Implicit Regularization of Random Feature Models

Arthur Jacot [*1] **Berfin Şimşek** [*1 2] **Francesco Spadaro** [1] **Clément Hongler** [1] **Franck Gabriel** [1]

Abstract

Random Feature (RF) models are used as efficient parametric approximations of kernel methods. We investigate, by means of random matrix theory, the connection between Gaussian RF models and Kernel Ridge Regression (KRR). For a Gaussian RF model with P features, N data points, and a ridge λ, we show that the average (i.e. expected) RF predictor is close to a KRR predictor with an *effective ridge* $\tilde{\lambda}$. We show that $\tilde{\lambda} > \lambda$ and $\tilde{\lambda} \searrow \lambda$ monotonically as P grows, thus revealing the *implicit regularization effect* of finite RF sampling. We then compare the risk (i.e. test error) of the $\tilde{\lambda}$-KRR predictor with the average risk of the λ-RF predictor and obtain a precise and explicit bound on their difference. Finally, we empirically find an extremely good agreement between the test errors of the average λ-RF predictor and $\tilde{\lambda}$-KRR predictor.

1. Introduction

In this paper, we consider the Random Feature (RF) model which is an approximation of Kernel Methods (Rahimi & Recht, 2008) which has seen many recent theoretical developements.

The conventional wisdom suggests that to ensure good generalization performance, one should choose a model class that is complex enough to learn the signal from the training data, yet simple enough to avoid fitting spurious patterns therein (Bishop, 2006). This view has been questioned by recent developments in machine learning. First, Zhang et al. (2016) observed that modern neural network models can perfectly fit randomly labeled training data, while still generalizing well. Second, the test error as a function of parameters exhibits a so-called 'double-descent' curve for many models including neural networks, random forests, and random feature models (Advani & Saxe, 2017; Spigler et al., 2018; Belkin et al., 2018; Mei & Montanari, 2019; Belkin et al., 2019; Nakkiran et al., 2019).

The above models share the feature that for fixed input, the learned predictor $\hat{f}$ is random: for neural networks, this is due to the random initialization of the parameters and/or to the stochasticity of the training algorithm; for random forests, to the random branching; for random feature models, to the sampling of random features. The somehow surprising generalization behavior of these models has recently been the subject of increasing attention. In general, the risk (i.e. test error) is a random variable with two sources of randomness: the usual one due to the sampling of the training set, and the second one due to the randomness of the model itself.

We consider the Random Feature (RF) model (Rahimi & Recht, 2008) with features sampled from a Gaussian Process (GP) and study the RF predictor $\hat{f}$ minimizing the regularized least squares error, isolating the randomness of the model by considering fixed training data points. RF models have been the subject of intense research activity: they are (randomized) approximations of Kernel Methods aimed at easing the computational challenges of Kernel Methods while being asymptotically equivalent to them (Rahimi & Recht, 2008; Yang et al., 2012; Sriperumbudur & Szabó, 2015; Yu et al., 2016). Unlike the asymptotic behavior, which is well studied, RF models with a finite number of features are much less understood.

1.1. Contributions

We consider a model of Random Features (RF) approximating a kernel method with kernel K. This model consists of P Gaussian features, sampled i.i.d. from a (centered) Gaussian process with covariance kernel K. For a given training set of size N, we study the distribution of the RF predictor $\hat{f}_\lambda^{(RF)}$ with ridge parameter $\lambda > 0$ (L^2 penalty on the parameters) and denote it by λ-RF. We show the following:

- The distribution of $\hat{f}_\lambda^{(RF)}$ is that of a mixture of Gaussian processes.

[*]Equal contribution [1]Chair of Statistical Field Theory, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland [2]Laboratory of Computational Neuroscience, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. Correspondence to: Arthur Jacot <arthur.jacot@epfl.ch>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

- The expected RF predictor is close to the $\tilde{\lambda}$-KRR (Kernel Ridge Regression) predictor for an effective ridge parameter $\tilde{\lambda} > 0$.
- The effective ridge $\tilde{\lambda} > \lambda$ is determined by the number of features P, the ridge λ and the Gram matrix of K on the dataset; $\tilde{\lambda}$ decreases monotonically to λ as P grows, revealing the implicit regularization effect of finite RF sampling. Conversely, when using random features to approximate a kernel method with a specific ridge λ^*, one should choose a smaller ridge $\lambda < \lambda^*$ to ensure $\tilde{\lambda}(\lambda) = \lambda^*$.
- The test errors of the expected λ-RF predictor and of the $\tilde{\lambda}$-KRR predictor $\hat{f}^{(K)}_{\tilde{\lambda}}$ are numerically found to be extremely close, even for small P and N.
- The RF predictor's concentration around its expectation can be explicitly controlled in terms of P and of the data; this yields in particular $\mathbb{E}[L(\hat{f}^{(RF)}_{\lambda})] = L(\hat{f}^{(K)}_{\tilde{\lambda}}) + \mathcal{O}(P^{-1})$ as $N, P \to \infty$ with a fixed ratio $\gamma = P/N$ where L is the MSE risk.

Since we compare the behavior of λ-RF and $\tilde{\lambda}$-KRR predictors on the same fixed training set, our result does not rely on any probabilistic assumption on the training data (in particular, we do not assume that our training data is sampled i.i.d.). While our proofs currently require the features to be Gaussian processes, we are confident that they could be generalized to a more general setting (Louart et al., 2017; Benigni & Péché, 2019).

1.2. Related works

Generalization of Random Features. The generalization behavior of Random Feature models has seen intense study in the Statistical Learning Theory framework. Rahimi & Recht (2009) find that $\mathcal{O}(N)$ features are sufficient to ensure the $\mathcal{O}(\frac{1}{\sqrt{N}})$ decay of the generalization error of Kernel Ridge Regression (KRR). Rudi & Rosasco (2017) improve on their result and show that $\mathcal{O}(\sqrt{N}\log N)$ features is actually enough to obtain the $\mathcal{O}(\frac{1}{\sqrt{N}})$ decay of the KRR error.

Hastie et al. (2019) use random matrix theory tools to compute the asymptotic risk when both $P, N \to \infty$ with $\frac{P}{N} \to \gamma > 0$. When the training data is sampled i.i.d. from a Gaussian distribution, the variance is shown to explode at $\gamma = 1$. In the same linear regression setup, Bartlett et al. (2019) establish general upper and lower bounds on the excess risk. Mei & Montanari (2019) prove that the double-descent (DD) curve also arises for random ReLU features, and adding a ridge suppresses the explosion around $\gamma = 1$.

Double-descent and the effect of regularization. For the cross-entropy loss, Neyshabur et al. (2014) observed that for two-layer neural networks the test error exhibits the double-descent (DD) curve as the network width increases (without regularizers, without early stopping). For MSE and hinge losses, the DD curve was observed also in multilayer networks on the MNIST dataset (Advani & Saxe, 2017; Spigler et al., 2018). Neal et al. (2018) study the variance due to stochastic training in neural networks and find that it increases until a certain width, but then decreases down to 0. Nakkiran et al. (2019) establish the DD phenomenon across various models including convolutional and recurrent networks on more complex datasets (e.g. CIFAR-10, CIFAR-100).

Belkin et al. (2018; 2019) find that the DD curve is not peculiar to neural networks and observe the same for random Fourier features and decision trees. In Geiger et al. (2019), the DD curve for neural networks is related to the variance associated with the random initialization of the Neural Tangent Kernel (Jacot et al., 2018); as a result, ensembling is shown to suppress the DD phenomenon in this case, and the test error stays constant in the overparameterized regime. Recent theoretical work (d'Ascoli et al., 2020) study the same setting and derive formulas for the asymptotic error, relying on the so-called replica method.

General Wishart Matrices. Our theoretical analysis relies on the study of the spectrum of the so-called general Wishart matrices of the form $W\Sigma W^T$ (for $N \times N$ matrix Σ and $P \times N$ matrix W with i.i.d. standard Gaussian entries) and in particular their Stieltjes transform $m_P(z) = \frac{1}{P}\mathrm{Tr}\left(W\Sigma W^T - zI_P\right)^{-1}$. A number of asymptotic results (Silverstein, 1995; Bai & Wang, 2008) about the spectrum and Stieltjes transform of such matrices can be understood using the asymptotic freeness of $W^T W$ and Σ (Gabriel, 2015; Speicher, 2017). In this paper, we provide non-asymptotic variants of these results for an arbitrary matrix Σ (which in our setting is the kernel Gram matrix); the proofs in our setting are detailed in the Supp. Mat.

1.3. Outline

The rest of this paper is organized as follows:

- In Section 2, the setup (linear regression, Gaussian RF model, λ-RF predictor, and λ-KRR predictor) is introduced.
- In Section 3, preliminary results on the distribution of the λ-RF model are provided: the RF predictors are Gaussian mixtures (Proposition 3.1) and the $\lambda \searrow 0$-RF model is unbiased in the overparameterized regime (Corollary 3.2). Graphical illustrations of the RF predictors in various regimes are presented (Figure 1).
- In Section 4, the first main theorem is stated (Theorem 4.1): the average (expected) λ-RF predictor is close to the $\tilde{\lambda}$-KRR predictor for an explicit $\tilde{\lambda} > \lambda$. As a con-

sequence (Corollary 4.3), the test errors of these two predictors are close. Finally, numerical experiments show that the test errors are in fact virtually identical (Figure 2).

- In Section 5, the second main theorem is stated (Theorem 5.1): a bound on the variance of the λ-RF predictor is given, which show that it concentrates around the average λ-RF predictor. As a consequence, the test error of the λ-RF predictor is shown to be close to that of the $\tilde{\lambda}$-KRR predictor (Corollary 5.2). The ridgeless $\lambda \searrow 0$ case is then investigated (Section 5.2): a lower bound on the variance of the λ-RF predictor is given, suggesting an explanation for the double-descent curve in the ridgeless case.
- In Section 6, we summarize our results and discuss potential implications and extensions.

2. Setup

Linear regression is a parametric model consisting of linear combinations

$$f_\theta = \frac{1}{\sqrt{P}}\left(\theta_1\phi^{(1)} + \cdots + \theta_P\phi^{(P)}\right)$$

of (deterministic) features $\phi^{(1)}, \dots, \phi^{(P)} : \mathbb{R}^d \to \mathbb{R}$. We consider an arbitrary training dataset (X, y) with $X = [x_1, ..., x_N] \in \mathbb{R}^{d\times N}$ and $y = [y_1, \dots, y_N] \in \mathbb{R}^N$, where the labels could be noisy observations. For a ridge parameter $\lambda > 0$, the linear estimator corresponds to the parameters $\hat{\theta} = [\hat{\theta}_1, \dots, \hat{\theta}_P] \in \mathbb{R}^P$ that minimize the (regularized) Mean Square Error (MSE) functional $\hat{L}_\lambda$ defined by

$$\hat{L}_\lambda(f_\theta) = \frac{1}{N}\sum_{i=1}^{N}(f_\theta(x_i) - y_i)^2 + \frac{\lambda}{N}\|\theta\|^2. \qquad (1)$$

The *data matrix* F is defined as the $N \times P$ matrix with entries $F_{ij} = \frac{1}{\sqrt{P}}\phi^{(j)}(x_i)$. The minimization of (1) can be rewritten in terms of F as

$$\hat{\theta} = \mathrm{argmin}_\theta \|F\theta - y\|^2 + \lambda\|\theta\|^2. \qquad (2)$$

The optimal solution $\hat{\theta}$ is then given by

$$\hat{\theta} = F^T\left(FF^T + \lambda I_N\right)^{-1} y \qquad (3)$$

and the optimal predictor $\hat{f} = f_{\hat{\theta}}$ by

$$\hat{f}(x) = \frac{1}{\sqrt{P}}\sum_{j=1}^{P}\phi^{(j)}(x)F^T_{:,j}\left(FF^T + \lambda I_N\right)^{-1} y. \qquad (4)$$

In this paper, we consider linear models of *Gaussian random features* associated with a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. We take $\phi^{(j)} = f^{(j)}$, where $f^{(1)}, \dots, f^{(P)}$ are sampled i.i.d. from a Gaussian Process of zero mean (i.e. $\mathbb{E}[f^{(j)}(x)] = 0$ for all $x \in \mathbb{R}^d$) and with covariance K (i.e. $\mathbb{E}[f^{(j)}(x)f^{(j)}(x')] = K(x, x')$ for all $x, x' \in \mathbb{R}^d$). In our setup, the optimal parameter $\hat{\theta}$ still satisfies (3) where F is now a random matrix. The associated predictor, called λ-RF predictor, is then given by

Definition 2.1 (Random Feature Predictor)**.** *Consider a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, a ridge $\lambda > 0$, and random features $f^{(1)}, \dots, f^{(P)}$ sampled i.i.d. from a centered Gaussian Process of covariance K. Let $\hat{\theta}$ be the optimal solution to (1) taking $\phi^{(j)} = f^{(j)}$. The Random Feature predictor with ridge λ is the random function $\hat{f}^{(RF)}_\lambda : \mathbb{R}^d \to \mathbb{R}$ defined by*

$$\hat{f}^{(RF)}_\lambda(x) = \frac{1}{\sqrt{P}}\sum_{j=1}^{P}\hat{\theta}_j f^{(j)}(x). \qquad (5)$$

The λ-RF can be viewed as an approximation of kernel ridge predictors: observing from (4) that $\hat{f}^{(RF)}_\lambda$ only depends on the scalar product $K_P(x, x') = \frac{1}{P}\sum_{j=1}^{P} f^{(j)}(x)f^{(j)}(x')$ between datapoints, we see that as $P \to \infty$, $K_P \to K$ and hence $\hat{f}^{(RF)}_\lambda$ converges (Rahimi & Recht, 2008) to a kernel predictor with ridge λ (Schölkopf et al., 1998), which we call λ-KRR predictor.

Definition 2.2 (Kernel Predictor)**.** *Consider a kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and a ridge $\lambda > 0$. The Kernel Predictor is the function $\hat{f}^{(K)}_\lambda : \mathbb{R}^d \to \mathbb{R}$*

$$\hat{f}^{(K)}_\lambda(x) = K(x, X)(K(X, X) + \lambda \mathrm{I}_N)^{-1} y$$

where $K(X, X)$ is the $N \times N$ matrix of entries $(K(X, X))_{ij} = K(x_i, x_j)$ and $K(\,\cdot\,, X) : \mathbb{R}^d \to \mathbb{R}^N$ is the map $(K(x, X))_i = K(x, x_i)$.

2.1. Bias-Variance Decomposition.

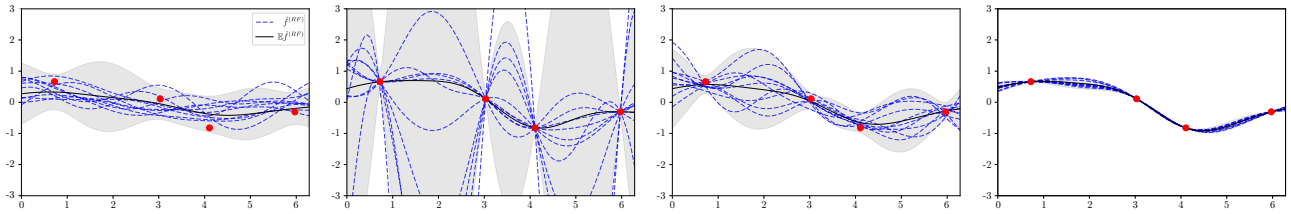
Let us assume that there exists a true regression function $f^* : \mathbb{R}^d \to \mathbb{R}$ and a data generating distribution $\mathcal{D}$ on $\mathbb{R}^d$. The risk of a predictor $f : \mathbb{R}^d \to \mathbb{R}$ is measured by the MSE defined as

$$L(f) = \mathbb{E}_{\mathcal{D}}\left[(f(x) - f^*(x))^2\right].$$

Let π denote the joint distribution of the i.i.d. sample $f^{(1)}, ..., f^{(P)}$ from the centered Gaussian process with covariance kernel K. The risk of $\hat{f}^{(RF)}_\lambda$ can be decomposed into a bias-variance form as

$$\mathbb{E}_\pi\left[L(\hat{f}^{(RF)}_\lambda)\right] = L\left(\mathbb{E}_\pi[\hat{f}^{(RF)}_\lambda]\right) + \mathbb{E}_{\mathcal{D}}\left[\mathrm{Var}_\pi(\hat{f}^{(RF)}_\lambda(x))\right].$$

This decomposition into the risk of the *average* RF predictor and of the $\mathcal{D}$-expectation of its variance will play a crucial



(a) $P = 2, \lambda = 10^{-4}$ (b) $P = 4, \lambda = 10^{-4}$ (c) $P = 4, \lambda = 0.1$ (d) $P = 100, \lambda = 10^{-4}$

Figure 1. Distribution of the RF Predictor. Red dots represent a sinusoidal dataset $y_i = \sin(x_i)$ for $N = 4$ points x_i in $[0, 2\pi)$. For selected P and λ, we sample ten RF predictors (blue dashed lines) and compute empirically the average RF predictor (black lines) with ± 2 standard deviations intervals (shaded regions).

role in the next sections. This is in contrast with the classical bias-variance decomposition in Geman et al. (1992)

$$\mathbb{E}_{\mathcal{D}^{\otimes N}}[L(f)] = L(\mathbb{E}_{\mathcal{D}^{\otimes N}}[f]) + \mathbb{E}_{\mathcal{D}}[\mathrm{Var}_{\mathcal{D}^{\otimes N}}[f(x)]]$$

where $\mathcal{D}^{\otimes N}$ denotes the joint distribution on $x_1, ..., x_N$, sampled i.i.d. from $\mathcal{D}$. Note that in our decomposition no probabilistic assumption is made on the data, which is fixed.

2.2. Additional Notation

In this paper, we consider a fixed dataset (X, y) with distinct data points and a kernel K (i.e. a positive definite symmetric function $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$). We denote by $\|y\|_{K^{-1}}$ the inverse kernel norm of the labels defined as $y^T(K(X,X))^{-1}y$.

Let UDU^T be the spectral decomposition of the kernel matrix $K(X,X)$, with $D = \mathrm{diag}(d_1, \dots, d_N)$. Let $D^{\frac{1}{2}} = \mathrm{diag}(\sqrt{d_1}, \dots, \sqrt{d_N})$ and set $K^{\frac{1}{2}} = UD^{\frac{1}{2}}U^T$. The law of the (random) data matrix F is now that of $\frac{1}{\sqrt{P}}K^{\frac{1}{2}}W^T$ where W is a $P \times N$ matrix of i.i.d. standard Gaussian entries, so that $\mathbb{E}[FF^T] = K(X,X)$.

We will denote by $\gamma = \frac{P}{N}$ the parameter-to-datapoint ratio: the *underparameterized regime* corresponds to $\gamma < 1$, while the *overparameterized regime* corresponds to $\gamma \geq 1$. In order to stress the dependence on the ratio parameter γ, we write $\hat{f}^{(RF)}_{\lambda,\gamma}$ instead of $\hat{f}^{(RF)}_{\lambda}$.

3. First Observations

The distribution of the RF predictor features a variety of behaviors depending on γ and λ, as displayed in Figure 1. In the underparameterized regime $P < N$, sample RF predictors induce some *implicit regularization* and do not interpolate the dataset (1a); at the interpolation threshold $P = N$, RF predictors interpolate the dataset but the variance explodes when there is no ridge (1b), however adding some ridge suppresses variance explosion (1c); in the overparameterized regime $P \geq N$ with large P, the variance vanishes thus the RF predictor converges to its average (1d). We will investigate the average RF predictor (solid lines) in detail in Section 4 and study its variance in Section 5.

We start by characterizing the distribution of the RF predictor as a Gaussian mixture:

Proposition 3.1. *Let $\hat{f}^{(RF)}_{\lambda,\gamma}(x)$ be the random features predictor as in (5) and let $\hat{y} = F\hat{\theta}$ be the prediction vector on training data, i.e. $\hat{y}_i = \hat{f}^{(RF)}_{\lambda,\gamma}(x_i)$. The process $\hat{f}^{(RF)}_{\lambda,\gamma}$ is a mixture of Gaussians: conditioned on F, we have that $\hat{f}^{(RF)}_{\lambda,\gamma}$ is a Gaussian process. The mean and covariance of $\hat{f}^{(RF)}_{\lambda,\gamma}$ conditioned on F are given by*

$$\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}(x)|F] = K(x,X)K(X,X)^{-1}\hat{y}, \tag{6}$$

$$\mathrm{Cov}[\hat{f}^{(RF)}_{\lambda,\gamma}(x), \hat{f}^{(RF)}_{\lambda,\gamma}(x')|F] = \frac{\|\hat{\theta}\|^2}{P}\tilde{K}(x,x'), \tag{7}$$

with $\tilde{K}(x,x') = K(x,x') - K(x,X)K(X,X)^{-1}K(X,x')$ denoting the posterior covariance kernel.

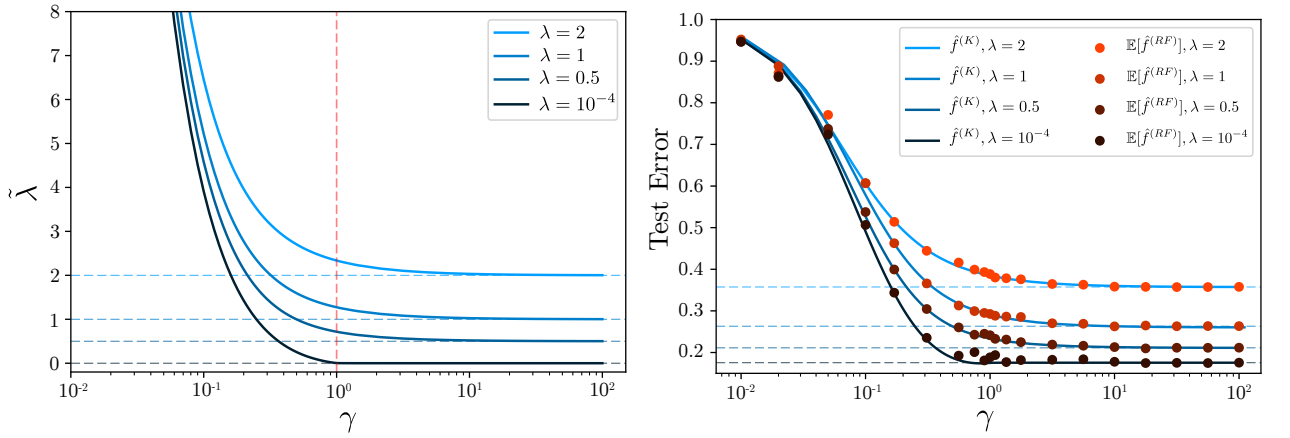
The proof of Proposition 3.1 relies on the fact that $f^{(j)}$ conditioned on $\left(f^{(j)}(x_i)\right)_{i=1,\dots,N}$ is a Gaussian Process.

Note that (6) and (7) depend on λ and P through $\hat{y}$ and $\|\hat{\theta}\|^2$; in fact, as the proof shows, these identities extend to the ridgeless case $\lambda \searrow 0$. For the ridgeless case, when one is in the overparameterized regime ($P \geq N$), one can (with probability one) fit the labels y and hence $\hat{y} = y$:

Corollary 3.2. *When $P \geq N$, the average ridgeless RF predictor is equivalent to the ridgeless KRR predictor*

$$\mathbb{E}\left[\hat{f}^{(RF)}_{\lambda\searrow 0,\gamma}(x)\right] = K(x,X)K(X,X)^{-1}y = \hat{f}^{(K)}_{\lambda\searrow 0}(x).$$

This corollary shows that in the overparameterized case, the ridgeless RF predictor is an unbiased estimator of the ridgeless kernel predictor. The difference between the expected loss of ridgeless RF predictor and that of the ridgeless KRR predictor is hence equal to the variance of the RF predictor. As will be demonstrated in this article, outside of this specific regime, a systematic bias appears, which reveals an implicit regularizing effect of random features.



(a) Evolution of $\tilde{\lambda}$

(b) Average λ-RF predictor vs. $\tilde{\lambda}$-KRR

Figure 2. Comparison of the test errors of the average λ-RF predictor and the $\tilde{\lambda}$-KRR predictor. We train the RF predictors on $N = 100$ MNIST data points where K is the RBF kernel, i.e. $K(x, x') = \exp\left(-\|x - x'\|^2/\ell\right)$. We approximate the average λ-RF on 100 random test points for various ridges λ. In **(a)**, given γ and λ, the effective ridge $\tilde{\lambda}$ is computed numerically using (9). In **(b)**, the test errors of the $\tilde{\lambda}$-KRR predictor (blue lines) and the empirical average of the λ-RF predictor (red dots) agree perfectly.

4. Average Predictor

In this section, we study the average RF predictor $\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}]$. As shown by Corollary 3.2 above, in the ridgeless over-parmeterized regime, the RF predictor is an unbiased estimator of the ridgeless kernel predictor. However, in the presence of a non-zero ridge, we see the following *implicit regularization effect*: the average λ-RF predictor is close to the $\tilde{\lambda}$-KRR predictor for an effective ridge $\tilde{\lambda} > \lambda$ (in other words, sampling a finite number P of features amounts to taking a greater kernel ridge $\tilde{\lambda}$).

Theorem 4.1. *For $N, P > 0$ and $\lambda > 0$, we have*

$$\left|\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}(x)] - \hat{f}^{(K)}_{\tilde{\lambda}}(x)\right| \leq \frac{c\sqrt{K(x,x)}\,\|y\|_{K^{-1}}}{P} \tag{8}$$

where the effective ridge $\tilde{\lambda}(\lambda, \gamma) > \lambda$ is the unique positive number satisfying

$$\tilde{\lambda} = \lambda + \frac{\tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^{N}\frac{d_i}{\tilde{\lambda} + d_i}, \tag{9}$$

and where $c > 0$ depends on λ, γ, and $\frac{1}{N}\mathrm{Tr}K(X, X)$ only.

Proof. (Sketch; see Supp. Mat. for details) Set $A_\lambda = F(F^TF + \lambda I_P)^{-1}F^T$. The vector of the predictions on the training set is given by $\hat{y} = A_\lambda y$ and the expected predictor is given by

$$\mathbb{E}\left[\hat{f}^{(RF)}_{\lambda,\gamma}(x)\right] = K(x, X)K(X, X)^{-1}\mathbb{E}\left[A_\lambda\right] y.$$

By a change of basis, we may assume the kernel Gram matrix to be diagonal, i.e. $K(X, X) = \mathrm{diag}(d_1, \ldots, d_N)$. In this basis $\mathbb{E}\left[A_\lambda\right]$ turns out to be diagonal too. For each $i = 1, \ldots, N$ we can isolate the contribution of the i-th row of F: by the Sherman-Morrison formula, we have $(A_\lambda)_{ii} = \frac{d_i g_i}{1 + d_i g_i}$, where

$$g_i = \frac{1}{P}W_i^T(F_{(i)}^TF_{(i)} + \lambda \mathrm{I}_P)^{-1}W_i,$$

with W_i denoting the i-th column of $W = \sqrt{P}F^TK^{-\frac{1}{2}}$ and $F_{(i)}$ being obtained by removing the i-th row of F. The g_i's are all within $\mathcal{O}(1/\sqrt{P})$ distance to the Stieltjes transform

$$m_P(-\lambda) = \frac{1}{P}\mathrm{Tr}\left(F^TF + \lambda \mathrm{I}_P\right)^{-1}.$$

By a fixed point argument, the Stieltjes transform $m_P(-\lambda)$ is itself within $\mathcal{O}(1/\sqrt{P})$ distance to the deterministic value $\tilde{m}(-\lambda)$, where $\tilde{m}$ is the unique positive solution to

$$\gamma = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i\tilde{m}(z)}{1 + d_i\tilde{m}(z)} - \gamma z\tilde{m}(z).$$

(The detailed proof in the Supp. Mat. uses non-asymptotic variants of arguments found in (Bai & Wang, 2008); the constants in the $\mathcal{O}$ bounds are in particular made explicit).

As a consequence, from the above results, we obtain

$$\mathbb{E}\left[(A_\lambda)_{ii}\right] = \mathbb{E}\left[\frac{d_i g_i}{1 + d_i g_i}\right] \approx \frac{d_i\tilde{m}}{1 + d_i\tilde{m}} = \frac{d_i}{\tilde{\lambda} + d_i},$$

revealing the effective ridge $\tilde{\lambda} = {}^1/_{\tilde{m}(-\lambda)}$.

This implies that $\mathbb{E}[A_\lambda] \approx K(X,X)(K(X,X)+\tilde{\lambda}\mathrm{I}_N)^{-1}$ and

$$\mathbb{E}\left[\hat{f}^{(RF)}_{\lambda,\gamma}(x)\right] \approx K(x,X)(K(X,X)+\tilde{\lambda}\mathrm{I}_N)^{-1}y = \hat{f}^{(K)}_{\tilde{\lambda}}(x),$$

yielding the desired result. □

Note that asymptotic forms of equations similar to the ones in the above proof appear in different settings (Dobriban & Wager, 2018; Mei & Montanari, 2019; Liu & Dobriban, 2020), related to the study of the Stieltjes transform of the product of asymptotically free random matrices.

While the above theorem does not make assumptions on P, N, and K, the case of interest is when the right hand side $\frac{cK(x,x)\|y\|_{K^{-1}}}{P}$ is small. The constant $c > 0$ is uniformly bounded whenever γ and λ are bounded away from 0 and $\frac{1}{N}\mathrm{Tr}K(X,X)$ is bounded from above. As a result, to bound the right hand side of (8), the two quantities we need to bound are $T = \frac{1}{N}\mathrm{Tr}K(X,X)$ and $\|y\|_{K^{-1}}$.

- The boundedness of T is guaranteed for kernels that are translation-invariant, i.e. of the form $K(x,y) = k(\|x-y\|)$: in this case, one has $T = k(0)$.
- If we assume $\mathbb{E}_{\mathcal{D}}[K(x,x)] < \infty$ (as is commonly done in the literature (Rudi & Rosasco, 2017)), T converges to $\mathbb{E}_{\mathcal{D}}[K(x,x)]$ as $N \to \infty$ (assuming i.i.d. data points).
- For $\|y\|_{K^{-1}}$, under the assumption that the labels are of the form $y_i = f^*(x_i)$ for a true regression function f^* lying in Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ of the kernel K (Schölkopf et al., 1998), we have $\|y\|_{K^{-1}} \leq \|f^*\|_{\mathcal{H}}$.

Our numerical experiments in Figure (2b) show excellent agreement between the test error of the expected λ-RF predictor and the one of the $\tilde{\lambda}$-KRR predictor suggesting that the two functions are indeed very close, even for small N, P.

Thanks to the implicit definition of the effective ridge $\tilde{\lambda}$ (which depends on λ, γ, N and on the eigenvalues d_i of $K(X,X)$) we obtain the following:

Proposition 4.2. *The effective ridge $\tilde{\lambda}$ satisfies the following properties:*

1. *for any $\gamma > 0$, we have $\lambda < \tilde{\lambda}(\lambda,\gamma) \leq \lambda + \frac{1}{\gamma}T$;*
2. *the function $\gamma \mapsto \tilde{\lambda}(\lambda,\gamma)$ is decreasing;*
3. *for $\gamma > 1$, we have $\tilde{\lambda} \leq \frac{\gamma}{\gamma-1}\lambda$;*
4. *for $\gamma < 1$, we have $\tilde{\lambda} \geq \frac{1-\sqrt{\gamma}}{\sqrt{\gamma}}\min_i d_i$.*

The above proposition shows the implicit regularization effect of the RF model: sampling fewer features (i.e. decreasing γ) increases the effective ridge $\tilde{\lambda}$.

Furthermore, as $\lambda \to 0$ (ridgeless case), the effective ridge $\tilde{\lambda}$ behave as follows:

- in the overparameterized regime ($\gamma > 1$), $\tilde{\lambda}$ goes to 0;
- in the underparameterized regime ($\gamma < 1$), $\tilde{\lambda}$ goes to a limit $\tilde{\lambda}_0 > 0$.

These observations match the profile of $\tilde{\lambda}$ in Figure (2a).

Remark. When $\lambda \searrow 0$, the constant c in our bound (8) explodes (see Supp. Mat.). As a result, this bound is not directly useful when $\lambda = 0$. However, we know from Corollary 3.2 that in the ridgeless overparametrized case ($\gamma > 1$), the average RF predictor is equal to the ridgeless KRR predictor. In the underparametrized case ($\gamma < 1$), our numerical experiments suggest that the ridgeless RF predictor is an excellent approximation of the $\tilde{\lambda}_0$-KRR predictor.

4.1. Effective Dimension

The effective ridge $\tilde{\lambda}$ is closely related to the so-called effective dimension appearing in statistical learning theory. For a linear (or kernel) model with ridge λ, the *effective dimension* $\mathcal{N}(\lambda) \leq N$ is defined as $\sum_{i=1}^{N}\frac{d_i}{\lambda+d_i}$ (Zhang, 2003; Caponnetto & De Vito, 2007). It allows one to measure the effective complexity of the Hilbert space in the presence of a ridge.

For a given $\lambda > 0$, the effective ridge $\tilde{\lambda}$ introduced in Theorem 4.1 is related to the effective dimension $\mathcal{N}(\tilde{\lambda})$ by

$$\mathcal{N}(\tilde{\lambda}) = P\left(1 - \frac{\lambda}{\tilde{\lambda}}\right).$$

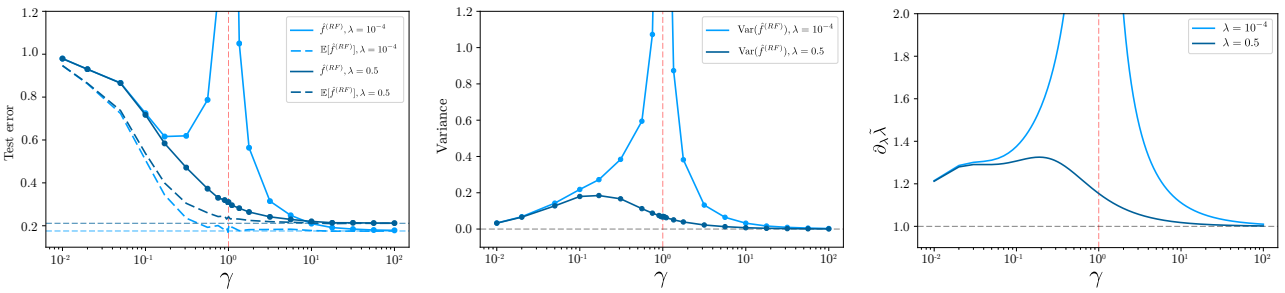
In particular, we have that $\mathcal{N}(\tilde{\lambda}) \leq \min(N,P)$: this shows that the choice of a finite number of features corresponds to an automatic lowering of the effective dimension of the related kernel method.

Note that in the ridgeless underparameterized case ($\lambda \searrow 0$ and $\gamma < 1$), the effective dimension $\mathcal{N}(\tilde{\lambda})$ equals precisely the number of features P.

4.2. Risk of the Average Predictor

A corollary of Theorem 4.1 is that the loss of the expected RF predictor is close to the loss of the KRR predictor with ridge $\tilde{\lambda}$:

Corollary 4.3. *If $\mathbb{E}_{\mathcal{D}}[K(x,x)] < \infty$, we have that the difference of errors $\delta_E = \left|L(\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}]) - L(\hat{f}^{(K)}_{\tilde{\lambda}})\right|$ is*



(a) Ridgeless vs. Ridge (b) Variance of λ-RF (c) Evolution of $\partial_\lambda \tilde{\lambda}$

Figure 3. Average test error of the ridgeless vs. ridge λ-RF predictors. In **(a)**, the average test errors of the ridgeless and the ridge RF predictors (solid lines) and the effect of ensembling (dashed lines) for $N = 100$ MNIST data points. In **(b)**, the variance of the RF predictors and in **(c)**, the evolution of $\partial_\lambda \tilde{\lambda}$ in the ridgeless and ridge cases. The experimental setup is the same as in Figure 2.

bounded from above by

$$\delta_E \leq \frac{C\,\|y\|_{K^{-1}}}{P}\left(2\sqrt{L\left(\hat{f}^{(K)}_{\tilde{\lambda}}\right)} + \frac{C\,\|y\|_{K^{-1}}}{P}\right),$$

where C is given by $c\sqrt{\mathbb{E}_{\mathcal{D}}[K(x,x)]}$, with c the constant appearing in (8) above.

As a result, δ_E can be bounded in terms of $\lambda, \gamma, T, \|y\|_{K^{-1}}$, which are discussed above, and of the kernel generalization error $L(f^{(K)}_{\tilde{\lambda}})$. Such a generalization error can be controlled in a number of settings as N grows: in (Caponnetto & De Vito, 2007; Marteau-Ferey et al., 2019), for instance, the loss is shown to vanish as $N \to \infty$. Figure (2b) shows that the two test losses are indeed very close.

5. Variance

In the previous sections, we analyzed the loss of the expected predictor $\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}]$. In order to analyze the expected loss of the RF predictor $\hat{f}^{(RF)}_{\lambda,\gamma}$, it remains to control the variance of the RF predictor: this follows from the bias-variance decomposition

$$\mathbb{E}\left[L(\hat{f}^{(RF)}_{\lambda,\gamma})\right] = L\left(\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}]\right) + \mathbb{E}_{\mathcal{D}}\left[\mathrm{Var}(\hat{f}^{(RF)}_{\lambda,\gamma}(x))\right],$$

introduced in Section 2.1.

The variance $\mathrm{Var}\left(\hat{f}^{(RF)}_{\lambda,\gamma}(x)\right)$ of the RF predictor can itself be written as the sum

$$\mathrm{Var}\left(\mathbb{E}\left[\hat{f}^{(RF)}_{\lambda,\gamma}(x) \mid F\right]\right) + \mathbb{E}\left[\mathrm{Var}\left(\hat{f}^{(RF)}_{\lambda,\gamma}(x) \mid F\right)\right].$$

By Proposition 3.1, we have

$$\mathbb{E}\left[\hat{f}^{(RF)}_{\lambda,\gamma}(x) \mid F\right] = K(x,X)K(X,X)^{-1}\hat{y}$$

$$\mathrm{Var}\left(\hat{f}^{(RF)}_{\lambda,\gamma}(x) \mid F\right) = \frac{\|\hat{\theta}\|^2}{P}\tilde{K}(x,x).$$

5.1. RF Predictor Concentration

The following theorem allows us to bound both terms:

Theorem 5.1. *There are constants $c_1, c_2 > 0$ depending on λ, γ, T only such that*

$$\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right) \leq \frac{c_1 K(x,x)\|y\|^2_{K^{-1}}}{P}$$

$$\left|\mathbb{E}\|[\hat{\theta}\|^2] - \partial_\lambda\tilde{\lambda} y^T M_{\tilde{\lambda}} y\right| \leq \frac{c_2\|y\|^2_{K^{-1}}}{P},$$

where $\partial_\lambda\tilde{\lambda}$ is the derivative of $\tilde{\lambda}$ with respect to λ and for $M_{\tilde{\lambda}} = K(X,X)(K(X,X)+\tilde{\lambda}I_N)^{-2}$. As a result

$$\mathrm{Var}\left(\hat{f}^{(RF)}_{\lambda,\gamma}(x)\right) \leq \frac{c_3 K(x,x)\|y\|^2_{K^{-1}}}{P},$$

where $c_3 > 0$ depends on λ, γ, T.

Putting the pieces together, we obtain the following bound on the difference $\Delta_E = |\mathbb{E}[L(\hat{f}^{(RF)}_{\lambda,\gamma})] - L(\hat{f}^{(K)}_{\tilde{\lambda}})|$ between the expected RF loss and the KRR loss:

Corollary 5.2. *If $\mathbb{E}_{\mathcal{D}}[K(x,x)] < \infty$, we have*

$$\Delta_E \leq \frac{C_1\|y\|_{K^{-1}}}{P}\left(\sqrt{L(\hat{f}^{(K)}_{\tilde{\lambda}})} + C_2\|y\|_{K^{-1}}\right).$$

where C_1 and C_2 depend on λ, γ, T and $\mathbb{E}_{\mathcal{D}}[K(x,x)]$ only.

5.2. Double Descent Curve

We now investigate the neighborhood of the frontier $\gamma = 1$ between the under- and overparameterized regimes, known empirically to exhibit a double descent curve, where the test error explodes at $\gamma = 1$ (i.e. when $P \approx N$) as exhibited in Figure 3.

Thanks to Theorem 5.1, we get a lower bound on the variance of $\hat{f}^{(RF)}_{\lambda,\gamma}$:

Corollary 5.3. *There exists $c_4 > 0$ depending on λ, γ, T only such that* $\mathrm{Var}(\hat{f}^{(RF)}_{\lambda,\gamma}(x))$ *is bounded from below by*

$$\partial_\lambda \tilde{\lambda} \frac{y^T M_{\tilde{\lambda}} y}{P} \tilde{K}(x,x) - \frac{c_4 K(x,x) \|y\|^2_{K^{-1}}}{P^2}.$$

If we assume the second term of Corollary 5.3 to be negligible, then the only term which depends on P is $\partial_\lambda \tilde{\lambda} \frac{y^T M_{\tilde{\lambda}} y}{P}$. The derivative $\partial_\lambda \tilde{\lambda}$ has an interesting behavior as a function of λ and γ:

Proposition 5.4. *For $\gamma > 1$, as $\lambda \to 0$, the derivative $\partial_\lambda \tilde{\lambda}$ converges to $\frac{\gamma}{\gamma - 1}$. As $\lambda\gamma \to \infty$, we have $\partial_\lambda \tilde{\lambda}(\lambda, \gamma) \to 1$.*

The explosion of $\partial_\lambda \tilde{\lambda}$ in $(\gamma = 1, \lambda = 0)$ is displayed in Figure (3c).

Corollary 5.3 can be used to explain the double-descent curve numerically observed for small $\lambda > 0$. It is natural to assume that in this case $\partial_\lambda \tilde{\lambda} \gg 1$ around $\gamma = 1$, dominating the lower bound in Corollary 5.3. In turn, by Proposition 5.4 this implies that the variance of $\hat{f}^{(RF)}$ gets large. Finally, by the bias-variance decomposition, we obtain a sharp increase of the test error around $\gamma = 1$, which is in line with the results of (Hastie et al., 2019; Mei & Montanari, 2019).

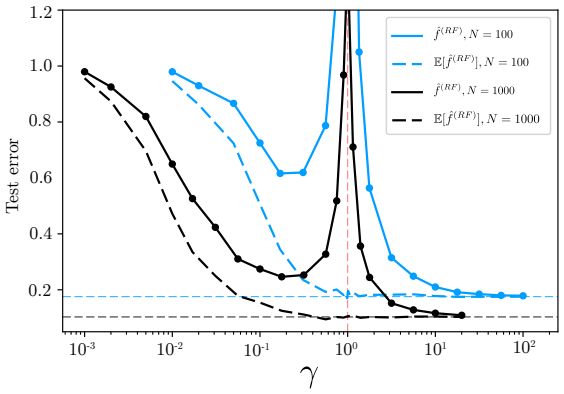
6. Conclusion

In this paper, we have identified the implicit regularization arising from the finite sampling of Random Features (RF): using a Gaussian RF model with ridge parameter $\lambda > 0$ (λ-RF) and feature-to-datapoints ratio $\gamma = \frac{P}{N}$ is essentially equivalent to using a Kernel Ridge Regression with effective ridge $\tilde{\lambda} > \lambda$ ($\tilde{\lambda}$-KRR) which we characterize explicitly. More precisely, we have shown the following:

- The expectation of the λ-RF predictor is very close to the $\tilde{\lambda}$-KRR predictor (Theorem 4.1).
- The λ-RF predictor concentrates around its expectation when λ is bounded away from zero (Theorem 5.1); this implies in particular that the test errors of the λ-RF and $\tilde{\lambda}$-KRR predictors are close to each other (Corollary 5.2).

Both theorems are proven using tools from random matrix theory, in particular finite-size results on the concentration of the Stieltjes transform of general Wishart matrix models. While our current proofs require the assumption that the RF model is Gaussian, it seems natural to postulate that the results and the proofs extend to more general setups, along the lines of (Louart et al., 2017; Benigni & Péché, 2019).

Our numerical verifications on the expected λ-RF predictor and the $\tilde{\lambda}$-KRR predictor have shown that both are in excellent agreement. This shows in particular that in order to use RF predictors to approximate KRR predictors with a given ridge, one should choose both the number of features and the explicit ridge appropriately.



(a) $N = 100$ vs. $N = 1000$

Figure 4. Average test error of the λ-RF predictor for two values of N and $\lambda = 10^{-4}$. For $N = 1000$, the test error is naturally lower and the cusp at $\gamma = 1$ is narrower than for $N = 100$. The experimental setup is the same as in Figure 2.

Finally, we investigate the ridgeless limit case $\lambda \searrow 0$. In this case, we see a sharp transition at $\gamma = 1$: in the overparameterized regime $\gamma > 1$, the effective ridge goes to zero, while in the underparameterized regime $\gamma < 1$, it converges to a positive value. At the interpolation threshold $\gamma = 1$, the variance of the λ-RF explodes, leading to the double descent curve emphasized in (Advani & Saxe, 2017; Spigler et al., 2018; Belkin et al., 2018; Nakkiran et al., 2019). We investigate this numerically and prove a lower bound yielding a plausible explanation for this phenomenon.

Thanks and Acknowledgements

The authors would like to thank Andrea Montanari, Song Mei, Lénaïc Chizat and Alessandro Rudi for the helpful discussions. Clément Hongler acknowledges support from the ERC SG CONSTAMIS grant, the NCCR SwissMAP grant, the Minerva Foundation, the Blavatnik Family Foundation, and the Latsis foundation.

References

Advani, M. S. and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017. URL http://arxiv.org/abs/1710.03667.

Bai, Z. and Wang, Z. Large sample covariance matrices without independence structures in columns. *Statistica Sinicia*, 18:425–442, 2008.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019. URL http://arxiv.org/abs/1906.11300.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018. URL http://arxiv.org/abs/1812.11118.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019. URL http://arxiv.org/abs/1903.07571.

Benigni, L. and Péché, S. Eigenvalue distribution of nonlinear models of random matrices. *arXiv preprint arXiv:1904.03090*, 2019. URL http://arxiv.org/abs/1904.03090.

Bishop, C. M. *Pattern recognition and machine learning*. springer, 2006.

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

d'Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. Double trouble in double descent: Bias and variance (s) in the lazy regime. *arXiv preprint arXiv:2003.01054*, 2020.

Dobriban, E. and Wager, S. High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.*, 46(1):247–279, 02 2018. doi: 10.1214/17-AOS1549. URL https://doi.org/10.1214/17-AOS1549.

Gabriel, F. Combinatorial theory of permutation-invariant random matrices ii: Cumulants, freeness and Levy processes. *arXiv preprint arXiv:1507.02465*, 2015. URL http://arxiv.org/abs/1507.02465.

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019. URL http://arxiv.org/abs/1901.01608.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019. URL http://arxiv.org/abs/1903.08560.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

Liu, S. and Dobriban, E. Ridge regression: Structure, cross-validation, and sketching. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklRwaEKwB.

Louart, C., Liao, Z., and Couillet, R. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28, 02 2017. doi: 10.1214/17-AAP1328.

Marteau-Ferey, U., Ostrovskii, D., Bach, F., and Rudi, A. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. *CoRR*, abs/1902.03046, 2019. URL http://arxiv.org/abs/1902.03046.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019. URL http://arxiv.org/abs/1908.05355.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019. URL http://arxiv.org/abs/1912.02292.

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., and Mitliagkas, I. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018. URL http://arxiv.org/abs/1810.08591.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014. URL http://arxiv.org/abs/1412.6614.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

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

Rudi, A. and Rosasco, L. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pp. 3215–3225, 2017.

Schölkopf, B., Smola, A., and Müller, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

Silverstein, J. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331 – 339, 1995. ISSN 0047-259X. doi: https://doi.org/10.1006/jmva.1995.1083. URL http://www.sciencedirect.com/science/article/pii/S0047259X85710834.

Speicher, R. Free probability and random matrices. In *Free Probability and Random Matrices*, 2017.

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. A jamming transition from under-to over-parametrization affects loss landscape and generalization. *arXiv preprint arXiv:1810.09665*, 2018. URL http://arxiv.org/abs/1810.09665.

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

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

Yu, F. X. X., Suresh, A. T., Choromanski, K. M., Holtmann-Rice, D. N., and Kumar, S. Orthogonal random features. In *Advances in Neural Information Processing Systems*, pp. 1975–1983, 2016.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016. URL http://arxiv.org/abs/1611.03530.

Zhang, T. Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems*, pp. 471–478, 2003.