Towards a Unified Analysis of Random Fourier Features

Zhu Li [1] **Jean-François Ton** [1] **Dino Oglic** [2] **Dino Sejdinovic** [1]

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 expected 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.

1. Introduction

Kernel methods are one of the pillars of machine learning (Schölkopf & 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 & De Vito, 2007; Steinwart & 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 $\mathcal{H}$ 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 & Recht, 2007; Mahoney & Drineas, 2009; Bach, 2013; Alaoui & Mahoney, 2015; Rudi et al., 2015; Rudi & 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 & 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 a Fourier transform of a bounded positive measure, called 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 kernel 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 & Recht, 2007; Sriperumbudur & Szabó, 2015; Sutherland & Schneider, 2015, and references therein) and bases performance guarantees on the accuracy of this approximation. However, all of these works require $\Omega(n)$

[1]Department of Statistics, University of Oxford, United Kingdom [2]Department of Informatics, King's College London, United Kingdom. Correspondence to: Zhu Li <zhu.li@stats.ox.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

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 & 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 due to the Lipschitz continuity requirement, the analysis does not apply to kernel ridge regression (KRR), one of the most commonly used kernel methods. 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 empirical risk of kernel ridge regression and show that it is possible to use $o(n)$ features and have the empirical risk of the linear ridge regression estimator based on random Fourier features close to the empirical risk of the original kernel estimator, also relying on a modification to the sampling distribution. However, this result is for the empirical risk only, does not provide any expected risk convergence rates, 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 & Rosasco (2017), where it is shown that $\Omega(\sqrt{n}\log n)$ features suffices for an optimal $O(1/\sqrt{n})$ learning error in a minimax sense (Caponnetto & 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 & 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, we provide, to the best of our knowledge, the sharpest results on the number of features required. In particular, we show that already with $\Omega(\sqrt{n}\log d_{\mathbf{K}}^{\lambda})$ features, we incur no loss of learning accuracy in kernel ridge regression, 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 addition, $\Omega(1/\lambda)$ features is sufficient to ensure $O(\sqrt{\lambda})$ expected risk rate in kernel support vector machines and kernel logistic regression.
- In the case of a modified data-dependent sampling distribution, 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 $O(\lambda)$ rate in kernel ridge regression ($O(\sqrt{\lambda})$ in kernel support vector machines and kernel logistic regression).
- 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. Utilizing these samples one can significantly reduce the computation time during the in sample prediction and testing stages, $O(n)$ and $O(\log n \log\log n)$, respectively. We also include a proof that gives a trade-off between the computational cost and the expected risk of the algorithm, showing that the statistical efficiency can be preserved while provably reducing the required computational cost.

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 a 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

$$k(x,y) = \int_{\mathcal{V}} e^{-2\pi i v^T(x-y)} d\tau(v) = \int_{\mathcal{V}} \left(e^{-2\pi i v^T x}\right)\left(e^{-2\pi i v^T y}\right)^* p(v)dv, \tag{1}$$

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

$$k(x,y) = \int_{\mathcal{V}} z(v,x)z(v,y)p(v)dv, \tag{2}$$

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 the random Fourier features approach is to approximate the kernel function by its Monte-Carlo estimate

$$\tilde{k}(x,y) = \frac{1}{s}\sum_{i=1}^{s} z(v_i,x)z(v_i,y), \tag{3}$$

with reproducing kernel Hilbert space $\tilde{\mathcal{H}}$ (not necessarily contained in the reproducing kernel Hilbert space $\mathcal{H}$ corresponding to the kernel function k) 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 [1]:

$$f(x) = \int_{\mathcal{V}} g(v)z(v,x)p(v)dv \qquad (\forall x \in \mathcal{X}) \tag{4}$$

where $g \in L_2(d\tau)$ is a real-valued function such that $\|g\|^2_{L_2(d\tau)} < \infty$ and $\|f\|_{\mathcal{H}}$ is equal to the minimum of $\|g\|_{L_2(d\tau)}$, 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 jth row.

As the latter approximation is simply a Monte Carlo estimate, one could also pick 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 a 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 jth row.

Let $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{K}}_q$ be the 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

$$\tilde{\mathbf{K}} = {}^1\!/_s\,\mathbf{Z}\mathbf{Z}^T \qquad \wedge \qquad \tilde{\mathbf{K}}_q = {}^1\!/_s\,\mathbf{Z}_q\mathbf{Z}_q^T.$$

If we now denote the jth column of $\mathbf{Z}$ by $\mathbf{z}_{v_j}(\mathbf{x})$ and the jth column of $\mathbf{Z}_q$ by $\mathbf{z}_{q,v_j}(\mathbf{x})$, then the following equalities can be derived easily from Eq. (3):

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

An importance weighted density function based on the notion of *ridge leverage scores* is introduced in Alaoui & Mahoney (2015) for landmark selection in the Nyström method (Nyström, 1930; Smola & Schölkopf, 2000; Williams & Seeger, 2001). For landmarks selected using that sampling strategy, Alaoui & Mahoney (2015) establish a sharp convergence rate of the low-rank estimator based on the Nyström method. This result motivates the pursuit of a similar notion for random Fourier features. Indeed, Bach (2017b) propose 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) propose the ridge leverage function with respect to a fixed input dataset, i.e.,

$$l_\lambda(v) = p(v)\mathbf{z}_v(\mathbf{x})^T(\mathbf{K}+n\lambda\mathbf{I})^{-1}\mathbf{z}_v(\mathbf{x}). \tag{5}$$

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):

$$l_\lambda(v) \leq p(v)\frac{z_0^2}{\lambda} \qquad \text{with}$$
$$\int_{\mathcal{V}} l_\lambda(v)dv = \mathrm{Tr}\big[\mathbf{K}(\mathbf{K}+n\lambda\mathbf{I})^{-1}\big] \coloneqq d^\lambda_{\mathbf{K}}.$$

The quantity $d^\lambda_{\mathbf{K}}$ 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 & De Vito, 2007) or the *number of effective degrees of freedom* (Bach, 2013; Hastie, 2017).

We can now observe that $q^*(v) = l_\lambda(v)/d^\lambda_{\mathbf{K}}$ is a probability density function. In Avron et al. (2017), it has been established that sampling according to $q^*(v)$ requires fewer Fourier features 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*.

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

3. Theoretical Analysis

In this section, we provide a unified analysis of the generalization properties of learning with random Fourier features. We start with a bound for learning with the mean squared error loss function and then extend our results to problems with Lipschitz continuous loss functions. Before presenting the results, we 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 $\rho(x, y) = \rho_{\mathcal{X}}(x)\rho(y \mid x)$ a probability measure 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 the distribution ρ, known only through the sample. The distribution $\rho_{\mathcal{X}}$ is called the marginal distribution of a data-generating process. 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[2] $f \colon \mathcal{X} \to \mathcal{Y}$ 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 (Scholkopf & Smola, 2001)

$$\hat{f}^{\lambda} := \operatorname*{arg\,min}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{L}(y_i, (\mathbf{K}\alpha)_i) + \lambda \alpha^T \mathbf{K} \alpha, \quad (6)$$

where $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$ with $\alpha \in \mathbb{R}^n$, $\mathbf{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ 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 describe ρ is measured by the expected risk (Caponnetto & De Vito, 2007)

$$\mathcal{E}(\hat{f}^{\lambda}) = \int_{\mathcal{X} \times \mathcal{Y}} \mathbf{L}(y, \hat{f}^{\lambda}(x)) d\rho(x, y).$$

Henceforth, we assume[3] that there exists $f_{\mathcal{H}} \in \mathcal{H}$ such that for all $f \in \mathcal{H}, \mathcal{E}(f_{\mathcal{H}}) \leq \mathcal{E}(f)$.

3.1. Learning with the Squared Error Loss

In this section, we consider learning with the squared error loss, i.e., $\mathbf{L}(y, f(x)) = (y - f(x))^2$. For this particular loss function, the optimization problem from Eq. (6) is known as *kernel ridge regression*. The problem can be reduced to solving a linear system $(\mathbf{K} + n\lambda\mathbf{I})\alpha = Y$, 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 millions of examples. More specifically, for a vector of observed labels Y the goal is to find a hypothesis $\tilde{\mathbf{f}}_x = \mathbf{Z}_q \beta$ 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. It turns out that $\|\tilde{f}\|_{\tilde{\mathcal{H}}}^2$ can be upper bounded by $s\|\beta\|_2^2$, where s is the number of sampled features (Appendix B). Hence, the learning problem with random Fourier features and the squared error loss can be cast as

$$\beta_{\lambda} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^s} \frac{1}{n}\|Y - \mathbf{Z}_q\beta\|_2^2 + \lambda s \|\beta\|_2^2. \quad (7)$$

This is a linear ridge regression problem in the space of Fourier features and the optimal hypothesis is given by $f_{\beta}^{\lambda} = \mathbf{Z}_q \beta_{\lambda}$, where $\beta_{\lambda} = (\mathbf{Z}_q^T \mathbf{Z}_q + n\lambda \mathbf{I})^{-1} \mathbf{Z}_q^T Y$. Since $\mathbf{Z}_q \in \mathbb{R}^{n \times s}$, the computational and space complexities are $O(s^3 + ns^2)$ and $O(ns)$. 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 expected risk and the choice of s.

3.1.1. Worst Case Analysis

In this section, we assume that the unit ball of the reproducing kernel Hilbert space contains the hypothesis $f_{\mathcal{H}}$ and provide a bound on the required number of random Fourier features with respect to the worst case minimax rate of the corresponding kernel ridge regression problem. The following theorem (a proof can be found in Appendix C) gives a general result while taking into account both the number of features s and a sampling strategy for selecting them.

Theorem 1. *Assume a kernel function k has a decomposition as in Eq. (2) and let $|y| \leq y_0$ be bounded with $y_0 > 0$. Denote with $\lambda_1 \geq \cdots \geq \lambda_n$ the eigenvalues of the kernel matrix $\mathbf{K}$ and assume the regularization parameter satisfies $0 \leq n\lambda \leq \lambda_1$. Let $\tilde{l} : \mathcal{V} \to \mathbb{R}$ be a measurable function such that $\tilde{l}(v) \geq l_{\lambda}(v)$ $(\forall v \in \mathcal{V})$ and $d_{\tilde{l}} = \int_{\mathcal{V}} \tilde{l}(v) dv < \infty$. Suppose $\{v_i\}_{i=1}^s$ are sampled independently from the probability density function $q(v) = \tilde{l}(v)/d_{\tilde{l}}$. If the unit ball of $\mathcal{H}$ contains the optimal hypothesis $f_{\mathcal{H}}$ and*

$$s \geq 5 d_{\tilde{l}} \log {}^{(16 d_{\mathbf{K}}^{\lambda})}/_{\delta},$$

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

$$\mathcal{E}(f_{\beta}^{\lambda}) - \mathcal{E}(f_{\mathcal{H}}) \leq 2\lambda + O({}^{1}/_{\sqrt{n}}) + \mathcal{E}(\hat{f}^{\lambda}) - \mathcal{E}(f_{\mathcal{H}}). \quad (8)$$

Theorem 1 expresses the trade-off between the computational and statistical efficiency through the regularization

[2] Throughout the paper, we assume (without loss of generality) that our hypothesis space is the unit ball in a reproducing kernel Hilbert space $\mathcal{H}$, i.e., $\|f\|_{\mathcal{H}} \leq 1$. This is a pretty standard assumption, characteristic to the analysis of random Fourier features (e.g., see Rudi & Rosasco, 2017)

[3] The existence of $f_{\mathcal{H}}$ depends on the complexity of $\mathcal{H}$ which is related to the conditional distribution $\rho(y|x)$ and the marginal distribution $\rho_{\mathcal{X}}$. For more details please refer to Caponnetto & De Vito (2007) and Rudi & Rosasco (2017).

parameter λ, the effective dimension of the problem $d^{\lambda}_{\mathbf{K}}$, and the normalization constant of the sampling distribution $d_{\tilde{l}}$. The regularization parameter can be considered as some function of the number of training examples (Caponnetto & De Vito, 2007; Rudi & Rosasco, 2017) and we use its decay rate as the sample size increases to quantify the complexity of the target regression function $f_\rho(x) = \int y d\rho(y \mid x)$. In particular, Caponnetto & De Vito (2007) have shown that the minimax risk convergence rate for kernel ridge regression is $O(1/\sqrt{n})$. Setting $\lambda \propto 1/\sqrt{n}$, we observe that the estimator f^{λ}_{β} attains the worst case minimax rate of kernel ridge regression. As a consequence of Theorem 1, we have the following bounds on the number of required features for the two strategies: *leverage weighted* and *plain RFF*.

Corollary 1. *If the probability density function from Theorem 1 is the empirical ridge leverage score distribution $q^*(v)$, then the upper bound on the risk from Eq. (8) holds for all $s \geq 5 d^{\lambda}_{\mathbf{K}} \log(16 d^{\lambda}_{\mathbf{K}})/\delta$.*

Theorem 1 and Corollary 1 have several implications on the choice of λ and s. First, we could pick $\lambda \in O(n^{-1/2})$ that implies the worst case minimax rate for kernel ridge regression (Caponnetto & De Vito, 2007; Rudi & Rosasco, 2017; Bartlett et al., 2005) and observe that in this case s is proportional to $d^{\lambda}_{\mathbf{K}} \log d^{\lambda}_{\mathbf{K}}$. As $d^{\lambda}_{\mathbf{K}}$ is determined by the learning problem (i.e., the marginal distribution $\rho_{\mathcal{X}}$), we can consider several different cases. In the best case (e.g., the Gaussian kernel with a sub-Gaussian marginal distribution $\rho_{\mathcal{X}}$), the eigenvalues of $\mathbf{K}$ exhibit a geometric/-exponential decay, i.e., $\lambda_i \propto R_0 r^i$ (R_0 is some constant). From Bach (2017b), we know that $d^{\lambda}_{\mathbf{K}} \leq \log(R_0/\lambda)$, implying $s \geq \log^2 n$. Hence, significant savings can be obtained with $O(n \log^4 n + \log^6 n)$ computational and $O(n \log^2 n)$ storage complexities of linear ridge regression over random Fourier features, as opposed to $O(n^3)$ and $O(n^2)$ 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 $d^{\lambda}_{\mathbf{K}} \leq (R_0/\lambda)^{1/(2t)}$ and $s \geq n^{1/(4t)} \log n$. Hence, even in this case a substantial saving in computational cost can be achieved.

Furthermore, in the worst case with λ_i very close to $R_0 i^{-1}$, our bound implies that $s \geq \sqrt{n} \log n$ features is sufficient, recovering the result from Rudi & Rosasco (2017).

Corollary 2. *If the probability density function from Theorem 1 is the spectral measure $p(v)$ from Eq. (2), then the upper bound on the risk from Eq. (8) holds for all $s \geq 5 z_0^2/\lambda \log \frac{16 d^{\lambda}_{\mathbf{K}}}{\delta}$.*

Corollary 2 addresses plain random Fourier features and states that if s is chosen to be greater than $\sqrt{n} \log d^{\lambda}_{\mathbf{K}}$ and $\lambda \propto 1/\sqrt{n}$ then the minimax risk convergence rate is guaranteed. 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 & Rosasco (2017) where $s \geq \sqrt{n} \log n$. For the other two cases, we derive $s \geq \sqrt{n} \log n$ and recover the results from Rudi & Rosasco (2017).

3.1.2. Refined Analysis

In this section, we provide a more refined analysis with expected risk convergence rates faster than $\mathcal{O}(1/\sqrt{n})$, depending on the spectrum decay of the kernel function and/or the complexity of the target regression function.

Theorem 2. *Suppose that the conditions from Theorem 1 apply and let*

$$s \geq 5 d_{\tilde{l}} \log(16 d^{\lambda}_{\mathbf{K}})/\delta.$$

Then, for all $D > 1$ and $\delta \in (0, 1)$, with probability $1 - \delta$, the excess risk of f^{λ}_{β} can be upper bounded as

$$\mathcal{E}(f^{\lambda}_{\beta}) - \mathcal{E}(f_{\mathcal{H}}) \leq 2\hat{r}^*_{\mathcal{H}} + 2\lambda^{D/(D-1)} + O(1/n) + \mathcal{E}(\hat{f}^{\lambda}) - \mathcal{E}(f_{\mathcal{H}}). \quad (9)$$

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

$$\hat{r}^*_{\mathcal{H}} \leq min_{0 \leq h \leq n} \Big(\frac{h}{n} * \frac{e_4}{n^2 \lambda^2} + \sqrt{\frac{1}{n} \sum_{i>h} \hat{\lambda}_i} \Big), \quad (10)$$

where $e_4 > 0$ is a constant and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$.

Theorem 2 covers a wide range of cases and can provide sharper risk convergence rates. In particular, note that $\hat{r}^*_{\mathcal{H}}$ is of order $O(1/\sqrt{n})$, which happens when the spectrum decays approximately as $1/n$ and $h = 0$. In this case, the excess risk converges with the rate $O(1/\sqrt{n})$, which corresponds to the considered worst case minimax rate.

On the other hand, if the eigenvalues decay exponentially, then setting $h = \lceil \log n \rceil$ implies that $\hat{r}^*_{\mathcal{H}} \leq O(\log n/n)$. Furthermore, setting $\lambda \propto \log n/n$, we can show that the excess risk converges at a much faster rate of $O(\log n/n)$.

In the best case, when the kernel function has only finitely many positive eigenvalues, we have that $\hat{r}^*_{\mathcal{H}} \leq O(1/n)$ by letting h be any fixed value larger than the number of positive eigenvalues. In this case, we obtain the fastest rate of $O(1/n)$ for the regularization parameter $\lambda \propto 1/n$.

3.2. Learning with a Lipschitz Continuous Loss

We next consider kernel methods with Lipschitz continuous loss, 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

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

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

Theorem 3. *Suppose that all the assumptions from Theorem 1 apply to the setting with a Lipschitz continuous loss. If*

$$s \geq 5d_{\tilde{l}} \log {}^{(16d^{\lambda}_{\mathbf{K}})}/_{\delta},$$

then for all $\delta \in (0,1)$, with probability $1-\delta$, the expected risk of g^{λ}_{β} can be upper bounded as

$$\mathcal{E}(g^{\lambda}_{\beta}) \leq \mathcal{E}(g_{\mathcal{H}}) + \sqrt{2\lambda} + O({}^{1}/_{\sqrt{n}}). \qquad (11)$$

This theorem, similar to Theorem 1, describes the relationship between s and $\mathcal{E}(g^{\lambda}_{\beta})$ in the Lipschitz continuous loss case. However, a key difference here is that the expected risk can only be upper bounded by $\sqrt{\lambda}$, requiring $\lambda \propto {}^{1}/_{n}$ in order to preserve the convergence properties of the risk. Corollaries 3 and 4 provide bounds for the cases of leverage weighted and plain RFF, respectively.

Corollary 3. *If the probability density function from Theorem 3 is the empirical ridge leverage score distribution $q^{*}(v)$, then the upper bound on the risk from Eq. (11) holds for all $s \geq 5d^{\lambda}_{\mathbf{K}} \log {}^{(16d^{\lambda}_{\mathbf{K}})}/_{\delta}$.*

In the three considered cases for the effective dimension of the problem $d^{\lambda}_{\mathbf{K}}$, Corollary 3 states that the statistical efficiency is preserved if the leverage weighted RFF strategy is used with $s \geq \log^2 n$, $s \geq n^{1/(2t)} \log n$, and $s \geq n \log n$, respectively. Again, significant computational savings can be achieved if the eigenvalues of the kernel matrix $\mathbf{K}$ have either a geometric/exponential or a polynomial decay.

Corollary 4. *If the probability density function from Theorem 3 is the spectral measure $p(v)$ from Eq. (2), then the upper bound on the risk from Eq. (11) holds for all $s \geq 5{}^{z_0^2}/_{\lambda} \log {}^{(16d^{\lambda}_{\mathbf{K}})}/_{\delta}$.*

Corollary 4 states that $n \log n$ features are required to attain $O(n^{-1/2})$ convergence rate of the expected risk with plain RFF, recovering results from Rahimi & Recht (2009). Similar to the analysis in the squared error loss case, Theorem 3 together with Corollaries 3 and 4 allows theoretically motivated trade-offs between the statistical and computational efficiency of the estimator g^{λ}_{β}.

3.3. A Fast Approximation of Leverage Weighted RFF

As discussed in Sections 3.1 and 3.2, 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.

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 λ
Output: set of features $\{(v_1, p_1), \cdots, (v_l, p_l)\}$ with l and each p_i computed as in lines 3–4
1: sample s features $\{v_1, \ldots, v_s\}$ from $p(v)$
2: create a feature matrix $\mathbf{Z}_s$ such that the ith 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 real number p_i such that p_i is equal to the ith diagonal element of the matrix
$$\mathbf{Z}_s^T \mathbf{Z}_s ((1/s)\mathbf{Z}_s^T \mathbf{Z}_s + n\lambda I)^{-1}$$
4: $l \leftarrow \sum_{i=1}^{s} p_i$ and $M \leftarrow \{(v_i, p_i/l)\}_{i=1}^{s}$
5: sample l features from M using the multinomial distribution given by the vector $(p_1/l, \cdots, p_s/l)$

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 $l \ll s$ 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 $\mathbf{Z}_l \in \mathbb{R}^{n \times l}$ such that the ith row of $\mathbf{Z}_l$ is given by $\mathbf{z}_{x_i}(\mathbf{v}) = [\sqrt{{}^{l}/_{p_1}} z(v_1, x_i), \cdots, \sqrt{{}^{l}/_{p_l}} z(v_l, x_i)]^T$.

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 $O(ns^2)$ and inverting $\mathbf{Z}_s^T \mathbf{Z}_s + n\lambda I$ costs only $O(s^3)$. Hence, for $s \ll n$, the overall runtime is only $O(ns^2 + s^3)$. 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 cost of $O(n \times s)$.

The following theorem gives the convergence rate for the expected risk of Algorithm 1 in the kernel ridge regression setting (a proof can be found in Appendix E).

Theorem 4. *Suppose the conditions from Theorem 1 apply to 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. (2) and the regularization parameter λ. Denote with $\tilde{f}_l^{\lambda^*}$ the ridge regression estimator obtained using a regularization parameter λ^* and a set of random Fourier features $\{v_i\}_{i=1}^{l}$ returned by Algorithm 1. If*

$$s \geq 7{}^{z_0^2}/_{\lambda} \log {}^{(16d^{\lambda}_{\mathbf{K}})}/_{\delta} \quad \text{and} \quad l \geq 5d^{\lambda^*}_{\mathbf{K}} \log {}^{(16d^{\lambda^*}_{\mathbf{K}})}/_{\delta},$$

then for all $\delta \in (0,1)$, with probability $1-\delta$, the expected

risk of $\tilde{f}_l^{\lambda^*}$ can be upper bounded as

$$\mathcal{E}(\tilde{f}_l^{\lambda^*}) \leq \mathcal{E}(f_{\mathcal{H}}) + 2\lambda + 2\lambda^* + O(1/\sqrt{n}).$$

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

Theorem 4 bounds the expected 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 $\tilde{f}_l^{\lambda^*}$ converges with the minimax rate to the hypothesis $f_{\mathcal{H}} \in \mathcal{H}$ for $l \in \Omega(\log n \cdot \log\log n)$. 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.

The latter result can also be generalized to kernel support vector machines and logistic regression. The convergence rate of the expected risk, however, is at a slightly slower $O(\sqrt{\lambda} + \sqrt{\lambda^*})$ rate due to the difference in the employed loss function (see Section 3.2).

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 $O(\sqrt{n}\log n)$ to $O(\log n \cdot \log\log n)$. 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 experiment section below.

4. Numerical Experiments

In this section, we report the results of our numerical experiments (on both simulated and real-world datasets) aimed at validating our theoretical results and demonstrating the utility of Algorithm 1. We first verify our results through a simulation experiment. Specifically, we consider a spline kernel of order r where $k_{2r}(x,y) = 1 + \sum_{i=1}^{\infty} \frac{1}{m^{2r}} \cos 2\pi m(x-y)$ (also considered by Bach, 2017b; Rudi & Rosasco, 2017). If the marginal distribution of X is uniform on $[0,1]$, we can show that $k_{2r}(x,y) = \int_0^1 z(v,x)z(v,y)q^*(v)dv$, where $z(v,x) = k_r(v,x)$ and $q^*(v)$ is also uniform on $[0,1]$. 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 1 and Corollary 1, 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 $O(n^{-1/2})$, or at $O(n^{-1/3})$ if $\lambda \propto n^{-1/3}$. Figure 1 demonstrates this with 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. We use four datasets from Chang & Lin (2011) and Dheeru & Karra Taniskidou (2017) for this purpose, including two for regression and two for classification: `CPU`, `KINEMATICS`, `COD-RNA` and `COVTYPE`. Except `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/or prediction error.

5. Discussion

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 ridge leverage scores and demonstrated its superiority in both theoretical and empirical analyses.

For kernel ridge regression, Avron et al. (2017) and Rudi & 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 *empir-*

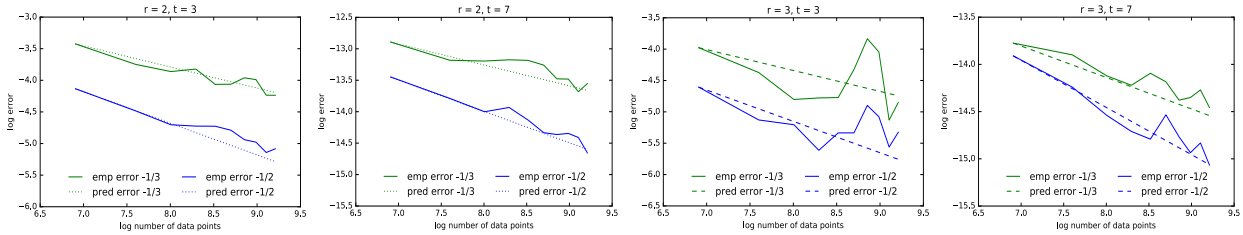

Figure 1: The log-log plot of the theoretical and simulated risk convergence rates, averaged over 100 repetitions.

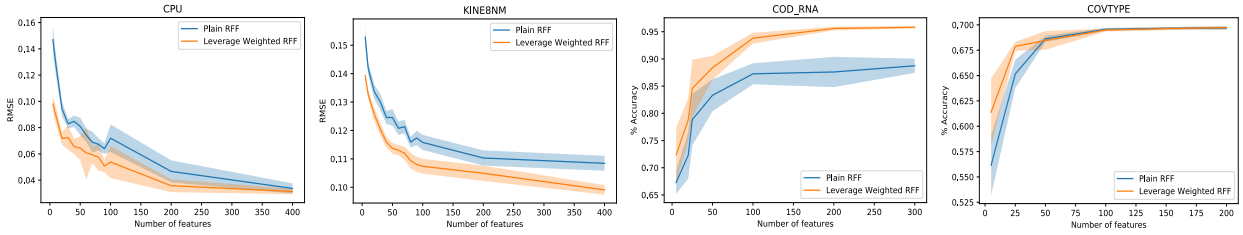

Figure 2: Comparison of leverage weighted and plain RFFs, with weights computed according to Algorithm 1.

ical risk. The authors of that work have also proposed a modified data-dependent sampling distribution and demonstrated that a further reduction on the number of random Fourier features is possible for leverage weighted sampling. However, their results do not provide a convergence rate for the *expected 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 & Rosasco (2017) give a comprehensive study of the generalization properties of random Fourier features for kernel ridge regression by bounding the expected 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 expected 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 & 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 & 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 & Recht (2009) and Bach (2017b) have studied this problem setting and obtained that $\Omega(n)$ features are needed to ensure $O(1/\sqrt{n})$ expected 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. We 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 estimator. We further test our findings on several different real-world datasets and verify this empirically. 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; Zhang et al., 2018), as well as a comparison to the Nyström method.

Acknowledgment: We thank Fadhel Ayed, Qinyi Zhang and Anthony Caterini for fruitful discussion on some of the results as well as for proofreading of this paper. 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

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

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning*, pp. 253–262, 2017.

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

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

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

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

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

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

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

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

Chang, C.-C. and Lin, C.-J. 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.

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

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

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

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

Oglic, D. and Gärtner, T. Greedy feature construction. In *Advances in Neural Information Processing Systems 29*, pp. 3945–3953. Curran Associates, Inc., 2016.

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

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

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

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

Rudi, A., Camoriano, R., and Rosasco, L. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pp. 1657–1665, 2015.

Rudi, A., Carratino, L., and Rosasco, L. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pp. 3891–3901, 2017.

Rudi, A., Calandriello, D., Carratino, L., and Rosasco, L. On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems*, pp. 5672–5682, 2018.

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

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

Scholkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Schölkopf, B., Tsuda, K., and Vert, J.-P. *Kernel methods in computational biology*. MIT press, 2004.

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

Sriperumbudur, B. and Szabó, Z. Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems*, pp. 1144–1152, 2015.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

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

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

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

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

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

Zhang, J., May, A., Dao, T., and Ré, C. Low-precision random fourier features for memory-constrained kernel approximation. *arXiv preprint arXiv:1811.00155*, 2018.

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