional to the computed scores (Algorithm 1, line 5). The output of the algorithm can be compactly represented via the feature matrix $\mathbf{Z}_m \in \mathbb{R}^{n \times m}$ such that the i-th row of $\mathbf{Z}_m$ is given by $\mathbf{z}_{x_i}(\mathbf{v}) = [\sqrt{m/p_1} z(v_1, x_i), \cdots, \sqrt{m/p_m} z(v_m, x_i)]^T$.

The computational cost of Algorithm 1 is dominated by the operations in step 3. As $\mathbf{Z}_s \in \mathbb{R}^{n \times s}$, the multiplication of matrices $\mathbf{Z}_s^T \mathbf{Z}_s$ costs $\mathcal{O}(ns^2)$ and inverting $\mathbf{Z}_s^T \mathbf{Z}_s + n\lambda I$ costs only $\mathcal{O}(s^3)$. Hence, for $s \ll n$, the overall runtime is only $\mathcal{O}(ns^2 + s^3)$. Moreover, $\mathbf{Z}_s^T \mathbf{Z}_s = \sum_{i=1}^n \mathbf{z}_{x_i}(\mathbf{v}) \mathbf{z}_{x_i}(\mathbf{v})^T$ and it is possible to store only the rank-one matrix $\mathbf{z}_{x_i}(\mathbf{v}) \mathbf{z}_{x_i}(\mathbf{v})^T$ into the memory. Thus, the algorithm only requires to store an $s \times s$ matrix and can avoid storing $\mathbf{Z}_s$, which would incur a storage cost of $\mathcal{O}(n \times s)$.

The following theorem gives the convergence rate for the learning risk of Algorithm 1 in the kernel ridge regression setting.

Algorithm 1 APPROXIMATE LEVERAGE WEIGHTED RFF

```
Input: sample of examples ${(x_i, y_i)}_{i=1}^n$, shift-invariant kernel function $k$, and regularization parameter $\lambda$
Output: set of features ${(v_1, p_1), \cdots, (v_m, p_m)}$ with $m$ and each $p_i$ computed as in lines 3–4
1: sample a pool of $s$ random Fourier features ${v_1, \dots, v_s}$ from $p(v)$
2: create a feature matrix $\mathbf{Z}_s$ such that the $i$-th row of $\mathbf{Z}_s$ is
       $[z(v_1, x_i), \cdots, z(v_s, x_i)]^T$
3: associate with each feature $v_i$ a positive real number $p_i$ such that $p_i$ is equal to the $i$-th diagonal element
   of matrix
       $\mathbf{Z}_s^T \mathbf{Z}_s ((1/s)\mathbf{Z}_s^T \mathbf{Z}_s + n\lambda I)^{-1}$
4: $m \leftarrow \sum_{i=1}^s p_i$ and $M \leftarrow {(v_i, p_i/m)}_{i=1}^s$
5: sample $\lceil m \rceil$ features from set $M$ using the multinomial distribution given by vector $(p_1/m, \cdots, p_s/m)$
```

Theorem 20 *Suppose that Assumption A.1 holds and consider the regression problem defined with a shift-invariant kernel k, a sample of examples $\{(x_i, y_i)\}_{i=1}^n$, and a regularization parameter λ. Let s be the number of random Fourier features in the pool of features from Algorithm 1, sampled using the spectral measure $p(\cdot)$ from Eq. (3) and the regularization parameter λ. Denote with $\tilde{f}_m^{\lambda^*}$ the ridge regression estimator obtained using a regularization parameter λ^* and a set of random Fourier features $\{v_i\}_{i=1}^m$ returned by Algorithm 1. If*

$$s \;\geq\; \frac{7z_0^2}{\lambda} \log \frac{(16 d_{\mathbf{K}}^{\lambda})}{\delta} \quad \text{and} \quad m \;\geq\; 5 d_{\mathbf{K}}^{\lambda^*} \log \frac{(16 d_{\mathbf{K}}^{\lambda^*})}{\delta} \;,$$

then for all $\delta \in (0, 1)$, with probability $1 - \delta$, the learning risk of $\tilde{f}_m^{\lambda^*}$ can be upper bounded by

$$\mathbb{E}[l_{\tilde{f}_m^{\lambda^*}}] \;\leq\; \mathbb{E}[l_{f_{\mathcal{H}}}] + 4\lambda + 4\lambda^* + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

Moreover, this upper bound holds for $m \in \Omega(\frac{s}{n\lambda})$.

Theorem 20 bounds the learning risk of the ridge regression estimator over random features generated by Algorithm 1. We can now observe that using the minimax choice of the regularization parameter for kernel ridge regression $\lambda, \lambda^* \propto n^{-1/2}$, the number of features that Algorithm 1 needs to sample from the spectral measure of the kernel k is $s \in \Omega(\sqrt{n} \log n)$. Then, the ridge regression estimator $\tilde{f}_m^{\lambda^*}$ converges with the minimax rate to the hypothesis $f_{\mathcal{H}} \in \mathcal{H}$ for $m \in \Omega(\log n \cdot \log \log n)$.

This is a significant improvement compared to the spectral measure sampling (plain RFF), which requires $\Omega(n^{3/2})$ features for in-sample training and $\Omega(\sqrt{n} \log n)$ for out-of-sample test predictions.

Theorem 21 provides a convergence bound for kernel support vector machines and logistic regression. Compared to the previous result, the convergence rate of the learning risk, however, is at a slower $\mathcal{O}(\sqrt{\lambda} + \sqrt{\lambda^*})$ rate due to the difference in the employed loss function (see also Section 3.2).

Theorem 21 *Suppose that Assumption B.1 holds and consider the learning problem with a Lipschitz continuous loss function, a shift-invariant kernel k, a sample of examples $\{(x_i, y_i)\}_{i=1}^n$, and a regularization parameter λ. Let s be the number of random Fourier features in the pool of features* *from Algorithm 1, sampled using the spectral measure $p(\cdot)$ from Eq. (3) and the regularization parameter λ. Denote with $\tilde{g}_m^{\lambda^*}$ the estimator obtained using a regularization parameter λ^* and a set of random Fourier features $\{v_i\}_{i=1}^m$ returned by Algorithm 1. If*

$$s \;\geq\; \frac{5z_0^2}{\lambda} \log \frac{(16d_{\mathbf{K}}^{\lambda})}{\delta} \quad \textit{and} \quad m \;\geq\; 5d_{\mathbf{K}}^{\lambda^*} \log \frac{(16d_{\mathbf{K}}^{\lambda^*})}{\delta},$$

then for all $\delta \in (0,1)$, with probability $1-\delta$, the learning risk of $\tilde{g}_m^{\lambda^*}$ can be upper bounded by

$$\mathbb{E}[l_{\tilde{g}_m^{\lambda^*}}] \;\leq\; \mathbb{E}[l_{g_{\mathcal{H}}}] + \sqrt{2\lambda} + \sqrt{2\lambda^*} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

We conclude by pointing out that the proposed algorithm provides an interesting new trade-off between the computational cost and prediction accuracy. In particular, one can pay an upfront cost (same as plain RFF) to compute the leverage scores, re-sample significantly fewer features and employ them in the training, cross-validation, and prediction stages. This can reduce the computational cost for predictions at test points from $\Omega(\sqrt{n}\log n)$ to $\Omega(\log n \cdot \log\log n)$. Moreover, in the case where the amount of features with approximated leverage scores utilized is the same as in plain RFF, the prediction accuracy would be significantly improved, as demonstrated in our experiments.

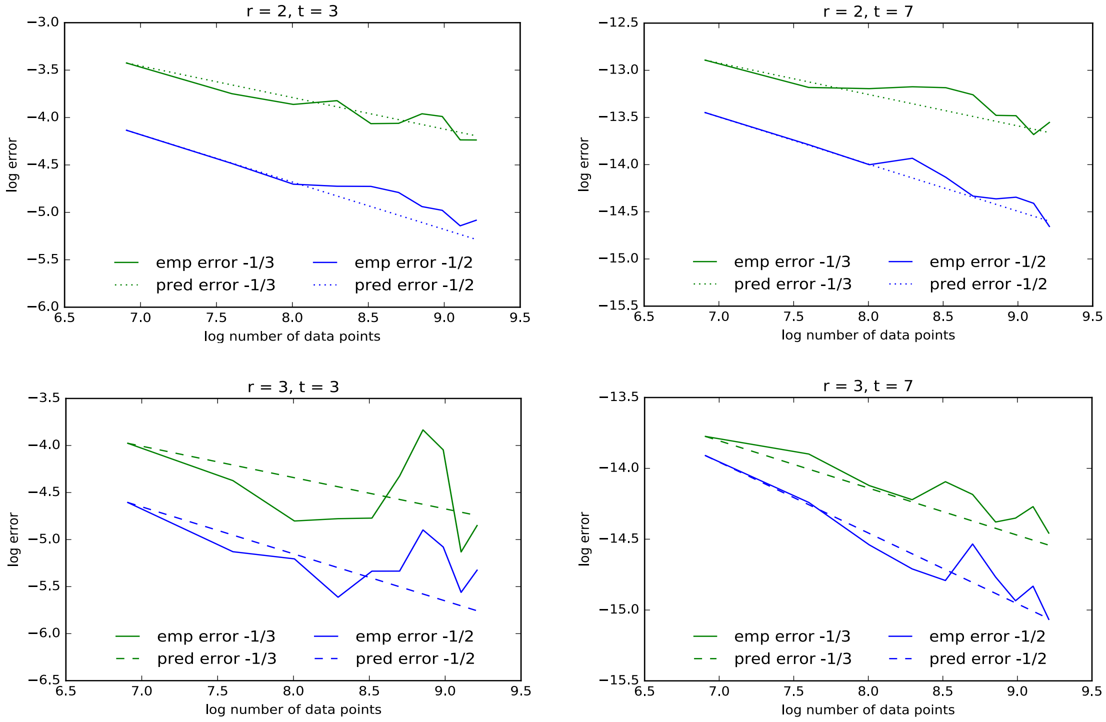


Figure 1: The log-log plot of the theoretical and simulated risk convergence rates, averaged over 100 repetitions.

5. Numerical Experiments

In this section, we report the results of our numerical experiments (on both simulated and real-world datasets) aimed at validating the theoretical results and demonstrating the utility of Algorithm 1. In

# OF FEATURES	PLAIN RFF	LEVERAGE WEIGHTED RFF
$1,000$	0.13 ± 0.06	0.04 ± 0.01
$50,000$	0.04 ± 0.02	–

Table 10: The table summarizes our experiment on the synthetic dataset and illustrates the effectiveness of the proposed algorithm (i.e., leverage weighted RFF) relative to plain RFF sampling. The reported numbers are the root mean squared error (RMSE) along with a corresponding confidence interval.

the first experiment, the goal is to show that our bounds for the ridge regression setting are tight by demonstrating empirically that the observed error rates follow closely the provided learning risk bounds. The second experiment deals with the effectiveness of the proposed algorithm relative to the plain random Fourier features sampling scheme, evaluated on four benchmark datasets typically used for this type of problems. The third and final experiment aims at demonstrating the utility of the leverage weighted features in a simulated experiment designed such that an effective approximation of the target hypothesis requires a small number of random features which are located in the tails of the spectral measure corresponding to the selected shift-invariant kernel.

We use a simulated experiment to verify the sharpness of our theoretical results. More specifically, we consider a spline kernel of order r where $k_{2r}(x, y) = 1 + \sum_{m>0} \frac{1}{m^{2r}} \cos 2\pi m(x - y)$ (also considered by Bach, 2017b; Rudi and Rosasco, 2017). If the marginal distribution of X is uniform on $[0, 1]$, one can show that $k_{2r}(x, y) = \int_0^1 z(v, x)z(v, y)p(v)dv$, where $z(v, x) = k_r(v, x)$ and $p(v)$ is uniform on $[0, 1]$. Moreover, one can show that the optimal weighted sampling distribution $q^*(v)$ is the same as $p(v)$, which allows us to use weighted RFF sampling strategy. We let y be a Gaussian random variable with mean $f(x) = k_t(x, x_0)$ (for some $x_0 \in [0, 1]$) and variance σ^2. We sample features according to $q^*(v)$ to estimate f and compute the excess risk. By Theorem 9 and Corollary 10, if the number of features is proportional to $d_{\mathbf{K}}^{\lambda}$ and $\lambda \propto n^{-1/2}$, we should expect the excess risk converging at $\mathcal{O}(n^{-1/2})$, or at $\mathcal{O}(n^{-1/3})$ if $\lambda \propto n^{-1/3}$. Figure 1 demonstrates that this is indeed the case for different values of r and t.

Next, we make a comparison between the performances of leverage weighted (computed according to Algorithm 1) and plain RFF on real-world datasets. In particular, we use four datasets from Chang and Lin (2011) and Dheeru and Karra Taniskidou (2017) for this purpose, including two for regression and two for classification: `CPU`, `KINEMATICS`, `COD-RNA` and `COVTYPE`. Apart from `KINEMATICS`, the other three datasets were used in Yang et al. (2012) to investigate the difference between the Nyström method and plain RFF. We use the ridge regression and SVM package from Pedregosa et al. (2011) as a solver to perform our experiments. We evaluate the regression tasks using the root mean squared error and the classification ones using the average percentage of misclassified examples. The Gaussian/RBF kernel is used for all the datasets with hyper-parameter tuning via 5-fold inner cross validation. We have repeated all the experiments 10 times and reported the average test error for each dataset. Figure 2 compares the performances of leverage weighted and plain RFF. In regression tasks, we observe that the upper bound of the confidence interval for the root mean squared error corresponding to leverage weighted RFF is below the lower bound of the confidence interval for the error corresponding to plain RFF. Similarly, the lower bound of the confidence interval for the classification accuracy of leverage weighted RFF is (most of the time) higher than the upper bound on the confidence interval for plain RFF. This indicates that leverage weighted RFFs perform statistically significantly better than plain RFFs in terms of the learning accuracy and prediction error.

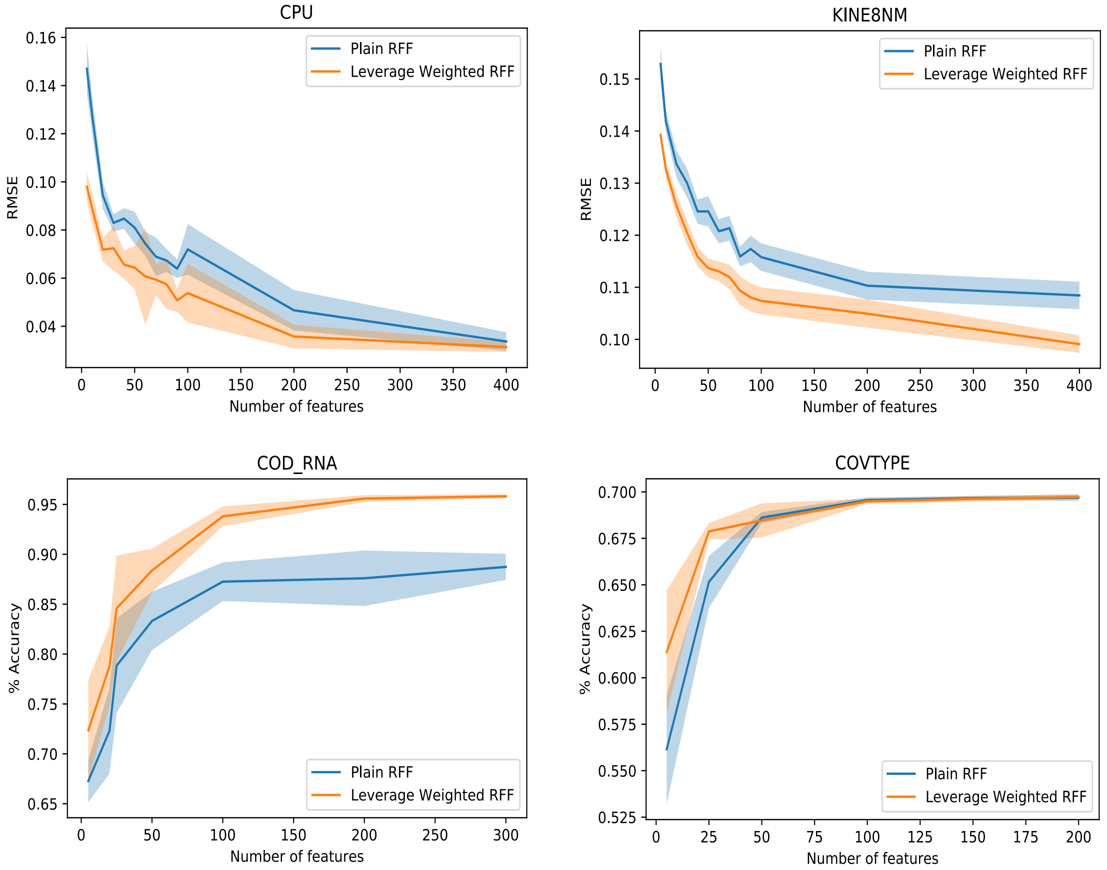


Figure 2: Comparison of leverage weighted and plain RFFs, with weights computed according to Algorithm 1.

In the final experiment, we would like to show that the proposed algorithm can significantly reduce the number of required features without loss of statistical efficiency. As it is challenging to find an appropriate real-world dataset for this illustration, we design a synthetic regression problem on our own. The main idea is to define a target regression function as a linear model directly in the space of random Fourier features. We select the target features such that they are in the tail of the spectral measure of the kernel that defines our hypothesis space. This ensures that plain RFF strategy will require a large number of features to describe the target hypothesis. We evaluate the effectiveness of our algorithm relative to plain RFFs and construct the described synthetic dataset as follows: we first generate samples w^* from a multimodal Gaussian distribution where the modes are at $(-2,-2), (-2,2), (2,-2)$, and $(2,2)$. Moreover, each of the modes has a diagonal covariance matrix of 0.5. These samples are going to be our frequencies for a RFF mapping. Next, we sample our covariates x from $\mathcal{N}(0, 5 * I)$. In order to generate our response variables, we map the covariates x through a RFF map where the frequencies are given by samples w^*. We then randomly sample regression weights α_r from $\mathcal{N}(0,1)$. Hence, the data generating process can be specified by:

$$y = \alpha_r^T \phi_{w^*}(x) + \epsilon \; ,$$

where $\epsilon \sim \mathcal{N}(0,\sigma)$ and ϕ_{w^*} is a RFF map with w^* as the frequencies. By setting up our data generating process in this way, we are able to systematically investigate how well the proposed algorithm works. In particular, we consider learning the above described hypothesis using RFFs that correspond to a Gaussian kernel. Such a kernel corresponds to a uni-modal Gaussian distribution in the frequency domain and we will show that leverage weighted sampling is capable of selecting a sub-set of plain RFF sampled from that distribution, which covers the modes of the multimodal distribution that characterizes the data generating process.

In our simulations, we have opted for the following setting: $50,000$ data points denoted with x in the data-generating process, 400 features/frequencies w^* that define the target hypothesis y, and the additive noise variance parameter $\sigma = 0.1$. We then run plain RFF as well as our leverage weighted RFF on this dataset. For plain RFF, we run the experiments with $1,000$ and $50,000$ frequencies/features, while with our leverage weighted RFF we only use $1,000$ features which have been selected from a pool consisting of $10,000$ plain random RFFs (i.e., a sub-sample from the original $50,000$ features). We carefully cross-validate both methods across a grid of hyper-parameters and report the results in Table 10. The results confirm our theoretical findings and illustrate that learning with $1,000$ leverage weighted RFFs is as effective as learning with a complete pool of $50,000$ plain RFFs.

6. Proofs

6.1 Proof of Proposition 8

Proposition 8 *Assume that the reproducing kernel Hilbert space $\mathcal{H}$ with kernel k admits a decomposition as in Eq. (3) and denote by $\tilde{\mathcal{H}} := \{\tilde{f} \mid \tilde{f} = \sum_{i=1}^{s} \alpha_i z(v_i, \cdot), \alpha_i \in \mathbb{R}\}$ the reproducing kernel Hilbert space with kernel $\tilde{k}$ (see Eq. 4). Then, for all $\tilde{f} \in \tilde{\mathcal{H}}$ it holds that $\|\tilde{f}\|^2_{\tilde{\mathcal{H}}} \leq s\|\alpha\|^2_2$.*

Proof Let us define a space of functions as

$$\mathcal{H}_1 := \{f \mid f(x) = \alpha z(v, x), \alpha \in \mathbb{R}\} .$$

We now show that $\mathcal{H}_1$ is a reproducing kernel Hilbert space with kernel defined as $k_1(x, y) = (1/s)z(v, x)z(v, y)$, where s is a constant. Define a map $M : \mathbb{R} \to \mathcal{H}_1$ such that $M\alpha = \alpha z(v, \cdot), \forall \alpha \in \mathbb{R}$. The map M is a bijection, i.e. for any $f \in \mathcal{H}_1$ there exists a unique $\alpha_f \in \mathbb{R}$ such that $M^{-1}f = \alpha_f$. Now, we define an inner product on $\mathcal{H}_1$ as

$$\langle f, g\rangle_{\mathcal{H}_1} = \langle \sqrt{s}M^{-1}f, \sqrt{s}M^{-1}g\rangle_{\mathbb{R}} = s\alpha_f\alpha_g .$$

It is easy to show that this is a well defined inner product and, thus, $\mathcal{H}_1$ is a Hilbert space.

For any instance y, $k_1(\cdot, y) = (1/s)z(v, \cdot)z(v, y) \in \mathcal{H}_1$, since $(1/s)z(v, y) \in \mathbb{R}$ by definition. Take any $f \in \mathcal{H}_1$ and observe that

$$\begin{aligned}\langle f, k_1(\cdot, y)\rangle_{\mathcal{H}_1} &= \langle \sqrt{s}M^{-1}f, \sqrt{s}M^{-1}k_1(\cdot, y)\rangle_{\mathbb{R}} \\ &= s\langle \alpha_f, 1/sz(v, y)\rangle_{\mathbb{R}} \\ &= \alpha_f z(v, y) = f(y) .\end{aligned}$$

Hence, we have demonstrated the reproducing property for $\mathcal{H}_1$ and $\|f\|^2_{\mathcal{H}_1} = s\alpha_f^2$.

Now, suppose we have a sample of features $\{v_i\}_{i=1}^s$. For each v_i, we define the reproducing kernel Hilbert space

$$\mathcal{H}_i := \{f \mid f(x) = \alpha z(v_i, x), \alpha \in \mathbb{R}\}$$

with the kernel $k_i(x, y) = (1/s)z(v_i, x)z(v_i, y)$. Denoting with

$$\tilde{\mathcal{H}} = \oplus_{i=1}^s \mathcal{H}_i = \{\tilde{f} : \tilde{f} = \sum_{i=1}^s f_i, f_i \in \mathcal{H}_i\}$$

and using the fact that the direct sum of reproducing kernel Hilbert spaces is another reproducing kernel Hilbert space (Berlinet and Thomas-Agnan, 2011), we have that $\tilde{k}(x, y) = \sum_{i=1}^s k_i(x, y) = (1/s) \sum_{i=1}^s z(v_i, x)z(v_i, y)$ is the kernel of $\tilde{\mathcal{H}}$ and that the squared norm of $\tilde{f} \in \tilde{\mathcal{H}}$ is defined as

$$\min_{f_i \in \mathcal{H}_i \mid \tilde{f} = \sum_{i=1}^s f_i} \sum_{i=1}^s \|f_i\|_{\mathcal{H}_i}^2 =$$

$$\min_{\alpha_i \in \mathbb{R} \mid \tilde{f} = \sum_{i=1}^s \alpha_i z(v_i, \cdot)} \sum_{i=1}^s s\alpha_i^2 = \min_{\alpha_i \in \mathbb{R} \mid \tilde{f} = \sum_{i=1}^s \alpha_i z(v_i, \cdot)} s\|\alpha\|_2^2 .$$

Hence, we have that $\|\tilde{f}\|_{\tilde{\mathcal{H}}}^2 \leq s\|\alpha\|_2^2$. ■

6.2 Proof of Theorem 9

To prove Theorem 9, we need two auxiliary results formulated in Lemmas 22 and 24 (the proofs are provided in Appendices B and C, respectively). More specifically, Lemma 22 is a general result that gives an upper bound on the approximation error between any function $f \in \mathcal{H}$ and its estimator based on random Fourier features. As discussed in Section 2, we would like to approximate a function $f \in \mathcal{H}$ at observation points using $\tilde{f} \in \tilde{\mathcal{H}}$, with preferably as small function norm $\|\tilde{f}\|_{\tilde{\mathcal{H}}}$ as possible. As such, the estimation of $\mathbf{f}_x = [f(x_1), \ldots, f(x_n)]^T$ with $\tilde{\mathbf{f}}_x = [\tilde{f}(x_1), \ldots, \tilde{f}(x_n)]^T$ can be formulated via the following optimization problem:

$$\min_{\beta} \frac{1}{n}\|\mathbf{f}_x - \mathbf{Z}_q\beta\|_2^2 + \lambda s\|\beta\|_2^2 .$$

Lemma 22 *Suppose that Assumption A.1 holds and that the conditions on sampling measure $\tilde{l}$ from Theorem 9 apply as well. If*

$$s \geq 5d_{\tilde{l}} \log \frac{16d_{\mathbf{K}}^{\lambda}}{\delta} ,$$

then for all $\delta \in (0, 1)$ and any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$, with probability greater than $1 - \delta$, the following holds

$$\min_{\beta} \left\{\frac{1}{n}\|\mathbf{f}_x - \mathbf{Z}_q\beta\|_2^2 + \lambda s\|\beta\|_2^2\right\} \leq 2\lambda .$$

For the sake of brevity, we will henceforth use $\tilde{\mathcal{H}}$ to denote the hypothesis space corresponding to this optimization problem. Then, the latter bound can be written as:

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \inf_{\|\tilde{f}\|_{\tilde{\mathcal{H}}} \leq \sqrt{2}} \frac{1}{n}\|\mathbf{f}_x - \tilde{\mathbf{f}}_x\|_2^2 \leq 2\lambda .$$

Remark 23 *We note here that in the strict sense the hypothesis space $\tilde{\mathcal{H}}$ is contained in the interior of the ball of radius $\sqrt{2}$, centered at the origin (see also Proposition 8). To simplify our presentation and avoid carrying the cumbersome optimization problem for learning with random Fourier features (e.g., formally given in Lemma 22), we make a minor adjustment/violation in notation and refer to the whole ball of radius $\sqrt{2}$ as the hypothesis space of random Fourier features.*

The following lemma is important for demonstrating the risk convergence rate and its proof is given in Appendix C. Recall that we have defined $\hat{f}^\lambda$ as the empirical estimator for the kernel ridge regression problem in Eq. (1).

Lemma 24 *Denote the in-sample prediction of $\hat{f}^\lambda$ with*

$$\hat{\mathbf{f}}_x^\lambda = [\hat{f}^\lambda(x_1), \ldots, \hat{f}^\lambda(x_n)]^T \tag{14}$$

and let $\{v_i\}_{i=1}^s$ be independent samples selected according to a probability density function $q(v)$, which define the feature matrix $\mathbf{Z}_q$ and the corresponding reproducing kernel Hilbert space $\tilde{\mathcal{H}}$. Let $\tilde{\beta}^\lambda$ be the solution to the following optimization problem

$$\tilde{\beta}^\lambda := \min_\beta \frac{1}{n} \|\hat{\mathbf{f}}_x^\lambda - \mathbf{Z}_q \beta\|_2^2 + \lambda s \|\beta\|_2^2$$

and denote the in-sample prediction of the resulting hypothesis with $\tilde{\mathbf{f}}_x^\lambda = \mathbf{Z}_q \tilde{\beta}^\lambda$. Then, we have

$$\frac{1}{n} \langle Y - \hat{\mathbf{f}}_x^\lambda, \hat{\mathbf{f}}_x^\lambda - \tilde{\mathbf{f}}_x^\lambda \rangle \leq \lambda \ .$$

Equipped with Lemma 22 and Lemma 24, we are now ready to prove Theorem 9.

Theorem 9 *Suppose that Assumption A.1 holds and let $\tilde{l} : \mathcal{V} \to \mathbb{R}$ be a measurable function such that $\tilde{l}(v) \geq l_\lambda(v)$ ($\forall v \in \mathcal{V}$) with $d_{\tilde{l}} = \int_{\mathcal{V}} \tilde{l}(v) dv < \infty$. Suppose also that $\{v_i\}_{i=1}^s$ are sampled independently from the probability density function $q(v) = \tilde{l}(v)/d_{\tilde{l}}$. If*

$$s \ \geq \ 5 d_{\tilde{l}} \log \frac{16 d_{\mathbf{K}}^\lambda}{\delta} \ ,$$

then for all $\delta \in (0,1)$, with probability $1-\delta$, the excess risk of $\tilde{f}_\beta^\lambda$ can be upper bounded by

$$\mathbb{E}[l_{\tilde{f}_\beta^\lambda}] - \mathbb{E}[l_{f_{\mathcal{H}}}] \ \leq \ 4\lambda + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) + \mathbb{E}[l_{\hat{f}^\lambda}] - \mathbb{E}[l_{f_{\mathcal{H}}}] \ . \tag{8}$$

Proof The proof relies on the decomposition of the learning risk of $\mathbb{E}[l_{\tilde{f}_\beta^\lambda}]$ as follows

$$\begin{aligned}
\mathbb{E}[l_{\tilde{f}_\beta^\lambda}] = {} & \mathbb{E}[l_{\tilde{f}_\beta^\lambda}] - \mathbb{E}_n[l_{\tilde{f}_\beta^\lambda}] && (15)\\
& + \mathbb{E}_n[l_{\tilde{f}_\beta^\lambda}] - \mathbb{E}_n[l_{\hat{f}^\lambda}] && (16)\\
& + \mathbb{E}_n[l_{\hat{f}^\lambda}] - \mathbb{E}[l_{\hat{f}^\lambda}] && (17)\\
& + \mathbb{E}[l_{\hat{f}^\lambda}] - \mathbb{E}[l_{f_{\mathcal{H}}}] && (18)\\
& + \mathbb{E}[l_{f_{\mathcal{H}}}] \ .
\end{aligned}$$

For (15), the bound is based on the Rademacher complexity of the reproducing kernel Hilbert space $\tilde{\mathcal{H}}$, where $\tilde{\mathcal{H}}$ corresponds to the approximated kernel $\tilde{k}$. We can upper bound the Rademacher complexity of this hypothesis space using Lemma 2. More specifically, as $l(y, f(x))$ is the squared error loss function with y and $f(x)$ both bounded, we have that l is a Lipschitz continuous function with some constant $L > 0$. Hence,

$$\begin{aligned}
(15) &\le R_n(l \circ \tilde{\mathcal{H}}) + \sqrt{\frac{8\log(2/\delta)}{n}} \\
&\le \sqrt{2}L\frac{1}{n}\mathbb{E}_X[\sqrt{\mathrm{Tr}(\tilde{\mathbf{K}})}] + \sqrt{\frac{8\log(2/\delta)}{n}} \\
&\le \sqrt{2}L\frac{1}{n}\sqrt{\mathbb{E}_X[\mathrm{Tr}(\tilde{\mathbf{K}})]} + \sqrt{\frac{8\log(2/\delta)}{n}} \\
&\le \sqrt{2}L\frac{1}{n}\sqrt{nz_0^2} + \sqrt{\frac{8\log(2/\delta)}{n}} \\
&= \frac{\sqrt{2}Lz_0}{\sqrt{n}} + \sqrt{\frac{8\log(2/\delta)}{n}} \in \mathcal{O}\left(\frac{1}{\sqrt{n}}\right), \qquad (19)
\end{aligned}$$

where in the first inequality we applied Lemma 3 to $\tilde{\mathcal{H}}$, which is a reproducing kernel Hilbert space contained in the ball of radius $\sqrt{2}$ centered at the origin. Moreover, the bound on $R_n(l \circ \tilde{\mathcal{H}})$ relies on the Lipschitz composition property of Rademacher complexity (Bartlett and Mendelson, 2002). For (17), a similar reasoning can be applied to the unit ball in the reproducing kernel Hilbert space $\mathcal{H}$.

For (16), we recall that $\tilde{\mathbf{f}}_\beta^\lambda = \mathbf{Z}\beta_\lambda$ where β_λ is the solution of the following optimization problem

$$\beta_\lambda := \min_\beta \frac{1}{n}\|Y - \mathbf{Z}\beta\|_2^2 + \lambda s\|\beta\|_2^2 \,.$$

We now observe that

$$\begin{aligned}
\mathbb{E}_n[l_{\tilde{f}_\beta^\lambda}] - \mathbb{E}_n[l_{\hat{f}^\lambda}] &= \frac{1}{n}\|Y - \tilde{\mathbf{f}}_\beta^\lambda\|_2^2 - \frac{1}{n}\|Y - \hat{\mathbf{f}}_x^\lambda\|_2^2 \\
&= \frac{1}{n}\|Y - \mathbf{Z}\beta_\lambda\|_2^2 - \frac{1}{n}\|Y - \hat{\mathbf{f}}_x^\lambda\|_2^2 \\
&\le \min_\beta \left\{\frac{1}{n}\|Y - \mathbf{Z}\beta\|_2^2 + \lambda s\|\beta\|_2^2\right\} - \frac{1}{n}\|Y - \hat{\mathbf{f}}_x^\lambda\|_2^2 \\
&= \min_\beta \left\{\frac{1}{n}\|Y - \hat{\mathbf{f}}_x^\lambda\|_2^2 + \frac{1}{n}\|\hat{\mathbf{f}}_x^\lambda - \mathbf{Z}\beta\|_2^2 + \frac{2}{n}\langle Y - \hat{\mathbf{f}}_x^\lambda, \hat{\mathbf{f}}_x^\lambda - \mathbf{Z}\beta\rangle + \lambda s\|\beta\|_2^2\right\} \\
&\quad -\frac{1}{n}\|Y - \hat{\mathbf{f}}_x^\lambda\|_2^2 \\
&= \frac{1}{n}\min_\beta \left\{\|\hat{\mathbf{f}}_x^\lambda - \mathbf{Z}\beta\|_2^2 + 2\langle Y - \hat{\mathbf{f}}_x^\lambda, \hat{\mathbf{f}}_x^\lambda - \mathbf{Z}\beta\rangle + \lambda sn\|\beta\|_2^2\right\} \\
&\le \frac{1}{n}\left\{\|\hat{\mathbf{f}}_x^\lambda - \mathbf{Z}\tilde{\beta}^\lambda\|_2^2 + \lambda sn\|\tilde{\beta}^\lambda\|_2^2 + 2\langle Y - \hat{\mathbf{f}}_x^\lambda, \hat{\mathbf{f}}_x^\lambda - \mathbf{Z}\tilde{\beta}^\lambda\rangle\right\} \\
&\le \frac{1}{n}\|\hat{\mathbf{f}}_x^\lambda - \mathbf{Z}\tilde{\beta}^\lambda\|_2^2 + \lambda s\|\tilde{\beta}^\lambda\|_2^2 + 2\lambda \quad \text{(by Lemma 24)} \\
&\le 4\lambda \,. \quad \text{(by Lemma 22)}
\end{aligned}$$

For the last inequality, we observe that $\tilde{\beta}^\lambda$ is the solution of the following optimization problem

$$\tilde{\beta}^\lambda = \min_\beta \frac{1}{n}\|\hat{\mathbf{f}}_x^\lambda - \mathbf{Z}\beta\|_2^2 + \lambda s\|\beta\|_2^2 \,.$$

Recall also that we have defined $\hat{f}^\lambda$ and $\hat{\mathbf{f}}_x^\lambda$ in Eq. (1) and Eq. (14), respectively. Now, observe that $\hat{f}^\lambda \in \mathcal{H}$ and in combination with Lemma 22 one obtains the upper bound on Eq. (16).

Finally, we combine the three results and derive

$$\mathbb{E}[l_{\tilde{f}_\beta^\lambda}] - \mathbb{E}[l_{f_\mathcal{H}}] \leq 4\lambda + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) + \mathbb{E}[l_{\hat{f}^\lambda}] - \mathbb{E}[l_{f_\mathcal{H}}]\,. \tag{20}$$

■

6.3 Proof of Theorem 12

To prove Theorem 12, we rely on the notion of local Rademacher complexity instead of the global one. More specifically, the bound in Theorem 9 is not sharp because when analysing Eqs. (15) and (17), we used the global Rademacher complexity that accounts for the whole reproducing kernel Hilbert space. As empirically optimal hypotheses are typically concentrated around $f_\mathcal{H}$, we could just analyse the local space around it. In particular, we can apply Lemma 6 to Eqs. (15) and (17).

To this end, we define the transformed function class as $l_\mathcal{H} := \{(x, y) \mapsto l(f(x), y) \mid f \in \mathcal{H}\}$, for any reproducing kernel Hilbert space $\mathcal{H}$ and a loss function l. We now would like to apply Lemma 6 to the function class $l_\mathcal{H}$. First, it is easy to see that $\mathbb{E}[l_f^2] \leq B\mathbb{E}[l_f]$ for some constant B since l_f is bounded. Now if we assume that there exists a sub-root function $\hat{\psi}_n(r)$ such that it satisfies:

$$\hat{\psi}_n(r) \geq c_1 \hat{R}_n\{l_f \in \operatorname{star}(l_\mathcal{H}, 0) \mid \mathbb{E}_n[l_f^2] \leq r\} + \frac{c_2}{n}\log\frac{1}{\delta}\,,$$

then with high probability, we have

$$\mathbb{E}[l_f] \leq \frac{D}{D-1}\mathbb{E}_n[l_f] + \frac{6D}{B}\hat{r}^* + \frac{c_3}{n}\log\frac{1}{\delta}\,,$$

where r^* is the fixed point of $\hat{\psi}_n(r)$.

Hence, our job now is to find a proper $\hat{\psi}_n(r)$ such that we can compute its fixed point r^*. To this end, we define $\hat{f} = \inf_{f\in\mathcal{H}} \mathbb{E}_n[l_f] = \inf_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2$ for a given training sample $\{(x_i, y_i)\}_{i=1}^n$. We observe that for all $l_f \in l_\mathcal{H}$ it holds that

$$\begin{aligned}
\mathbb{E}_n[l_f^2] &\geq (\mathbb{E}_n[l_f])^2 \quad (x^2 \text{ is convex}) \\
&\geq (\mathbb{E}_n[l_f])^2 - (\mathbb{E}_n[l_{\hat{f}}])^2 \\
&\geq 2\mathbb{E}_n[l_{\hat{f}}]\,\mathbb{E}_n[l_f - l_{\hat{f}}] \quad (\text{since } a^2 - b^2 \geq 2b(a-b), \forall a, b \geq 0) \\
&\geq \frac{2}{B}\mathbb{E}_n[l_{\hat{f}}]\,\mathbb{E}_n[(f-\hat{f})^2]\,. \quad (\text{Lemma 30 in Appendix D})
\end{aligned} \tag{21}$$

Hence, to obtain a lower bound on $\mathbb{E}_n[l_f^2]$ expressed solely in terms of $\mathbb{E}_n[(f-\hat{f})^2]$, we need to find a lower bound of $\mathbb{E}_n[l_{\hat{f}}]$. Since $\mathbb{E}_n[l_{\hat{f}}] = \frac{1}{n}\sum_{i=1}^n (\hat{f}(x_i) - y_i)^2$, we have $\mathbb{E}_P\left[\mathbb{E}_n[l_{\hat{f}}]\right] \geq \mathbb{E}[l_{f_\mathcal{H}}] \geq \sigma^2$, where we recall σ^2 is the variance of ϵ defined in Assumption A.1. In addition, for each labeled example (x_i, y_i), $l(\hat{f}(x_i), y_i)$ is bounded and i.i.d. Applying Hoeffding lemma, we can see that for all $\delta \in (0,1)$, with probability greater than $1-\delta$, $\mathbb{E}_n[l_{\hat{f}}]$ is lower bounded. Denote such a lower bound with some constant e_0. Hence, with probability greater than $1-\delta$, Eq. (21) becomes

$$\mathbb{E}_n[l_f^2] \geq \frac{e_1}{B}\mathbb{E}_n[(f-\hat{f})^2] =: e_2\mathbb{E}_n[(f-\hat{f})^2] \ .$$

As a result of this, we have the following inequality for the two function classes

$$\{l_f \in \mathrm{star}(l_{\mathcal{H}},0) \mid \mathbb{E}_n[l_f^2] \leq r\} \subseteq \{l_f \in \mathrm{star}(l_{\mathcal{H}},0) \mid \mathbb{E}_n[(f-\hat{f})^2] \leq \frac{r}{e_2}\} \ .$$

Recall that for a function class $\mathcal{H}$, we denote its empirical Rademacher complexity by $\hat{R}_n(\mathcal{H})$. Then, we have the following inequality

$$\begin{aligned}
&\hat{R}_n\{l_f \in \mathrm{star}(l_{\mathcal{H}},0) \mid \mathbb{E}_n[l_f^2] \leq r\} \leq \\
&\hat{R}_n\{l_f \in \mathrm{star}(l_{\mathcal{H}},0) \mid \mathbb{E}_n[(f-\hat{f})^2] \leq \frac{r}{e_2}\} = \\
&\hat{R}_n\{l_f - l_{\hat{f}} \mid \mathbb{E}_n[(f-\hat{f})^2] \leq \frac{r}{e_2} \ \wedge \ l_f \in \mathrm{star}(l_{\mathcal{H}},0)\} \leq \\
&L\hat{R}_n\{f-\hat{f} \mid \mathbb{E}_n[(f-\hat{f})^2] \leq \frac{r}{e_2} \ \wedge \ f \in \mathcal{H}\} \leq \\
&L\hat{R}_n\{f-g \mid \mathbb{E}_n[(f-g)^2] \leq \frac{r}{e_2} \ \wedge \ f,g \in \mathcal{H}\} \leq \\
&2L\hat{R}_n\{f \in \mathcal{H} \mid \mathbb{E}_n[f^2] \leq \frac{1}{4}\frac{r}{e_2}\} = \\
&2L\hat{R}_n\{f \in \mathcal{H} \mid \mathbb{E}_n[f^2] \leq e_3 r\} \ ,
\end{aligned} \tag{22}$$

where the last inequality was proved [3] in Bartlett et al. (2005, Corollary 6.7). Now, since $\mathcal{H}$ is a reproducing kernel Hilbert space with kernel k, applying Lemma 7 gives an upper bound of Eq. (22). We can then derive the following lemma which gives us the proper sub-root function $\hat{\psi}_n$. The lemma is proved in Appendix E.

Lemma 25 *Assume $\{x_i,y_i\}_{i=1}^n$ is an independent sample from a probability measure P defined on $\mathcal{X}\times\mathcal{Y}$, where $\mathcal{Y}$ has a bounded range. Let k be a positive definite kernel with the reproducing kernel Hilbert space $\mathcal{H}$ and let $\hat{\lambda}_1 \geq \cdots, \geq \hat{\lambda}_n$ be the eigenvalues of the normalized kernel Gram-matrix. Denote the squared error loss function by $l(f(x),y) = (f(x)-y)^2$ and fix $\delta \in (0,1)$. If*

$$\hat{\psi}_n(r) = 2Lc_1\left(\frac{2}{n}\sum_{i=1}^n \min\{r,\hat{\lambda}_i\}\right)^{1/2} + \frac{c_2}{n}\log(\frac{1}{\delta}) \ ,$$

then for all $l_f \in l_{\mathcal{H}}$ and $D>1$, with probability $1-\delta$,

$$\mathbb{E}[l_f] \leq \frac{D}{D-1}\mathbb{E}_n[l_f] + \frac{6D}{B}\hat{r}^* + \frac{c_3}{n}\log(\frac{1}{\delta}) \ .$$

Moreover, the fixed point $\hat{r}^*$ defined with $\hat{r}^* = \hat{\psi}_n(\hat{r}^*)$ can be upper bounded by

$$\hat{r}^* \leq \min_{0\leq h\leq n}\left(e_0\frac{h}{n} + \sqrt{\frac{1}{n}\sum_{i>h}\hat{\lambda}_i}\right) \ ,$$

where e_0 is a constant.

[3] The results follows from the first three lines in the proof of Corollary 6.7.

We are now ready to deliver the proof of Theorem 12.

Theorem 12 *Suppose that Assumption A.1 holds and that the conditions on sampling measure $\tilde{l}$ from Theorem 9 apply to this setting. If*

$$s \;\geq\; 5d_{\tilde{l}} \log \frac{16d_{\mathbf{K}}^{\lambda}}{\delta}$$

then for all $D > 1$ and $\delta \in (0, 1)$, with probability $1 - \delta$, the excess risk of $\tilde{f}_{\hat{\beta}}^{\lambda}$ can be bounded by

$$\mathbb{E}[l_{\tilde{f}_{\hat{\beta}}^{\lambda}}] - \mathbb{E}[l_{f_{\mathcal{H}}}] \;\leq\; \frac{12D}{B} \hat{r}_{\mathcal{H}}^{*} + 4\frac{D}{D-1}\lambda + \mathcal{O}\left(\frac{1}{n}\right) + \mathbb{E}[l_{\hat{f}^{\lambda}}] - \mathbb{E}[l_{f_{\mathcal{H}}}] \,. \tag{9}$$

Furthermore, denoting the eigenvalues of the normalized kernel matrix $(1/n)\mathbf{K}$ with $\{\hat{\lambda}_i\}_{i=1}^{n}$, we have that

$$\hat{r}_{\mathcal{H}}^{*} \leq \min_{0 \leq h \leq n} \left(e_0 \frac{h}{n} + \sqrt{\frac{1}{n} \sum_{i>h} \hat{\lambda}_i} \right) \,, \tag{10}$$

where $B, e_0 > 0$ are some constant and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$.

Proof We decompose $\mathbb{E}[l_{\tilde{f}_{\hat{\beta}}^{\lambda}}]$ with $D > 1$ as follows:

$$\mathbb{E}[l_{\tilde{f}_{\hat{\beta}}^{\lambda}}] - \mathbb{E}[l_{f_{\mathcal{H}}}] \leq \mathbb{E}[l_{\tilde{f}_{\hat{\beta}}^{\lambda}}] - \frac{D}{D-1}\mathbb{E}_n[l_{\tilde{f}_{\hat{\beta}}^{\lambda}}] \tag{23}$$

$$+\frac{D}{D-1}(\mathbb{E}_n[l_{\tilde{f}_{\hat{\beta}}^{\lambda}}] - \mathbb{E}_n[l_{\hat{f}^{\lambda}}]) \tag{24}$$

$$+\frac{D}{D-1}\mathbb{E}_n[l_{\hat{f}^{\lambda}}] - \mathbb{E}[l_{\hat{f}^{\lambda}}] \tag{25}$$

$$+\mathbb{E}[l_{\hat{f}^{\lambda}}] - \mathbb{E}[l_{f_{\mathcal{H}}}] \,. \tag{26}$$

We have already demonstrated that

$$\text{Eq. } (24) \leq 4\frac{D}{D-1}\lambda \,.$$

For Eq. (23), we can derive an upper bound using Lemma 25. Similarly, we can upper bound Eq. (25) by interchaning the positions of empirical and expected risk functions in Lemma 25. The proof of the latter result resembles that of Lemma 25 and is a direct consequence of the second part of Theorem 4.1 in Bartlett et al. (2005). However, note that $\tilde{f}_{\hat{\beta}}^{\lambda}$ and $\hat{f}^{\lambda}$ belong to different reproducing kernel Hilbert spaces. As a result, we have

$$\text{Eq. } (23) \leq \frac{6D}{B}\hat{r}_{\tilde{\mathcal{H}}}^{*} + \mathcal{O}(1/n) \,,$$

$$\text{Eq. } (25) \leq \frac{6D}{B}\hat{r}_{\mathcal{H}}^{*} + \mathcal{O}(1/n) \,.$$

Now, combining these inequalities together we deduce

$$\begin{aligned}\mathbb{E}[l_{\tilde{f}^\lambda_\beta}] - \mathbb{E}[l_{f_\mathcal{H}}] &\leq \frac{6D}{B}\hat{r}^*_{\tilde{\mathcal{H}}} + \frac{6D}{B}\hat{r}^*_{\mathcal{H}} + 4\frac{D}{D-1}\lambda + \mathcal{O}(1/n) \\ &+\mathbb{E}[l_{\hat{f}^\lambda}] - \mathbb{E}[l_{f_\mathcal{H}}] \\ &\leq \frac{12D}{B}\hat{r}^*_{\mathcal{H}} + 4\frac{D}{D-1}\lambda + \mathcal{O}(1/n) \\ &+\mathbb{E}[l_{\hat{f}^\lambda}] - \mathbb{E}[l_{f_\mathcal{H}}] \,.\end{aligned}$$

The last inequality holds because the eigenvalues of the Gram-matrix for the reproducing kernel Hilbert space $\tilde{\mathcal{H}}$ decay faster than the eigenvalues of $\mathcal{H}$. Consequently, we have that $\hat{r}^*_{\tilde{\mathcal{H}}} \leq \hat{r}^*_{\mathcal{H}}$.

Now, Lemma 25 implies that

$$\hat{r}^*_{\mathcal{H}} \leq \min_{0\leq h\leq n}\Big(e_0\frac{h}{n} + \sqrt{\frac{1}{n}\sum_{i>h}\hat{\lambda}_i}\Big) \,. \tag{27}$$

There are two cases that merit a discussion here. First, if the eigenvalues of $\mathbf{K}$ decay exponentially then setting $h = \lceil \log n \rceil$ implies that

$$\hat{r}^*_{\mathcal{H}} \leq O\left(\frac{\log n}{n}\right) \,.$$

Now, according to Caponnetto and De Vito (2007)

$$\mathbb{E}[l_{\hat{f}^\lambda}] - \mathbb{E}[l_{f_\mathcal{H}}] \in O\left(\frac{\log n}{n}\right),$$

and, thus, if we set $\lambda \propto \frac{\log n}{n}$ then the learning risk rate can be upper bounded by

$$\mathbb{E}[l_{\tilde{f}^\lambda_\beta}] - \mathbb{E}[l_{f_\mathcal{H}}] \in O\left(\frac{\log n}{n}\right).$$

On the other hand, if $\mathbf{K}$ has finitely many non-zero eigenvalues (t), then setting $h \geq t$ implies that

$$\hat{r}^*_{\mathcal{H}} \in O\left(\frac{1}{n}\right) \,.$$

Moreover, in this case, $\mathbb{E}[l_{\hat{f}^\lambda}] - \mathbb{E}[l_{f_\mathcal{H}}] \in \mathcal{O}(\frac{1}{n})$ and setting $\lambda \propto \frac{1}{n}$, we deduce that

$$\mathbb{E}[l_{\tilde{f}^\lambda_\beta}] - \mathbb{E}[l_{f_\mathcal{H}}] \leq O\left(\frac{1}{n}\right) \,.$$

■

6.4 Proof of Theorem 15

Theorem 15 *Suppose that Assumption B.1 holds and that the conditions on sampling measure $\tilde{l}$ from Theorem 9 apply to the setting with a Lipschitz continuous loss. If*

$$s \;\geq\; 5 d_{\tilde{l}} \log \frac{16 d_{\mathbf{K}}^{\lambda}}{\delta}$$

then for all $\delta \in (0,1)$, with probability $1-\delta$, the learning risk of g_{β}^{λ} can be upper bounded by

$$\mathbb{E}[l_{g_{\beta}^{\lambda}}] \;\leq\; \mathbb{E}[l_{g_{\mathcal{H}}}] + \sqrt{2\lambda} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right). \tag{11}$$

Proof The proof is similar to Theorem 9. In particular, we decompose the expected learning risk as

$$\begin{aligned} \mathbb{E}[l_{g_{\beta}^{\lambda}}] = \mathbb{E}[l_{g_{\beta}^{\lambda}}] - \mathbb{E}_n[l_{g_{\beta}^{\lambda}}] \quad & (28)\\ +\mathbb{E}_n[l_{g_{\beta}^{\lambda}}] - \mathbb{E}_n[l_{g_{\mathcal{H}}}] \quad & (29)\\ +\mathbb{E}_n[l_{g_{\mathcal{H}}}] - \mathbb{E}[l_{g_{\mathcal{H}}}] \quad & (30)\\ +\mathbb{E}[l_{g_{\mathcal{H}}}] \, . & \end{aligned}$$

Now, (28) and (30) can be upper bounded similar to Theorem 9, through the Rademacher complexity bound from Lemma 3. For (29), we have

$$\begin{aligned} & \mathbb{E}_n[l_{g_{\beta}^{\lambda}}] - \mathbb{E}_n[l_{g_{\mathcal{H}}}] = \\ & \frac{1}{n}\sum_{i=1}^{n} l(y_i, g_{\beta}^{\lambda}(x_i)) - \frac{1}{n}\sum_{i=1}^{n} l(y_i, g_{\mathcal{H}}(x_i)) = \\ & \frac{1}{n}\inf_{\|g_\beta\|}\sum_{i=1}^{n} l(y_i, g_{\beta}(x_i)) - \frac{1}{n}\sum_{i=1}^{n} l(y_i, g_{\mathcal{H}}(x_i)) \\ & \leq \inf_{\|g_\beta\|}\frac{1}{n}\sum_{i=1}^{n} |g_\beta(x_i) - g_{\mathcal{H}}(x_i)| \\ & \leq \inf_{\|g_\beta\|}\sqrt{\frac{1}{n}\sum_{i=1}^{n} |g_\beta(x_i) - g_{\mathcal{H}}(x_i)|^2} \\ & \leq \sup_{\|g\|}\inf_{\|g_\beta\|}\sqrt{\frac{1}{n}\|g - g_\beta\|_2^2} \\ & \leq \sqrt{2\lambda}\, . \end{aligned}$$

■

6.5 Proof of Theorem 19

To prove Theorem 19, we adopt a similar strategy to the proof of Theorem 12. In particular, we utilize the properties of the local Rademacher complexity by applying Lemma 6 to the decomposition of the learning risk in the Lipschitiz continuous loss case, namely Eqs. (28) and (30). In order to do that, we need two steps. The first step is to find a proper sub-root function $\hat{\psi}_n(r)$. The second step is to find the fixed point of $\hat{\psi}_n(r)$. Hence, the following is devoted to solving these two problems.

Theorem 19 *Suppose that Assumptions B.1-2 hold and that the conditions on sampling measure $\tilde{l}$ from Theorem 9 apply to the setting with a Lipschitz continuous loss. If*

$$s \;\geq\; 5d_{\tilde{l}} \log \frac{16 d_{\mathbf{K}}^{\lambda}}{\delta}$$

then for all $D > 1$ and $\delta \in (0,1)$ with probability greater than $1-\delta$, we have

$$\mathbb{E}[l_{g_{\beta}^{\lambda}}] \leq \frac{12D}{B}\hat{r}_{\mathcal{H}}^{*} + \frac{D}{D-1}\sqrt{2\lambda} + \mathcal{O}(1/n) + \mathbb{E}[l_{g_{\mathcal{H}}}] \,. \qquad (12)$$

Furthermore, denoting the eigenvalues of the normalized kernel matrix $(1/n)\mathbf{K}$ with $\{\hat{\lambda}_i\}_{i=1}^{n}$, we have that $\hat{r}_{\mathcal{H}}^{*}$ can be upper bounded by

$$\hat{r}_{\mathcal{H}}^{*} \leq \min_{0\leq h\leq n}\left(b_0\frac{h}{n} + \sqrt{\frac{1}{n}\sum_{i>h}\hat{\lambda}_i}\right) \,, \qquad (13)$$

where B and b_0 are some constants.

Proof First, recall that we have defined the transformed function class as $g_{\mathcal{H}} := \{(x,y) \mapsto g(f(x),y) \mid g \in \mathcal{H}\}$ for a Lipschitz continuous loss function l and $\hat{g} = \inf_{g\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} l(g(x_i),y_i)$. We observe that for any $l_g \in l_{\mathcal{H}}$, we have

$$\begin{aligned}
\mathbb{E}_n[l_g^2] &\geq (\mathbb{E}_n[l_g])^2 \;\; (x^2 \text{ is convex})\\
&\geq (\mathbb{E}_n[l_g])^2 - (\mathbb{E}_n[l_{\hat{g}}])^2\\
&\geq 2\mathbb{E}_n[l_{\hat{g}}]\,\mathbb{E}_n[l_g - l_{\hat{g}}] \;\; (\text{since } a^2 - b^2 \geq 2b(a-b), \forall a,b \geq 0)\\
&\geq \frac{2}{B}\mathbb{E}_n[l_{\hat{g}}]\mathbb{E}_n[(g-\hat{g})^2] \,. \;\; (\text{Assumption B.4}) \qquad (31)
\end{aligned}$$

Following the reasoning for Eq. (21) in Section 6.3, we can see that for all $\delta \in (0,1)$, with probability greater than $1-\delta$, the term $\mathbb{E}_n[l_{\hat{g}}]$ can be lower bounded by a constant, denoted with b_0. Hence, Eq. (31) becomes

$$\mathbb{E}_n[l_g^2] \geq b_1\mathbb{E}_n[(g-\hat{g})^2] \,.$$

Similar to Section 6.3, we have

$$\{l_g \in l_{\mathcal{H}} \mid \mathbb{E}_n[l_g^2] \leq r\} \subseteq \{l_g \in l_{\mathcal{H}} \mid \mathbb{E}_n[(g-\hat{g})^2] \leq \frac{r}{b_1}\} \,.$$

This further implies that

$$\hat{R}_n\{l_g \in l_{\mathcal{H}} \mid \mathbb{E}_n[l_g^2] \leq r\} \leq 2L\hat{R}_n\{g \in \mathcal{H} \mid \mathbb{E}_n[g^2] \leq b_2 r\} \,,$$

where we recall L is the Lipschitz constant of the loss function l. By appealing to Lemma 7, we obtain an upper bound on $\hat{R}_n\{g \in \mathcal{H} \mid \mathbb{E}_n[g^2] \leq b_2 r\}$. Applying Lemma 25 to the function class $l_{\mathcal{H}}$, we have that for all $l_g \in l_{\mathcal{H}}$ and $D > 1$, with probability greater than $1-\delta$,

$$\mathbb{E}[l_g] \leq \frac{D}{D-1}\mathbb{E}_n[l_g] + \frac{12D}{B}\hat{r}^* + \frac{c_1}{n}\log(\frac{1}{\delta}) \ . \tag{32}$$

Moreover, the fixed point $\hat{r}^*$ can be upper bounded by

$$\hat{r}^* \leq \min_{0\leq h\leq n}\left(b_0\frac{h}{n} + \sqrt{\frac{1}{n}\sum_{i>h}\hat{\lambda}_i}\right) ,$$

where b_0 is a constant.

Having this in mind, we turn to the risk decomposition:

$$\begin{aligned}\mathbb{E}[l_{g_\beta^\lambda}] = {} & \mathbb{E}[l_{g_\beta^\lambda}] - \frac{D}{D-1}\mathbb{E}_n[l_{g_\beta^\lambda}] & (33)\\ & + \frac{D}{D-1}\mathbb{E}_n[l_{g_\beta^\lambda}] - \frac{D}{D-1}\mathbb{E}_n[l_{g_{\mathcal{H}}}] & (34)\\ & + \frac{D}{D-1}\mathbb{E}_n[l_{g_{\mathcal{H}}}] - \mathbb{E}[l_{g_{\mathcal{H}}}] & (35)\\ & + \mathbb{E}[l_{g_{\mathcal{H}}}] \ . \end{aligned}$$

The term in Eq. (34) can be upper bounded by $\frac{D}{D-1}\sqrt{2\lambda}$ using the results from Section 6.4. As $g_\beta^\lambda \in \tilde{\mathcal{H}}$, we can upper bound Eq. (33) using the result from Eq. (32) adjusted to the reproducing kernel Hilbert space $\tilde{\mathcal{H}}$. We repeat the same procedure for Eq. (35) using Eq. (32) applied to $\mathcal{H}$. Now, combing these three auxiliary results, we obtain that with probability greater than $1-\delta$,

$$\mathbb{E}[l_{g_\beta^\lambda}] \leq \frac{12D}{B_0}\hat{r}^*_{\mathcal{H}} + \frac{D}{D-1}\sqrt{2\lambda} + \mathcal{O}(1/n) + \mathbb{E}[l_{g_{\mathcal{H}}}] \ , \tag{36}$$

where $\hat{r}^*_{\mathcal{H}}$ can be upper bounded by

$$\hat{r}^*_{\mathcal{H}} \leq \min_{0\leq h\leq n}\left(b_0\frac{h}{n} + \sqrt{\frac{1}{n}\sum_{i>h}\hat{\lambda}_i}\right) \ . \tag{37}$$

■

6.6 Proof of Theorem 20

Theorem 20 *Suppose that Assumption A.1 holds and consider the regression problem defined with a shift-invariant kernel k, a sample of examples $\{(x_i, y_i)\}_{i=1}^n$, and a regularization parameter λ. Let s be the number of random Fourier features in the pool of features from Algorithm 1, sampled using the spectral measure $p(\cdot)$ from Eq. (3) and the regularization parameter λ. Denote with $\tilde{f}_m^{\lambda^*}$ the ridge regression estimator obtained using a regularization parameter λ^* and a set of random Fourier features $\{v_i\}_{i=1}^m$ returned by Algorithm 1. If*

$$s \ \geq \ \frac{7z_0^2}{\lambda}\log\frac{(16d_{\mathbf{K}}^{\lambda})}{\delta} \quad \textit{and} \quad m \ \geq \ 5d_{\mathbf{K}}^{\lambda^*}\log\frac{(16d_{\mathbf{K}}^{\lambda^*})}{\delta} \ ,$$

then for all $\delta \in (0,1)$, with probability $1-\delta$, the learning risk of $\tilde{f}_m^{\lambda^*}$ can be upper bounded by

$$\mathbb{E}[l_{\tilde{f}_m^{\lambda^*}}] \;\leq\; \mathbb{E}[l_{f_\mathcal{H}}] + 4\lambda + 4\lambda^* + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

Moreover, this upper bound holds for $m \in \Omega(\frac{s}{n\lambda})$.

Proof Suppose the examples $\{x_i, y_i\}_{i=1}^n$ are independent and identically distributed and that the kernel k can be decomposed as in Eq. (3). Let $\{v_i\}_{i=1}^s$ be an independent sample selected according to $p(v)$. Then, using these s features we can approximate the kernel as

$$\begin{aligned}\tilde{k}(x,y) &= \frac{1}{s}\sum_{i=1}^{s} z(v_i,x)z(v_i,y) \\ &= \int_V z(v,x)z(v,y)d\hat{P}(v)\ ,\end{aligned} \tag{38}$$

where $\hat{P}$ is the empirical measure on $\{v_i\}_{i=1}^s$. Denote the reproducing kernel Hilbert space associated with kernel $\tilde{k}$ by $\tilde{\mathcal{H}}$ and suppose that kernel ridge regression was performed with the approximate kernel $\tilde{k}$. From Theorem 9 and Corollary 11, it follows that if

$$s \geq \frac{7z_0^2}{\lambda}\log\frac{16d_{\mathbf{K}}^{\lambda}}{\delta}\ ,$$

then for all $\delta \in (0,1)$, with probability $1-\delta$, the risk convergence rate of the kernel ridge regression estimator based on random Fourier features can be upper bounded by

$$\mathbb{E}[l_{f_\alpha^\lambda}] \leq 4\lambda + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) + \mathbb{E}[l_{f_\mathcal{H}}]\ . \tag{39}$$

Note that in Eq. (39) we have used the fact that $\mathbb{E}[l_{f_\mathcal{H}}]$ differs with $\mathbb{E}[l_{\hat{f}^\lambda}]$ by at most $\mathcal{O}(1/\sqrt{n})$. Let $f_{\tilde{\mathcal{H}}}$ be the function in the reproducing kernel Hilbert space $\tilde{\mathcal{H}}$ achieving the minimal risk, i.e., $\mathbb{E}[l_{f_{\tilde{\mathcal{H}}}}] = \inf_{f\in\tilde{\mathcal{H}}}\mathbb{E}[l_f]$. We now treat $\tilde{k}$ as the actual kernel that can be decomposed via the expectation with respect to the empirical measure in Eq. (38) and re-sample features from the set $\{v_i\}_{i=1}^s$, but this time the sampling is performed using the optimal ridge leverage scores. As $\tilde{k}$ is the actual kernel, it follows from Eq. (6) that the leverage function in this case can be defined by

$$l_\lambda(v) = p(v)\mathbf{z}_v(\mathbf{x})^T(\tilde{\mathbf{K}} + n\lambda I)^{-1}\mathbf{z}_v(\mathbf{x})\ .$$

Now, observe that

$$l_\lambda(v_i) = p(v_i)[\mathbf{Z}_s^T(\tilde{\mathbf{K}} + n\lambda I)^{-1}\mathbf{Z}_s]_{ii}$$

where $[A]_{ii}$ denotes the i-th diagonal element of matrix A. As $\tilde{\mathbf{K}} = (1/s)\mathbf{Z}_s\mathbf{Z}_s^T$, then the Woodbury inversion lemma implies that

$$l_\lambda(v_i) = p(v_i)[\mathbf{Z}_s^T\mathbf{Z}_s(\frac{1}{s}\mathbf{Z}_s^T\mathbf{Z}_s + n\lambda I)^{-1}]_{ii}\ .$$

If we let $l_\lambda(v_i) = p_i$, then the optimal distribution for $\{v_i\}_{i=1}^s$ is multinomial with individual probabilities $q(v_i) = p_i/(\sum_{j=1}^s p_j)$. Hence, we can re-sample m features according to $q(v)$ and perform linear ridge regression using the sampled leverage weighted features. Denoting this estimator with $\tilde{f}_m^{\lambda^*}$ and the corresponding number of degrees of freedom with $d_{\tilde{\mathbf{K}}}^{\lambda} = \mathrm{Tr}\tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda)^{-1}$, we deduce (using Theorem 9 and Corollary 10)

$$\mathbb{E}[l_{\tilde{f}_m^{\lambda^*}}] \leq 4\lambda^* + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) + \mathbb{E}[l_{f_{\tilde{\mathcal{H}}}}] \,, \tag{40}$$

with the number of features $l \propto d_{\tilde{\mathbf{K}}}^{\lambda}$, and we again used the fact that $\mathbb{E}[l_{f_{\tilde{\mathcal{H}}}}]$ differs from $\mathbb{E}[l_{\tilde{f}_m^{\lambda^*}}]$ by at most $\mathcal{O}(1/\sqrt{n})$.

As $f_{\tilde{\mathcal{H}}}$ is the function achieving the minimal risk over $\tilde{\mathcal{H}}$, we can conclude that $\mathbb{E}[l_{f_{\tilde{\mathcal{H}}}}] \leq \mathbb{E}[l_{f_{\tilde{\alpha}}^{\lambda}}]$. Now, combining Eqs. (39) and (40), we obtain the final bound on $\mathbb{E}[l_{\tilde{f}_m^{\lambda^*}}]$. ■

Conclusion

We have investigated the generalization properties of learning with random Fourier features in the context of different kernel methods: kernel ridge regression, support vector machines, and kernel logistic regression. In particular, we have given generic bounds on the number of features required for consistency of learning with two sampling strategies: *leverage weighted* and *plain random Fourier features*. The derived convergence rates account for the complexity of the target hypothesis and the structure of the reproducing kernel Hilbert space with respect to the marginal distribution of a data-generating process. In addition to this, we have also proposed an algorithm for fast approximation of empirical leverage scores and demonstrated its superiority in both theoretical and empirical analyses.

For kernel ridge regression, Avron et al. (2017) and Rudi and Rosasco (2017) have extensively analyzed the performance of learning with random Fourier features. In particular, Avron et al. (2017) have shown that $o(n)$ features are enough to guarantee a good estimator in terms of its *empirical risk*. The authors of that work have also proposed a modified data-dependent sampling distribution and demonstrated that a further reduction in the number of random Fourier features is possible for leverage weighted sampling. However, their results do not provide a convergence rate for the *learning risk* of the estimator which could still potentially imply that computational savings come at the expense of statistical efficiency. Furthermore, the modified sampling distribution can only be used in the 1D Gaussian kernel case. While Avron et al. (2017) focus on bounding the empirical risk of an estimator, Rudi and Rosasco (2017) give a comprehensive study of the generalization properties of random Fourier features for kernel ridge regression by bounding the learning risk of an estimator. The latter work for the first time shows that $\Omega(\sqrt{n}\log n)$ features are sufficient to guarantee the (kernel ridge regression) minimax rate and observes that further improvements to this result are possible by relying on a data-dependent sampling strategy. However, such a distribution is defined in a complicated way and it is not clear how one could devise a practical algorithm by sampling from it. While in our analysis of learning with random Fourier features we also bound the learning risk of an estimator, the analysis is not restricted to kernel ridge regression and covers other kernel methods such as support vector machines and kernel logistic regression. In addition to this, our derivations are much simpler compared to Rudi and Rosasco (2017) and provide sharper bounds in some cases. More specifically, we have demonstrated that $\Omega(\sqrt{n}\log\log n)$ features are sufficient to attain the minimax rate in the case where eigenvalues of the Gram matrix have a geometric/exponential decay. In other cases, we have recovered the results from Rudi and Rosasco (2017). Another important difference with respect to this work is that we consider a data-dependent sampling distribution based on empirical ridge leverage scores, showing that it can further reduce the number of features and in this way provide a more effective estimator.

In addition to the squared error loss, we also investigate the properties of learning with random Fourier features using the Lipschitz continuous loss functions. Both Rahimi and Recht (2009) and Bach (2017b) have studied this problem setting and obtained that $\Omega(n)$ features are needed to ensure $\mathcal{O}(1/\sqrt{n})$ learning risk convergence rate. Moreover, Bach (2017b) has defined an optimal sampling distribution by referring to the leverage score function based on the integral operator and shown that the number of features can be significantly reduced when the eigenvalues of a Gram matrix exhibit a fast decay. The $\Omega(n)$ requirement on the number of features is too restrictive and precludes any computational savings. Also, the optimal sampling distribution is typically intractable. In our analysis, through assuming the realizable case, we have demonstrated that for the first time, $\mathcal{O}(\sqrt{n})$ features are possible to guarantee $\mathcal{O}(\frac{1}{\sqrt{n}})$ risk convergence rate. In extreme cases, where the complexity of target function is small, constant features is enough to guarantee fast risk convergence. Moreover, we also provide a much simpler form of the empirical leverage score distribution and demonstrate that the number of features can be significantly smaller than n, without incurring any loss of statistical efficiency.

Having given risk convergence rates for learning with random Fourier features, we provide a fast and practical algorithm for sampling them in a data-dependent way, such that they approximate the ridge leverage score distribution. In the kernel ridge regression setting, our theoretical analysis demonstrates that, compared to spectral measure sampling, significant computational savings can be achieved while preserving the statistical properties of the estimators. Furthermore, we verify our findings empirically on simulated and real-world datasets. An interesting extension of our empirical analysis would be a thorough and comprehensive comparison of the proposed leverage weighted sampling scheme to other recently proposed data-dependent strategies for selecting good features (e.g., Rudi et al., 2018), as well as a comparison to the Nyström method.

Acknowledgments: We would like to thank Fadhel Ayed, Qinyi Zhang, and Anthony Caterini for fruitful discussions on some of the results as well as for proofreading of this paper. We would also like to extend our thanks to the anonymous reviewers for their insightful comments and suggestions, which have helped us to improve the quality of the paper and its presentation. This work was supported by the EPSRC and MRC through the OxWaSP CDT programme (EP/L016710/1). Dino Oglic was supported in part by EPSRC grant EP/R012067/1. Zhu Li was supported in part by Huawei UK.

References

Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning*, pages 253–262, 2017.

Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017a.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017b.

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

Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Salomon Bochner. Vorlesungen über Fouriersche Integrale. In *Akademische Verlagsgesellschaft*, 1932.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Trevor J Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random fourier features. In *International Conference on Machine Learning*, pages 3905–3914. PMLR, 2019.

Michael W Mahoney and Petros Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

Shahar Mendelson. Improving the sample complexity using global data. *IEEE transactions on Information Theory*, 48(7):1977–1991, 2002.

Evert J. Nyström. Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Mathematica*, 1930.

Dino Oglic and Thomas Gärtner. Greedy feature construction. In *Advances in Neural Information Processing Systems 29*, pages 3945–3953, 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2007.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pages 1313–1320, 2009.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3218–3228, 2017.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3891–3901, 2017.

Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco. On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems*, pages 5672–5682, 2018.

Walter Rudin. *Fourier analysis on groups*. Courier Dover Publications, 2017.

Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press, 2001.

Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Kernel methods in computational biology*. MIT press, 2004.

Alexander J. Smola and Bernhard Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

Bharath Sriperumbudur and Zoltán Szabó. Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems*, pages 1144–1152, 2015.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Yitong Sun, Anna Gilbert, and Ambuj Tewari. But how does it work in theory? linear svm with random features. In *Advances in Neural Information Processing Systems*, pages 3379–3388, 2018.

Dougal J Sutherland and Jeff Schneider. On the error of random Fourier features. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 862–871. AUAI Press, 2015.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Alexander B Tsybakov et al. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, 2001.

Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random Fourier features: A theoretical and empirical comparison. In *Advances in neural information processing systems*, pages 476–484, 2012.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16 (1):3299–3340, 2015.

Appendix A. Bernstein Inequality

The next lemma is the matrix Bernstein inequality, cited from (Avron et al., 2017, Lemma 27) which is a restatement of Corollary 7.3.3 in Tropp (2015) with some fix in the typos.

Lemma 26 *(Bernstein inequality, Tropp, 2015, Corollary 7.3.3) Let* $\mathbf{R}$ *be a fixed* $d_1 \times d_2$ *matrix over the set of complex/real numbers. Suppose that* $\{\mathbf{R}_1, \cdots, \mathbf{R}_n\}$ *is an independent and identically distributed sample of* $d_1 \times d_2$ *matrices such that*

$$\mathbb{E}[\mathbf{R}_i] = \mathbf{R} \qquad \textit{and} \qquad \|\mathbf{R}_i\|_2 \leq L \, ,$$

where $L > 0$ *is a constant independent of the sample. Furthermore, let* $\mathbf{M}_1, \mathbf{M}_2$ *be semidefinite upper bounds for the matrix-valued variances*

$$\begin{aligned} \mathrm{Var}_1[\mathbf{R}_i] &\preceq \mathbb{E}[\mathbf{R}_i \mathbf{R}_i^T] \preceq \mathbf{M}_1 \\ \mathrm{Var}_2[\mathbf{R}_i] &\preceq \mathbb{E}[\mathbf{R}_i^T \mathbf{R}_i] \preceq \mathbf{M}_2 \, . \end{aligned}$$

Let $m = \max(\|\mathbf{M}_1\|_2, \|\mathbf{M}_2\|_2)$ *and* $d = \frac{Tr(\mathbf{M}_1) + Tr(\mathbf{M}_2)}{m}$*. Then, for* $\epsilon \geq \sqrt{m/n} + 2L/3n$*, we can bound*

$$\bar{\mathbf{R}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{R}_i$$

around its mean using the concentration inequality

$$P(\|\bar{\mathbf{R}}_n - \mathbf{R}\|_2 \geq \epsilon) \leq 4d \exp\left(\frac{-n\epsilon^2/2}{m + 2L\epsilon/3} \right) .$$

Appendix B. Proof of Lemma 22

The following two lemmas are required for our proof of Lemma 22, presented subsequently.

Lemma 27 *Suppose that the assumptions from Lemma 22 hold and let* $\epsilon \geq \sqrt{\frac{m}{s}} + \frac{2L}{3s}$ *with constants* m *and* L *(see the proof for explicit definition). If the number of features*

$$s \geq d_{\tilde{l}}(\frac{1}{\epsilon^2} + \frac{2}{3\epsilon}) \log \frac{16 d_{\mathbf{K}}^{\lambda}}{\delta} \, ,$$

then for all $\delta \in (0, 1)$*, with probability greater than* $1 - \delta$*,*

$$-\epsilon \mathbf{I} \preceq (\mathbf{K} + n\lambda \mathbf{I})^{-\frac{1}{2}} (\tilde{\mathbf{K}} - \mathbf{K}) (\mathbf{K} + n\lambda \mathbf{I})^{-\frac{1}{2}} \preceq \epsilon \mathbf{I} \, .$$

Proof Following the derivations in Avron et al. (2017), we utilize the matrix Bernstein concentration inequality to prove the result. More specifically, we observe that

$$\begin{aligned} &(\mathbf{K} + n\lambda \mathbf{I})^{-\frac{1}{2}} \tilde{\mathbf{K}} (\mathbf{K} + n\lambda \mathbf{I})^{-\frac{1}{2}} = \\ &\frac{1}{s} \sum_{i=1}^{s} (\mathbf{K} + n\lambda \mathbf{I})^{-\frac{1}{2}} \mathbf{z}_{q,v_i}(\mathbf{x}) \mathbf{z}_{q,v_i}(\mathbf{x})^T (\mathbf{K} + n\lambda \mathbf{I})^{-\frac{1}{2}} = \\ &\frac{1}{s} \sum_{i=1}^{s} \mathbf{R}_i =: \bar{\mathbf{R}}_s \, , \end{aligned}$$

with

$$\mathbf{R}_i = (\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{q,v_i}(\mathbf{x})\mathbf{z}_{q,v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} .$$

Now, observe that

$$\mathbf{R} = \mathbb{E}[\mathbf{R}_i] = (\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{K}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} .$$

The operator norm of $\mathbf{R}_i$ is equal to

$$\|(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{q,v_i}(\mathbf{x})\mathbf{z}_{q,v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\|_2 .$$

As $\mathbf{z}_{q,v_i}(\mathbf{x})\mathbf{z}_{q,v_i}(\mathbf{x})^T$ is a rank one matrix, we have that the operator norm of this matrix is equal to its trace, i.e.,

$$\begin{aligned}
&\|\mathbf{R}_i\|_2 = \\
&\text{Tr}((\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{q,v_i}(\mathbf{x})\mathbf{z}_{q,v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}) = \\
&\frac{p(v_i)}{q(v_i)}\text{Tr}((\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{v_i}(\mathbf{x})\mathbf{z}_{v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}) = \\
&\frac{p(v_i)}{q(v_i)}\text{Tr}(\mathbf{z}_{v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{z}_{v_i}(\mathbf{x})) = \\
&\frac{l_\lambda(v_i)}{q(v_i)} =: L_i \quad \text{and} \quad L_q \coloneqq \sup_i L_i .
\end{aligned}$$

Observe that $L_q = \sup_i L_i = \sup_i \frac{l_\lambda(v_i)}{q(v_i)} \leq \sup_i \frac{\tilde{l}(v_i)}{q(v_i)} = d_{\tilde{l}}$. On the other hand,

$$\begin{aligned}
&\mathbf{R}_i\mathbf{R}_i^T = \\
&(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{q,v_i}(\mathbf{x})\mathbf{z}_{q,v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{z}_{q,v_i}(\mathbf{x}) \\
&\cdot \mathbf{z}_{q,v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} = \\
&\frac{p(v_i)l_\lambda(v_i)}{q^2(v_i)}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{v_i}(\mathbf{x})\mathbf{z}_{v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} \preceq \\
&\frac{\tilde{l}(v_i)}{q(v_i)}\frac{p(v_i)}{q(v_i)}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{v_i}(\mathbf{x})\mathbf{z}_{v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} = \\
&d_{\tilde{l}}\frac{p(v_i)}{q(v_i)}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{v_i}(\mathbf{x})\mathbf{z}_{v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} .
\end{aligned}$$

From the latter inequality, we obtain that

$$\mathbb{E}[\mathbf{R}_i\mathbf{R}_i^T] \preceq d_{\tilde{l}}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{K}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} =: \mathbf{M}_1 .$$

We also have the following two equalities

$$\begin{aligned}
m &= \|\mathbf{M}_1\|_2 = d_{\tilde{l}}\frac{\lambda_1}{\lambda_1 + n\lambda} =: d_{\tilde{l}}d_1 \\
d &= \frac{2\,\text{Tr}(\mathbf{M}_1)}{m} = 2\frac{\lambda_1 + n\lambda}{\lambda_1}d_{\mathbf{K}}^\lambda = 2d_1^{-1}d_{\mathbf{K}}^\lambda .
\end{aligned}$$

We are now ready to apply the matrix Bernstein concentration inequality (Tropp, 2015, Corollary 7.3.3). More specifically, for $\epsilon \geq \sqrt{m/s} + 2L/3s$ and for all $\delta \in (0, 1)$, with probability $1 - \delta$, we have that

$$\begin{aligned} \mathrm{P}(\|\bar{\mathbf{R}}_s - \mathbf{R}\|_2 \geq \epsilon) &\leq 4d \exp\left(\frac{-s\epsilon^2/2}{m + 2L\epsilon/3}\right) \\ &\leq 8d_1^{-1} d_{\mathbf{K}}^{\lambda} \exp\left(\frac{-s\epsilon^2/2}{d_{\tilde{l}} d_1 + d_{\tilde{l}} 2\epsilon/3}\right) \\ &\leq 16 d_{\mathbf{K}}^{\lambda} \exp\left(\frac{-s\epsilon^2}{d_{\tilde{l}}(1 + 2\epsilon/3)}\right) \leq \delta \; . \end{aligned}$$

In the third line, we have used the assumption that $n\lambda \leq \lambda_1$ and, consequently, $d_1 \in [1/2, 1)$. ■

Remark 28 *We note here that the two considered sampling strategies lead to two different results. In particular, if we let $\tilde{l}(v) = l_\lambda(v)$ then $q(v) = l_\lambda(v)/d_{\mathbf{K}}^{\lambda}$, i.e., we are sampling proportional to the ridge leverage scores. Thus, the leverage weighted random Fourier features sampler requires*

$$s \geq d_{\mathbf{K}}^{\lambda}\left(\frac{1}{\epsilon^2} + \frac{2}{3\epsilon}\right) \log \frac{16 d_{\mathbf{K}}^{\lambda}}{\delta} \; . \tag{41}$$

Alternatively, we can opt for the plain random Fourier feature sampling strategy by taking $\tilde{l}(v) = z_0^2 p(v)/\lambda$, with $l_\lambda(v) \leq z_0^2 p(v)/\lambda$. Then, plain random Fourier features sampling requires

$$s \geq \frac{z_0^2}{\lambda}\left(\frac{1}{\epsilon^2} + \frac{2}{3\epsilon}\right) \log \frac{16 d_{\mathbf{K}}^{\lambda}}{\delta} \; . \tag{42}$$

Thus, the leverage weighted random Fourier features sampling scheme can dramatically change the number of features required to achieve a predefined approximation error in the operator norm.

Lemma 29 *Let $f \in \mathcal{H}$, where $\mathcal{H}$ is the reproducing kernel Hilbert space associated with a kernel k. Recall we have assumed that $\|f\|_{\mathcal{H}} \leq 1, \forall f$ and $\mathbf{f}_x = [f(x_1), \cdots, f(x_n)]^T$. Let $\mathbf{K}$ be the Gram-matrix of the kernel k given by the provided set of instances. Then,*

$$\mathbf{f}_x^T \mathbf{K}^{-1} \mathbf{f}_x \leq 1 \; .$$

Proof Recall that a function $f \in \mathcal{H}$ can be expressed as:

$$f(x) = \int_{\mathcal{V}} g(v) z(v, x) p(v) dv \qquad (\forall x \in \mathcal{X}) \; , \tag{43}$$

where $g \in L_2(d\tau)$ is a real-valued function with $\|f\|_{\mathcal{H}}$ equal to the minimum of $\|g\|_{L_2(d\tau)}$, over all possible decompositions of f. For a vector $\mathbf{a} \in \mathbb{R}^n$, we have that

$$\begin{aligned}
\mathbf{a}^T \mathbf{f}_x \mathbf{f}_x^T \mathbf{a} &= \Big(\mathbf{f}_x^T \mathbf{a}\Big)^2 = \Big(\sum_{i=1}^n a_i f(x_i)\Big)^2 \\
&= \Big(\sum_{i=1}^n a_i \int_{\mathcal{V}} g(v) z(v, x_i) d\tau(v)\Big)^2 \\
&= \Big(\int_{\mathcal{V}} g(v) \mathbf{z}_v(\mathbf{x})^T \mathbf{a} \, d\tau(v)\Big)^2 \\
&\leq \int_{\mathcal{V}} g(v)^2 d\tau(v) \int_{\mathcal{V}} (\mathbf{z}_v(\mathbf{x})^T \mathbf{a})^2 \, d\tau(v) \\
&= \int_{\mathcal{V}} \mathbf{a}^T \mathbf{z}_v(\mathbf{x}) \mathbf{z}_v(\mathbf{x})^T \mathbf{a} \, d\tau(v) \\
&= \mathbf{a}^T \int_{\mathcal{V}} \mathbf{z}_v(\mathbf{x}) \mathbf{z}_v(\mathbf{x})^T \, d\tau(v) \, \mathbf{a} \\
&= \mathbf{a}^T \mathbf{K} \mathbf{a} \, .
\end{aligned}$$

The third equality is due to the fact that, for all $f \in \mathcal{H}$, we have that $f(x) = \int_{\mathcal{V}} g(v) z(v, x) p(v) dv$ ($\forall x \in \mathcal{X}$) and

$$\|f\|_{\mathcal{H}} = \min_{\left\{g \;|\; f(x) = \int_{\mathcal{V}} g(v) z(v,x) p(v) dv\right\}} \|g\|_{L_2(d\tau)} \, .$$

The first inequality, on the other hand, follows from the Cauchy-Schwarz inequality. The bound implies that $\mathbf{f}_x \mathbf{f}_x^T \preceq \mathbf{K}$ and, consequently, we derive $\mathbf{f}_x^T \mathbf{K}^{-1} \mathbf{f}_x \leq 1$. ■

Now we are ready to prove Lemma 22.

Lemma 22 *Suppose that Assumption A.1 holds and that the conditions on sampling measure $\tilde{l}$ from Theorem 9 apply as well. If*

$$s \geq 5 d_{\tilde{l}} \log \frac{16 d_{\mathbf{K}}^{\lambda}}{\delta} \, ,$$

then for all $\delta \in (0, 1)$ and any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$, with probability greater than $1 - \delta$, the following holds

$$\min_{\beta} \left\{ \frac{1}{n} \|\mathbf{f}_x - \mathbf{Z}_q \beta\|_2^2 + \lambda s \|\beta\|_2^2 \right\} \leq 2\lambda \, .$$

For the sake of brevity, we will henceforth use $\tilde{\mathcal{H}}$ to denote the hypothesis space corresponding to this optimization problem. Then, the latter bound can be written as:

$$\sup\nolimits_{\|f\|_{\mathcal{H}} \leq 1} \inf\nolimits_{\|\tilde{f}\|_{\tilde{\mathcal{H}}} \leq \sqrt{2}} \frac{1}{n} \|\mathbf{f}_x - \tilde{\mathbf{f}}_x\|_2^2 \leq 2\lambda \, .$$

Proof For any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$, we write the following optimization problem:

$$\frac{1}{n} \|\mathbf{f}_x - \mathbf{Z}_q \beta\|_2^2 + s\lambda \|\beta\|_2^2 \, . \tag{44}$$

The minimizer can be computed as:

$$\begin{aligned}\beta &= \frac{1}{s}(\frac{1}{s}\mathbf{Z}_q^T\mathbf{Z}_q + n\lambda\mathbf{I})^{-1}\mathbf{Z}_q^T\mathbf{f}_x \\ &= \frac{1}{s}\mathbf{Z}_q^T(\frac{1}{s}\mathbf{Z}_q\mathbf{Z}_q^T + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \\ &= \frac{1}{s}\mathbf{Z}_q^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \ ,\end{aligned}$$

where the second equality follows from the Woodbury inversion lemma.

Substituting β into Eq. (44), we transform the first part as

$$\begin{aligned}\frac{1}{n}\|\mathbf{f}_x - \mathbf{Z}_q\beta\|_2^2 &= \frac{1}{n}\|\mathbf{f}_x - \frac{1}{s}\mathbf{Z}_q\mathbf{Z}_q^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x\|_2^2 \\ &= \frac{1}{n}\|\mathbf{f}_x - \tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x\|_2^2 \\ &= \frac{1}{n}\|n\lambda(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x\|_2^2 \\ &= n\lambda^2\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-2}\mathbf{f}_x \ .\end{aligned}$$

On the other hand, the second part can be transformed as

$$\begin{aligned}s\lambda\|\beta\|_2^2 &= s\lambda\frac{1}{s^2}\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{Z}_q\mathbf{Z}_q^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \\ &= \lambda\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \\ &= \lambda\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}(\tilde{\mathbf{K}} + n\lambda\mathbf{I})(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x - n\lambda^2\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-2}\mathbf{f}_x \\ &= \lambda\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x - n\lambda^2\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-2}\mathbf{f}_x \ .\end{aligned}$$

Now, summing up the first and the second part, we deduce

$$\begin{aligned}&\frac{1}{n}\|\mathbf{f}_x - \mathbf{Z}_q\beta\|_2^2 + s\lambda\|\beta\|_2^2 = \lambda\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \\ &= \lambda\mathbf{f}_x^T(\mathbf{K} + n\lambda\mathbf{I} + \tilde{\mathbf{K}} - \mathbf{K})^{-1}\mathbf{f}_x \\ &= \lambda\mathbf{f}_x^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\left(\mathbf{I} + (\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}(\tilde{\mathbf{K}} - \mathbf{K})(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\right)^{-1}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{f}_x \ .\end{aligned}$$

From Lemma 27, it follows that when

$$s \geq d_{\tilde{l}}(\frac{1}{\epsilon^2} + \frac{2}{3\epsilon})\log\frac{16d_{\mathbf{K}}^{\lambda}}{\delta}$$

then $(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}(\tilde{\mathbf{K}} - \mathbf{K})(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} \succeq -\epsilon\mathbf{I}$.

We can now upper bound the objective function as follows (with $\epsilon = 1/2$):

$$\begin{aligned}\lambda\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \quad &\leq \lambda\mathbf{f}_x^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}(1-\epsilon)^{-1}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{f}_x \\ &= (1-\epsilon)^{-1}\lambda\mathbf{f}_x^T(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \leq (1-\epsilon)^{-1}\lambda\mathbf{f}_x^T\mathbf{K}^{-1}\mathbf{f}_x \leq 2\lambda \ ,\end{aligned}$$

where in the last inequality we have used Lemma 29. Therefore, we prove the first inequality.

For the second claim, we have that

$$\begin{aligned} s\|\beta\|_2^2 &= \mathbf{f}_x^T(\tilde{\mathbf{K}}+n\lambda\mathbf{I})^{-1}\mathbf{f}_x - n\lambda\mathbf{f}_x^T(\tilde{\mathbf{K}}+n\lambda\mathbf{I})^{-2}\mathbf{f}_x \\ &\leq \mathbf{f}_x^T(\tilde{\mathbf{K}}+n\lambda\mathbf{I})^{-1}\mathbf{f}_x \leq (1-\epsilon)^{-1}\mathbf{f}_x^T\mathbf{K}^{-1}\mathbf{f}_x \leq 2 \; . \end{aligned}$$

Hence, the squared norm of our approximated function is bounded by $\|\tilde{f}\|_{\tilde{\mathcal{H}}}^2 \leq s\|\beta\|_2^2 \leq 2$. As such, problem (44) can now be written as $\min_\beta (1/n)\|\mathbf{f}_x - \tilde{\mathbf{f}}_x\|_2^2$ subject to $\|\tilde{f}\|_{\tilde{\mathcal{H}}}^2 \leq s\|\beta\|_2^2 \leq 2$, which is equivalent to

$$\inf_{\|\tilde{f}\|_{\tilde{\mathcal{H}}}\leq\sqrt{2}} \frac{1}{n}\|\mathbf{f}_x - \tilde{\mathbf{f}}_x\|_2^2 \; ,$$

and we have shown that this can be upper bounded by 2λ. Since we are approximating any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$, this can further be written as

$$\sup_{\|f\|_{\mathcal{H}}\leq 1} \inf_{\|\tilde{f}\|_{\tilde{\mathcal{H}}}\leq\sqrt{2}} \frac{1}{n}\|\mathbf{f}_x - \tilde{\mathbf{f}}_x\|_2^2 \leq 2\lambda \; .$$

■

Appendix C. Proof of Lemma 24

Lemma 24 *Denote the in-sample prediction of $\hat{f}^\lambda$ with*

$$\hat{\mathbf{f}}_x^\lambda = [\hat{f}^\lambda(x_1), \ldots, \hat{f}^\lambda(x_n)]^T \tag{14}$$

and let $\{v_i\}_{i=1}^s$ be independent samples selected according to a probability density function $q(v)$, which define the feature matrix $\mathbf{Z}_q$ and the corresponding reproducing kernel Hilbert space $\tilde{\mathcal{H}}$. Let $\tilde{\beta}^\lambda$ be the solution to the following optimization problem

$$\tilde{\beta}^\lambda := \min_\beta \frac{1}{n}\|\hat{\mathbf{f}}_x^\lambda - \mathbf{Z}_q\beta\|_2^2 + \lambda s\|\beta\|_2^2$$

and denote the in-sample prediction of the resulting hypothesis with $\tilde{\mathbf{f}}_x^\lambda = \mathbf{Z}_q\tilde{\beta}^\lambda$. Then, we have

$$\frac{1}{n}\langle Y - \hat{\mathbf{f}}_x^\lambda, \hat{\mathbf{f}}_x^\lambda - \tilde{\mathbf{f}}_x^\lambda\rangle \leq \lambda \; .$$

Proof By definition, $\tilde{\mathbf{f}}_x^\lambda$ has the format as $\tilde{\mathbf{f}}_x^\lambda = \mathbf{Z}_q\tilde{\beta}^\lambda$, where $\tilde{\beta}^\lambda \in \mathbb{R}^s$. In addition, definition of $\tilde{f}^\lambda$ can be reparametrized by the following optimization problem:

$$\tilde{\beta}^\lambda := \min_{\tilde{\beta}} \frac{1}{n}\|\hat{\mathbf{f}}_x^\lambda - \mathbf{Z}_q\tilde{\beta}\|_2^2 + s\lambda\|\tilde{\beta}\| \; . \tag{45}$$

This gives the closed-form solution of $\tilde{\beta}^\lambda = \frac{1}{s}\mathbf{Z}_q^T(\frac{1}{s}\mathbf{Z}_q\mathbf{Z}_q^T + n\lambda\mathbf{I})^{-1}\hat{\mathbf{f}}_x^\lambda$. As a result, we have

$$\tilde{\mathbf{f}}_x^\lambda = \frac{1}{s}\mathbf{Z}_q\mathbf{Z}_q^T(\frac{1}{s}\mathbf{Z}_q\mathbf{Z}_q^T + n\lambda\mathbf{I})^{-1}\hat{\mathbf{f}}_x^\lambda = \tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda I)^{-1}\hat{\mathbf{f}}_x^\lambda \; .$$

Now recall $\hat{\mathbf{f}}_x^\lambda$ is the in-sample prediction of the KRR estimator $\hat{f}^\lambda$, so it can be written as $\hat{\mathbf{f}}_x^\lambda = \mathbf{K}(\mathbf{K}+n\lambda I)^{-1}Y$. As a result, we have the following:

$$\begin{aligned}\frac{1}{n}\langle Y-\hat{\mathbf{f}}_x^\lambda, \hat{\mathbf{f}}_x^\lambda - \tilde{\mathbf{f}}_x^\lambda\rangle &= \frac{1}{n}\langle Y-\hat{\mathbf{f}}_x^\lambda, \hat{\mathbf{f}}_x^\lambda - \tilde{\mathbf{K}}(\tilde{\mathbf{K}}+n\lambda I)^{-1}\hat{\mathbf{f}}_x^\lambda\rangle \\ &= \frac{1}{n}\langle Y-\mathbf{K}(\mathbf{K}+n\lambda I)^{-1}Y, (I-\tilde{\mathbf{K}}(\tilde{\mathbf{K}}+n\lambda I)^{-1})\hat{\mathbf{f}}_x^\lambda\rangle \\ &= \frac{1}{n}Y^T(I-\mathbf{K}(\mathbf{K}+n\lambda I)^{-1})(I-\tilde{\mathbf{K}}(\tilde{\mathbf{K}}+n\lambda I)^{-1})\hat{\mathbf{f}}_x^\lambda \\ &\leq \frac{1}{n}Y^T(I-\mathbf{K}(\mathbf{K}+n\lambda I)^{-1})\hat{\mathbf{f}}_x^\lambda \qquad (46) \\ &= \lambda Y^T(\mathbf{K}+n\lambda I)^{-1}\hat{\mathbf{f}}_x^\lambda \\ &= \lambda Y^T(\mathbf{K}+n\lambda I)^{-1}\mathbf{K}\mathbf{K}^{-1}\hat{\mathbf{f}}_x^\lambda \\ &= \lambda \hat{\mathbf{f}}_x^{\lambda T}\mathbf{K}^{-1}\hat{\mathbf{f}}_x^\lambda \leq \lambda\ .\end{aligned}$$

Note that in Eq. (46), we have used the fact that

$$\|I-\tilde{\mathbf{K}}(\tilde{\mathbf{K}}+n\lambda I)^{-1}\|_2 \leq 1\ .$$

For the last inequality, since $\hat{f}^\lambda \in \mathcal{H}$, we employ Lemma 29. ■

Appendix D. Property of Squared Error Loss

In this section, we state the property of square loss function.

Lemma 30 *(Bartlett et al., 2005, Section 5.2) Let l be the squared error loss function and $\mathcal{H}$ a convex and uniformly bounded hypothesis space. Assume that for every probability distribution P in a class of data-generating distributions, there is an $f_\mathcal{H} \in \mathcal{H}$ such that $\mathbb{E}[l_{f_\mathcal{H}}] = \inf_{f\in\mathcal{H}} \mathbb{E}[l_f]$. Then, there exists a constant $B \geq 1$ such that for all $f \in \mathcal{H}$ and for every probability distribution P*

$$\mathbb{E}[(f-f_\mathcal{H})^2] \leq B\mathbb{E}[l_f - l_{f_\mathcal{H}}]\ . \qquad (47)$$

Appendix E. Proof of Lemma 25

Lemma 25 *Assume $\{x_i, y_i\}_{i=1}^n$ is an independent sample from a probability measure P defined on $\mathcal{X}\times\mathcal{Y}$, where $\mathcal{Y}$ has a bounded range. Let k be a positive definite kernel with the reproducing kernel Hilbert space $\mathcal{H}$ and let $\hat{\lambda}_1 \geq \cdots, \geq \hat{\lambda}_n$ be the eigenvalues of the normalized kernel Gram-matrix. Denote the squared error loss function by $l(f(x), y) = (f(x)-y)^2$ and fix $\delta \in (0,1)$. If*

$$\hat{\psi}_n(r) = 2Lc_1\left(\frac{2}{n}\sum_{i=1}^n \min\{r, \hat{\lambda}_i\}\right)^{1/2} + \frac{c_2}{n}\log(\frac{1}{\delta})\ ,$$

then for all $l_f \in l_\mathcal{H}$ and $D > 1$, with probability $1-\delta$,

$$\mathbb{E}[l_f] \leq \frac{D}{D-1}\mathbb{E}_n[l_f] + \frac{6D}{B}\hat{r}^* + \frac{c_3}{n}\log(\frac{1}{\delta})\ .$$

Moreover, the fixed point $\hat{r}^*$ defined with $\hat{r}^* = \hat{\psi}_n(\hat{r}^*)$ can be upper bounded by

$$\hat{r}^* \leq \min_{0 \leq h \leq n} \left(e_0 \frac{h}{n} + \sqrt{\frac{1}{n} \sum_{i>h} \hat{\lambda}_i} \right) ,$$

where e_0 is a constant.

Proof It is easy to see that $\mathbb{E}[l_f^2] \leq \mathbb{E}[l_f]$. Hence, we can apply Lemma 6 to function class $l_{\mathcal{H}}$ and obtain that for all $l_f \in l_{\mathcal{H}}$

$$\mathbb{E}[l_f] \leq \frac{D}{D-1} \mathbb{E}_n[l_f] + \frac{6D}{B} \hat{r}^* + \frac{c_3}{n} \log(\frac{1}{\delta}) ,$$

as long as there is a sub-root function $\hat{\psi}_n(r)$ such that

$$\hat{\psi}_n(r) \geq c_1 \hat{R}_n \{ l_f \in star(l_{\mathcal{H}}, 0) \mid \mathbb{E}_n[l_f^2] \leq r \} + \frac{c_2}{n} \log(\frac{1}{\delta}) . \tag{48}$$

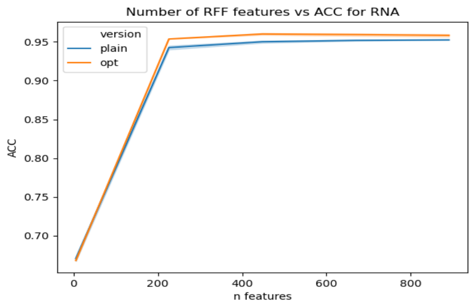
We have previously demonstrated that

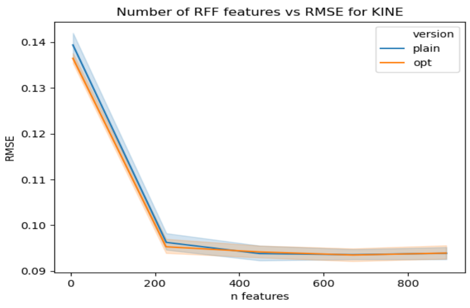
$$\begin{aligned} & c_1 \hat{R}_n \{ l_f \in star(l_{\mathcal{H}}, 0) \mid \mathbb{E}_n[l_f^2] \leq r \} + \frac{c_2}{n} \log(\frac{1}{\delta}) \\ \leq \; & 2 c_1 L \hat{R}_n \left\{ f \in \mathcal{H} \mid \mathbb{E}_n[f^2] \leq e_1 r \right\} + \frac{c_2}{n} \log(\frac{1}{\delta}) \\ \leq \; & 2 c_1 L \left(\frac{2}{n} \sum_{i=1}^{n} \min \left\{ e_1 r, \hat{\lambda}_i \right\} \right)^{1/2} + \frac{c_2}{n} \log(\frac{1}{\delta}) . \quad \text{(by Lemma 7)} \end{aligned} \tag{49}$$

Hence, if we choose $\hat{\psi}_n(r)$ to be equal to the right hand side of Eq. (49), then $\hat{\psi}_n(r)$ is a sub-root function that satisfies Eq. (48). Now, the upper bound on the fixed point $\hat{r}^*$ follows from Corollary 6.7 in Bartlett et al. (2005). ■

Appendix F. Additional Experiments with more features

We have also added extra experiments where we use more features for the experiments that have not yet converged i.e. `KINEMATICS` and `COD-RNA`. In the below we see that only when we increase the number of features up to 1000 we are able to attain comparable performance.


Number of RFF features vs ACC for RNA
version
plain
opt
ACC
n features


Number of RFF features vs RMSE for KINE
version
plain
opt
RMSE
n features