Information and Inference: A Journal of the IMA (2024) **13**, iaad054
https://doi.org/10.1093/imaiai/iaad054

Conditioning of random Fourier feature matrices: double descent and generalization error

ZHIJUN CHEN
Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA

AND

HAYDEN SCHAEFFER*
Department of Mathematics, University of California, Los Angeles, Los Angeles, CA 90095, USA
*Corresponding author: hayden@math.ucla.edu

[Received on 4 March 2023; revised on 9 October 2023; accepted on 18 November 2023]

We provide high-probability bounds on the condition number of random feature matrices. In particular, we show that if the complexity ratio N/m, where N is the number of neurons and m is the number of data samples, scales like $\log^{-1}(N)$ or $\log(m)$, then the random feature matrix is well-conditioned. This result holds without the need of regularization and relies on establishing various concentration bounds between dependent components of the random feature matrix. Additionally, we derive bounds on the restricted isometry constant of the random feature matrix. We also derive an upper bound for the risk associated with regression problems using a random feature matrix. This upper bound exhibits the double descent phenomenon and indicates that this is an effect of the double descent behaviour of the condition number. The risk bounds include the underparameterized setting using the least squares problem and the overparameterized setting where using either the minimum norm interpolation problem or a sparse regression problem. For the noiseless least squares or sparse regression cases, we show that the risk decreases as m and N increase. The risk bound matches the optimal scaling in the literature and the constants in our results are explicit and independent of the dimension of the data.

Keywords: random feature models; conditioning; restricted isometry property; double descent.

1. Introduction

Random feature methods are essentially two-layer shallow networks whose single hidden layer is randomized and not trained [32,33]. They can be used for various learning tasks, including regression, interpolation, kernel-based classification, etc. Studying their behaviour not only results in a deeper understanding of kernel machines but could also yield insights into more complex models, such as very wide or deep neural networks.

Consider $\boldsymbol{x} \in \mathbb{R}^d$ to be the input vector and $\boldsymbol{W} = [\boldsymbol{\omega}_{i,j}] \in \mathbb{R}^{d\times N}$ to be a random weight matrix, where each column of the matrix $\boldsymbol{W}$ is sampled from some prescribed probability density $\rho(\boldsymbol{\omega})$. Then the random feature map is defined by $\phi(\boldsymbol{W}^T\boldsymbol{x}) \in \mathbb{C}^N$, where $\phi : \mathbb{R} \to \mathbb{C}$ is some Lipschitz activation function applied element-wise to the input. The resulting two-layer network with N neurons is defined by $\boldsymbol{y} = \boldsymbol{C}^T\phi(\boldsymbol{W}^T\boldsymbol{x}) \in \mathbb{C}^{d_{out}}$, where the final layer $\boldsymbol{C} \in \mathbb{C}^{N\times d_{out}}$ needs to be trained for a given problem. The analysis of these models with respect to the dimensionality, stability, testing error, etc. often depends on the extreme singular values (or the condition number) of the matrix $\boldsymbol{A} := \phi(\boldsymbol{X}^T\boldsymbol{W}) \in \mathbb{C}^{m\times N}$, where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots \boldsymbol{x}_m] \in \mathbb{R}^{d\times m}$ is a collection of m random samples of the input variable. In particular, for regression or interpolation, knowledge of the condition number of $\boldsymbol{A}$ can yield control

over the error as well as reveal the landscape of the testing error as a function of the model complexity ratio N/m.

The testing error (or risk) of random feature models have been of recent interest [39,17,20,26,34,36], in particular, with a goal of quantifying the number of features needed to obtain a given learning rate. In [34], it was shown that the random feature model yields a test error of $\mathcal{O}(N^{-\frac{1}{2}} + m^{-\frac{1}{2}})$ when trained on Lipschitz loss functions. Thus if $m \asymp N$ then the generalization error is $\mathcal{O}(N^{-\frac{1}{2}})$ for large N. In [35], a comparison between the test error using kernel ridge regression with both the full kernel matrix and a column subsampled kernel matrix (related to Nyström's method) suggests that the number of columns (and rows) should be roughly the square root of the number of original samples in order to obtain their learning rate. The squared loss case was considered in [20,36], providing risk bounds when the random feature model is trained via ridge regression. In [36], it was shown that for f in an RKHS, using $N = \mathcal{O}(\sqrt{m}\,\log m)$ is sufficient to achieve a test error of $\mathcal{O}(m^{-\frac{1}{2}})$ with a random feature method. The results in [35,36] use similar technical assumptions on the kernel and the second moment operator, e.g. a certain spectral decay rate on second moment operator, which may be difficult to verify. Note that for the ridge regression results, the penalty parameter $\lambda = \lambda(m)$ often must remain positive and thus such results do not include interpolation. In [39], the risk was analysed for a regularized model with the target function in an RKHS, noting that to achieve $N^{-1} + m^{-\frac{1}{2}}$ one must place explicit assumptions on the decay rate of the spectrum of the kernel operator, see also [3]. In [26], an investigation of the kernel ridge regression and the random feature ridge regression problems in terms of the N, m and d was given. The main result suggests that an 'optimal' choice for the complexity ratio N/m depends on an algebraic scaling in N, m and d.

In the overparameterized setting, minimum norm regression problems are often used, either by minimizing the ℓ^2 or ℓ^1 norm with a data fitting constraint. The ℓ^1 problem can be used to obtain a low complexity random feature model in the highly overparameterized setting and was proposed and studied in [17] (see also [40] for a related algorithmic approach). By solving the ℓ^1 basis pursuit denoising problem with a random feature matrix, test error bounds were obtained as a function of the sparsity, N, m and d. It is worth noting that in the dense case, i.e. when the sparsity equals N and $m \asymp N^2$ up to log terms, the test error scales like $N^{-1} + m^{-\frac{1}{2}}$ up to log terms, thus the method achieves the upper bounds provided by other methods but utilizes a different computational approach.

The min-norm interpolation problem (or the ridgeless limit as $\lambda \to 0^+$) has received recent attention for random feature methods [4,5,8,18,21,26,27,38]. In [4,38], the test error for general ridge and ridgeless regression problems is shown to be controlled by the condition number, or the related notion of effective rank, of the random design matrix. It was shown in [26] that the min-norm interpolators are optimal among all kernel methods in the overparameterized regime.

In some settings, overparameterized random feature models can produce lower test error even though they have large complexity. However, choosing N and m only using the risk bounds could lead to an insufficient picture of the parameter landscape for random feature models or neural network. In modern learning, the graph of the test error as a function of the model's complexity (i.e. relative number of neurons) has two valleys. The first is when the number of parameters is relatively small; this is the underparameterized regime where the model has small complexity and small risk. As the complexity reaches the interpolation threshold ($N = m$), the risk tends to peak and then begins to decay again as the number of parameters continues to increase well into the overparameterized regime. This is referred to as the double descent phenomenon [1,5–7] and has had several recent results, specifically, in the characterization of the test error as a function of the complexity ratio [1,2,5–7,19,26,27]. A direct analysis of the double descent curves for random feature regression [26,27] showed that in the overparameterized

regime, taking $N/m = \mathcal{O}(1)$ can lead to suboptimal results which can be corrected if $N/m \to \infty$. To achieve near optimal test errors, they suggest setting $N \asymp m^{1+\delta}$ for small $\delta > 0$.

1.1 *Some related work*

Some of the earlier results in the literature on random matrix theory for some related problems focused on random kernel matrices in the form $K_{i,j} = \phi(\boldsymbol{x}_i^T \boldsymbol{x}_j)$, specifically, characterizing the spectrum of square matrices depending on one random variable $\boldsymbol{x}$ [10,12,13]. In [9,18,28–30], the asymptotic behaviour of asymmetric rectangular random matrices $\boldsymbol{A} = \phi(\boldsymbol{X}^T \boldsymbol{W})$ was studied, in particular, to quantify the limiting distribution on the spectrum of the Gram matrix $\boldsymbol{A}\boldsymbol{A}^*$. In [25] the asymptotic behaviour of the resolvent operator of the Gram matrix was studied, where $\boldsymbol{W}$ is random but $\boldsymbol{X}$ is deterministic. Using concentration arguments, [23,25] analysed the spectrum of the Gram matrix of random feature maps in the high-dimensional setting where N grows like m. Precise asymptotics for the random feature ridge regression model in the large m, N and d limit and a characteristic of the test error as a function of the dimensional parameters were given in [24]. They also showed the existence of a phase transition with respect to N/m at the interpolation threshold. In [22], the authors studied the multiple-descent curve for the scaling $d = m^{\alpha}$, where $\alpha \in (0, 1)$. They observed non-monotonic behaviour in the test error as a function of the sample size m. The results in [22] rely on the restricted lower isometry of the kernels, showing that (with high probability) the empirical kernel matrix has a certain number of non-zero eigenvalues with a lower bound.

1.2 *Our contributions*

In this work, we provide high-probability bounds on the conditioning of the random feature matrix $\boldsymbol{A} = \phi(\boldsymbol{X}^T \boldsymbol{W})$ for N neurons, m samples and with dimension d. In particular, when the N/m scales like $\log(m)$ (overparametrized) or $\log^{-1}(N)$ (underparameterized), the singular values of $\boldsymbol{A}$ concentrate around 1 and thus with high probability the condition number of the random feature matrix is small. In addition, the guarantees break down in the interpolation region where we prove that the system is ill-conditioned. When $N = m$, as the number of features (or equivalently the number of samples) increases the conditioning at the interpolation regime worsens. This transition between the conditioning states coincides with the phase transition in the double descent curves (the risk) for random feature regression. As an additional application of our results, we show that the condition number can be used to obtain state-of-the-art test error bounds for a particular class of target functions. We highlight some comparisons with other works below.

- Our theoretical results are related to other random matrix results, such as [9,18,25,28–30]; however, we quantify the behaviour of the random feature matrix for both finite and large N and m and provide both sample and feature complexity bounds. Our results give a simple scaling between N and m that yields well-conditioned random feature matrices.
- We connect the conditioning of the random matrix to the double descent curves and provide another characterization of the underparameterized and overparameterized regions, compared with [26,27]. Specifically, our results show that the complexity ratio N/m needs to be only logarithmically far from the interpolation threshold to have small risk, rather than algebraically far. Our results support similar conclusions to [26,27].
- We provide non-asymptotic bounds, i.e. N, m and d are finite, and the bounds explicitly relate the scaling between these quantities. In our work, the weights and data can be unbounded. In [27], the

focus is on the case for $N, m, d \to \infty$ where the weights and data are assumed to be uniformly distributed and bounded. In addition, our work also includes the sparse recovery regime (specifically the restricted isometry property (RIP) condition), which has not been studied in previous random feature works.

- The proofs give explicit values for the constants used in our bounds and thus our results hold up to additive terms, for example, compared with the results of [36] which are based on overall rates. Our bounds improve for high-dimensional data, since the constants are independent of the dimension d.
- The analysis in [17] established the structure of the random feature matrix using a coherence estimate (for sparse regression in the overparameterize regime). A coherence bound leads to a non-optimal quadratic scaling (up to log terms) between m and N (or the sparsity s). *Our results directly estimate the restricted isometry constant thus yielding a linear scaling (up to log terms) between m and N (or s). A complete proof is given in C and is one of the key theoretical contributions of this work.* This difference allows for a more precise characterization of the underparameterized and overparameterized regimes, i.e. it shows that the complexity ratio only needs to scale by log factors rather than polynomials. Our results also match the transition point $N = m$ seen in [24,26,27] and the linear scaling considered in [30]. Our RIP result allows for the use of random feature models in sparse regression which is novel in this field.
- We establish risk bounds for random feature regression without the need for assumptions on the spectrum of the second moment operator or kernel [35,36]. We also improve the bounds in the sparse setting [17]. Similar results hold for other conditioning-based methods, such as greedy solvers. Our risk bounds include the effects of noise on the measurements, with the assumption that the noise is bounded (or bounded with high probability, e.g. normally distributed).

1.3 *Notation*

Let $\mathbb{R}$ be the set of all real number and $\mathbb{C}$ be the set of all complex number where $i = \sqrt{-1}$ denotes the imaginary unit. Define the set $[N]$ to be all natural numbers smaller than N, i.e. $\{1, 2, \ldots, N\}$. Throughout the paper, we denote vectors or matrices with bold letters, and denote the identity matrix of size $n \times n$ by $\boldsymbol{I}_n$. For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^d$, the inner product is denoted by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{j=1}^{d} x_j \bar{y}_j$, where $\boldsymbol{x} = [x_1, \ldots, x_d]^T$ and $\boldsymbol{y} = [y_1, \ldots, y_d]^T$. For a vector $\boldsymbol{x} \in \mathbb{C}^d$, we denote by $\|\boldsymbol{x}\|_p$ the ℓ^p-norm of $\boldsymbol{x}$ and for a matrix $\boldsymbol{A} \in \mathbb{C}^{m \times N}$ the (induced) p-norm is written as $\|\boldsymbol{A}\|_p$. For a matrix $\boldsymbol{A} \in \mathbb{C}^{m \times N}$, its transpose is denoted as $\boldsymbol{A}^T$ and its conjugate transpose is denoted as $\boldsymbol{A}^*$. Lastly, let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. If $\boldsymbol{A} \in \mathbb{C}^{m \times N}$ is full rank, then the pseudo-inverse is $\boldsymbol{A}^\dagger = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}$ if $m \leq N$ or $\boldsymbol{A}^\dagger = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*$ if $m \geq N$. The condition number with respect to the 2-norm of a matrix $\boldsymbol{A}$ is defined by $\kappa(\boldsymbol{A}) := \|\boldsymbol{A}\|_2 \|\boldsymbol{A}^\dagger\|_2$. Denote the k-th eigenvalue of the matrix $\boldsymbol{M}$ by $\lambda_k(\boldsymbol{M})$.

2. Empirical behaviour

The double descent phenomena appears in both the condition number for random matrices and the risk associated with the corresponding regression problem. In Fig. 1, we examine the behaviour of the risk and the condition number of the Gram matrix as a function of the complexity N/m, where N is the column size, i.e. the number of features, and m is the row size, i.e. the number of samples. Let d be the dimension of each sample. We set $d = 3$ and $m = 100$, and vary the number of features $N \in [500]$.

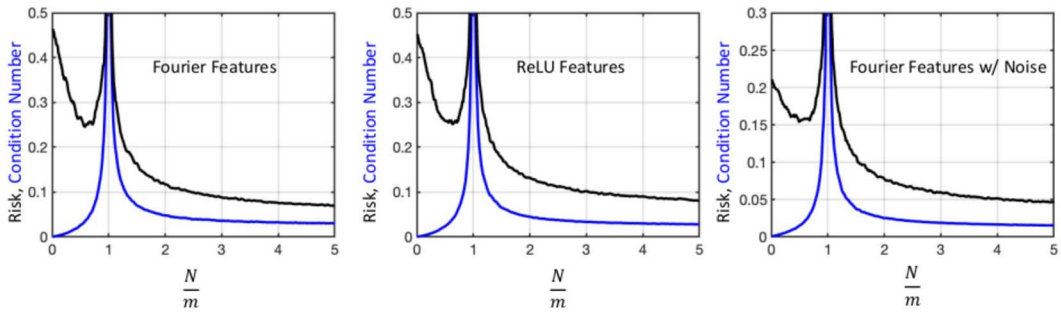

FIG. 1. **Double Descent of the Condition Number and Risk**. The condition number and risk curves are plotted as a function of the complexity ratio N/m. The curves are rescaled for visualization purposes, where the units are multiples of 2.9×10^3 for the risk and 4.8×10^4 for the condition number. Each of the graphs display the double descent phenomena corresponding to the same complexity regions. In each plot, the global minima of the risk is obtained in the overparameterized regime $N/m > 1$ where the linear system remains well-conditioned as N grows.

The weights of the random feature matrix are sampled from $\mathcal{N}(0, 0.1)$ and the 100 training points are sampled from $\mathcal{N}(0, 1)$. The target function for the ℓ^2 regression problem is linear $f(\boldsymbol{x}) = \boldsymbol{b}^T\boldsymbol{x}$, where the components of $\boldsymbol{b} \in \mathbb{R}^d$ are randomly sampled from the uniform distribution $U[0, 1]^d$, which differs from the randomness used in the weights. The outputs on the training samples have either zero noise or additive gaussian noise with 10% SNR (as indicated on the graphs). The training is done using the psuedo-inverse, i.e. the least squares problem or the min-norm interpolator depending on the complexity ratio. The empirical risk is measured on 1000 testing samples from $\mathcal{N}(0, 1)$. For each N, the risk and condition numbers are averaged over 10 random trials. Each curve in Fig. 1 is rescaled so that the values are within the same range, noting that the minimum value of the condition number is 1 before rescaling. In each case, the double descent of the condition number and the empirical risk coincides with the same complexity regimes. Specifically, both peak at the interpolation threshold $N/m = 1$ with large values obtained within a certain width of this threshold, which was also observed for this problem in [11]. It is worth noting that in these three experiments, using Fourier features [32–34] with and without noise or using ReLU features, the global minima of the risk is obtained in the overparameterized regime, i.e. when $N/m > 1$.

In Fig. 2, the probability density function associated with the singular values of the normalized random feature matrix in the underparameterized and overparameterized regimes is plotted using different complexity scales N/m. The curves in Fig. 2 are normalized to have their maxima equal to 1 and are estimated using a smooth density approximation over 10 trials. In both plots, the red curves indicate the interpolation threshold $N/m = 1$ and are the transition boundaries between the two regimes (i.e. between the two plots). The weights and data are sampled from $\mathcal{N}(0, 1)$. In this experiment we set $d = 50$ and $m = 3d = 150$, and vary the number of features N based on the logarithmic scalings indicated on the graphs. In Fig. 2, the corresponding linear system is well conditioned when the support of the probability density is bounded away from the endpoints of the interval. Empirically, this is the case (with high probability) for the log and log-cubed scaling.

The choice of scalings in the plots is to provide empirical support for the theoretical results in Section 3, specifically, that a logarithmic scaling is sufficient to obtain a bound on the extreme singular values. In addition, there is an observable symmetry between the (estimated) probability density functions in the underparameterized and overparameterized regimes using the same scaling (with m and

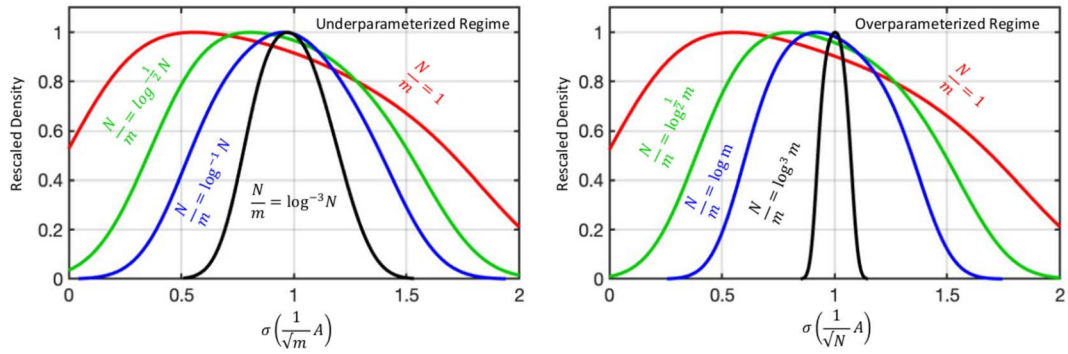

FIG. 2. **Probability Density Functions for the Singular Values:** The (normalized) smooth estimations of the probability density functions associated with the singular values of the normalized random feature matrix in the underparameterized (left) and overparameterized (right) regimes are plotted. The red curves indicate the interpolation case $N/m = 1$ and are the transition boundaries between the two plots.

N switched). The symmetry is not exact, partly due to rounding in order to obtain integer values for N, but the plots in Fig. 2 show that the overparameterized and underparameterized systems are well-conditioned when using the same functional scaling.

3. Condition number of random feature maps and insights

In this section, we discuss the bounds on the extreme singular values of the random feature matrix $\boldsymbol{A} = \phi(\boldsymbol{X}^T \boldsymbol{W})$ where the elements of the matrices $\boldsymbol{W}$ and $\boldsymbol{X}$ are randomly drawn from normal distributions.

3.1 *Main result*

Before discussing the results, we state the main assumption that is used throughout the paper. Specifically, the data and weights are normally distributed and the random feature matrix is built using Fourier features. This assumption is not necessary for the conclusions in this work; however, they simplify expressions for the constants and the parameter scalings.

ASSUMPTION 1. The data samples $\{\boldsymbol{x}_j\}_{j\in[m]}$ are drawn from $\mathcal{N}(\boldsymbol{0}, \gamma^2 \boldsymbol{I}_d)$ and the set of random feature weights $\{\boldsymbol{\omega}_k\}_{k\in[N]}$ are drawn from $\mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d)$. The random feature matrix $\boldsymbol{A} \in \mathbb{C}^{m\times N}$ is defined component-wise by $a_{j,k} = \phi(\boldsymbol{x}_j, \boldsymbol{\omega}_k)$, where $\phi(\boldsymbol{x}, \boldsymbol{\omega}) = \exp(i\langle \boldsymbol{x}, \boldsymbol{\omega}\rangle)$.

THEOREM 1. [**Conditioning in Three Regimes**] Assume that the data $\{\boldsymbol{x}_j\}_{j\in[m]}$, weights $\{\boldsymbol{\omega}_k\}_{k\in[N]}$ and random feature matrix $\boldsymbol{A} \in \mathbb{C}^{m\times N}$ satisfy Assumption 1.

(a) (**Underparameterized Regime** $m > N$) If the following conditions hold:

$$m \geq C\eta^{-2} N \log\left(\frac{2N}{\delta}\right) \tag{3.1}$$

$$\sqrt{\delta}\,\eta\,(4\gamma^2\sigma^2 + 1)^{\frac{d}{4}} \geq N, \tag{3.2}$$

for some $\eta, \delta > 0$, then with probability at least $1 - 2\delta$ we have

$$\left| \lambda_k \left(\frac{1}{m} \boldsymbol{A}^* \boldsymbol{A} \right) - 1 \right| \leq 3\eta,$$

for all $k \in [N]$.

(b) (**Overparameterized Regime** $m < N$) If the following conditions hold:

$$N \geq C\eta^{-2} m \log\left(\frac{2m}{\delta}\right)$$
$$\sqrt{\delta}\, \eta\, (4\gamma^2\sigma^2 + 1)^{\frac{d}{4}} \geq m,$$

for some $\eta, \delta > 0$, then with probability at least $1 - 2\delta$ we have

$$\left| \lambda_k \left(\frac{1}{N} \boldsymbol{A} \boldsymbol{A}^* \right) - 1 \right| \leq 3\eta,$$

for all $k \in [m]$.

(c) (**Interpolation Threshold**) If $m = N \geq 2$, then the eigenvalues satisfy

$$\mathbb{E}\, \lambda_{\min} \left(\frac{1}{N} \boldsymbol{A}^* \boldsymbol{A} \right) \leq \left(1 - \frac{1}{N}\right)^{\frac{1}{2}} \frac{1}{(4\gamma^2\sigma^2 + 1)^{d/4}} + \frac{1}{N}$$
$$\mathbb{E} \lambda_{\max} \left(\frac{1}{N} \boldsymbol{A}^* \boldsymbol{A} \right) \geq 2 - \frac{1}{N},$$

where the expectation is with respect to $\{\boldsymbol{x}_j\}_{j \in [m]}$ and $\{\boldsymbol{\omega}_k\}_{k \in [N]}$.

The universal constant $C > 0$ is no greater than 4 if $\min\{m, N\} > 25$.

The proof of Theorem 1 is given in Section 5.

Theorem 1 provides a characterization of the behaviour of the condition number of the random feature matrix as a function of complexity N/m, with respect to any dimension d. If the parameter η is chosen to satisfy $\eta < 1/3$, then the eigenvalues of the normalized Gram matrices are strictly between 0 and 2. Moreover, in the extreme cases, when $m \gg N$ or $m \ll N$, then η can be small and the eigenvalues of the normalized Gram matrices will concentrate near 1. If the random feature matrix is square and we assume that the complexity bound $m = N \leq (4\gamma^2\sigma^2 + 1)^{d/4}$ holds, then by applying Markov's inequality the minimum eigenvalue of the matrix $N^{-1}\boldsymbol{A}^*\boldsymbol{A}$ satisfies

$$\mathbb{P}\left(\lambda_{\min} \left(\frac{1}{N} \boldsymbol{A}^* \boldsymbol{A} \right) \geq N^{-\frac{1}{2}} \right) \leq 2N^{-\frac{1}{2}}.$$

On the other hand, the largest eigenvalue of $N^{-1}\boldsymbol{A}^*\boldsymbol{A}$ must be larger than 1, since columns of this matrix are of norm 1. Therefore, at the interpolation threshold, the Gram matrix becomes ill-conditioned.

Altogether, Theorem 1 shows that the conditioning of the random feature matrix exhibits the double descent phenomena.

One interesting consequence of Theorem 1 is that the random feature matrix does not need regularization to be well-conditioned. This was also observed in [31] for random matrices $\boldsymbol{A}$, where $a_{i,j}$ are drawn randomly from some probability distribution and for kernel matrices $\boldsymbol{A} = K(\boldsymbol{X}^T\boldsymbol{X})$, i.e. symmetric systems.

REMARK 1. The symmetry in the conditions for the overparameterized and underparameterized regimes is inherited from the assumption that the weights and samples use the same distribution (with different variances). The results can be extended to uniform or sub-gaussian random sampling for $\boldsymbol{x}$ or $\boldsymbol{\omega}$ and other feature maps, which will lead to similar overall conditions but break this type of symmetry.

REMARK 2. In several of the theoretical results, the smaller of the dimensional parameter (N, m or sparsity s) must be less than $\sqrt{\delta}\,\eta\,(4\gamma^2\sigma^2+1)^{\frac{d}{4}}$. If we set the parameters to $\delta = 0.01$, $\eta = 1/4$ and $\gamma, \sigma = 2$, then for $d \geq 44$ the complexity limitation is on the exascale. If one wants the inner product $\langle \boldsymbol{x}, \boldsymbol{\omega}\rangle$ to be order 1 (for example, dividing the variance d), then after rescaling we get $\left(4\gamma^2\sigma^2/d+1\right)^{d/4}$, which is like $\exp(\gamma^2\sigma^2)$ for large d. Thus in the rescaled case in high dimensions, if $\gamma\sigma = \sqrt{46}$, then the complexity limitation is on the exascale. On the other hand, for small $\gamma\sigma$ the complexity bounds can place a severe limitation on the smaller of the dimensional parameters.

4. Generalization error using the condition number

With the control on the condition number, we can provide bounds on the generalization error (or risk) associated with regression problems using random feature matrices in both the underparameterized and overparameterized settings. We show that the double descent phenomena on the risk is a consequence of the double descent phenomena on the condition number.

4.1 *Set-up and notation*

The analysis for the generalization error will follow the set-up from [17,32–34]. For a probability density ρ (associated with the random weights $\boldsymbol{\omega}$) in $\mathbb{R}^d$ and an activation function $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$, a function $f : \mathbb{R}^d \to \mathbb{C}$ has finite ρ-norm with respect to ϕ if it belongs to the class [32]

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

The regression problem is to find an approximation of a target function $f \in \mathcal{F}(\phi,\rho)$, that is, given a set of random weights $\{\boldsymbol{\omega}_k\}_{k\in[N]}$, we train

$$f^\sharp(\boldsymbol{x}) = \sum_{k=1}^{N} c_k^\sharp\, \phi(\boldsymbol{x},\boldsymbol{\omega}_k) \tag{4.1}$$

using $(\boldsymbol{x}_j, y_j)$, where $y_j = f(\boldsymbol{x}_j) + e_j$ and e_j is the measurement noise, for $j \in [m]$. Let $\boldsymbol{A} \in \mathbb{C}^{m\times N}$ be the matrix with $a_{j,k} = \phi(\boldsymbol{x}_j,\boldsymbol{\omega}_k)$ for $j \in [m]$ and $k \in [N]$; the training problem is equivalent to finding $\boldsymbol{c} \in \mathbb{C}^N$ such that $\boldsymbol{A}\boldsymbol{c} \approx \boldsymbol{y}$, where $\boldsymbol{y} = [y_1,\ldots,y_m]^T \in \mathbb{C}^m$.

ASSUMPTION 2. The measurement noise $\{e_j\}_{j\in[m]}$ is bounded by some constant E, i.e. $\|e_j\|_\infty \leq E$.

Assumption 2 can be extended to random noise. For example, if $e_j \sim \mathcal{N}(\mathbf{0}, \nu^2 \boldsymbol{I}_d)$ and $m \geq 2\log\left(\delta^{-1}\right)$ then with probability at least $1-\delta$ the noise is bounded by $|e_j| \leq 2\nu$ for all $j \in [m]$. This does not change our technical arguments except for an additional δ term in the probability bounds.

4.2 *Underparameterized regime, least squares problem*

In the underparameterized (overdetermined) regime where we have more data samples than features, the coefficient $\boldsymbol{c}^\sharp$ are trained via the least squares problem:

$$\boldsymbol{c}^\sharp \in \underset{\boldsymbol{c}\in\mathbb{R}^N}{\operatorname{argmin}} \|\boldsymbol{A}\boldsymbol{c} - \boldsymbol{y}\|_2,$$

where $\boldsymbol{c}^\sharp = [c_1^\sharp, \ldots, c_N^\sharp]^T$ and the (trained) approximation is given by

$$f^\sharp(\boldsymbol{x}) = \sum_{k=1}^{N} c_k^\sharp \, \phi(\boldsymbol{x}, \boldsymbol{\omega}_k). \tag{4.2}$$

Using the condition number of $\boldsymbol{A}$, we can control the risk

$$R(f^\sharp) := \|f - f^\sharp\|^2_{L^2(d\mu)} = \int_{\mathbb{R}^d} \left|f(x) - f^\sharp(x)\right|^2 \, d\mu(x),$$

with high probability, where μ is defined by the sampling measure for the data x (see Assumption 1).

THEOREM 2. [**Least Squares Risk Bound** $m > N$] For any $f \in \mathcal{F}(\phi, \rho)$, let the data $\{\boldsymbol{x}_j\}_{j\in[m]}$, weights $\{\boldsymbol{\omega}_k\}_{k\in[N]}$ and random feature matrix $\boldsymbol{A} \in \mathbb{C}^{m\times N}$ satisfy Assumption 1, and let the noise satisfy Assumption 2. If for some $\eta, \delta > 0$ the following conditions hold:

$$m \geq C\eta^{-2} N \log\left(\frac{2N}{\delta}\right)$$
$$\sqrt{\delta}\eta \left(4\gamma^2\sigma^2 + 1\right)^{\frac{d}{4}} \geq N,$$

where $C > 0$ is a universal constant independent of the dimension d; then with probability at least $1 - 8\delta$ the following risk bound holds:

$$R(f^\sharp) \leq 16\, K(\eta) \left(1 + Nm^{-\frac{1}{2}} \log^{\frac{1}{2}}\left(\frac{1}{\delta}\right)\right) (\epsilon^2 \|f\|^2_\rho + E^2), \tag{4.3}$$

where $K(\eta) := \frac{1+3\eta}{1-3\eta}$ and

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

The proof of Theorem 2 is given in Appendix D. Note that $K(\eta)$ is an upper bound for the condition number of the matrix $m^{-1}\boldsymbol{A}^*\boldsymbol{A}$ using Theorem 1.

REMARK 3. In the noiseless case, the risk bound that appears in Theorem 2 scales like $\mathcal{O}\left(N^{-1} + m^{-\frac{1}{2}}\right)$, where the order hides the dimension, parameters and constants.

4.3 *Overparameterized regime, min-norm interpolator*

In the overparameterized (underdetermined) regime, we have more features than data, and thus popular models for training the coefficient vector $\boldsymbol{c}^\sharp$ rely on minimizing $\|\boldsymbol{c}^\sharp\|_p$ under the constraint that $\boldsymbol{A}\boldsymbol{c}^\sharp - \boldsymbol{y}$ is zero or small. First, we consider training $\boldsymbol{c}^\sharp$ using the min-norm interpolation problem:

$$\boldsymbol{c}^\sharp \in \underset{\boldsymbol{A}\boldsymbol{c}=\boldsymbol{y}}{\operatorname{argmin}} \ \|\boldsymbol{c}\|_2. \tag{4.4}$$

The solution is given by $\boldsymbol{c}^\sharp = \boldsymbol{A}^\dagger\boldsymbol{y}$ where the psuedoinverse $\boldsymbol{A}^\dagger = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}$, noting that the inverse of the Gram matrix exists with high probability under the conditions of Theorem 1. This is also referred to as the ridgeless case, since $\boldsymbol{c}^\sharp = \boldsymbol{A}^\dagger\boldsymbol{y}$ can be viewed as the ridgeless limit of the coefficients, $\boldsymbol{c}^\sharp_\lambda$, which are obtained by the ridge regression problem

$$\boldsymbol{c}^\sharp_\lambda = \underset{\boldsymbol{c}\in\mathbb{R}^N}{\operatorname{argmin}} \ \frac{1}{m}\|\boldsymbol{A}\boldsymbol{c} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{c}\|_2^2, \tag{4.5}$$

i.e. the vector obtained when $\lambda \to 0^+$. With $f^\sharp$ defined as (4.2), we have the following result analogous to Theorem 2.

THEOREM 3. [**Min-Norm Risk Bound** $m < N$] For any $f \in \mathcal{F}(\phi, \rho)$, let the data $\{\boldsymbol{x}_j\}_{j\in[m]}$, weights $\{\boldsymbol{\omega}_k\}_{k\in[N]}$ and random feature matrix $\boldsymbol{A} \in \mathbb{C}^{m\times N}$ satisfy Assumption 1, and let the noise satisfy Assumption 2. If for some $\eta, \delta > 0$ the following conditions hold:

$$N \geq C\eta^{-2}m\log\left(\frac{2m}{\delta}\right)$$

$$\sqrt{\delta}\eta\left(4\gamma^2\sigma^2 + 1\right)^{\frac{d}{4}} \geq m,$$

where $C > 0$ is a universal constant independent of the dimension d, then there exists a constant $\tilde{C} > 0$ such that the following risk bound holds:

$$R(f^\sharp) \leq \tilde{C} \log^{\frac{1}{2}}\left(\frac{1}{\delta}\right)\left(m^{-\frac{1}{2}} + K(\eta)\, m^{\frac{1}{2}}\, \epsilon^2\right) \|f\|_\rho^2 + \tilde{C} m^{\frac{1}{2}}\, K(\eta)\, \log^{\frac{1}{2}}\left(\frac{1}{\delta}\right) E^2,$$

with probability at least $1 - 8\delta$, where $K(\eta) := \frac{1+3\eta}{1-3\eta}$ and

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

The proof of Theorem 3 is given in Appendix D. In this case, $K(\eta)$ is an upper bound for the condition number of the matrix $N^{-1}\boldsymbol{A}\boldsymbol{A}^*$ using Theorem 1.

REMARK 4. In the noiseless case, the risk bound that appears in Theorem 3 scales like $\mathcal{O}\left(m^{-\frac{1}{2}}\right)$, where the order hides the dimension, parameters and constants.

4.4 *Overparameterized regime, sparse regression*

Next, we consider the sparse regression setting, where we minimize the ℓ^1 norm of the coefficients. The goal is to obtain a subset of the features which maintain an accurate approximation to the target function without needing to use the entire feature space, i.e. to obtain low complexity random feature models (see [17]). This is implicitly related to methods for pruning overparameterized networks and the *lottery ticket hypothesis* [15]. Specifically, it is conjectured that there are smaller subnetworks of overparameterized randomly trained networks which are accurate representation of the target function with lower complexity. Our results also provide an additional motivation based on the structure of the transform of f, i.e. the decay of $\alpha(\boldsymbol{\omega})$.

We consider the ℓ^1 basis pursuit denoising problem [14]

$$\boldsymbol{c}^\sharp \in \underset{\|\boldsymbol{A}\boldsymbol{c}-\boldsymbol{y}\|_2 \leq \xi\sqrt{m}}{\operatorname{argmin}} \|\boldsymbol{c}\|_1, \tag{4.6}$$

where ξ is a parameter related to the noise level (taking into account model inaccuracy as well). The constraint in (4.6) is inexact compared with the interpolation condition used in (4.4). In addition, one cannot guarantee the existence of exact sparse solutions of $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{y}$ in the presence of approximation error and measurement noise.

In this setting, we modify the approximation $f^\sharp$ to incorporate a pruning step as done in [17]. Suppose that $\boldsymbol{c}^\sharp$ is a solution of (4.6); we define $f^\sharp$ by

$$f^\sharp(\boldsymbol{x}) = \sum_{k \in \mathcal{S}^\sharp} c_k^\sharp \phi(\boldsymbol{x}, \boldsymbol{\omega}_k), \tag{4.7}$$

where $\mathcal{S}^\sharp$ is the index set of the s-largest absolute entries of $\boldsymbol{c}^\sharp$. This guarantees that the approximation only depends on at most $s < N$ random features. For any vector $\boldsymbol{c}$, we denote by $\vartheta_{s,p}(\boldsymbol{c})$ the ℓ^p-error of

the best s-term approximation to $\boldsymbol{c}$, i.e.

$$\vartheta_{s,p}(\boldsymbol{c}) := \inf\{\|\boldsymbol{c} - \tilde{\boldsymbol{c}}\|_p : \tilde{\boldsymbol{c}} \text{ is } s \text{ sparse}\}.$$

This is a measure of the compressibility of a vector $\boldsymbol{c}$ with respect to the ℓ^p norm and will appear in the risk bounds as a source of error from the sparsification.

To measure the conditioning of the system with respect to sparse recovery, we define the restricted isometry constant of a matrix $\boldsymbol{A} \in \mathbb{C}^{m\times N}$ using the notation from [14]. For an integer $s \leq N$, the s-th RIP constant of $\boldsymbol{A}$, denoted by $\delta_s = \delta_s(\boldsymbol{A})$, is the smallest $\delta \geq 0$ such that

$$(1-\delta)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq (1+\delta)\|\boldsymbol{x}\|_2^2$$

holds for all s-sparse $\boldsymbol{x}$. We have the following estimate of the s-RIP constant of the random feature matrix $\boldsymbol{A}$.

THEOREM 4. [**Estimate on Restricted Isometry Constants**] Assume that the data $\{\boldsymbol{x}_j\}_{j\in[m]}$, weights $\{\boldsymbol{\omega}_k\}_{k\in[N]}$ and random feature matrix $\boldsymbol{A} \in \mathbb{C}^{m\times N}$ satisfy Assumption 1. For $\eta_1, \eta_2, \delta \in (0, 1/2)$ and some integer $s \geq 1$, if

$$\begin{aligned} m &\geq C_1 \eta_1^{-2}\, s\, \log(\delta^{-1}) \\ m &\geq C_2 \eta_2^{-2}\, s\, \log^3(s)\log(N) \\ \sqrt{\delta}\,\eta_1\,(4\gamma^2\sigma^2+1)^{\frac{d}{4}} &\geq N, \end{aligned}$$

where C_1 and C_2 are universal positive constants, then with probability at least $1-2\delta$, the s-RIP constant $\delta_s\left(\frac{1}{\sqrt{m}}A\right)$ is bounded by $f(\eta_1,\eta_2)$, where

$$f(\eta_1,\eta_2) := 3\eta_1 + \eta_2^2 + \sqrt{2}\eta_2.$$

With the above theorem, we can derive the following result.

THEOREM 5. [**Sparse Regression Risk Bounds**] For any $f \in \mathcal{F}(\phi,\rho)$, let the data $\{\boldsymbol{x}_j\}_{j\in[m]}$, weights $\{\boldsymbol{\omega}_k\}_{k\in[N]}$ and random feature matrix $\boldsymbol{A} \in \mathbb{C}^{m\times N}$ satisfy Assumption 1, and let the noise satisfy Assumption 2. Let $c^\sharp$ be obtained from (4.6) with $\xi = \sqrt{2(\epsilon^2\|f\|_\rho + E^2)}$. If the following conditions hold:

$$\begin{aligned} m &\geq Cs\log^3(2s)\log(N) \\ \frac{\sqrt{\delta}}{10}\left(4\gamma^2\sigma^2+1\right)^{\frac{d}{4}} &\geq N \\ \delta &\geq N^{-\log^3(2s)}, \end{aligned}$$

then with probability at least $1 - 8\delta$ the following risk bound holds:

$$R(f^\sharp) \le C' \left(1 + Nm^{-\frac{1}{2}} \log^{\frac{1}{2}}\left(\frac{1}{\delta}\right)\right) (\epsilon^2 \|f\|_\rho^2 + E^2) + C'' \left(1 + Nm^{-\frac{1}{2}} s^{-1} \log^{\frac{1}{2}}\left(\frac{1}{\delta}\right)\right) \vartheta_{s,1}(\boldsymbol{c}^\star)^2,$$

where $f^\sharp$ is defined as (4.7), $\boldsymbol{c}^\star$ is the vector whose components are defined by $\boldsymbol{c}_k^\star = \frac{1}{N}\frac{\alpha(\boldsymbol{\omega}_k)}{\rho(\boldsymbol{\omega}_k)}$, for $k \in [N]$ and

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

The constants $C, C', C'' > 0$ are universal constants independent of the dimension d.

The proofs are given in Appendix D. Note that in the worst case setting:

$$\vartheta_{s,1}(\boldsymbol{c}^\star) \le \left(1 - \frac{s}{N}\right) \|f\|_\rho.$$

REMARK 5. The sparsity parameter s in Theorem 5 can be defined by the user since it is obtained by the pruning step used in (4.7). When $s = N$, $\vartheta_{N,1}(\boldsymbol{c}^\star) = 0$ and thus Theorem 3 and Theorem 5 generally agree in terms of the risk's dependency on $\epsilon^2\|f\|_\rho^2$ and E^2. On the other hand, if $\alpha(\boldsymbol{\omega})$ decays much faster than $\rho(\boldsymbol{\omega})$ or has small support within the support set of $\rho(\boldsymbol{\omega})$, then the vector $\boldsymbol{c}^\star$ will be compressible and the sparse regression problem would be better.

5. Conditioning in three regimes

In this section, we prove Theorem 1. The proofs for Theorem 1 Part (a) and (b) depend on the matrix Bernstein inequality Lemma A.2, and Part (c) follows from a more direct estimate on the expectation of the eigenvalues of the Gram matrix.

Proof of Theorem 1. To prove Part (a), we bound each of the following three terms:

$$\begin{aligned}\left|\lambda_k\left(\frac{1}{m}\boldsymbol{A}^*\boldsymbol{A}\right) - 1\right| &\le \left\|\frac{1}{m}\boldsymbol{A}^*\boldsymbol{A} - \boldsymbol{I}_N\right\|_2 \\ &\le \frac{1}{m}\left\|\sum_{\ell=1}^m (\boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)\right\|_2 \\ &\quad + \frac{1}{m}\left\|\sum_{\ell=1}^m (\mathbb{E}_{\boldsymbol{x}}\boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)\right\|_2 + \|\boldsymbol{L}\|_2, \end{aligned} \tag{5.1}$$

by $\eta > 0$, where $\boldsymbol{X}_\ell$ is the ℓ-th column of $\boldsymbol{A}^*$ and $\boldsymbol{L} := \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*) - \boldsymbol{I}_N$.

First, $\boldsymbol{L}$ is a symmetric matrix with $L_{j,j} = 0$ and

$$L_{j,k} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}[\exp(i\langle \boldsymbol{x}_\ell, \boldsymbol{\omega}_k - \boldsymbol{\omega}_j\rangle)] = (2\gamma^2\sigma^2 + 1)^{-\frac{d}{2}},$$

for $j, k \in [N], j \neq k$. Let $\boldsymbol{z}$ be a unit vector, then by Hölder's inequality

$$\begin{aligned}
|\langle \boldsymbol{L}\boldsymbol{z}, \boldsymbol{z}\rangle| = \left| \sum_{\substack{j,k=1 \\ j\neq k}}^{N} \left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{\frac{d}{2}} z_j \overline{z_k} \right| &\leq \left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{\frac{d}{2}} \|z\|_1^2 \\
&\leq N \left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{\frac{d}{2}} \\
&\leq N \left(\frac{1}{4\gamma^2\sigma^2+1}\right)^{\frac{d}{4}} \\
&\leq \sqrt{\delta}\eta,
\end{aligned}$$

where assumption (3.2) was used in the last inequality. Therefore, the third term in (5.1) is bounded by $\|\boldsymbol{L}\|_2 \leq \sqrt{\delta}\eta \leq \eta$.

Next, define the random variable Z by

$$Z(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N) := \|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\|_2.$$

Since $\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)$ depends only on $\{\boldsymbol{\omega}_k\}_{k\in[N]}$, the random variable Z is independent of data $\{\boldsymbol{x}_j\}_{j\in[m]}$. Note that Z is the norm of a dependent random matrix, and thus we cannot guarantee an exponential concentration rate and instead we apply Markov's inequality. The second moment of Z is bounded by

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\omega}}(Z(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N)^2) &= \mathbb{E}_{\boldsymbol{\omega}} \|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\|_2^2 \\
&\leq \mathbb{E}_{\boldsymbol{\omega}} \|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\|_F^2 \\
&= \sum_{\substack{j,k=1 \\ j\neq k}} \mathbb{E}_{\boldsymbol{\omega}} \left| \exp\left(-\frac{\gamma^2}{2}\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_2^2\right) - \left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{\frac{d}{2}} \right|^2 \\
&= \sum_{\substack{j,k=1 \\ j\neq k}} \left(\mathbb{E}_{\boldsymbol{\omega}} \exp\left(-\gamma^2\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_2^2\right) - \left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{d} \right) \\
&\leq N^2 \left(\frac{1}{4\gamma^2\sigma^2+1}\right)^{\frac{d}{2}}.
\end{aligned}$$

By Markov's inequality and the complexity assumption (3.2),

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\omega}}(Z(\boldsymbol{\omega}) \geq \eta) &\leq \mathbb{P}_{\boldsymbol{\omega}}\left(Z(\boldsymbol{\omega}) \geq \frac{N}{\sqrt{\delta}}\left(\frac{1}{4\gamma^2\sigma^2+1}\right)^{\frac{d}{4}}\right)\\ &\leq \frac{\delta(4\gamma^2\sigma^2+1)^{\frac{d}{2}}}{N^2}\,\mathbb{E}_{\boldsymbol{\omega}}(Z(\boldsymbol{\omega})^2)\\ &\leq \delta.\end{aligned}$$

This implies that with probability at least $1-\delta$ with respect to the draw of $\{\boldsymbol{\omega}_k\}_{k\in[N]}$, we have $\|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)\|_2 \leq \eta$.

Next, we condition the remaining term in (5.1) on the draw of $\{\boldsymbol{\omega}_k\}_{k\in[N]}$ where we have shown that $\|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)\|_2 \leq \eta$. Specifically, define the random matrices $\{\boldsymbol{Y}_\ell\}_{\ell\in[m]}$ by

$$\boldsymbol{Y}_\ell = \boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*),$$

where each $\boldsymbol{Y}_\ell$ depends only on $\{\boldsymbol{x}_\ell\}_{\ell\in[m]}$, conditioned on the given draw of $\{\boldsymbol{\omega}_k\}_{k\in[N]}$. The matrix $\boldsymbol{Y}_\ell$ satisfies $(\boldsymbol{Y}_\ell)_{j,j} = 0$ and $(\boldsymbol{Y}_\ell)_{j,k} = \exp(i\langle\boldsymbol{x}_\ell, \boldsymbol{\omega}_k - \boldsymbol{\omega}_j\rangle) - \exp(-\gamma^2\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_2^2/2)$ for $j,k \in [N], j \neq k$. The norms are bounded by

$$\|\boldsymbol{Y}_\ell\|_2 \leq \max_{j\in[N]} \sum_{\substack{k=1\\k\neq j}}^{N} |e^{i\langle\boldsymbol{x}_\ell,\boldsymbol{\omega}_k-\boldsymbol{\omega}_j\rangle} - e^{-\frac{\gamma^2}{2}\|\boldsymbol{\omega}_k-\boldsymbol{\omega}_j\|_2^2}| \leq 2N \quad \ell \in [m],$$

using Gershgorin's theorem. By the previous results, we have

$$\begin{aligned}\|\mathbb{E}_{\boldsymbol{x}}\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*\|_2 &\leq \|\mathbb{E}_{\boldsymbol{x}}\boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*\|_2 + \|\mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*\|_2\\ &\leq \eta + \|\boldsymbol{I}_N + \boldsymbol{L}\|_2\\ &\leq 2\eta + 1\end{aligned}$$

and thus

$$\begin{aligned}\left\|\sum_{\ell=1}^{m}\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{Y}_\ell^2)\right\|_2 &\leq \sum_{\ell=1}^{m}\left\|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{Y}_\ell^2)\right\|_2\\ &\leq \sum_{\ell=1}^{m}\left\|N\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*) - (\mathbb{E}_{\boldsymbol{x}}\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)^2\right\|_2\\ &\leq \sum_{\ell=1}^{m}\left(N\|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)\|_2 + \|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)\|_2^2\right)\\ &\leq m[N(1+2\eta) + (1+2\eta)^2],\end{aligned}$$

noting that $\boldsymbol{X}_\ell^* \boldsymbol{X}_\ell = N$. Applying Lemma A.2 (by setting the parameters in the theorem's statement to $\sigma^2 = m\left(\frac{5}{3}N + \frac{25}{9}\right)$ and $K = 2N$ and assuming that $\eta \le \frac{1}{3}$) yields

$$\mathbb{P}_{\boldsymbol{x}}\left(\frac{1}{m}\left\|\sum_{\ell=1}^{m}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\right\|_2 \ge \eta\right) = \mathbb{P}_{\boldsymbol{x}}\left(\left\|\sum_{\ell=1}^{m}\boldsymbol{Y}_\ell\right\|_2 \ge m\eta\right)$$
$$\le 2N\exp\left(-\frac{m\eta^2}{\frac{10}{3}N + \frac{50}{9} + \frac{4N\eta}{3}}\right).$$

If $m \ge 4N\eta^{-2}\log(2N/\delta)$ (assuming $N > 25$), then with probability at least $1-\delta$ with respect to the draw of $\{\boldsymbol{x}_j\}_{j\in[m]}$, we have $m^{-1}\|\sum_{\ell=1}^m \boldsymbol{X}_\ell \boldsymbol{X}_\ell^* - \sum_{\ell=1}^m \mathbb{E}_{\boldsymbol{x}}\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*\|_2 \le \eta$.

Altogether, (5.1) is bounded by

$$\left|\lambda_k\left(\frac{1}{m}\boldsymbol{A}^*\boldsymbol{A}\right) - 1\right| \le 3\eta,$$

with probability at least $(1-\delta)^2 \ge 1 - 2\delta$ if the conditions in the previous steps are satisfied.

For Part (b), note that the samples $\{\boldsymbol{x}_j\}_{j\in[m]}$ and the weights $\{\boldsymbol{\omega}_k\}_{k\in[N]}$ take real values and their distributions are symmetric about $\boldsymbol{0}$. Therefore, the feature matrix $\boldsymbol{A} = [a_{j,k}]$ where $a_{j,k} := \exp(i\langle \boldsymbol{x}_j, \boldsymbol{\omega}_k\rangle)$ and its conjugate transpose $\boldsymbol{A}^* = [\overline{a}_{k,j}]$ where $\overline{a}_{k,j} = \exp(-i\langle \boldsymbol{\omega}_k, \boldsymbol{x}_j\rangle)$ have the same distribution. Arguing similarly as in the proof of Part (a) proves Part (b) as well. Essentially, one can consider the system where the meaning of $\boldsymbol{\omega}$ and $\boldsymbol{x}$ is switched.

For Part (c), consider the case when $m = N$. The matrix $N^{-1}\boldsymbol{A}^*\boldsymbol{A}$ can be written as the following rank-1 decomposition:

$$\frac{1}{N}\boldsymbol{A}^*\boldsymbol{A} = \frac{1}{N}\sum_{\ell=1}^{N}\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*,$$

where $\boldsymbol{X}_\ell$ is the ℓ-th column of $\boldsymbol{A}^*$. Since the Gram matrix, $N^{-1}\boldsymbol{A}^*\boldsymbol{A}$, is Hermitian we can use the Rayleigh quotient to bound the maximum eigenvalue from below by

$$\lambda_{\max}\left(\frac{1}{N}\boldsymbol{A}^*\boldsymbol{A}\right) \ge \frac{1}{N}\langle \boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v}, \boldsymbol{v}\rangle \quad \text{for all } \boldsymbol{v} \in \mathbb{C}^N \text{ with } \|\boldsymbol{v}\|_2 = 1.$$

Thus setting $\boldsymbol{v} = \frac{1}{\sqrt{N}}\boldsymbol{X}_1$ yields

$$\begin{aligned}\lambda_{\max}\left(\frac{1}{N}\boldsymbol{A}^*\boldsymbol{A}\right) &\geq \frac{1}{N^2}\sum_{\ell=1}^{N}\boldsymbol{X}_1^*\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*\boldsymbol{X}_1\\ &= \frac{1}{N^2}\left(N^2+\sum_{\ell=2}^{N}\boldsymbol{X}_1^*\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*\boldsymbol{X}_1\right)\\ &= 1+\frac{1}{N^2}\sum_{\ell=2}^{N}\sum_{j,k=1}^{N}\exp(i\langle \boldsymbol{x}_1-\boldsymbol{x}_\ell,\boldsymbol{\omega}_j-\boldsymbol{\omega}_k\rangle)\\ &= 1+\frac{(N-1)N}{N^2}+\frac{1}{N^2}\sum_{\ell=2}^{N}\sum_{\substack{j,k=1\\ j\neq k}}^{N}\exp(i\langle \boldsymbol{x}_1-\boldsymbol{x}_\ell,\boldsymbol{\omega}_j-\boldsymbol{\omega}_k\rangle).\end{aligned}$$

Taking the expectation on both sides yields

$$\mathbb{E}\lambda_{\max}\left(\frac{1}{N}\boldsymbol{A}^*\boldsymbol{A}\right) \geq 2-\frac{1}{N}+\frac{(N-1)^2}{N}\left(\frac{1}{\sqrt{4\gamma^2\sigma^2+1}}\right)^d \geq 2-\frac{1}{N},$$

where we used the characteristic function of the normal distribution.

Similarly, for the smallest eigenvalue, we have

$$\lambda_{\min}\left(\frac{1}{N}\boldsymbol{A}^*\boldsymbol{A}\right) \leq \frac{1}{N}\langle \boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v},\boldsymbol{v}\rangle \quad \text{for all } \boldsymbol{v}\in\mathbb{C}^N \text{ with } \|\boldsymbol{v}\|_2 = 1.$$

Since $\{\boldsymbol{X}_\ell\}_{\ell\in[N]}$ are vectors in n-dimensional space $\mathbb{C}^N$, we can find a unit vector $\boldsymbol{u} = [u_1,\dots,u_N]^T \in \mathbb{C}^N$ such that $\boldsymbol{u}$ is orthogonal to span $\left(\{\boldsymbol{X}_\ell\}_{\ell\in[N-1]}\right)$ and thus

$$\begin{aligned}\lambda_{\min}\left(\frac{1}{N}\boldsymbol{A}^*\boldsymbol{A}\right) &\leq \frac{1}{N}\langle \boldsymbol{A}^*\boldsymbol{A}\boldsymbol{u},\boldsymbol{u}\rangle\\ &\leq \frac{1}{N}\boldsymbol{u}^*\boldsymbol{X}_N\boldsymbol{X}_N^*\boldsymbol{u}\\ &= \frac{1}{N}\sum_{j,k=1}^{N}\overline{u_j}u_k\exp(i\langle \boldsymbol{x}_N,\boldsymbol{\omega}_k-\boldsymbol{\omega}_j\rangle).\end{aligned}$$

The vector $\boldsymbol{u}$ depends on $\{\boldsymbol{X}_\ell\}_{\ell\in[N-1]}$, thus the components u_j for $j\in[N]$ are random variables depending on $\{\boldsymbol{x}_j\}_{j\in[N-1]}$ and weights $\{\boldsymbol{\omega}_k\}_{\ell\in[N]}$ but are independent of $\boldsymbol{x}_N$. By taking the expectation on both sides

and applying the Fubini's theorem, we bound the expectation on the minimum eigenvalue by

$$\begin{aligned}
\mathbb{E}\lambda_{\min}\left(\frac{1}{N}\boldsymbol{A}^*\boldsymbol{A}\right) &\leq \frac{1}{N} + \frac{1}{N}\,\mathbb{E}\sum_{j\neq k}\overline{u_j}u_k\exp(i\langle \boldsymbol{x}_N, \boldsymbol{\omega}_k - \boldsymbol{\omega}_j\rangle)\\
&= \frac{1}{N} + \frac{1}{N}\,\mathbb{E}_{\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_N,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{N-1}}\sum_{j\neq k}\overline{u_j}u_k\mathbb{E}_{\boldsymbol{x}_N}[\exp(i\langle \boldsymbol{x}_N, \boldsymbol{\omega}_k - \boldsymbol{\omega}_j\rangle)]\\
&= \frac{1}{N} + \frac{1}{N}\,\mathbb{E}_{\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_N,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{N-1}}\sum_{j\neq k}\overline{u_j}u_k\exp\left(-\frac{\gamma^2}{2}\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_2^2\right)\\
&\leq \frac{1}{N} + \frac{1}{N}\mathbb{E}_{\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_N}\sqrt{\sum_{j\neq k}\exp\left(-\gamma^2\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_2^2\right)}\\
&\leq \frac{1}{N} + \frac{1}{N}\sqrt{\sum_{j\neq k}\mathbb{E}_{\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_N}\exp\left(-\gamma^2\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_2^2\right)}\\
&\leq \frac{1}{N} + \left(1 - \frac{1}{N}\right)^{\frac{1}{2}}\left(\frac{1}{4\gamma^2\sigma^2 + 1}\right)^{\frac{d}{4}},
\end{aligned}$$

where in the fourth line we use the Cauchy–Schwarz inequality to bound

$$\begin{aligned}
\sum_{j\neq k}\overline{u_j}u_k\exp\left(-\frac{\gamma^2}{2}\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_2^2\right) &\leq \sqrt{\sum_{j\neq k}|u_j|^2|u_k|^2}\sqrt{\sum_{j\neq k}\exp(-\gamma^2\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_2^2)}\\
&\leq \sqrt{\|u\|_2^4}\sqrt{\sum_{j\neq k}\exp(-\gamma^2\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_2^2)}\\
&\leq \sqrt{\sum_{j\neq k}\exp(-\gamma^2\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_2^2)},
\end{aligned}$$

and in the fifth line we use the Jensen's inequality. This completes the proof. □

6. Discussion

We analyse the double descent phenomenon [6] for random feature regression by relating it to the condition number of asymmetric rectangular random matrices and deriving high-probability bounds on the extreme singular values. The technical arguments rely on random matrix theory, specifically, deriving a concentration bound on eigenvalues of the Gram matrix (or the restricted isometry constants) for the various complexity setting. The bounds improve on previous results in the literature and give a refined picture of the test error landscape as function of the complexity. In the interpolation regime, we directly derive a lower bound on the condition number, showing that the linear system becomes ill-conditioned when $N = m$. We provide risk bounds which are controlled by the conditioning of the random feature matrix, thereby relating the generalization error with the conditioning of the system. The

risk bounds include the least squares, min-norm interpolation and sparse regression problems. While our analysis focused on Fourier features with normally distributed weights and samples in order to provide neater bounds, the proofs could be extended to other features and probability distributions. For example, if we change the feature map, the constants in our results would need to include the L^∞ norm of the new activation function. Additionally, using other probability distributions (e.g. uniform or subgaussian) would lead to changes in the constants and slight changes in the scalings. When $\boldsymbol{W}$ and $\boldsymbol{X}$ are sampled from different distributions, the symmetry between the underparameterized and overparameterized results breakdown but the main conclusions should still hold (i.e. constants may change, but the overall scaling should remain valid).

The connections between random feature models and fully trained neural networks are rich, and the precise translation of our results in this setting is left as a future discussion. For deep networks, weight initialization and normalization can help to avoid the issue of vanishing gradients. That is, if the Jacobian of the hidden layers are close to being an isometry then the network is *stable* in the dynamical sense [16,29,37,41]. Based on our results, since the spectrum of random feature maps concentrate near 1, one could show that randomization helps to provide dynamic stability within each layer of a deep neural network.

A consequence of the theory presented in this work is that the random feature matrix does not need regularization to be well-conditioned. This is in align with the results in previous works. Thus the risk bounds in the noiseless case can decrease to zero without adding a penalty to the training problem. However, it may be the case that regularization improves generalization in the presence of noise and outliers. We leave a more detailed analysis of unconstrained regularized methods (for example, the ridge regression problem) with noisy data for future work.

Funding

Air Force Office of Scientific Research Multidisciplinary University Research Initiative [FA9550-21-1-0084]; National Science Foundation Division of Mathematical Sciences [1752116].

Data availability statement

No new data were generated or analysed in support of this research.

References

1. Advani, M. S., Saxe, A. M. & Sompolinsky, H. (2020) High-dimensional dynamics of generalization error in neural networks. *Neural Netw.*, **132**, 428–446.
2. Ba, J., Erdogdu, M., Suzuki, T., Denny W. & Zhang, T. (2020) Generalization of two-layer neural networks: An asymptotic viewpoint. In *International conference on learning representations*. https://openreview.net/forum?id=H1gBsgBYwH.
3. Bach, F. (2017) On the equivalence between kernel quadrature rules and random feature expansions. *J. Mach. Learn. Res.*, **18**, 714–751.
4. Bartlett, P. L., Long, P. M., Lugosi, G. & Tsigler, A. (2020) Benign overfitting in linear regression. *Proc. Natl. Acad.Sci.*, **117**, 30063–30070.
5. Belkin, M., Hsu, D., Ma, S. & Mandal, S. (2019) Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci.*, **116**, 15849–15854.
6. Belkin, M., Hsu, D. & Ji, X. (2020) Two models of double descent for weak features. *SIAM J. Math. Data Sci.*, **2**, 1167–1180.

7. Belkin, M., Ma, S. & Mandal, S. (2018) To understand deep learning we need to understand kernel learning. In Dy, J. & Krause, A. (eds). *Proceedings of the 35th International Conference on Machine Learning*. **80**, 541–549. PMLR.
8. Belkin, M., Rakhlin, A. & Tsybakov, A. B. (2019) Does data interpolation contradict statistical optimality? In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. (Chaudhuri, K. & Sugiyama, M. eds.) **89**, 1611–1619. PMLR.
9. Benigni, L. & Péché, S. (2019) Eigenvalue distribution of nonlinear models of random matrices. *Electron. J. Probab.*, **26**, 1–37. The Institute of Mathematical Statistics and the Bernoulli Society.
10. Cheng, X. & Singer, A. (2013) The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, **02**, 1350010.
11. Chao, M., Lei, W. & Weinan, E. (2020) The slow deterioration of the generalization error of the random feature model. *Mathematical and Scientific Machine Learning*. *Mathematical and Scientific Machine Learning*. pp. 373–389. PMLR.
12. El Karoui, N. (2010) The spectrum of kernel random matrices. *Ann. Stat.*, **38**, 1–50.
13. Fan, Z. & Montanari, A. (2019) The spectral norm of random inner-product kernel matrices. *Probab. Theory Relat. Fields*, **173**, 27–85.
14. Foucart, S. & Rauhut, H. (2013) *A Mathematical Introduction to Compressive Sensing*. Birkhäuser: Applied and Numerical Harmonic Analysis.
15. Frankle, J. & Carbin, M. (2019) The lottery ticket hypothesis: finding sparse, trainable neural networks. *International Conference on Learning Representations*. https://openreview.net/forum?id=rJl-b3RcF7.
16. Haber, E. & Ruthotto, L. (2017) Stable architectures for deep neural networks. *Inverse Probl.*, **34**, 014004.
17. Hashemi, A., Schaeffer, H., Shi, R., Topcu, U., Tran, G. & Ward, R. (2023) Generalization bounds for sparse random feature expansions. *Appl. Comput. Harmon. Anal.*, **62**, 310–330. Elsevier.
18. Hastie, T., Montanari, A., Rosset, S. & Tibshirani, R. J. (2022) Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Stat.*, **50**, 949. NIH Public Access.
19. Kan, K., Nagy, J. G. & Ruthotto, L. (2020) Avoiding the double descent phenomenon of random feature models using hybrid regularization. arXiv preprint arXiv:2012.06667.
20. Li, Z., Ton, J.-F., Oglic, D. & Sejdinovic, D. (2019) Towards a unified analysis of random fourier features. In *International conference on machine learning*, pp 3905–3914. PMLR.
21. Liang, T. & Rakhlin, A. (2020) Just interpolate: kernel "ridgeless" regression can generalize. *Ann. Stat.*, **48**, 1329–1347.
22. Liang, T., Rakhlin, A. & Zhai, X. (2020) On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pp 2683–2711. PMLR.
23. Liao, Z. & Couillet, R. (2018) On the spectrum of random features maps of high dimensional data. *International Conference on Machine Learning*. pp. 3063–3071. PMLR.
24. Liao, Z., Couillet, R. & Mahoney, M. W. (2020) A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. *Adv. Neural. Inf. Process. Syst*., **33**, 13939–13950.
25. Louart, C., Liao, Z. & Couillet, R. (2018) A random matrix approach to neural networks. *Ann. Appl. Probab.*, **28**, 1190–1248.
26. Mei, S., Misiakiewicz, T. & Montanari, A. (2021) Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Appl. Comput. Harmon. Anal.*, **59**, 3–84.
27. Mei, S. & Montanari, A. (2022) The generalization error of random features regression: precise asymptotics and double descent curve. *Commun. Pure Appl. Math.*, **75**, 667–766. Wiley Online Library.
28. Pastur, L. (2020) On random matrices arising in deep neural networks. Gaussian case. arXiv preprint arXiv:2001.06188.
29. Pennington, J., Schoenholz, S.S. & Ganguli, S. (2017) Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp 4788–4798.
30. Pennington, J. & Worah, P. (2019) Nonlinear random matrix theory for deep learning. *J. Stat. Mech.: Theory Exp.*, **2019**, 124005.

31. Poggio, T., Kur, G. & Banburski, A. (2019) Double descent in the condition number. arXiv preprint arXiv:1912.06190.
32. Rahimi, A. & Recht, B. (2007) Random features for large-scale kernel machines. *Advances in neural information processing systems*, **20**.
33. Rahimi, A. and Recht, B. (2008) Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp 555–561. IEEE.
34. Rahimi, A. & Recht, B. (2008) Weighted sums of random kitchen sinks: replacing minimization with randomization in learning. *Advances in neural information processing systems*, **21**.
35. Rudi, A., Camoriano, R. & Rosasco, L. (2015) Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, **28**.
36. Rudi, A. and Rosasco, L.. (2017) Generalization properties of learning with random features. In *dvances in Neural Information Processing Systems*. (Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. & Garnett, R. eds.) **30**, pp 3218–3228. Curran Associates, Inc.
37. Sun, Y., Zhang, L. & Schaeffer, H. (2020) Neupde: Neural network based ordinary and partial differential equations for modeling time-dependent data. *Mathematical and Scientific Machine Learning*. PMLR, pp. 352–372.
38. Tsigler, A. & Bartlett, P. L. (2023) Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, **24**, 1–76.
39. Weinan, E., Ma, C., Wojtowytsch, S. & Wu, L. (2020) Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *CSIAM Transactions on Applied Mathematics*, **1**, 561–615.
40. Yen, I. E.-H., Lin, T.-W., Lin, S.-D., Ravikumar, P. K. & Dhillon, I. S. (2014) Sparse random feature algorithm as coordinate descent in hilbert space. *Advances in Neural Information Processing Systems*, **27**.
41. Zhang, L. & Schaeffer, H. (2020) Forward stability of resnet and its variants. *J. Math. Imaging Vision*, **62**, 328–351.

A. Key Lemmata

In this section we present some lemmata that are used in the proof of our main results.

We first introduce a characterization of RIP constant and some notation. For a subset $S \subset [N]$, let $\boldsymbol{A}_S$ denote the submatrix of $\boldsymbol{A}$ consisting of only the columns indexed by the set S. The s-th RIP constant $\delta_s(\boldsymbol{A})$ can be characterized by the maximization problem

$$\delta_s(\boldsymbol{A}) = \max_{S\subset[N],\mathrm{card}(S)=s} \|\boldsymbol{A}_S^*\boldsymbol{A}_S - \boldsymbol{I}_s\|_2.$$

To measure the sparsity, we denote $\|z\|_0$ to be the number of non-zero entries of the vector z. Let $D_{s,N}$ be the union of all unit balls of dimension s in the ambient dimension N i.e.

$$D_{s,N} := \{z \in \mathbb{C}^N : \|z\|_2 \le 1, \|z\|_0 \le s\},$$

define the following seminorm on $\mathbb{C}^{N\times N}$:

$$|||\boldsymbol{B}||| := \sup_{z\in D_{s,N}} |\langle \boldsymbol{B}z, z\rangle|,$$

and thus the RIP constant $\delta_s(\boldsymbol{A}) = \left|\left|\left|\boldsymbol{A}^*\boldsymbol{A} - \boldsymbol{I}_N\right|\right|\right|$ [14].

The following lemma is crucial to the proof of Theorem 4. This is an improvement over Lemma 12.36 in [14].

LEMMA A.1. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_m$ be vectors in $\mathbb{C}^N$ with $\|\boldsymbol{X}_\ell\|_\infty \leq 1$ for all $\ell \in [m]$. For $p, q \geq 1$ such that $1/p + 1/q = 1$, we have

$$\mathbb{E}\left|\left|\left|\sum_{\ell=1}^m \epsilon_\ell \boldsymbol{X}_\ell \boldsymbol{X}_\ell^*\right|\right|\right| \leq C s^{\frac{p-1}{2p}} m^{\frac{1}{2q}} \sqrt{qs \log^2(s) \log\left(3 + \frac{N}{q}\right)} \left(\left|\left|\left|\sum_{\ell=1}^m \boldsymbol{X}_\ell \boldsymbol{X}_\ell^*\right|\right|\right|\right)^{\frac{1}{2p}}, \tag{A.1}$$

where ϵ_ℓ for $\ell \in [m]$ are independent Rademacher random variables and $C > 0$ is a universal constant.

The remaining lemmata are used throughout this work.

LEMMA A.2. [Corollary 8.15 from [14]] Let $\{\boldsymbol{X}_\ell\}_{\ell \in [m]}$ by a set of $\mathbb{C}^{N \times N}$ independent mean-zero self-adjoint random matrices, assume that $\|X_\ell\|_2 \leq K$ almost surely for all $\ell \in [m]$ and define

$$\sigma^2 := \left\|\sum_{\ell=1}^m \mathbb{E}(\boldsymbol{X}_\ell^2)\right\|.$$

Then, for all $t > 0$,

$$\mathbb{P}\left(\left\|\sum_{\ell=1}^m \boldsymbol{X}_\ell\right\|_2 \geq t\right) \leq 2N \exp\left(-\frac{t^2}{2\sigma^2 + \frac{2Kt}{3}}\right).$$

LEMMA A.3. [Theorem 8.42 and Remark 8.43(c) from [14]] Let $\mathcal{F}$ be a countable set of functions $f : \mathbb{C}^N \to \mathbb{R}$. Let X_ℓ for $\ell \in [m]$ be independent random vector in $\mathbb{C}^N$ such that $\mathbb{E}f(X_\ell) = 0$ and $f(X_\ell) \leq K$ a.s. for all $\ell \in [m]$ and for all $f \in \mathcal{F}$ for some constant $K > 0$ and let

$$Z = \sup_{f \in \mathcal{F}} \left|\sum_{\ell=1}^m f(X_\ell)\right|.$$

Let σ_ℓ^2 be non-zero such that $\mathbb{E}[\left(f(X_\ell)\right)^2] \leq \sigma_\ell^2$ for all $f \in \mathcal{F}$ and $\ell \in [m]$. Then, for all $t > 0$,

$$\mathbb{P}\left(Z \geq \mathbb{E}Z + t\right) \leq \exp\left(\frac{-t^2/2}{\sum_{\ell=1}^m \sigma_\ell^2 + 2K\mathbb{E}Z + tK/3}\right).$$

LEMMA A.4. [Theorem 9.24 in [17]] Suppose that the $2s$-RIP constant of the matrix $\boldsymbol{A} \in \mathbb{C}^{m \times N}$ satisfies

$$\delta_{2s}(\boldsymbol{A}) \leq \frac{4}{\sqrt{41}},$$

then for any vector $\boldsymbol{c}^* \in \mathbb{C}^N$ satisfying $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{c}^* + \boldsymbol{e}$ with $\boldsymbol{e} \leq \xi\sqrt{m}$, a minimizer $\boldsymbol{c}^\sharp$ of the BP problem (4.6) approximates the vector $\boldsymbol{c}^*$ with the error bounds

$$\|\boldsymbol{c}^* - \boldsymbol{c}^\sharp\|_1 \leq C'\kappa_{s,1}(\boldsymbol{c}^*) + C\sqrt{s}\xi$$
$$\|\boldsymbol{c}^* - \boldsymbol{c}^\sharp\|_2 \leq C'\frac{\kappa_{s,1}(\boldsymbol{c}^*)}{\sqrt{s}} + C\xi,$$

where $C, C' > 0$ are some constants. If we let $\mathcal{S}^\sharp$ be the index set of the s largest (in magnitude) entries of $\boldsymbol{c}^\sharp$, then we have

$$\|\boldsymbol{c}^* - \boldsymbol{c}^\sharp|_{\mathcal{S}^\sharp}\|_2 \leq C'\frac{\kappa_{s,1}(\boldsymbol{c}^*)}{\sqrt{s}} + C\xi + 4\vartheta_{s,2}(\boldsymbol{c}^*).$$

LEMMA A.5. [Lemma 8 in [17]] Suppose that $\boldsymbol{x}_j$ for $j \in [m]$ are sampled from the normal distribution $\mathcal{N}(\boldsymbol{0}, \gamma^2\boldsymbol{I}_d)$. Let $\mathbb{B}^d(R)$ be the ℓ^2-ball centred at $\boldsymbol{0}$ with radius $R > 0$. Then for $\delta \in (0, 1)$, the probability of $\boldsymbol{x}_j \in \mathbb{B}^d(R)$ for all $j \in [m]$ is at least $1 - \delta$ provided that

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

LEMMA A.6. [Lemma 9 in [17]] Fix confidence parameter $\delta > 0$ and accuracy parameter $\epsilon > 0$. Suppose $f \in \mathcal{F}(\phi, \rho)$ where $\phi(\boldsymbol{x}, \boldsymbol{\omega}) = \exp(i\langle\boldsymbol{x}, \boldsymbol{\omega}\rangle)$ and ρ is a probability distribution with finite second moment used for sampling the random weights $\boldsymbol{\omega}$. Consider a set $\mathcal{X} \subset \mathbb{R}^d$ with diameter $R = \sup_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|$. If the following holds:

$$N \geq \frac{4}{\epsilon^2}\left(1 + 4R\sqrt{\mathbb{E}\|\boldsymbol{\omega}\|_2^2} + \sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)}\right)^2,$$

then with probability at least $1 - \delta$ with respect to the draw of $\boldsymbol{\omega}_k$ for $k \in [N]$, the following holds:

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x}) - f^\star(\boldsymbol{x})| \leq \epsilon\|f\|_\rho,$$

where

$$f^\star(\boldsymbol{x}) = \sum_{k=1}^N c_k^\star \exp(i\langle\boldsymbol{x}, \boldsymbol{\omega}_k\rangle), \quad \text{with} \quad c_k^\star := \frac{\alpha(\boldsymbol{\omega}_k)}{N\rho(\boldsymbol{\omega}_k)}. \tag{A.2}$$

LEMMA A.7. [Lemma 1 in [17]] Fix the confidence parameter $\delta > 0$ and accuracy parameter $\epsilon > 0$. Suppose $f \in \mathcal{F}(\phi, \rho)$ where $\phi(\boldsymbol{x}, \boldsymbol{\omega}) = \exp(i\langle\boldsymbol{x}, \boldsymbol{\omega}\rangle)$ and ρ is a probability distribution with finite second

moment used for sampling the random weights $\boldsymbol{\omega}$. Suppose

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

then with probability at least $1-\delta$ with respect to the draw of $\boldsymbol{\omega}_j$ for $j \in [N]$, the following holds:

$$\|f - f^\star\|_{L^2(d\mu)} \leq \epsilon \|f\|_\rho,$$

where $f^\star$ is defined as in (A.2).

B. Proof of Lemma A.1

The proof of Lemma A.1 relies on Dudley's inequality. First, we introduce some notation. A stochastic process is a collection $\{X_t\}_{t\in T}$ of random variables indexed by some set T. The process $\{X_t\}_{t\in T}$ is called centerd if $\mathbb{E}X_t = 0$ for all $t \in T$. The pseudometric d associated with $\{X_t\}_{t\in T}$ is defined as

$$d(s,t) = \sqrt{\mathbb{E}|X_s - X_t|^2}, \quad s,t \in T.$$

A centred process is called subgaussian if

$$\mathbb{E}\exp(\theta(X_s - X_t)) \leq \exp(\theta^2 d(s,t)^2/2), \quad s,t \in T,$$

for some $\theta > 0$. For a set T, the covering number $\mathcal{N}(T,d,\delta)$ is the smallest integer N such that there exists a set $A \subset T$ with cardinality N and $\min_{s\in A} d(t,s) \leq \delta$ for all $t \in T$.

LEMMA A.8. [Dudley's inequality, Theorem 8.23 in [14]] Let $\{X_t\}_{t\in T}$ be a centred subgaussian process with associated pseudometric d. Then for any $t_0 \in T$,

$$\mathbb{E}\sup_{t\in T} X_t \leq 4\sqrt{2}\int_0^{\Delta(T)/2} \sqrt{\log(\mathcal{N}(T,d,u))}du$$

$$\mathbb{E}\sup_{t\in T} |X_t| \leq 4\sqrt{2}\int_0^{\Delta(T)/2} \sqrt{\log(2\mathcal{N}(T,d,u))}du,$$

where $\Delta(T) = \sup_{t\in T}\sqrt{\mathbb{E}|X_t|^2}$.

Proof of Lemma A.1. By the definition of $|||\cdot|||$,

$$E := \mathbb{E}\left|\left|\left|\sum_{\ell=1}^m \epsilon_\ell \boldsymbol{X}_\ell \boldsymbol{X}_\ell^*\right|\right|\right| = \mathbb{E}\sup_{z\in D_{s,N}}\left|\sum_{\ell=1}^m \epsilon_\ell |\langle \boldsymbol{X}_\ell, z\rangle|^2\right|.$$

To bound the expectation of the supremum of this Rademacher process, we will apply Lemma A.8. We need to estimate the covering number $\mathcal{N}(D_{s,N},d,t)$ with d being the pseudometric associated with the stochastic process.

For $z_1, z_2 \in D_{s,N}$,

$$\begin{aligned} d(z_1,z_2) &= \sqrt{\sum_{\ell=1}^{m} \left(|\langle \boldsymbol{X}_\ell, z_1\rangle|^2 - |\langle \boldsymbol{X}_\ell, z_2\rangle|^2 \right)^2} \\ &= \sqrt{\sum_{\ell=1}^{m} \left((z_1 - z_2)^* \boldsymbol{X}_\ell \boldsymbol{X}_\ell^* (z_1 + z_2) \right)^2} \\ &= \sqrt{\sum_{\ell=1}^{m} \left| \langle \boldsymbol{X}_\ell, z_1 + z_2\rangle \right|^2 \left| \langle \boldsymbol{X}_\ell, z_1 - z_2\rangle \right|^2} \\ &\le \left(\sum_{\ell=1}^{m} |\langle \boldsymbol{X}_\ell, z_1 + z_2\rangle|^{2p} \right)^{\frac{1}{2p}} \left(\sum_{\ell=1}^{m} |\langle \boldsymbol{X}_\ell, z_1 - z_2\rangle|^{2q} \right)^{\frac{1}{2q}} \\ &\le 2 \sup_{z \in D_{s,N}} \left(\sum_{\ell=1}^{m} |\langle \boldsymbol{X}_\ell, z\rangle|^{2p} \right)^{\frac{1}{2p}} \left(\sum_{\ell=1}^{m} |\langle \boldsymbol{X}_\ell, z_1 - z_2\rangle|^{2q} \right)^{\frac{1}{2q}}, \end{aligned} \tag{B.1}$$

where we use Hölder's inequality in the fourth line with some $p, q \ge 1$ such that $1/p + 1/q = 1$. Note that for any $z \in D_{s,N}$ and $\boldsymbol{X}_\ell$ we have

$$|\langle \boldsymbol{X}_\ell, z\rangle| \le \|\boldsymbol{X}_\ell\|_\infty \|z\|_1 \le \sqrt{s}.$$

Therefore,

$$\sup_{z \in D_{s,N}} \left(\sum_{\ell=1}^{m} |\langle \boldsymbol{X}_\ell, z\rangle|^{2p} \right)^{\frac{1}{2p}} \le s^{\frac{p-1}{2p}} \sup_{z \in D_{s,N}} \left(\sum_{\ell=1}^{m} |\langle \boldsymbol{X}_\ell, z\rangle|^{2} \right)^{\frac{1}{2p}}.$$

Introduce the seminorm $\|\cdot\|_{X,q}$ on $\mathbb{C}^N$

$$\|z\|_{X,q} := \left(\sum_{\ell=1}^{m} |\langle \boldsymbol{X}_\ell, z\rangle|^{2q} \right)^{\frac{1}{2q}},$$

and let

$$R := \sup_{z \in D_{s,N}} \left(\sum_{\ell=1}^{m} |\langle \boldsymbol{X}_\ell, z\rangle|^{2} \right)^{\frac{1}{2p}} = \left(\left\| \sum_{\ell=1}^{m} \boldsymbol{X}_\ell \boldsymbol{X}_\ell^* \right\| \right)^{\frac{1}{2p}},$$

we have $d(z_1, z_2) \leq 2Rs^{\frac{p-1}{2p}} \|z_1 - z_2\|_{X,q}$ for all $z_1, z_2 \in D_{s,N}$. The quantity $\Delta(D_{s,N})$ is bounded by

$$\Delta(D_{s,N}) = \sup_{z \in D_{s,N}} d(z, \mathbf{0}) \leq 2Rs^{\frac{p-1}{2p}} \sup_{z \in D_{s,N}} \|z\|_{X,q} \leq 2Rs^{\frac{p-1}{2p}} m^{\frac{1}{2q}} \sqrt{s}.$$

Apply Lemma A.8 we have

$$\begin{aligned} E &\leq 4\sqrt{2} \int_0^{\Delta(D_{s,N})/2} \sqrt{\log(2\mathcal{N}(D_{s,N}, d, u))} du \\ &\leq 8\sqrt{2}Rs^{\frac{p-1}{2p}} \int_0^{m^{\frac{1}{2q}}\sqrt{s}} \sqrt{\log(2\mathcal{N}(D_{s,N}, \|\cdot\|_{X,q}, t))} dt. \end{aligned} \tag{B.2}$$

Now we estimate the covering number $\mathcal{N}(D_{s,N}, \|\cdot\|_{X,q}, t)$. We first consider small t. For any $z_1, z_2 \in D_{s,N}$, $z_1 - z_2$ is at most $2s$-sparse, therefore,

$$|\langle \boldsymbol{X}_\ell, z_1 - z_2 \rangle| \leq \sqrt{2s} \|z_1 - z_2\|_2.$$

This shows that $\|\cdot\|_{X,q} \leq m^{\frac{1}{2q}} \sqrt{2s} \|\cdot\|_2$. Proposition C.3 from [14] then gives

$$\begin{aligned} \mathcal{N}(D_{s,N}, \|\cdot\|_{X,q}, t) &\leq \sum_{S \subset [N], \mathrm{card}(S) = s} \mathcal{N}(B^s_{\|\cdot\|_2}, \|\cdot\|_2, \frac{t}{m^{\frac{1}{2q}}\sqrt{2s}}) \\ &\leq \binom{N}{s} \left(1 + \frac{3m^{\frac{1}{2q}}\sqrt{s}}{t}\right)^{2s} \leq \left(\frac{eN}{s}\right)^s \left(1 + \frac{3m^{\frac{1}{2q}}\sqrt{s}}{t}\right)^{2s}, \end{aligned} \tag{B.3}$$

where $B^s_{\|\cdot\|_2}$ is the unit ℓ^2-ball in $\mathbb{C}^s$. Thus,

$$\log(2\mathcal{N}(D_{s,N}, \|\cdot\|_{X,q}, t)) \leq 2s \left(\log\left(\frac{eN}{s}\right) + \log\left(1 + \frac{3m^{\frac{1}{2q}}\sqrt{s}}{t}\right)\right). \tag{B.4}$$

For large t, if we define the norm $|\cdot|_1$ on $\mathbb{C}^N$ as

$$|z|_1 = \sum_{j=1}^N |\mathrm{Re}(z_j)| + |\mathrm{Im}(z_j)|.$$

Then it follows from Hölder's inequality that

$$D_{s,N} \subset \sqrt{2s} B^N_{|\cdot|_1} = \{z \in \mathbb{C}^N : |z|_1 \leq \sqrt{2s}\}.$$

We then use Maurey's empirical method to estimate $\mathcal{N}(B^N_{|\cdot|_1}, \|\cdot\|_{X,q}, t)$. Note that the set $B^N_{|\cdot|_1}$ is the convex hull of $\{\pm \boldsymbol{e}_j, \pm i\boldsymbol{e}_j : j \in [N]\}$. If we enumerate these vectors as $\{\boldsymbol{v}_j\}_{j\in[4N]}$, then for any $\boldsymbol{z} \in B^N_{|\cdot|_1}$, there are non-negative $\{\lambda_j\}_{j\in[4N]}$ with $\sum_{j=1}^{4N} \lambda_j = 1$ such that $\boldsymbol{z} = \sum_{j=1}^{4N} \lambda_j \boldsymbol{v}_j$. For this $\boldsymbol{z}$, we define a random variable $\boldsymbol{Y}$ which takes value $\boldsymbol{v}_j$ with probability λ_j and let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M$ be i.i.d. copies, where M is a positive integer to be determined. Since $\mathbb{E}\boldsymbol{Y} = \boldsymbol{z}$, if we let

$$\boldsymbol{y} = \frac{1}{M}\sum_{k=1}^{M} \boldsymbol{Y}_k, \tag{B.5}$$

by symmetrization we have

$$\mathbb{E}\|\boldsymbol{y} - \boldsymbol{z}\|_{X,q} = \frac{1}{M}\mathbb{E}\left\|\sum_{k=1}^{M}(\boldsymbol{Y}_k - \mathbb{E}\boldsymbol{Y}_k)\right\|_{X,q} \le \frac{2}{M}\mathbb{E}\left\|\sum_{k=1}^{M}\epsilon_k \boldsymbol{Y}_k\right\|_{X,q}.$$

Here we use the fact that the seminorm $\|\cdot\|_{X,q}$ is convex. By the definition of $\|\cdot\|_{X,q}$, we have

$$\begin{aligned}
\mathbb{E}\left\|\sum_{k=1}^{M}\epsilon_k \boldsymbol{Y}_k\right\|_{X,q} &= \mathbb{E}_{\epsilon,\boldsymbol{Y}}\left(\sum_{\ell=1}^{m}\left|\sum_{k=1}^{M}\epsilon_k\langle \boldsymbol{X}_\ell, \boldsymbol{Y}_k\rangle\right|^{2q}\right)^{\frac{1}{2q}} \\
&\le \mathbb{E}_{\boldsymbol{Y}}\left(\sum_{\ell=1}^{m}\mathbb{E}_{\epsilon}\left|\sum_{k=1}^{M}\epsilon_k\langle \boldsymbol{X}_\ell, \boldsymbol{Y}_k\rangle\right|^{2q}\right)^{\frac{1}{2q}} \\
&\le 2^{\frac{1}{2q}}e^{-\frac{1}{2}}\sqrt{2q}\,\mathbb{E}_{\boldsymbol{Y}}\left(\sum_{\ell=1}^{m}\|(\langle \boldsymbol{X}_\ell, \boldsymbol{Y}_k\rangle)_{k=1}^{M}\|_2^{2q}\right)^{\frac{1}{2q}} \\
&\le C_1 m^{\frac{1}{2q}}\sqrt{q}\sqrt{M},
\end{aligned}$$

where we use Fubini's theorem and Jensen's inequality in the second line, Khintchine's inequality (Corollary 8.7 in [14]) in the third line and the fact that $|\langle \boldsymbol{X}_\ell, \boldsymbol{Y}_k\rangle| \le \|X_\ell\|_\infty \le 1$ in the fourth line. The constant C_1 is universal. We now have the bound

$$\mathbb{E}\|\boldsymbol{y} - \boldsymbol{z}\|_{X,q} \le \frac{C_2}{\sqrt{M}} m^{\frac{1}{2q}}\sqrt{q},$$

for some universal constant C_2. Therefore, for this $\boldsymbol{z} \in B^N_{|\cdot|_1}$, we can find a $\boldsymbol{y}$ of the form (B.5) such that $\|\boldsymbol{z} - \boldsymbol{y}\|_{X,q} \le C_2 m^{\frac{1}{2q}}\sqrt{q}/\sqrt{M}$. Thus for any $t > 0$, if

$$M \ge \frac{C_3 m^{\frac{1}{q}} q}{t^2},$$

for some universal C_3, then $\|\boldsymbol{z}-\boldsymbol{y}\|_{X,q} \le t$. Let

$$M = \left\lfloor \frac{C_3 m^{\frac{1}{q}} q}{t^2} \right\rfloor,$$

then the set

$$\mathcal{Y} = \left\{ \boldsymbol{y} = \frac{1}{M}\sum_{k=1}^{M} \boldsymbol{Y}_k : \boldsymbol{Y}_k \in \{\pm\boldsymbol{e}_1, \dots, \pm\boldsymbol{e}_N, \pm i\boldsymbol{e}_1, \dots, \pm i\boldsymbol{e}_N\} \right\}$$

has cardinality at most $\binom{M+4N-1}{M} < (e(1+4N/M))^M$ and is a t-covering for $B^N_{|\cdot|_1}$. Note that $t \le \sup_{\boldsymbol{z}\in D_{s,N}} \|\boldsymbol{z}\|_{X,q} \le m^{\frac{1}{2q}}\sqrt{s}$, we have

$$\begin{aligned}\log(2\mathcal{N}(D_{s,N}, \|\cdot\|_{X,q}, t)) &\le \log(2\mathcal{N}(\sqrt{2s}B^N_{|\cdot|_1}, \|\cdot\|_{X,q}, t)) \\ &\le \frac{C_4 m^{\frac{1}{q}} qs}{t^2} \log\left(3+\frac{N}{q}\right), \end{aligned} \tag{B.6}$$

for some universal constant C_4. With the estimate (B.4) and (B.6) and let $\alpha > 0$, the integral in (B.2) can be bounded by

$$\begin{aligned}&\int_0^{m^{\frac{1}{2q}}\sqrt{s}} \sqrt{\log(2\mathcal{N}(D_{s,N}, \|\cdot\|_{X,q}, t))}dt \\ \le &\int_0^{\alpha} \sqrt{2s\left(\log\left(\frac{eN}{s}\right) + \log\left(1+\frac{3m^{\frac{1}{2q}}\sqrt{s}}{t}\right)\right)}dt \\ &+ \sqrt{C_4 m^{\frac{1}{q}} s \log\left(3+\frac{N}{q}\right)} \int_{\alpha}^{m^{\frac{1}{2q}}\sqrt{s}} \frac{1}{t}dt \\ \le &\,\alpha\sqrt{2s}\left(\sqrt{\log\left(\frac{eN}{s}\right)} + \sqrt{\log\left(e+\frac{3em^{\frac{1}{2q}}\sqrt{s}}{\alpha}\right)}\right) \\ &+ \sqrt{C_4 m^{\frac{1}{q}} qs \log\left(3+\frac{N}{q}\right)} \log\left(\frac{m^{\frac{1}{2q}}\sqrt{s}}{\alpha}\right). \end{aligned}$$

Choose $\alpha = m^{\frac{1}{2q}}$ we have

$$\int_0^{m^{\frac{1}{2q}}\sqrt{s}} \sqrt{\log(2\mathcal{N}(D_{s,N}, \|\cdot\|_{X,q}, t))}dt \le C_5 m^{\frac{1}{2q}} \sqrt{qs\log^2(s)\log\left(3+\frac{N}{q}\right)},$$

for some universal constant C_5. Thus,

$$E \leq 8\sqrt{2}Rs^{\frac{p-1}{2p}} \int_0^{m^{\frac{1}{2q}}\sqrt{s}} \sqrt{\log(2\mathcal{N}(D_{s,N}, \|\cdot\|_{X,q}, t))}dt$$
$$\leq C_6 s^{\frac{p-1}{2p}} m^{\frac{1}{2q}} \sqrt{qs\log^2(s)\log\left(3+\frac{N}{q}\right)} \left(\left|\left|\left|\sum_{\ell=1}^m \boldsymbol{X}_\ell \boldsymbol{X}_\ell^*\right|\right|\right|\right)^{\frac{1}{2p}},$$

where C_6 is some universal constant. This concludes the proof. □

C. Proof of Theorem 4

Proof of Theorem 4. Parts of the proof follow from the general idea of the proof of Theorem 12.32 in [14]; however, many key steps differ since random feature matrices do not form orthogonal systems and since the elements of a random feature matrix are not independent.

Denote the ℓ-th column of $\boldsymbol{A}^*$ by $\boldsymbol{X}_\ell = [\overline{\phi(\boldsymbol{x}_\ell;\boldsymbol{\omega}_1)}, \ldots, \overline{\phi(\boldsymbol{x}_\ell;\boldsymbol{\omega}_N)}]^T \in \mathbb{C}^N$. Then $\boldsymbol{A}^*\boldsymbol{A} = \sum_{\ell=1}^m \boldsymbol{X}_\ell \boldsymbol{X}_\ell^*$ and the s-th RIP constant $\delta_s\left(\frac{1}{\sqrt{m}}\boldsymbol{A}\right)$ is bounded by

$$\delta_s\left(\frac{1}{\sqrt{m}}\boldsymbol{A}\right) = \left|\left|\left|\frac{1}{m}\sum_{\ell=1}^m \boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \boldsymbol{I}_N\right|\right|\right|$$
$$\leq \frac{1}{m}\left|\left|\left|\sum_{\ell=1}^m (\boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)\right|\right|\right|$$
$$+ \frac{1}{m}\left|\left|\left|\sum_{\ell=1}^m (\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)\right|\right|\right| + |||\boldsymbol{L}|||,$$

where $\boldsymbol{L} \in \mathbb{C}^{N\times N}$ is a matrix with $L_{j,j} = 0$ and $L_{j,k} = (2\gamma^2\sigma^2+1)^{-\frac{d}{2}}$ for $j,k \in [N], j \neq k$.

For any $\boldsymbol{z} \in D_{s,N}$

$$|\langle \boldsymbol{L}\boldsymbol{z}, \boldsymbol{z}\rangle| = \left|\sum_{\substack{j,k\in \mathrm{supp}(\boldsymbol{z})\\ j\neq k}} \left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{\frac{d}{2}} z_j\overline{z_k}\right| \leq \left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{\frac{d}{2}} \|\boldsymbol{z}\|_1^2$$
$$\leq s\left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{\frac{d}{2}},$$

and using the complexity assumption yields

$$|||\boldsymbol{L}||| \leq s(2\gamma^2\sigma^2+1)^{-\frac{d}{2}} \leq \eta_1.$$

Note that by the definition of $|||\cdot|||$, $|||\boldsymbol{B}||| \leq \|\boldsymbol{B}\|_2$ for any self-adjoint matrix $\boldsymbol{B}$. From the proof of Theorem 1, we have $\left|\left|\left|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\right|\right| \leq \eta_1$ with probability at least $1-\delta$ with respect to the draw of $\{\boldsymbol{\omega}_\ell\}_{\ell\in[N]}$ if the complexity assumption holds.

For the remaining arguments, we will use that $\left|\left|\left|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\right|\right| \leq \eta_1$ conditioned on the given draw of $\{\boldsymbol{\omega}_k\}_{k\in[N]}$. Note that $|||\cdot|||$ is a convex function (since it is a seminorm) and thus symmetrization yields

$$\mathbb{E}_{\boldsymbol{x}} \left|\left|\left| \sum_{\ell=1}^m (\boldsymbol{X}_\ell \boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)) \right|\right|\right| \leq 2\mathbb{E}_{\boldsymbol{x},\epsilon} \left|\left|\left| \sum_{\ell=1}^m \epsilon_\ell \boldsymbol{X}_\ell \boldsymbol{X}_\ell^* \right|\right|\right|,$$

where we introduce the Rademacher random variables ϵ_ℓ for $\ell \in [m]$ which are independent of data and weights. Applying Fubini's theorem and Lemma A.1, we obtain

$$\begin{aligned}
E :&= \frac{1}{m}\mathbb{E}_{\boldsymbol{x}} \left|\left|\left| \sum_{\ell=1}^m (\boldsymbol{X}_\ell \boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)) \right|\right|\right| \\
&\leq \frac{2}{m}\mathbb{E}_{\boldsymbol{x},\epsilon} \left|\left|\left| \sum_{\ell=1}^m \epsilon_\ell \boldsymbol{X}_\ell \boldsymbol{X}_\ell^* \right|\right|\right| \\
&\leq \frac{\tilde{C} s^{\frac{p-1}{2p}} \sqrt{qs\log^2(s)\log(3+N/q)}}{m^{1-\frac{1}{2p}}} \left(\left|\left|\left| \sum_{\ell=1}^m \boldsymbol{X}_\ell \boldsymbol{X}_\ell^* \right|\right|\right| \right)^{\frac{1}{2q}} \\
&= \frac{\tilde{C} s^{\frac{p-1}{2p}} \sqrt{qs\log^2(s)\log(3+N/q)}}{m^{1-\frac{1}{2p}-\frac{1}{2q}}} \left(\frac{1}{m} \left|\left|\left| \sum_{\ell=1}^m \boldsymbol{X}_\ell \boldsymbol{X}_\ell^* \right|\right|\right| \right)^{\frac{1}{2q}} \\
&\leq \tilde{C} s^{\frac{p-1}{2p}} \sqrt{\frac{qs\log^2(s)\log(3+N/q)}{m}} \\
&\qquad \times \sqrt{\mathbb{E}_{\boldsymbol{x}} \left|\left|\left| \frac{1}{m} \sum_{\ell=1}^m \left(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\right) \right|\right|\right| + 1 + 2\eta_1},
\end{aligned}$$

where we use the fact that $1/p + 1/q = 1$ in the fifth line and Jensen's inequality together with the estimate

$$\begin{aligned}
\left|\left|\left|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\right|\right|\right| &\leq \left|\left|\left|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\right|\right|\right| + \left|\left|\left|\mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*) - \boldsymbol{I}_N\right|\right|\right| + |||\boldsymbol{I}_N||| \\
&= \left|\left|\left|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\right|\right|\right| + |||\boldsymbol{L}||| + \left|\left|\left|\boldsymbol{I}_N\right|\right|\right| \\
&\leq 2\eta_1 + 1,
\end{aligned}$$

in the last inequality. Choosing $p = 1 + 1/\log(s), q = 1 + \log(s)$, we have $s^{\frac{p-1}{2p}} \leq s^{\frac{1}{2\log(s)}} < 2$, and the above inequality becomes

$$E \leq C\sqrt{\frac{s\log^3(s)\log(N)}{m}}\sqrt{E + 1 + 2\eta_1}.$$

If we set $D = C\sqrt{\frac{s\log^3(s)\log(N)}{m}}$ and assume $\eta_1 \leq 1/2$, then rearranging the terms yields

$$E \leq \frac{D^2}{2} + \sqrt{\frac{D^4}{4} + 2D^2} \leq D^2 + \sqrt{2}D.$$

Next we use Bernstein's inequality for suprema of an empirical process (Lemma A.3) to show that $m^{-1}\left|\left|\left|\sum_{\ell=1}^m (\boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*))\right|\right|\right|$ is small with high probability. By the definition of the seminorm $|||\cdot|||$, we have

$$\begin{aligned}\left|\left|\left|\sum_{\ell=1}^m (\boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*))\right|\right|\right| &= \sup_{z\in D_{s,N}}\left|\sum_{\ell=1}^m \langle(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*))z, z\rangle\right| \\ &= \sup_{z\in D_{s,N}^*}\left|\sum_{\ell=1}^m \langle(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*))z, z\rangle\right|,\end{aligned}$$

where $D_{s,N}^*$ is a countable dense subset of $D_{s,N}$, which exists since $D_{s,N}$ is the finite union of separable sets (unit balls in $\mathbb{C}^s$ as a subspace of $\mathbb{C}^N$ with its norm).

For $z \in D_{s,N}^*$, define the random variable $f_z(\boldsymbol{X}_\ell) = \langle(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*))z, z\rangle$. Note that $\mathbb{E}_{\boldsymbol{x}} f_z(\boldsymbol{X}_\ell) = 0$. To apply Lemma A.3, we must first show that $f_z(\boldsymbol{X}_\ell)$ is bounded and compute its variance. The random variables $f_z(\boldsymbol{X}_\ell)$ for $\ell \in [m]$ are uniformly bounded by $2s$ since

$$\begin{aligned}|f_z(\boldsymbol{X}_\ell)| &\leq \sum_{j,k\in S}\left|\exp(i\langle\boldsymbol{x}_\ell, \boldsymbol{\omega}_k - \boldsymbol{\omega}_j\rangle) - \exp\left(-\frac{\gamma^2}{2}\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_2^2\right)\right|\left|\bar{z}_j z_k\right| \\ &\leq 2\|z\|_1^2 \leq 2s,\end{aligned}$$

where $S = \operatorname{supp}(z)$. The variance of $f_z(\boldsymbol{X}_\ell)$ for $\ell \in [m]$ is bounded by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[f_z(\boldsymbol{X}_\ell)]^2 &= \mathbb{E}_{\boldsymbol{x}}|\langle(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*))z, z\rangle|^2 \\ &\leq \mathbb{E}_{\boldsymbol{x}}\|(\boldsymbol{X}_{\ell,S}\boldsymbol{X}_{\ell,S}^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*))z\|_2^2 \\ &= \mathbb{E}_{\boldsymbol{x}}[z^*\boldsymbol{X}_{\ell,S}\boldsymbol{X}_{\ell,S}^*\boldsymbol{X}_{\ell,S}\boldsymbol{X}_{\ell,S}^*z] - z^*\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)z \\ &\leq s\langle\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)z, z\rangle \\ &\leq s\left|\left|\left|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell\boldsymbol{X}_\ell^*)\right|\right|\right| \leq 2s,\end{aligned}$$

using the fact that $\|\boldsymbol{X}_{\ell,S}\|_2^2 = s$ and the estimate $\left|\left|\left|\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\right|\right|\right| \le 1 + 2\eta_1 \le 2$. If we set

$$D = C\sqrt{\frac{s\log^3(s)\log(N)}{m}} \le \eta_2, \tag{C.1}$$

for some $\eta_2 \in (0, 1/2)$, then

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{x}}\left(\left|\left|\left|\sum_{\ell=1}^m (\boldsymbol{X}_\ell \boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*))\right|\right|\right| \ge m(\eta_2^2 + \sqrt{2}\eta_2) + m\eta_1\right) \\
&\le \mathbb{P}_{\boldsymbol{x}}\left(\left|\left|\left|\sum_{\ell=1}^m (\boldsymbol{X}_\ell \boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*))\right|\right|\right| \ge \mathbb{E}_{\boldsymbol{x}}\left|\left|\left|\sum_{\ell=1}^m (\boldsymbol{X}_\ell \boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*))\right|\right|\right| + m\eta_1\right) \\
&\le \mathbb{P}_{\boldsymbol{x}}\left(\sup_{z\in D_{s,N}^*}\left|\sum_{\ell=1}^m f_z(\boldsymbol{X}_\ell)\right| \ge \mathbb{E}_{\boldsymbol{x}}\sup_{z\in D_{s,N}^*}\left|\sum_{\ell=1}^m f_z(\boldsymbol{X}_\ell)\right| + m\eta_1\right) \\
&\le \exp\left(\frac{-m\eta_1^2/2}{2s + 4s(\eta_2^2 + \sqrt{2}\eta_2) + 2s\eta/3}\right),
\end{aligned} \tag{C.2}$$

where in the last inequality we use Lemma A.3 with $K = 2s$, $\sum_{\ell=1}^m \sigma_\ell^2 \le 2ms$, $\mathbb{E}_{\boldsymbol{x}}(Z) \le m(\eta_2^2 + \sqrt{2}\eta_2)$ and $t = m\eta_1$. Thus,

$$\mathbb{P}_{\boldsymbol{x}}\left(\frac{1}{m}\left|\left|\left|\sum_{\ell=1}^m (\boldsymbol{X}_\ell \boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*))\right|\right|\right| \le \eta_2^2 + \sqrt{2}\eta_2 + \eta_1\right) > 1 - \delta,$$

if the following conditions hold (assuming $\eta_2^2 + \sqrt{2}\eta_2 \le 1$ and $\eta_1 \le 1/2$):

$$\begin{aligned}
m &\ge C_1 \eta_1^{-2} s\log(\delta^{-1}) \\
m &\ge C_2 \eta_2^{-2} s\log^3(s)\log(N).
\end{aligned}$$

The constants C_1 and C_2 are universal.

Altogether, with probability at least $1 - 2\delta$,

$$\begin{aligned}
\delta_s\left(\frac{1}{\sqrt{m}}\boldsymbol{A}\right) &\le \frac{1}{m}\left|\left|\left|\sum_{\ell=1}^m (\boldsymbol{X}_\ell \boldsymbol{X}_\ell^* - \mathbb{E}_{\boldsymbol{x}}\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\right|\right|\right| \\
&\quad + \frac{1}{m}\left|\left|\left|\sum_{\ell=1}^m (\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}}\boldsymbol{X}_\ell \boldsymbol{X}_\ell^*)\right|\right|\right| + |||\boldsymbol{L}||| \\
&\le \eta_2^2 + \sqrt{2}\eta_2 + 3\eta_1,
\end{aligned}$$

if the conditions in the theorem are satisfied, which completes the proof. □

D. Proofs of Generalization Theorems from Section 4

The proofs of Theorem 2, Theorem 3 and Theorem 5 follow a similar structure in which we show that each trained vector $\boldsymbol{c}^\sharp$ is close to the best ϕ approximation $\boldsymbol{c}^\star$ from [34,35]. To relate the coefficient vectors to the risk, we utilize the conditioning result, Theorem 1, over a finite set of samples to approximate the risk by a new discrete system, as done in Theorem 1 from [17]. Since the proofs of Theorem 2 and Theorem 3 are similar, we present them together.

Proof of Theorem 2 and Theorem 3. We will work directly with the L^2 norm, which can be decomposed into two parts:

$$\|f - f^\sharp\|_{L^2(d\mu)} \leq \|f - f^\star\|_{L^2(d\mu)} + \|f^\star - f^\sharp\|_{L^2(d\mu)},$$

by triangle inequality. The approximation $f^\sharp$ is defined in (4.2) and the best ϕ-based approximation $f^\star$ is defined by

$$f^\star(\boldsymbol{x}) = \sum_{k=1}^{N} c_k^\star \, \phi(\boldsymbol{x}, \boldsymbol{\omega}_k) = \frac{1}{N} \sum_{k=1}^{N} \frac{\alpha(\boldsymbol{\omega}_k)}{\rho(\boldsymbol{\omega}_k)} \, \phi(\boldsymbol{x}, \boldsymbol{\omega}_k), \tag{D.1}$$

with $c_k^\star = N^{-1} \frac{\alpha(\boldsymbol{\omega}_k)}{\rho(\boldsymbol{\omega}_k)}$ for all $k \in [N]$. If the condition on N in Lemma A.7 is satisfied, then with probability at least $1 - \delta$

$$\|f - f^\star\|_{L^2(d\mu)} \leq \epsilon \|f\|_\rho.$$

To bound $\|f^\star - f^\sharp\|_{L^2(d\mu)}$, we use McDiarmid's inequality and argue similarly as in Lemma 2 from [17]. Let $\{z_j\}_{j\in[m]}$ be i.i.d. random variables sampled from the distribution μ and independent from the $\{\boldsymbol{x}_j\}_{j\in[m]}$ and $\{\boldsymbol{\omega}_k\}_{k\in[N]}$. Thus $\{z_j\}_{j\in[m]}$ is also independent of $\boldsymbol{c}^\sharp$ and $\boldsymbol{c}^\star$. We define the random variable

$$v(z_1, \ldots, z_m) := \|f^\star - f^\sharp\|_{L^2(d\mu)}^2 - \frac{1}{m} \sum_{j=1}^{m} |f^\star(z_j) - f^\sharp(z_j)|^2.$$

Note that for any $j \in [m]$,

$$\mathbb{E}_z \left[|f^\star(z_j) - f^\sharp(z_j)|^2 \right] = \mathbb{E}_{z_1,\ldots,z_m} \left[|f^\star(z_j) - f^\sharp(z_j)|^2 \right] = \|f^\star - f^\sharp\|_{L^2(d\mu)}^2,$$

thus $\mathbb{E}_z[v] = 0$. Perturbing the k-th component of v yields

$$\begin{aligned} &|v(z_1, \ldots, z_k, \ldots, z_m) - v(z_1, \ldots, \tilde{z}_k, \ldots, z_m)| \\ &\leq \frac{1}{m} \left| |f^\star(z_k) - f^\sharp(z_k)|^2 - |f^\star(\tilde{z}_k) - f^\sharp(\tilde{z}_k)|^2 \right|. \end{aligned}$$

By Cauchy–Schwarz inequality, for any z we have

$$|f^\star(z) - f^\sharp(z)|^2 = \left|\sum_{k=1}^{N}(c_k^\star - c_k^\sharp)\phi(z,\omega_k)\right|^2 \le N\|c^\star - c^\sharp\|_2^2,$$

which holds since $|\phi(z,\omega)| = 1$. Thus the difference is bounded by

$$|v(z_1,\dots,z_k,\dots,z_m) - v(z_1,\dots,\tilde{z}_k,\dots,z_m)| \le \frac{2N}{m}\|c^\star - c^\sharp\|_2^2 =: \Delta. \tag{D.2}$$

Next, we apply McDiarmid's inequality $\mathbb{P}_z(v - \mathbb{E}_z[v] \ge t) \le \exp(-\frac{2t^2}{m\Delta^2})$, by setting

$$t = \Delta\sqrt{\frac{m}{2}\log\left(\frac{1}{\delta}\right)},$$

which implies that

$$\begin{aligned}\|f^\star - f^\sharp\|_{L^2(d\mu)}^2 &\le \frac{1}{m}\sum_{j=1}^{m}|f^\star(z_j) - f^\sharp(z_j)|^2 + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\|c^\star - c^\sharp\|_2^2\\ &= \frac{1}{m}\|\tilde{A}(c^\star - c^\sharp)\|_2^2 + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\|c^\star - c^\sharp\|_2^2,\end{aligned}$$

with probability at least $1-\delta$ with respect to draw of $\{z_j\}_{j\in[m]}$, and $\tilde{A} \in \mathbb{C}^{m\times N}$ is the random feature matrix with $\tilde{a}_{j,k} = \exp(i\langle z_j, \omega_k\rangle)$.

In the underparameterized regime where $m > N$, we apply Part (a) of Theorem 1 to $\tilde{A}$ and we obtain

$$\begin{aligned}\|f^\star - f^\sharp\|_{L^2(d\mu)}^2 &\le \frac{1}{m}\|\tilde{A}(c^\star - c^\sharp)\|_2^2 + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\|c^\star - c^\sharp\|_2^2\\ &= \left\|\frac{1}{\sqrt{m}}\tilde{A}(c^\star - c^\sharp)\right\|_2^2 + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\|c^\star - c^\sharp\|_2^2\\ &\le (1+3\eta)\left\|c^\star - c^\sharp\right\|_2^2 + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\|c^\star - c^\sharp\|_2^2\\ &\le \left(1 + 3\eta + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\right)\|c^\star - c^\sharp\|_2^2,\end{aligned}$$

with probability at least $1-3\delta$. To estimate $\|\boldsymbol{c}^\sharp - \boldsymbol{c}^\star\|_2$, we utilize the the pseudo-inverse $A^\dagger$ and define the residual $\boldsymbol{h} = \boldsymbol{A}\boldsymbol{c}^\star - \boldsymbol{y}$, so that

$$\begin{aligned}\|\boldsymbol{c}^\sharp - \boldsymbol{c}^\star\|_2 = \|\boldsymbol{A}^\dagger \boldsymbol{y} - \boldsymbol{c}^\star\|_2 &= \|\boldsymbol{A}^\dagger(\boldsymbol{A}\boldsymbol{c}^\star - \boldsymbol{h}) - \boldsymbol{c}^\star\|_2 \\ &\le \|(\boldsymbol{A}^\dagger\boldsymbol{A} - \boldsymbol{I})\boldsymbol{c}^\star\|_2 + \|\boldsymbol{A}^\dagger\|_2 \|\boldsymbol{h}\|_2.\end{aligned}$$

Using Hölder's inequality and Assumption 2 on $\|\boldsymbol{e}\|_2$, the residual is bounded by

$$\begin{aligned}\|\boldsymbol{h}\|_2 = \sqrt{\sum_{j=1}^m \left(f^\star(\boldsymbol{x}_j) - f(\boldsymbol{x}_j) - e_j\right)^2} &\le \sqrt{\sum_{j=1}^m \left(f^\star(\boldsymbol{x}_j) - f(\boldsymbol{x}_j)\right)^2} + \|\boldsymbol{e}\|_2 \\ &\le \sqrt{m} \sup_{\|\boldsymbol{X}\|_2 \le R} |f^\star(\boldsymbol{x}) - f(\boldsymbol{x})| + \sqrt{m}E \\ &\le \sqrt{m}(\epsilon \|f\|_\rho + E), \end{aligned} \tag{D.3}$$

with probability $1-2\delta$ if conditions in Lemma A.5 and Lemma A.6 are satisfied. Therefore, the bound becomes

$$\|\boldsymbol{c}^\sharp - \boldsymbol{c}^\star\|_2 \le \|(\boldsymbol{A}^\dagger\boldsymbol{A} - \boldsymbol{I})\boldsymbol{c}^\star\|_2 + \sqrt{m}\,\|\boldsymbol{A}^\dagger\|_2 \left(\epsilon\|f\|_\rho + E\right).$$

Note that when $m > N$, the first term $\|(\boldsymbol{A}^\dagger\boldsymbol{A} - \boldsymbol{I})\boldsymbol{c}^\star\|_2 = 0$, since $\boldsymbol{A}^\dagger\boldsymbol{A} = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{I}$. Using Theorem 1, the operator norm of the pseudo-inverse is bounded by

$$\|\boldsymbol{A}^\dagger\|_2 = \frac{1}{\sqrt{\lambda_{\min}(\boldsymbol{A}^*\boldsymbol{A})}} = \frac{1}{\sqrt{m}} \frac{1}{\sqrt{\lambda_{\min}(m^{-1}\boldsymbol{A}^*\boldsymbol{A})}} \le \frac{1}{\sqrt{m}} \frac{1}{\sqrt{1-3\eta}},$$

with probability at least $1-2\delta$ and thus

$$\begin{aligned}\|f - f^\sharp\|_{L^2(d\mu)} &\le \|f - f^\star\|_{L^2(d\mu)} + \|f^\star - f^\sharp\|_{L^2(d\mu)} \\ &\le \epsilon\|f\|_\rho + \|f^\star - f^\sharp\|_{L^2(d\mu)} \\ &\le \epsilon\|f\|_\rho + \left(1 + 3\eta + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\right)^{\frac{1}{2}} \|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp\|_2 \\ &\le 2\sqrt{K(\eta)}\left(1 + N^{\frac{1}{2}}m^{-\frac{1}{4}}\log^{\frac{1}{4}}\left(\frac{1}{\delta}\right)\right)(\epsilon\|f\|_\rho + E),\end{aligned}$$

with probability at least $1-8\delta$.

In the overparameterized regime where $m < N$, we apply Part (b) of Theorem 1 to $\tilde{\boldsymbol{A}}$ and the error of $\|f - f^\sharp\|_{L^2(d\mu)}^2$ is bounded by

$$\begin{aligned}
\|f^\star - f^\sharp\|_{L^2(d\mu)}^2 &\le \frac{1}{m}\|\tilde{\boldsymbol{A}}(\boldsymbol{c}^\star - \boldsymbol{c}^\sharp)\|_2^2 + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\,\|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp\|_2^2 \\
&= \frac{N}{m}\left\|\frac{1}{\sqrt{N}}\tilde{\boldsymbol{A}}(\boldsymbol{c}^\star - \boldsymbol{c}^\sharp)\right\|_2^2 + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\,\|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp\|_2^2 \\
&\le \frac{N}{m}\lambda_{\max}\left(\frac{1}{N}\boldsymbol{A}\boldsymbol{A}^*\right)\|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp\|_2^2 + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp\|_2^2 \\
&\le \left(\frac{N}{m}(1+3\eta) + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\right)\|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp\|_2^2,
\end{aligned}$$

with probability at least $1 - 3\delta$. For $\|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp\|_2$, we have

$$\|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp\|_2 \le \|(\boldsymbol{A}^\dagger\boldsymbol{A} - \boldsymbol{I})\boldsymbol{c}^\star\|_2 + \|\boldsymbol{A}^\dagger\|_2\|\boldsymbol{h}\|_2.$$

Note that when $m < N$, $\boldsymbol{I} - \boldsymbol{A}^\dagger\boldsymbol{A}$ is the projection onto the null space of $\boldsymbol{A}$ and thus its operator norm is bounded by 1 and

$$\begin{aligned}
\|\boldsymbol{c}^\sharp - \boldsymbol{c}^\star\|_2 \le \|\boldsymbol{c}^\star\|_2 + \|\boldsymbol{A}^\dagger\|_2\,\|\boldsymbol{h}\|_2 &\le \|\boldsymbol{c}^\star\|_2 + \frac{1}{\sqrt{N}}\,\frac{1}{\sqrt{\lambda_{\min}(N^{-1}\boldsymbol{A}\boldsymbol{A}^*)}}\,\|\boldsymbol{h}\|_2 \\
&\le \frac{1}{\sqrt{N}}\|f\|_\rho + \frac{\sqrt{m}}{\sqrt{N}}\frac{1}{\sqrt{1-3\eta}}(\epsilon\|f\|_\rho + E),
\end{aligned}$$

with probability at least $1 - 4\delta$, noting that $\left|\boldsymbol{c}_k^\star\right| = N^{-1}\left|\frac{\alpha(\boldsymbol{\omega}_k)}{\rho(\boldsymbol{\omega}_k)}\right| \le N^{-1}\|f\|_\rho$. In this regime, the L^2 generalization error is bounded by

$$\begin{aligned}
&\|f - f^\sharp\|_{L^2(d\mu)} \\
&\quad\le \|f - f^\star\|_{L^2(d\mu)} + \|f^\star - f^\sharp\|_{L^2(d\mu)} \\
&\quad\le \epsilon\|f\|_\rho + \|f^\star - f^\sharp\|_{L^2(d\mu)} \\
&\quad\le \epsilon\|f\|_\rho + \left(\frac{N}{m}(1+3\eta) + N\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}\right)^{\frac{1}{2}}\|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp\|_2 \\
&\quad\le \epsilon\|f\|_\rho \\
&\qquad + \left(m^{-\frac{1}{2}}\sqrt{1+3\eta} + m^{-\frac{1}{4}}\left(2\log\left(\frac{1}{\delta}\right)\right)^{\frac{1}{4}}\right)\left(\|f\|_\rho + m^{\frac{1}{2}}\frac{1}{\sqrt{1-3\eta}}(\epsilon\|f\|_\rho + E)\right)
\end{aligned}$$

$$\leq \epsilon \|f\|_\rho + 2m^{-\frac{1}{4}} \log^{\frac{1}{4}}\left(\frac{1}{\delta}\right) \|f\|_\rho + 2m^{\frac{1}{4}} \sqrt{K(\eta)}\, \log^{\frac{1}{4}}\left(\frac{1}{\delta}\right) \left(\epsilon \|f\|_\rho + E\right)$$

$$\leq \epsilon \|f\|_\rho + 2\log^{\frac{1}{4}}\left(\frac{1}{\delta}\right) \left(m^{-\frac{1}{4}} + \sqrt{K(\eta)}\, m^{\frac{1}{4}}\, \epsilon\right) \|f\|_\rho + 2m^{\frac{1}{4}} \sqrt{K(\eta)}\, \log^{\frac{1}{4}}\left(\frac{1}{\delta}\right) E,$$

with probability at least $1 - 8\delta$. □

Proof of Theorem 5. Following the proof of Theorem 2 and Theorem 3, we have the analogous estimate

$$\|f^\star - f^\sharp\|_{L^2(d\mu)}^2 \leq \frac{1}{m} \sum_{j=1}^m |f^\star(z_j) - f^\sharp(z_j)|^2 + \left(N \sqrt{\frac{2}{m} \log\left(\frac{1}{\delta}\right)}\right) \|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp|_{\mathcal{S}^\sharp}\|_2^2,$$

with probability at least $1 - \delta$. We apply Theorem 4 with s replaced by $2s$ and if the conditions of Theorem 4 are satisfied (for example, by setting the parameters to $\eta_1 = 0.1$ and $\eta_2 = 0.2$), then

$$\delta_{2s}\left(\frac{1}{\sqrt{m}}\boldsymbol{A}\right) \leq \frac{4}{\sqrt{41}} \quad \text{and} \quad \delta_{2s}\left(\frac{1}{\sqrt{m}}\tilde{\boldsymbol{A}}\right) \leq \frac{4}{\sqrt{41}},$$

with probability at least $1 - 2\delta$. Following the proof of Lemma 2 in [17], the first term is bounded by

$$m^{-\frac{1}{2}} \sqrt{\sum_{j=1}^m |f^\star(z_j) - f^\sharp(z_j)|^2} \leq 2\|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp|_{\mathcal{S}^\sharp}\|_2 + 2\vartheta_{s,2}(\boldsymbol{c}^\star) + \vartheta_{s,1}(\boldsymbol{c}^\star).$$

Then by the robust and stable recovery results, Lemma A.4, we have

$$\|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp|_{\mathcal{S}^\sharp}\|_2 \leq C' \frac{\vartheta_{s,1}(\boldsymbol{c}^\star)}{\sqrt{s}} + C\xi + 4\vartheta_{s,2}(\boldsymbol{c}^\star).$$

Combining these two results and Lemma A.7 yields (after possibly multiple redefinitions of the constants)

$$\begin{aligned}
&\|f - f^\sharp\|_{L^2(d\mu)} \\
&\quad \leq \epsilon \|f\|_\rho + 2\left(1 + N^{\frac{1}{2}} m^{-\frac{1}{4}} \log^{\frac{1}{4}}\left(\frac{1}{\delta}\right)\right) \|\boldsymbol{c}^\star - \boldsymbol{c}^\sharp|_{\mathcal{S}^\sharp}\|_2 + 2\vartheta_{s,2}(\boldsymbol{c}^\star) + \vartheta_{s,1}(\boldsymbol{c}^\star) \\
&\quad \leq \epsilon \|f\|_\rho + C\left(1 + N^{\frac{1}{2}} m^{-\frac{1}{4}} s^{-\frac{1}{2}} \log^{\frac{1}{4}}\left(\frac{1}{\delta}\right)\right) \vartheta_{s,1}(\boldsymbol{c}^\star) \\
&\qquad + C'\left(1 + N^{\frac{1}{2}} m^{-\frac{1}{4}} \log^{\frac{1}{4}}\left(\frac{1}{\delta}\right)\right) \vartheta_{s,2}(\boldsymbol{c}^\star) + C''\left(1 + N^{\frac{1}{2}} m^{-\frac{1}{4}} \log^{\frac{1}{4}}\left(\frac{1}{\delta}\right)\right) \xi \\
&\quad \leq C'\left(1 + N^{\frac{1}{2}} m^{-\frac{1}{4}} \log^{\frac{1}{4}}\left(\frac{1}{\delta}\right)\right) (\epsilon \|f\|_\rho + E) \\
&\qquad + C\left(1 + N^{\frac{1}{2}} m^{-\frac{1}{4}} s^{-\frac{1}{2}} \log^{\frac{1}{4}}\left(\frac{1}{\delta}\right)\right) \vartheta_{s,1}(\boldsymbol{c}^\star),
\end{aligned}$$

which concludes the proof. □