ON THE UNIVERSALITY OF THE DOUBLE DESCENT PEAK IN RIDGELESS REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

We prove a non-asymptotic distribution-independent lower bound for the expected mean squared generalization error caused by label noise in ridgeless linear regression. In contrast to previous works, our analysis applies to a broad class of input distributions with almost surely full-rank feature matrices, which allows us to cover various types of deterministic or random feature maps. Our lower bound implies that in the presence of label noise, ridgeless linear regression does not perform well around the interpolation threshold for any of these feature maps. We analyze the imposed assumptions in detail and provide a theory for analytic (random) feature maps. Using this theory, we can show that our assumptions are satisfied for input distributions with a (Lebesgue) density and a feature map given by a random deep neural network with analytic activation functions like sigmoid, tanh, softplus or GELU. As further examples, we show that feature maps from random Fourier features and polynomial kernels also satisfy our assumptions. We complement our theory with further experimental and analytic results.

1 INTRODUCTION

Seeking for a better understanding of the successes of deep learning, Zhang et al. (2016) pointed out that deep neural networks can achieve very good performance despite being able to fit random noise, which sparked the interest of many researchers in studying the performance of interpolating learning methods. Belkin et al. (2018) made a similar observation for kernel methods and showed that classical generalization bounds are unable to explain this phenomenon. Belkin et al. (2019a) observed a "double descent" phenomenon in various learning models, where the test error first decreases with increasing model complexity, then increases towards the "interpolation threshold" where the model is first able to fit the training data perfectly, and then decreases again in the "overparameterized" regime where the model capacity is larger than the training set. This phenomenon has also been discovered in several other works (Bös & Opper, 1997; Advani & Saxe, 2017; Neal et al., 2018; Spigler et al., 2019). Nakkiran et al. (2019) performed a large empirical study on deep neural networks and found that double descent can not only occur as a function of model capacity, but also as a function of the number of training epochs or as a function of the number of training samples.

Theoretical investigations of the double descent phenomenon have mostly focused on rather specialized unregularized ("ridgeless") or weakly regularized linear regression models. Such linear models can be described via i.i.d. samples $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d$, where the covariates $\boldsymbol{x}_i$ are mapped to feature representations $\boldsymbol{z}_i = \phi(\boldsymbol{x}_i) \in \mathbb{R}^p$ via a (potentially random) feature map ϕ, and (ridgeless) linear regression is then performed on the transformed samples $(\boldsymbol{z}_i, y_i)$. While linear regression with random features can be understood as a simplified model of fully trained neural networks, it is also interesting in its own right: For example, random Fourier features (Rahimi & Recht, 2008) and random neural network features (see e.g. Cao et al., 2018; Scardapane & Wang, 2017) have gained a notable amount of attention.

Existing theoretical investigations of double descent are usually limited in one or more of the following ways:

(1) They assume that the $\boldsymbol{z}_i$ (or a linear transformation thereof) have i.i.d. components. While this assumption facilitates the application of some random matrix theory results, it excludes most feature maps: For feature maps ϕ with $d < p$, the $\boldsymbol{z}_i$ will usually be concentrated on a d-dimensional submanifold of $\mathbb{R}^p$, and will therefore usually *not* have i.i.d. components.

The assumption of (centered) i.i.d. components is made by Hastie et al. (2019), while Advani & Saxe (2017) and Belkin et al. (2019b) even assume that the $\boldsymbol{z}_i$ follow a Gaussian distribution. Here, Advani & Saxe (2017) study the regularized case and in particular how the optimal regularization depends on the signal-to-noise ratio.

(2) They assume a (shallow) random feature model with fixed distribution of the $\boldsymbol{x}_i$, e.g. an isotropic Gaussian distribution or a uniform distribution on a sphere. Examples for this are the single-layer random neural network feature models by Hastie et al. (2019) in the unregularized case and by Mei & Montanari (2019); d'Ascoli et al. (2020) in the regularized case. A simple Fourier model with $d = 1$ has been studied by Belkin et al. (2019b).

(3) Their analysis only applies in a high-dimensional limit where $n, p \to \infty$ and $n/p \to \gamma$, where $\gamma \in (0, \infty)$ is a constant. This applies to all works mentioned in (1) and (2) except the model by Belkin et al. (2019b) where the $\boldsymbol{z}_i$ follow a standard Gaussian distribution.

Muthukumar et al. (2020) study the "fundamental price of interpolation" in the overparameterized regime, providing a lower bound for the generalization error under the assumption of subgaussian features or (suitably) bounded features. While these assumptions are weaker than (1) and partially weaker than (2), we discuss in Section 4 that our lower bound requires even weaker assumptions, is stronger, also applies to the underparameterized regime and is uniform across feature maps and input distributions.

Chen et al. (2020) show that in ridgeless linear regression, the feature distributions can be designed to control the locations of ascents and descents in the double descent curve for a "dimension-normalized" noise-induced generalization error. Our lower bound provides a fundamental limit to this "designability" of the generalization curve for methods that can interpolate with probability one in the overparameterized regime.

Our contributions are:

(1) In Section 4, we provide a non-asymptotic distribution-independent lower bound for the expected generalization error in ridgeless linear regression with (random) features that relies on significantly weaker assumptions than previous work. Our result shows that there is only limited potential to reduce the sensitivity of unregularized linear models to label noise via engineering better feature maps.

(2) In Section 5, we prove our assumptions, especially the strongest assumption that the feature matrix has full rank with probability one, for a large class of input distributions and feature maps including random deep neural networks. This analysis corrects and improves related flawed attempts in the literature (see Appendix E) and is also relevant for related work where similar assumptions are not investigated (e.g. Muthukumar et al., 2020).

(3) In Section 6 and Appendix C, we compare our lower bound to specific examples for which we provide new theoretical and experimental results, including finite-width Neural Tangent Kernels (Jacot et al., 2018).

Proofs for our statements can be found in the appendix and all experimental results can be reproduced using the supplemental code.

2 Basic Setting and Notation

Following Györfi et al. (2002), we consider the scenario where the samples $(\boldsymbol{x}_i, y_i)$ of a data set $D = ((\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)) \in (\mathbb{R}^p \times \mathbb{R})^n$ are sampled independently from a probability distribution P on $\mathbb{R}^d \times \mathbb{R}$, i.e. $D \sim P^n$.[1] We define

$$\boldsymbol{X} := \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}, \qquad \boldsymbol{y} := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n \ .$$

We also consider random variables $(\boldsymbol{x}, y) \sim P$ that are independent of D and denote the distribution of $\boldsymbol{x}$ by P_X. The (least squares) *population risk* of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$R_P(f) := \mathbb{E}_{\boldsymbol{x},y}(y - f(\boldsymbol{x}))^2 \ .$$

[1] Although many of our theorems apply to general domains $\boldsymbol{x}_i \in \mathcal{X}$ and not just $\mathcal{X} = \mathbb{R}^d$, we set $\mathcal{X} = \mathbb{R}^d$ for notational simplicity. We require $\mathcal{X} = \mathbb{R}^d$ whenever we assume that the distribution of the $\boldsymbol{x}_i$ has a Lebesgue density or work with analytic feature maps.

We assume $\mathbb{E}y^2 < \infty$. Then, R_P is minimized by the target function f_P^* given by

$$f_P^*(\boldsymbol{x}) = \mathbb{E}(y|\boldsymbol{x}) \ ,$$

we have $R_P(f_P^*) < \infty$, and the *excess risk* (a.k.a. generalization error) of a function f is

$$R_P(f) - R_P(f_P^*) = \mathbb{E}_{\boldsymbol{x}}(f(\boldsymbol{x}) - f_P^*(\boldsymbol{x}))^2 \ .$$

Notation For two symmetric matrices, we write $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite and $\boldsymbol{A} \succ \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive definite. For a symmetric matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n}$, we let $\lambda_1(\boldsymbol{S}) \geq \ldots \geq \lambda_n(\boldsymbol{S})$ be its eigenvalues in descending order. We denote the trace of $\boldsymbol{A}$ by $\operatorname{tr}(\boldsymbol{A})$ and the Moore-Penrose pseudoinverse of $\boldsymbol{A}$ by $\boldsymbol{A}^+$. For $\phi : \mathbb{R}^d \to \mathbb{R}^p$ and $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, we let $\phi(\boldsymbol{X}) \in \mathbb{R}^{n \times p}$ be the matrix with ϕ applied to each of the *rows* of $\boldsymbol{X}$ individually. For a set $\mathcal{A}$, we denote its indicator function by $\mathbb{1}_{\mathcal{A}}$. For a random variable $\boldsymbol{x}$, we say that $\boldsymbol{x}$ has a Lebesgue density if P_X can be represented by a probability density function (w.r.t. the Lebesgue measure). We denote the uniform distribution on a set $\mathcal{A}$, e.g. the unit sphere $\mathbb{S}^{p-1} \subseteq \mathbb{R}^p$, by $\mathcal{U}(\mathcal{A})$. We denote the normal (Gaussian) distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For $n \in \mathbb{N}$, we define $[n] \coloneqq \{1, \ldots, n\}$.

We review relevant matrix facts, e.g. concerning the Moore-Penrose pseudoinverse, in Appendix B.

3 Linear Regression With (Random) Features

The most general setting that we will consider in this paper is ridgeless linear regression in random features: Given a random variable $\boldsymbol{\theta}$ that is independent from the data set D and an associated random feature map $\phi_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^p$, we define the estimator

$$f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}) \coloneqq \phi_{\boldsymbol{\theta}}(\boldsymbol{x})^\top \phi_{\boldsymbol{\theta}}(\boldsymbol{X})^+ \boldsymbol{y} \ .$$

As a special case of this setting, the feature map $\phi_{\boldsymbol{\theta}}$ may be deterministic, in which case we drop the index $\boldsymbol{\theta}$. An even more specialized case is ordinary linear regression, where $d = p$ and $\phi_{\boldsymbol{\theta}} = \mathrm{id}$, yielding $f_{\boldsymbol{X},\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{X}^+ \boldsymbol{y}$.

As described in Hastie et al. (2019), the ridgeless linear regression parameter $\widehat{\boldsymbol{\beta}} \coloneqq \phi_{\boldsymbol{\theta}}(\boldsymbol{X})^+ \boldsymbol{y}$

- has minimal Euclidean norm among all parameters $\boldsymbol{\beta}$ minimizing $\|\phi_{\boldsymbol{\theta}}(\boldsymbol{X})\boldsymbol{\beta} - \boldsymbol{y}\|_2^2$,
- is the limit of gradient descent with sufficiently small step size on $L(\boldsymbol{\beta}) \coloneqq \|\phi_{\boldsymbol{\theta}}(\boldsymbol{X})\boldsymbol{\beta} - \boldsymbol{y}\|_2^2$ with initialization $\boldsymbol{\beta}^{(0)} \coloneqq \boldsymbol{0}$, and
- is the limit of ridge regression with regularization $\lambda > 0$ for $\lambda \searrow 0$:

$$\widehat{\boldsymbol{\beta}} = \lim_{\lambda \searrow 0} \phi_{\boldsymbol{\theta}}(\boldsymbol{X})^\top (\phi_{\boldsymbol{\theta}}(\boldsymbol{X})\phi_{\boldsymbol{\theta}}(\boldsymbol{X})^\top + \lambda \boldsymbol{I}_n)^{-1} \boldsymbol{y} \ . \tag{1}$$

For a fixed feature map ϕ, the kernel trick provides a correspondence between ridgeless linear regression with ϕ and ridgeless kernel regression with the kernel $k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \coloneqq \phi(\boldsymbol{x})^\top \phi(\tilde{\boldsymbol{x}})$ via

$$f_{\boldsymbol{X},\boldsymbol{y}}(\boldsymbol{x}) = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{X})^+ \boldsymbol{y} = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{X})^\top (\phi(\boldsymbol{X})\phi(\boldsymbol{X})^\top)^+ \boldsymbol{y} = k(\boldsymbol{x}, \boldsymbol{X}) k(\boldsymbol{X}, \boldsymbol{X})^+ \boldsymbol{y} \ , \tag{2}$$

where

$$k(\boldsymbol{x}, \boldsymbol{X}) \coloneqq \begin{pmatrix} k(\boldsymbol{x}, \boldsymbol{x}_1) \\ \vdots \\ k(\boldsymbol{x}, \boldsymbol{x}_n) \end{pmatrix}, \qquad k(\boldsymbol{X}, \boldsymbol{X}) \coloneqq \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \ldots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \ldots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{pmatrix} \ .$$

4 A Lower Bound

In this section, we prove our main theorem, which provides a non-asymptotic distribution-independent lower bound on the expected excess risk.

The expected excess risk $\mathbb{E}_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}} R_P(f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}) - R_P(f_P^*)$ can be decomposed into several different contributions (see e.g. d'Ascoli et al., 2020). In the following, we will focus on the contribution of label noise to the expected excess risk for the estimators $f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}$ considered in Section 3. Using a bias-variance decomposition with respect to $\boldsymbol{y}$, it is not hard to show that the label-noise-induced error provides a lower bound for the expected excess risk:

$$\mathcal{E}_{\text{Noise}} \coloneqq \mathbb{E}_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{x}} \left(f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}} f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}) \right)^2$$

$$
\begin{aligned}
&\leq \mathbb{E}_{\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{x}} \left[\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}} \left(f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}} f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}) \right)^2 + \left(\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}} f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}) - f_P^*(\boldsymbol{x}) \right)^2 \right] \\
&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{x}} \left(f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}) - f_P^*(\boldsymbol{x}) \right)^2 \\
&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}} (R_P(f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}) - R_P(f_P^*)) \,.
\end{aligned}
$$

For linear models as considered here, it is not hard to see that $\mathcal{E}_{\text{Noise}}$ does not depend on f_P^* and is equal to the expected excess risk in the special case $f_P^* \equiv 0$.

In the following, we first consider the setting where the feature map ϕ is deterministic. We will consider linear regression on $\boldsymbol{z} \coloneqq \phi(\boldsymbol{x})$ and $\boldsymbol{Z} \coloneqq \phi(\boldsymbol{X})$ and formulate our assumptions directly w.r.t. the distribution P_Z of $\boldsymbol{z}$, hiding the dependence on the feature map ϕ. While the distribution P_X of $\boldsymbol{x}$ is usually fixed and determined by the problem, the distribution P_Z can be actively influenced by choosing a suitable feature map ϕ. We will analyze in Section 5 how the assumptions on P_Z can be translated back to assumptions on P_X and assumptions on ϕ.

Remark 1. For typical feature maps, we have $p > d$, P_Z is concentrated on a d-dimensional submanifold of $\mathbb{R}^p$ and the components of $\boldsymbol{z}$ are not independent. A simple example is a polynomial feature map $\phi : \mathbb{R}^1 \to \mathbb{R}^p, x \mapsto (1, x, x^2, \ldots, x^{p-1})$. The imposed assumptions on P_Z should hence allow for such distributions on submanifolds and *not* require independent components. ◀

Definition 2. Assuming that $\mathbb{E}\|\boldsymbol{z}\|_2^2 < \infty$, i.e. (MOM) in Theorem 3 holds, we can define the (positive semidefinite) second moment matrix

$$
\boldsymbol{\Sigma} \coloneqq \mathbb{E}_{\boldsymbol{z} \sim P_Z} \left[\boldsymbol{z}\boldsymbol{z}^\top \right] \in \mathbb{R}^{p \times p} \,.
$$

If $\mathbb{E}\boldsymbol{z} = 0$, $\boldsymbol{\Sigma}$ is also the covariance matrix of $\boldsymbol{z}$. If $\boldsymbol{\Sigma}$ is invertible, i.e. (COV) in Theorem 3 holds, the rows $\boldsymbol{w}_i \coloneqq \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$ of the "whitened" data matrix $\boldsymbol{W} \coloneqq \boldsymbol{Z}\boldsymbol{\Sigma}^{-1/2}$ satisfy $\mathbb{E}\boldsymbol{w}_i\boldsymbol{w}_i^\top = \boldsymbol{I}_p$. ◀

With these preparations, we can now state our main theorem. Its assumptions and results will be discussed in Section 5 and Section 6, respectively.

Theorem 3 (Main result)**.** *Let $n, p \geq 1$. Assume that P and ϕ satisfy:*

(INT) $\mathbb{E}y^2 < \infty$ *and hence* $R_P(f_P^*) < \infty$,
(NOI) $\operatorname{Var}(y|\boldsymbol{z}) \geq \sigma^2$ *almost surely over* $\boldsymbol{z}$,
(MOM) $\mathbb{E}\|\boldsymbol{z}\|_2^2 < \infty$, *i.e.* $\boldsymbol{\Sigma}$ *exists and is finite,*
(COV) $\boldsymbol{\Sigma}$ *is invertible,*
(FRK) $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ *almost surely has full rank (i.e.* $\operatorname{rank} \boldsymbol{Z} = \min\{n, p\}$*).*

Then, for the ridgeless linear regression estimator $f_{\boldsymbol{Z},\boldsymbol{y}}(\boldsymbol{z}) = \boldsymbol{z}^\top \boldsymbol{Z}^+ \boldsymbol{y}$, the following holds:

$$
\text{If } p \geq n, \qquad \mathcal{E}_{\text{Noise}} \overset{(*)}{\geq} \sigma^2 \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) \overset{(**)}{\geq} \sigma^2 \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) \geq \sigma^2 \frac{n}{p+1-n} \,.
$$

$$
\text{If } p \leq n, \qquad \mathcal{E}_{\text{Noise}} \overset{(*)}{\geq} \sigma^2 \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) \overset{(***)}{=} \sigma^2 \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{W}^\top\boldsymbol{W})^{-1}) \geq \sigma^2 \frac{p}{n} \,.
$$

Here, the matrix inverses exist almost surely in the considered cases. Moreover, we have:

- *If (NOI) holds with equality, then $(*)$ holds with equality.*
- *If $n = p$ or $\boldsymbol{\Sigma} = \lambda \boldsymbol{I}_p$ for some $\lambda > 0$, then $(**)$ holds with equality.*

We can extend Theorem 3 to random features if it holds for almost all of the random feature maps:

Corollary 4 (Random features)**.** *Let $\boldsymbol{\theta} \sim P_\Theta$ be a random variable such that $\phi_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^p$ is a random feature map. Consider the random features regression estimator $f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{z}_{\boldsymbol{\theta}}^\top \boldsymbol{Z}_{\boldsymbol{\theta}}^+ \boldsymbol{y}$ with $\boldsymbol{z}_{\boldsymbol{\theta}} \coloneqq \phi_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $\boldsymbol{Z}_{\boldsymbol{\theta}} \coloneqq \phi_{\boldsymbol{\theta}}(\boldsymbol{X})$. If for P_Θ-almost all $\boldsymbol{\theta}_{\text{const}}$, the assumptions of Theorem 3 are satisfied for $\boldsymbol{z} = \boldsymbol{z}_{\boldsymbol{\theta}_{\text{const}}}$ and $\boldsymbol{Z} = \boldsymbol{Z}_{\boldsymbol{\theta}_{\text{const}}}$ (with the corresponding matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\text{const}}}$), then*

$$
\mathcal{E}_{\text{Noise}} \geq \begin{cases} \sigma^2 \frac{n}{p+1-n} & \text{if } p \geq n, \\ \sigma^2 \frac{p}{n} & \text{if } p \leq n. \end{cases}
$$

The main novelty in Theorem 3 is the explicit uniform lower bound: Statements similar to $(*)$, $(**)$ and $(***)$ have already been proven, see e.g. Hastie et al. (2019) and Theorem 1 in Muthukumar et al. (2020), but they have only been analyzed under significantly stronger assumptions, cf. Section 1. In

particular, in their Corollary 1, Muthukumar et al. (2020) provide a lower bound holding with high probability for $\boldsymbol{\varepsilon}^\top(\boldsymbol{W}\boldsymbol{W}^\top)^{-1}\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ is a noise vector independent of $\boldsymbol{W}$. Since $\mathbb{E}\boldsymbol{\varepsilon}^\top(\boldsymbol{W}\boldsymbol{W}^\top)^{-1}\boldsymbol{\varepsilon} = \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}\mathbb{E}_{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1})$, their lower bound yields a lower bound for $\mathbb{E}\operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1})$. However, the resulting lower bound only applies to the case $p \geq n$, and even in this case, it is weaker and requires stronger assumptions:

(1) Under the assumption that the $\boldsymbol{w}_i$ have a subgaussian norm bounded by a constant K, i.e. $\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \sup_{q\geq 1} q^{-1/2}(\mathbb{E}|\boldsymbol{v}^\top\boldsymbol{w}_i|^q)^{1/q} \leq K$, they obtain a lower bound of the form $c_K\sigma^2\frac{n}{p}$ with a constant $c_K > 0$ that is not explicitly specified and depends on K. This assumption is satisfied e.g. if the components of $\boldsymbol{w}_i$ are independent and all satisfy the assumption, which is not realistic, cf. Remark 1. As a special case, they explicitly provide c_K for Gaussian P_Z. In contrast, our lower bound is explicit, independent of bounds like K and has a much higher peak at $n = p$ of size $\sigma^2 n$ instead of $\sigma^2 c_K$.

(2) Under the assumption that $\|\boldsymbol{w}_i\|_2^2 \leq p$ almost surely, they obtain a lower bound of the form $c\sigma^2\frac{n}{p\log(n)}$. First of all, this lower bound converges to zero as $n = p \to \infty$. Moreover, thanks to $\mathbb{E}\|\boldsymbol{w}_i\|_2^2 = \mathbb{E}\operatorname{tr}(\boldsymbol{w}_i\boldsymbol{w}_i^\top) = \mathbb{E}\operatorname{tr}(\boldsymbol{I}_p) = p$, the assumption implies $\|\boldsymbol{w}_i\|_2^2 = p$ almost surely. Although we can sometimes guarantee constant $\|\boldsymbol{z}_i\|_2^2$, e.g. for certain random Fourier features, we cannot guarantee the same for $\boldsymbol{w}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{z}_i$ since $\boldsymbol{\Sigma}$ depends on the unknown input distribution P_X.

In particular, with their lower bound, it might still be possible to achieve good results around $n \approx p$ by choosing the feature map ϕ such that the $\boldsymbol{w}_i$ have a sufficiently large (or even infinite) subgaussian norm. Our result shows that this is not possible: Essentially the only possibility to avoid a large $\mathcal{E}_{\text{Noise}}$ for ridgeless linear regression around $n \approx p$ is to violate the property (FRK) that guarantees the ability to interpolate the data in the overparameterized case, see Section 5. Otherwise, in order to achieve $\mathcal{E}_{\text{Noise}} < \varepsilon\sigma^2$, $\varepsilon \ll 1$, it is necessary to make the model either strongly underparameterized ($p < \varepsilon n$) or strongly overparameterized ($p > n/\varepsilon$). As $(**)$ and $(***)$ indicate, the second moment matrix $\boldsymbol{\Sigma}$ can influence the result in the overparameterized but not in the underparameterized regime. We further discuss this in Remark G.1.

5 When are the Assumptions Satisfied?

In this section, we want to discuss the assumptions of Theorem 3 and provide different results helping to verify these assumptions for various input distributions and feature maps. The theory will be particularly nice for analytic feature maps, which we define now:

Definition 5 (Analytic function)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is called (real) analytic if for all $\boldsymbol{z} \in \mathbb{R}^d$, the Taylor series of f around $\boldsymbol{z}$ converges to f in a neighborhood of $\boldsymbol{z}$. A function $f : \mathbb{R}^d \to \mathbb{R}^p, \boldsymbol{z} \mapsto (f_1(\boldsymbol{z}), \ldots, f_p(\boldsymbol{z}))$ is called (real) analytic if $f_1, \ldots, f_p$ are analytic. ◀

Sums, products and compositions of analytic functions are analytic, cf. e.g. Section 2.2 in Krantz & Parks (2002). We will discuss examples of analytic functions later in this section.

Proposition 6 (Characterization of (COV) and (FRK))**.** *Consider the setting of Theorem 3 and let* $\mathrm{FRK}(n)$ *be the statement that (FRK) holds for* n*. Then,*

(i) Let $n \geq 1$*. Then,* $\mathrm{FRK}(n)$ *iff* $P(\boldsymbol{z} \in U) = 0$ *for all linear subspaces* $U \subseteq \mathbb{R}^p$ *of dimension* $\min\{n, p\} - 1$*.*

(ii) Let (MOM) hold. Then, (COV) holds iff $P(\boldsymbol{z} \in U) < 1$ *for all linear subspaces* $U \subseteq \mathbb{R}^p$ *of dimension* $p - 1$*.*

Assuming that (MOM) holds such that (COV) is well-defined, consider the following statements:

(a) $\mathrm{FRK}(p)$ *holds.*

(b) $\mathrm{FRK}(n)$ *holds for all* $n \geq 1$*.*

(c) (COV) holds.

(d) There exists a fixed deterministic matrix $\boldsymbol{X}_{\text{const}} \in \mathbb{R}^{p\times d}$ *such that* $\det(\phi(\boldsymbol{X}_{\text{const}})) \neq 0$*.*

We have (a) $\Leftrightarrow$ *(b)* $\Rightarrow$ *(c)* $\Rightarrow$ *(d). Furthermore, if* $\boldsymbol{x} \in \mathbb{R}^d$ *has a Lebesgue density and* ϕ *is analytic, then (a) – (d) are equivalent.*

With this in mind, we can characterize the assumptions now:

- The assumption (INT) is standard (see e.g. Section 1.6 in Györfi et al., 2002) and guarantees $R_P(f_P^*) < \infty$, such that the excess risk is well-defined.
- The assumption (NOI) is required to ensure the existence of sufficient label noise. Importantly, Lemma I.1 shows that (NOI), i.e. $\mathrm{Var}(y|\boldsymbol{z}) \geq \sigma^2$ almost surely over $\boldsymbol{z}$, holds if $\mathrm{Var}(y|\boldsymbol{x}) \geq \sigma^2$ almost surely over $\boldsymbol{x}$. All related works from Section 1 make the stronger assumption that the distribution of $y - \mathbb{E}(y|\boldsymbol{x})$ is independent of $\boldsymbol{x}$ or even Gaussian.
- The assumption (MOM) can be reformulated as $\mathbb{E}\|\boldsymbol{z}\|_2^2 = \mathbb{E}\|\phi(\boldsymbol{x})\|_2^2 = \mathbb{E}k(\boldsymbol{x}, \boldsymbol{x}) < \infty$. For example, if k or equivalently ϕ are bounded, or if ϕ is continuous and $\|\boldsymbol{x}\|_2$ is bounded, then (MOM) is satisfied. Such assumptions are frequently imposed (see e.g. Chapters 6, 7, 8, 10 in Györfi et al., 2002). In this sense, (MOM) is a standard assumption.
- The assumptions (COV) and $\mathrm{FRK}(n)$ are implied by $\mathrm{FRK}(p)$ and are equivalent to $\mathrm{FRK}(p)$ in the underparameterized case $p \leq n$. In the following, we will therefore focus on proving $\mathrm{FRK}(p)$ for various ϕ and P_X. In the case $p = n$, $\mathrm{FRK}(p)$ ensures that $f_{\boldsymbol{X},\boldsymbol{y}}$ almost surely interpolates the data, or equivalently that the kernel matrix $k(\boldsymbol{X}, \boldsymbol{X})$ almost surely has full rank. Importantly, $\mathrm{FRK}(p)$ is weaker than requiring a strictly positive definite kernel, since strictly positive definite kernels require $p = \infty$. Example D.1 shows that the assumption (FRK) in Theorem 3 cannot be removed.

Choosing the analytic function $\phi = \mathrm{id}$ in Proposition 6 yields a simple sufficient criterion: If $\boldsymbol{z}$ has a Lebesgue density, then (FRK) holds for all n. However, Proposition 6 is also very useful for other analytic feature maps, as we will see in the remainder of this section.

Remark 7. Suppose that $\phi \not\equiv 0$ is analytic, $\boldsymbol{x}$ has a Lebesgue density and (INT), (MOM) and (NOI) are satisfied. If (d) in Proposition 8 does not hold, there exists $\tilde{p} < p$ such that the lower bound from Theorem 3 holds with p replaced by $\tilde{p}$: Let $U := \mathrm{Span}\{\phi(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\}$. Since $\phi \not\equiv 0$, $\tilde{p} := \dim U \geq 1$. Moreover, (d) holds iff $\dim U = p$. Take any isometric isomorphism $\psi : U \to \mathbb{R}^{\tilde{p}}$ and define the feature map $\tilde{\phi} : \mathbb{R}^d \to \mathbb{R}^{\tilde{p}}, \boldsymbol{x} \mapsto \psi(\phi(\boldsymbol{x}))$. Then, $\tilde{\phi}$ is analytic since ψ is linear, and $\tilde{\phi}$ satisfies (d), hence Theorem 3 can be applied to $\tilde{\phi}$. However, ϕ and $\tilde{\phi}$ lead to the same kernel k since ψ is isometric, hence to the same estimator $f_{\boldsymbol{X},\boldsymbol{y}}$ by Eq. (2) and hence to the same $\mathcal{E}_{\mathrm{Noise}}$. ◀

Proposition 8 (Polynomial kernel)**.** *Let $m, d \geq 1$ and $c > 0$. For $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{R}^d$, define the polynomial kernel $k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) := (\boldsymbol{x}^\top \tilde{\boldsymbol{x}} + c)^m$. Then, there exists a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^p$, $p := \binom{m+d}{m}$, such that*

(a) $k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \phi(\boldsymbol{x})^\top \phi(\tilde{\boldsymbol{x}})$ for all $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{R}^d$, and
(b) if $\boldsymbol{x} \in \mathbb{R}^d$ has a Lebesgue density and we use $\boldsymbol{z} = \phi(\boldsymbol{x})$, then (FRK) is satisfied for all n.

Proposition 8 says that the lower bound from Theorem 3 holds for ridgeless kernel regression with the polynomial kernel with $p := \binom{m+d}{m}$ if $\boldsymbol{x}$ has a Lebesgue density and $\mathbb{E}\|\boldsymbol{z}\|_2^2 = \mathbb{E}k(\boldsymbol{x}, \boldsymbol{x}) = \mathbb{E}(\|\boldsymbol{x}\|_2^2 + c)^m < \infty$.

We can also extend our theory to analytic random feature maps:

Proposition 9 (Random feature maps)**.** *Consider feature maps $\phi_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^p$ with (random) parameter $\boldsymbol{\theta} \in \mathbb{R}^q$. Suppose the map $(\boldsymbol{\theta}, \boldsymbol{x}) \mapsto \phi_{\boldsymbol{\theta}}(\boldsymbol{x})$ is analytic and that $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are independent and have Lebesgue densities. If there exist fixed $\boldsymbol{\theta}_{\mathrm{const}} \in \mathbb{R}^q, \boldsymbol{X}_{\mathrm{const}} \in \mathbb{R}^{p \times d}$ with $\det(\phi_{\boldsymbol{\theta}_{\mathrm{const}}}(\boldsymbol{X}_{\mathrm{const}})) \neq 0$, then almost surely over $\boldsymbol{\theta}$, (FRK) holds for all n for $\boldsymbol{z} = \phi_{\boldsymbol{\theta}}(\boldsymbol{x})$.*

In Appendix C, we demonstrate that Proposition 9 can be used to computationally verify (FRK) for analytic random feature maps. Moreover, we can now prove (FRK) for certain random neural networks. Our analysis thus applies to a large class of feedforward neural networks where only the last layer is trained (and initialized to zero, such that GD converges to the Moore-Penrose-Pseudoinverse).

Theorem 10 (Random neural networks)**.** *Let $d, p, L \geq 1$ and let $\sigma : \mathbb{R} \to \mathbb{R}$ be analytic. Consider layer sizes $d_0 = d, d_1, \ldots, d_{L-1} \geq p$ and $d_L = p$. For weights $\boldsymbol{\theta} := (\boldsymbol{W}^{(0)}, \ldots, \boldsymbol{W}^{(L-1)})$ with $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d_{l+1} \times d_l}$, we define the feature map*

$$\phi_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^p, \boldsymbol{x} \mapsto \sigma(\boldsymbol{W}^{(L-1)} \sigma(\ldots \sigma(\boldsymbol{W}^{(0)} \boldsymbol{x}))) \ ,$$

where σ is applied element-wise.

If $\boldsymbol{\theta}$ and $\boldsymbol{x}$ have Lebesgue densities and σ is not a polynomial with less than p nonzero coefficients, (FRK) holds for all n and almost surely over $\boldsymbol{\theta}$. This statement also holds if $\phi_{\boldsymbol{\theta}}$ includes biases that are initialized from a probability distribution with a density.

The assumptions of Theorem 10 on $\boldsymbol{\theta}$ are satisfied by the classical initialization methods of He et al. (2015) and Glorot & Bengio (2010). Possible choices of σ are presented in Table 1. In Appendix E, we explain that while there exist theorems in the literature similar to Theorem 10 for single-layer networks, the proofs of these theorems or even the theorems themselves are incorrect.

Table 1: Some examples of analytic activation functions that are not polynomials. The CDF of the standard normal distribution is denoted by Φ. Other possible activation functions are $\sin$, $\cos$ or erf.

Activation function	Equation
Sigmoid	$\operatorname{sigmoid}(x) = 1/(1+e^{-x})$
Hyperbolic Tangent	$\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$
Softplus	$\operatorname{softplus}(x) = \log(1+e^x)$
RBF	$\operatorname{RBF}(x) = \exp(-\beta x^2)$
GELU (Hendrycks & Gimpel, 2016)	$\operatorname{GELU}(x) = x\Phi(x)$
SiLU (Elfwing et al., 2018)	$\operatorname{SiLU}(x) = x \operatorname{sigmoid}(x)$
Swish (Ramachandran et al., 2017)	$\operatorname{Swish}(x) = x \operatorname{sigmoid}(\beta x)$
Mish (Misra, 2019)	$\operatorname{Mish}(x) = x \tanh(\operatorname{softplus}(x))$

Under the assumptions of Theorem 10, if $\|\boldsymbol{x}\|_2$ is bounded, (MOM) holds for all $\boldsymbol{\theta}$. This follows since any analytic function ϕ is also continuous, and continuous functions preserve boundedness. However, the activation functions from Table 1 even satisfy $|\sigma(x)| \leq a|x| + b$ for some $a, b \geq 0$ and all $x \in \mathbb{R}$. In this case, it is not hard to see that (MOM) already holds for all $\boldsymbol{\theta}$ if $\mathbb{E}\|\boldsymbol{x}\|_2^2 < \infty$.

Theorem 10 does not hold for ReLU, ELU (Clevert et al., 2015), SELU (Klambauer et al., 2017) or other activation functions with a perfectly linear part. To see this, observe that with nonzero probability, all weights and inputs are positive. In this case, the feature map acts as a linear map from $\mathbb{R}^d$ to $\mathbb{R}^p$, and if $d < p = n$, the output matrix $\phi(\boldsymbol{X})$ cannot be invertible.

In Appendix J, we show that (FRK) is satisfied for random Fourier features if $\boldsymbol{x}$ has a Lebesgue density and the frequency distribution (i.e. the Fourier transform of the shift-invariant kernel) has a Lebesgue density.

6 Quality of the Lower Bound

In this section, we want to discuss to which extent the lower bound from Theorem 3 is suboptimal. To this end, we assume that (NOI) holds with equality, i.e. $\operatorname{Var}(y|\boldsymbol{z}) = \sigma^2$ almost surely over $\boldsymbol{z}$.

In their Lemma 3, Hastie et al. (2019) consider the case where $\boldsymbol{z}$ has i.i.d. entries with zero mean, unit variance and finite fourth moment. They then use the Marchenko-Pastur law to show in the limit $p, n \to \infty, p/n \to \gamma > 1$ that $\mathcal{E}_{\text{Noise}} \to \sigma^2 \frac{1}{\gamma - 1}$. In this overparameterized limit, our lower bound shows

$$\mathcal{E}_{\text{Noise}} \geq \sigma^2 \frac{n}{p+1-n} = \sigma^2 \frac{1}{p/n + 1/n - 1} \to \sigma^2 \frac{1}{\gamma - 1} \ ,$$

hence our overparameterized bound is asymptotically sharp. Our underparameterized bound is weaker, especially around $p = n$. Still, both bounds show that linear regression is sensitive to label noise for $p \approx n$ under the weak assumptions of Theorem 3. In order to better understand to which extent our lower bound is suboptimal, we explicitly compute $\mathcal{E}_{\text{Noise}}$ for some distributions:

Theorem 11. *Let $P_Z = \mathcal{U}(\mathbb{S}^{p-1})$. Then, P_Z satisfies the assumptions (MOM), (COV) and (FRK) for all n with $\boldsymbol{\Sigma} = \frac{1}{p}\boldsymbol{I}_p$. Moreover, for $n \geq p = 1$ or $p \geq n \geq 1$, we can compute*

$$\mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) = \begin{cases} \frac{1}{n} & \text{if } n \geq p = 1, \\ \frac{1}{p} & \text{if } p \geq n = 1, \\ \infty & \text{if } 2 \leq n \leq p \leq n+1, \\ \frac{n}{p-1-n} \cdot \frac{p-2}{p} & \text{if } 2 \leq n \leq n+2 \leq p. \end{cases}$$

The formulas in the next theorem have already been computed by Belkin et al. (2019b), but with a technical problem for $p \in \{n-1, n, n+1\}$ that we circumvent, see Appendix H.

Theorem 12. *Let $P_Z = \mathcal{N}(0, \boldsymbol{I}_p)$. Then, P_Z satisfies the assumptions (MOM), (COV) and (FRK) for all n with $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Moreover, for $n, p \geq 1$,*

$$\mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) = \begin{cases} \frac{n}{p-1-n} & \text{if } p \geq n+2, \\ \infty & \text{if } p \in \{n-1, n, n+1\}, \\ \frac{p}{n-1-p} & \text{if } p \leq n-2. \end{cases}$$

For $P_Z = \mathcal{N}(0, \boldsymbol{I}_p)$, Theorem 12 shows that the under- and overparameterized cases are symmetric. However, for $P_Z = \mathcal{U}(\mathbb{S}^{p-1})$, the hypothetical formula $\frac{p}{n-1-p} \cdot \frac{n-2}{n}$ does not agree with numerical estimates of $\mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+)$ for $p = 2, n \geq 4$.

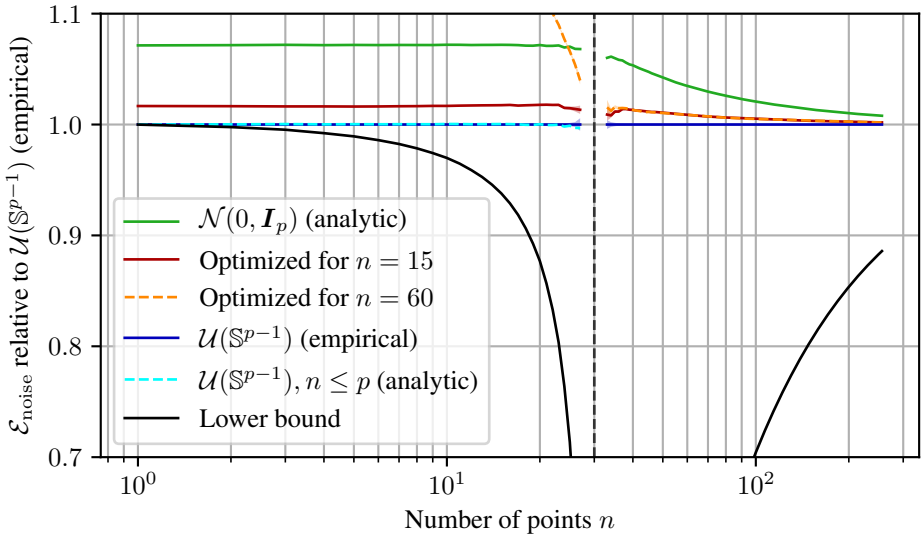

Figure 1: Various estimates and bounds for $\mathcal{E}_{\text{Noise}}$ relative to $\mathcal{E}_{\text{Noise}}$ for $P_Z = \mathcal{U}(\mathbb{S}^{p-1})$, using $\operatorname{Var}(y|\boldsymbol{z}) = 1$ and $p = 30$. The optimized curves correspond to multi-layer NN feature maps whose parameters were trained to minimize $\mathcal{E}_{\text{Noise}}$ for $n = 15$ or $n = 60$. More experiments and details on the setup can be found in Appendix C. We do not plot estimates for $n \in \{28, \ldots, 32\}$ since they have high estimation variances.

For $p \geq n+2$, we can relate our lower bound (Theorem 3), the result for the sphere (Theorem 11) and the result for the Gaussian distribution (Theorem 12) as follows:

$$\frac{n}{p+1-n} = \frac{n}{p-1-n} \cdot \frac{(p+1-n)-2}{p+1-n} \leq \frac{n}{p-1-n} \cdot \frac{p-2}{p} < \frac{n}{p-1-n} \,.$$

These values are also shown in Figure 1 together with empirical values of NN feature maps specifically optimized to minimize $\mathcal{E}_{\text{Noise}}$. Since $\boldsymbol{\Sigma}$ affects $\mathcal{E}_{\text{Noise}}$ only for $p > n$ (cf. Remark G.1), it is not surprising that the feature map optimized for $n = 60 > 30 = p$ performs badly in the overparameterized regime. The results in Figure 1 support the hypothesis that among all P_Z satisfying (MOM), (COV) and (FRK), $P_Z = \mathcal{U}(\mathbb{S}^{p-1})$ minimizes $\mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+)$. The plausibility of this hypothesis is further discussed in Remark G.3. We can prove the hypothesis for $n = 1$ or $p = 1$ since in this case, the results from Theorem 11 are equal to the lower bound from Theorem 3.

When using a continuous feature map $\phi : \mathbb{R}^d \to \mathbb{R}^p$ with $d \leq p-2$, it seems to be unclear at first whether the low $\mathcal{E}_{\text{Noise}}$ of $P_Z = \mathcal{U}(\mathbb{S}^{p-1})$ can even be achieved. We show in Proposition G.4 that this is possible using space-filling curve feature maps.

The results in this section and in Appendix C show that while our lower bound presumably does not capture the full height of the interpolation peak at $p \approx n$, it is quite sharp in the practically relevant regimes $p \gg n$ and $p \ll n$.

REFERENCES

Madhu S. Advani and Andrew M. Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv:1710.03667*, 2017.

Herbert Amann and Joachim Escher. *Analysis*. Springer, 2005.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arXiv:1802.01396*, 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv:1903.07571*, 2019b.

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013.

Vladimir I. Bogachev. *Measure Theory, Volume 2*, volume 1. Springer Science & Business Media, 2007.

Charles Bordenave and Djalil Chafaï. Around the circular law. *Probability surveys*, 9:1–89, 2012.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Siegfried Bös and Manfred Opper. Dynamics of training. In *Advances in Neural Information Processing Systems*, pp. 141–147, 1997.

Weipeng Cao, Xizhao Wang, Zhong Ming, and Jinzhu Gao. A review on neural networks with random weights. *Neurocomputing*, 275:278–287, 2018.

Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. *arXiv:2008.01036*, 2020.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv:1511.07289*, 2015.

Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. *arXiv:2003.01054*, 2020.

Rick Durrett. *Probability: theory and examples, 5th edition*, volume 49. Cambridge university press, 2019.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Catherine Forbes, Merran Evans, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. John Wiley & Sons, 2011.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

Gene H. Golub and Charles F. Van Loan. Matrix Computations. *The Johns Hopkins University Press, Baltimore, USA*, 1989.

Ping Guo. A vest of the pseudoinverse learning algorithm. *arXiv:1805.07828*, 2018.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media, 2002.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv:1903.08560*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv:1606.08415*, 2016.

Pablo A. Henriquez and Gonzalo A. Ruz. An empirical study of the hidden matrix rank for neural networks with random weights. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 883–888. IEEE, 2017.

Gao Huang, Guang-Bin Huang, Shiji Song, and Keyou You. Trends in extreme learning machines: A review. *Neural Networks*, 61:32–48, 2015.

Guang-Bin Huang. Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Transactions on Neural Networks*, 14(2):274–281, 2003.

Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in neural information processing systems*, pp. 971–980, 2017.

Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Springer Science & Business Media, 2002.

K. V. Mardia, J. T. Kent, and J. M. Bibby. Multivariate analysis. *Probability and mathematical statistics. Academic Press Inc.*, 1979.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv:1908.05355*, 2019.

Diganta Misra. Mish: A self regularized non-monotonic neural activation function. *arXiv:1908.08681*, 2019.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Preetum Nakkiran. More data can hurt for linear regression: Sample-wise double descent. *arXiv:1912.07242*, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv:1912.02292*, 2019.

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv:1810.08591*, 2018.

S. James Press. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. Courier Corporation, 2005.

Robert James Purser, Manuel de Pondeca, and Sei-Young Park. Construction of a Hilbert curve on the sphere with an isometric parameterization of area. *Office Note*, 2009.

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

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *arXiv:1710.05941*, 2017.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the Convergence of Adam and Beyond. In *International Conference on Learning Representations*, 2018.

Hans Sagan. *Space-filling curves*. Springer Science & Business Media, 2012.

Michael A. Sartori and Panos J. Antsaklis. A simple method to derive bounds on the size and to train multilayer neural networks. *IEEE transactions on neural networks*, 2(4):467–471, 1991.

Simone Scardapane and Dianhui Wang. Randomness in neural networks: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(2), 2017.

Wouter F. Schmidt, Martin A. Kraaijveld, and Robert PW Duin. Feed forward neural networks with random weights. In *International Conference on Pattern Recognition*, pp. 1–1. IEEE COMPUTER SOCIETY PRESS, 1992.

Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001, 2019.

Shin'ichi Tamura and Masahiko Tateishi. Capabilities of a four-layered feedforward neural network: four layers versus three. *IEEE Transactions on Neural Networks*, 8(2):251–255, 1997.

Dietrich von Rosen. Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics*, pp. 97–109, 1988.

Guorong Wang, Yimin Wei, Sanzheng Qiao, Peng Lin, and Yuzhuo Chen. *Generalized Inverses: Theory and Computations*, volume 53. Springer, 2018.

Bernard Widrow, Aaron Greenblatt, Youngsik Kim, and Dookun Park. The No-Prop algorithm: A new learning algorithm for multilayer neural networks. *Neural Networks*, 37:182–188, 2013.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530*, 2016.

A OVERVIEW

The appendix is structured as follows: In Appendix B, we provide an overview over some matrix identities used throughout the paper. We provide additional numerical experiments for various (random) feature maps in Appendix C. The counterexample given in Appendix D shows that the assumption (FRK) cannot be omitted from Theorem 3. In Appendix E, we provide an overview over full-rank theorems for random neural networks in the literature and explain why their proofs are incorrect. We then prove our main results in Appendix F before discussing consequences in Appendix G. Our exact results for $P_Z = \mathcal{U}(\mathbb{S}^{p-1})$ and $P_Z = \mathcal{N}(0, \boldsymbol{I}_p)$ are proven in Appendix H. In Appendix I, we give proofs for the theorems and propositions from Section 5. Finally, in Appendix J, we prove (FRK) for random Fourier features.

Whenever we prove theorems or propositions from the main paper (like Theorem 3) in the Appendix, we repeat their statement before the proof for convenience. In contrast, new theorems or propositions are numbered according to the section they are stated in, e.g. Proposition J.1.

B MATRIX ALGEBRA

In the following, we will present some facts about matrices that are relevant to this paper. For a general reference, we refer to textbooks on the subject (e.g. Golub & Van Loan, 1989; Bhatia, 2013).

Let $n, p \geq 1$ and let $k \coloneqq \min\{n, p\}$. The singular value decomposition (SVD) of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ is a decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ into orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{n\times k}, \boldsymbol{V} \in \mathbb{R}^{p\times k}$ with $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_k$ and a diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{k\times k}$ with non-negative diagonal elements $s_1(\boldsymbol{A}) \geq \ldots \geq s_k(\boldsymbol{A})$ called *singular values*.

For a given symmetric square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1(\boldsymbol{A}) \geq \ldots \geq \lambda_n(\boldsymbol{A})$, the trace satisfies

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} A_{ii} = \sum_{i=1}^{n} \lambda_i \ .$$

The trace is linear and invariant under cyclical permutations. We use this multiple times in arguments of the following type: If $\boldsymbol{v} \in \mathbb{R}^p$ and $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ are stochastically independent (e.g. because $\boldsymbol{A}$ is constant), we can write

$$\mathbb{E}_{\boldsymbol{v}} \boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} = \mathbb{E}_{\boldsymbol{v}} \operatorname{tr}(\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}) = \mathbb{E}_{\boldsymbol{v}} \operatorname{tr}(\boldsymbol{A}\boldsymbol{v}\boldsymbol{v}^\top) = \operatorname{tr}(\boldsymbol{A}\mathbb{E}_{\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^\top) \ .$$

Moreover, if $\boldsymbol{A} \succ 0$, then for all $i \in [n]$, $\lambda_i(\boldsymbol{A}) > 0$ and we have

$$\lambda_{n+1-i}(\boldsymbol{A}^{-1}) = \frac{1}{\lambda_i(\boldsymbol{A})} \ .$$

For a positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, the SVD and the eigendecomposition coincide as $\boldsymbol{\Sigma} = \boldsymbol{U} \operatorname{diag}(\lambda_1(\boldsymbol{\Sigma}), \ldots, \lambda_p(\boldsymbol{\Sigma}))\boldsymbol{U}^\top$ and we can define the inverse square root as

$$\boldsymbol{\Sigma}^{-1/2} \coloneqq \boldsymbol{U} \operatorname{diag}(\lambda_1(\boldsymbol{\Sigma})^{-1/2}, \ldots, \lambda_p(\boldsymbol{\Sigma})^{-1/2})\boldsymbol{U}^\top \succ 0 \ ,$$

which is the unique s.p.d. matrix satisfying $(\boldsymbol{\Sigma}^{-1/2})^2 = \boldsymbol{\Sigma}^{-1}$.

By the Courant-Fischer-Weyl theorem, two symmetric matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ with $\boldsymbol{A} \preceq \boldsymbol{B}$ satisfy

$$\lambda_i(\boldsymbol{A}) = \max_{\substack{\mathcal{V}\subseteq\mathbb{R}^d \text{ subspace} \\ \dim \mathcal{V} = i}} \min_{\substack{\boldsymbol{v}\in\mathcal{V} \\ \|\boldsymbol{v}\|_2=1}} \boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} \leq \max_{\substack{\mathcal{V}\subseteq\mathbb{R}^d \text{ subspace} \\ \dim \mathcal{V} = i}} \min_{\substack{\boldsymbol{v}\in\mathcal{V} \\ \|\boldsymbol{v}\|_2=1}} \boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{v} = \lambda_i(\boldsymbol{B}) \ .$$

Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$. The Moore-Penrose pseudoinverse $\boldsymbol{A}^+$ of $\boldsymbol{A}$ satisfies the following relations (see e.g. Section 1.1.1 in Wang et al., 2018):

- Suppose $\boldsymbol{A}$ has the SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, where $\boldsymbol{D} = \operatorname{diag}(s_1, \ldots, s_k)$, $k \coloneqq \min\{n, p\}$. Using the convention $1/0 \coloneqq 0$, we can write $\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{D}^+\boldsymbol{U}^\top$, where $\boldsymbol{D}^+ \coloneqq \operatorname{diag}(1/s_1, \ldots, 1/s_k)$.

- $\boldsymbol{A}^+ = (\boldsymbol{A}^\top \boldsymbol{A})^+ \boldsymbol{A}^\top = \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{A}^\top)^+$.
- If $\boldsymbol{A}$ is invertible, then $\boldsymbol{A}^+ = \boldsymbol{A}^{-1}$.
- $\boldsymbol{A}^+ (\boldsymbol{A}^+)^\top = (\boldsymbol{A}^\top \boldsymbol{A})^+$.

We will also use a basic fact on the Schur complement that, for example, is outlined in Appendix A.5.5 in Boyd & Vandenberghe (2004): If

$$0 \prec \boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} \in \mathbb{R}^{(m_1+m_2)\times(m_1+m_2)} ,$$

then $\boldsymbol{A}_{22} \succ 0$ and $\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} \succ 0$ and the top-left $m_1 \times m_1$ block of $\boldsymbol{A}^{-1}$ is given by $(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1}$.

C Experiments

In the following, we experimentally investigate $\mathcal{E}_{\text{Noise}}$ for different (random) feature maps. We will first give an overview over the plots and then explain the details of how they were generated and some implications. More details can be found in the provided code. All empirical curves show one estimated standard deviation of the mean estimator as a shaded area around the estimated mean, but this standard deviation is sometimes too small to be visible. We assume $\mathrm{Var}(y|\boldsymbol{z}) = 1$ almost surely over $\boldsymbol{z}$ such that $(*)$ in Theorem 3 holds with equality with $\sigma^2 = 1$.

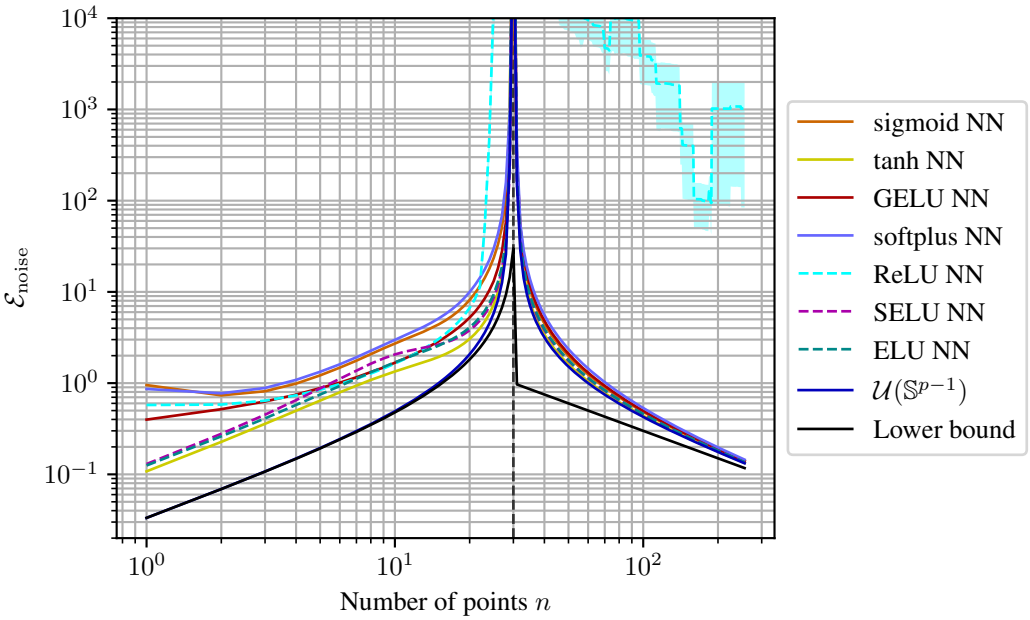

Figure C.1: Estimated $\mathcal{E}_{\text{Noise}}$ for random neural network feature maps (cf. Theorem 10) with different activation functions and $d_0 = d = 10, d_1 = d_2 = 256, d_3 = p = 30$. We include the results from $P_Z = \mathcal{U}(\mathbb{S}^{d-1})$ as comparison, cf. Section 6.

Figure C.1 shows $\mathcal{E}_{\text{Noise}}$ for random three-layer neural network feature maps with $p = 30$, different activation functions and varying n. Note that all neural networks produce higher $\mathcal{E}_{\text{Noise}}$ than $P_Z = \mathcal{U}(\mathbb{S}^{p-1})$. The effect of non-isotropic covariance matrices in the overparameterized regime can be clearly seen when comparing Figure C.1 to Figure C.2, where features have been whitened separately for each set of random parameters $\boldsymbol{\theta}$, cf. Remark G.1. Figure C.3 then shows $\mathcal{E}_{\text{Noise}}$ for $n = 30$ and varying p.

Note that double descent is usually plotted as a function of the "model complexity" p as in Belkin et al. (2019a), but varying p is only possible when we have a (random) feature map $\phi_{\boldsymbol{\theta}}^{(p)} : \mathbb{R}^d \to \mathbb{R}^p$ for each value of p. For the following NTK and polynomial kernels, there is no canonical way to

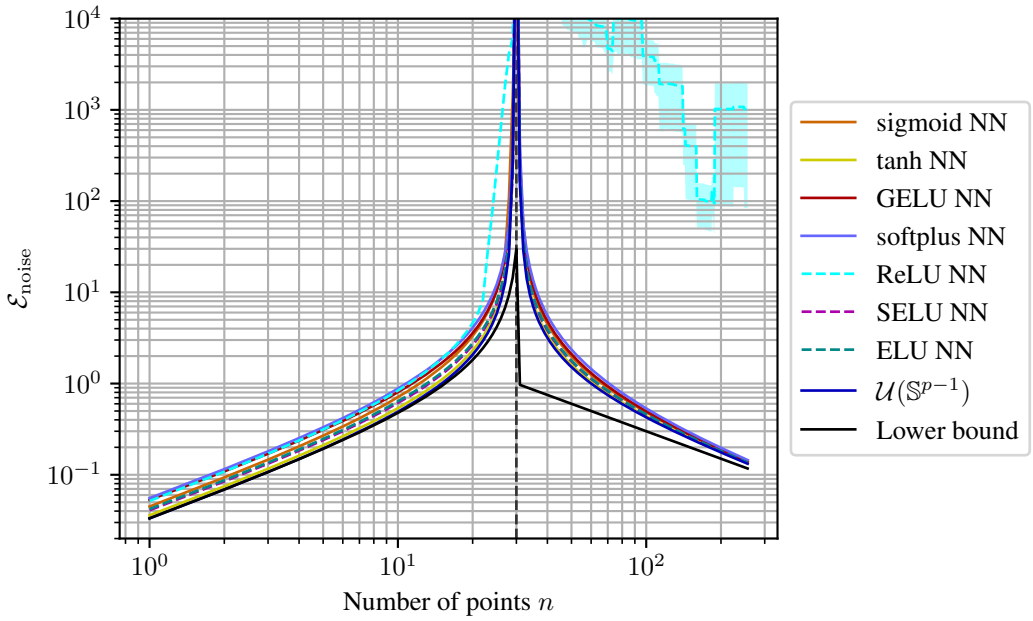

Figure C.2: As in Figure C.1 but with whitened features, i.e. using $\mathbb{E}\,\mathrm{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1})$ in the overparameterized case $n \leq p$, cf. Theorem 3 and Remark G.1.

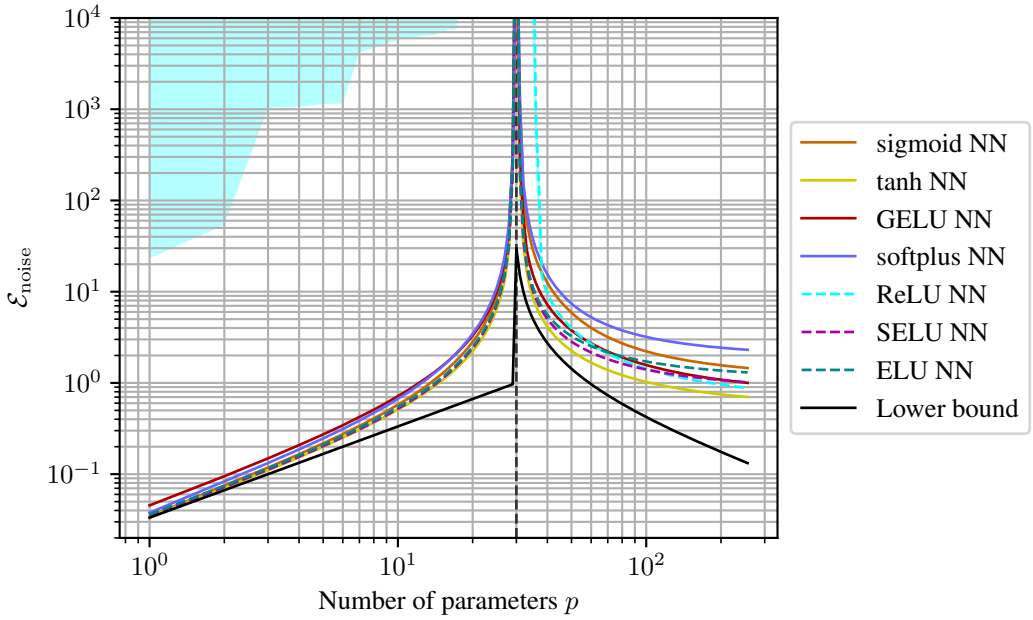

Figure C.3: As in Figure C.1 but with $n = 30$ fixed and varying $d_3 = p$.

define such a sequence of feature maps. For this reason, we will plot their results only with varying n. Double descent as a function of the number of samples n has for example been pointed out by Nakkiran et al. (2019) and Nakkiran (2019).

Figure C.4 shows $\mathcal{E}_{\text{Noise}}$ for various random finite-width Neural Tangent Kernels (NTKs), cf. Jacot et al. (2018). These results mostly exhibit larger $\mathcal{E}_{\text{Noise}}$ than the random NN feature maps from Figure C.1, perhaps because of correlations parameter gradients in different layers. However, this

comparison is not really fair since the NTK feature map uses a much smaller underlying neural network and the input dimension d is smaller.

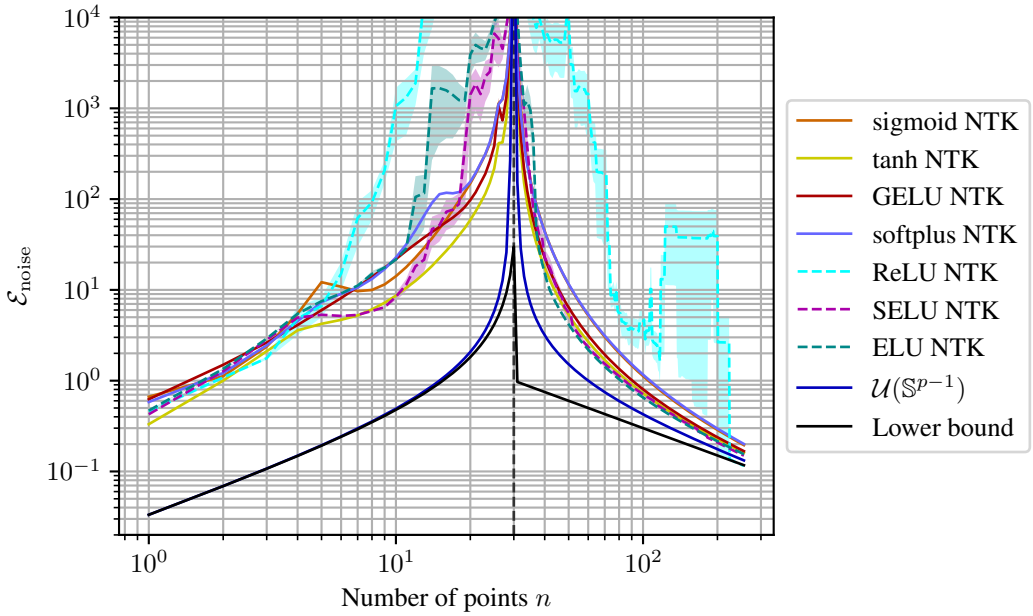

Figure C.4: Estimated $\mathcal{E}_{\text{Noise}}$ for Neural Tangent Kernel (NTK) feature maps given by various random neural networks (cf. Theorem 10) with $d_0 = d = 4, d_1 = 6, d_3 = 1$, resulting in $p = 4 \cdot 6 + 6 \cdot 1 = 30$. We include the empirical results from $P_Z = \mathcal{U}(\mathbb{S}^{d-1})$ as comparison, cf. Section 6.

Finally, Figure C.5 shows that linear regression with the polynomial kernel is quite sensitive to label noise. We use $p = 35$ for the polynomial kernel since there are no particularly interesting polynomial kernels with $p = 30$.

Neural Network feature maps For Figures C.1, C.2 and C.3, we use random neural network feature maps without biases as in Theorem 10 with $d_0 = d = 10, d_1 = d_2 = 256$ and $d_3 = p$. As the input distribution P_X, we use $\mathcal{N}(0, \boldsymbol{I}_d)$. We initialize the NN weights independently as $W_{ij}^{(l)} \sim \mathcal{N}(0, 1/V_l)$, where

$$V_l := \begin{cases} d_0 & \text{if } l = 0, \\ d_l \operatorname{Var}_{u \sim \mathcal{N}(0,1)}(\sigma(u)) & \text{if } l \geq 1. \end{cases}$$

Here, $\operatorname{Var}_{u \sim \mathcal{N}(0,1)}(\sigma(u))$ is approximated once by using with 10^4 samples for u. This initialization is similar (and for the ReLU activation essentially equivalent) to the initialization method by He et al. (2015). The initialization variances are chosen such that for fixed input $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_2 \approx 1$, the pre-activations in each layer are approximately $\mathcal{N}(0, 1)$-distributed.

NTK feature maps For Figure C.4, we use a NTK parameterization similar to the original one proposed by Jacot et al. (2018), but again with activation-dependent scaling. Our neural network is given by[2]

$$\tilde{\phi}_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^1, \boldsymbol{x} \mapsto \frac{1}{\sqrt{V_1}} \boldsymbol{W}^{(1)} \sigma\left(\frac{1}{\sqrt{V_0}} \boldsymbol{W}^{(0)}\right)$$

[2]For random NN feature maps as in Theorem 10, one can interpret the linear regression as being an extra layer on top of the neural network, and therefore the last layer of the feature map should contain an activation function. For NTK feature maps, one can instead interpret the linear regression as performing a "linearized" training of the whole NN, and the whole NN usually does not contain an activation function in the last layer.

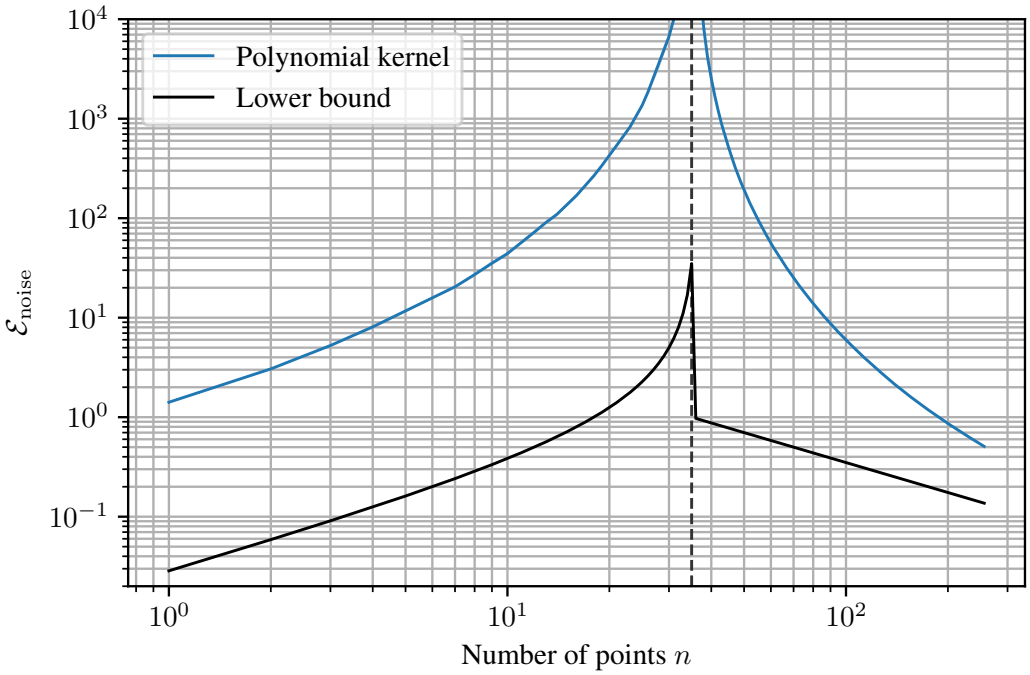

Figure C.5: Estimated $\mathcal{E}_{\text{Noise}}$ for the polynomial kernel (cf. Proposition 8) with $d = 3$, $m = 4$, $c = 1$, $P_X = \mathcal{N}(0, \boldsymbol{I}_d)$, resulting in $p = \binom{m+d}{m} = \binom{7}{4} = 35$.

with $d_0 = d = 4, d_1 = 6, d_2 = 1$ and $\boldsymbol{W}_{ij}^{(l)} \sim \mathcal{N}(0,1)$ i.i.d. Our input distribution is again $P_X = \mathcal{N}(0, \boldsymbol{I}_d)$. The NTK feature map is then defined as

$$\phi_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^p, \boldsymbol{x} \mapsto \frac{\partial}{\partial \boldsymbol{\theta}} \tilde{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}) \,,$$

where $p = 6 \cdot 4 + 1 \cdot 6 = 30$ is the number of parameters in $\boldsymbol{\theta}$. Note that moving the variances V_l outside of the $\boldsymbol{W}^{(l)}$ does not affect the forward pass but only the backward pass (i.e. the derivatives).

While we have not theoretically established the properties (FRK) and (COV) for such feature maps, we can do this experimentally for analytic activation functions like sigmoid, tanh, softplus and GELU: Since the random NTK feature map is a derivative of the analytic random NN feature map, it is also analytic. By Proposition 9, if (FRK) does not hold, then for every fixed $\boldsymbol{\theta}_{\text{const}}$, the range of $\phi_{\boldsymbol{\theta}_{\text{const}}}$ must be contained in a proper linear subspace of $\mathbb{R}^p$, and therefore $\boldsymbol{Z} = \phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ never has full rank for $n \geq p$. In this case, the singular value $s_p(\boldsymbol{Z})$ would be zero, and even accounting for numerical errors, the "inverse condition number" $\frac{s_p(\boldsymbol{Z})}{s_1(\boldsymbol{Z})}$ should at most be of the order of 64-bit float machine precision, i.e. around 10^{-16}. However, among 10^4 samples of $\boldsymbol{Z}$ for $n := 90 \geq 30 = p$, the maximum observed "inverse condition number" was greater than 10^{-3} for all of the activation functions sigmoid, tanh, softplus and GELU.[3] Hence, by Proposition 6 and Proposition 9, we can confidently conclude that (COV) and (FRK) hold for all n almost surely over $\boldsymbol{\theta}$ for this network size and these activation functions.

Estimation of $\mathcal{E}_{\text{Noise}}$ In order to estimate $\mathcal{E}_{\text{Noise}}$, we proceed as follows: Recall from Section 3 that ridgeless regression is the limit of ridge regression for $\lambda \searrow 0$. We use a small regularization of $\lambda = 10^{-12}$ in order to improve numerical stability. Also for numerical stability, we use a singular value decomposition (SVD) $\boldsymbol{Z} = \boldsymbol{U} \operatorname{diag}(s_1, \ldots, s_k) \boldsymbol{V}^\top$ with $k := \min\{n, p\}$ as in Appendix B. The regularized approximation of $\boldsymbol{Z}^+$ is then

$$\boldsymbol{Z}^+ \approx (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{Z}^\top = \boldsymbol{V} \operatorname{diag}\left(\frac{s_1}{s_1^2 + \lambda}, \ldots, \frac{s_k}{s_k^2 + \lambda}\right) \boldsymbol{U}^\top \,. \tag{3}$$

[3] We use $n > p$ since this usually improves the "inverse condition number" of $\boldsymbol{Z}$.

We can then estimate

$$\operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) = \operatorname{tr}(\boldsymbol{Z}^+ (\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma}) \approx \operatorname{tr}\left(\boldsymbol{V} \operatorname{diag}\left(\frac{s_1^2}{(s_1^2+\lambda)^2}, \dots, \frac{s_k^2}{(s_k^2+\lambda)^2}\right) \boldsymbol{V}^\top \boldsymbol{\Sigma}\right). \quad (4)$$

We can then obtain $m = 10^4$ (non-ReLU NNs and Polynomial Kernel) or $m = 10^5$ (all other empirical results) sampled estimates for $\operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+)$ in order to obtain a Monte-Carlo estimate of $\mathbb{E} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+)$ by repeating the following procedure m times:

(1) Sample a random parameter $\boldsymbol{\theta}$.
(2) Sample random matrices $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $\tilde{\boldsymbol{X}} \in \mathbb{R}^{l \times d}$, $l = 10^4$, with i.i.d. P_X-distributed rows.
(3) Compute $\boldsymbol{Z} \coloneqq \phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ and $\tilde{\boldsymbol{Z}} \coloneqq \phi_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}})$.
(4) Compute the estimate $\boldsymbol{\Sigma} \coloneqq \frac{1}{l} \tilde{\boldsymbol{Z}}^\top \tilde{\boldsymbol{Z}}$.
(5) Compute a regularized estimate of $\operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+)$ using the SVD of $\boldsymbol{Z}$ and Eq. (4).

For performance reasons, we make the following modification of step (2) and (5): Since we perform the computation for all $n \in [256]$, we sample $\boldsymbol{X} \in \mathbb{R}^{256 \times d}$ and then, for all $n \in [256]$, perform the computation for n using the first n rows of $\boldsymbol{Z}$. Hence, the estimates for different n are not independent. In Figure C.3, we perform an analogous optimization for p by taking the first $p \in [256]$ of the $d_3 = 256$ output features in $\boldsymbol{Z}$.

Curious ReLU results The curves for the ReLU NNs and ReLU NTKs in the underparameterized regime $p \leq n$ may seem odd. The locally almost constant "plateaus" are presumably an artefact of the non-independent estimates for different n as explained in the last paragraph. As discussed in Section 5, networks with ReLU, ELU or SELU activations do not satisfy (FRK). It seems plausible that, since both "halves" of the ReLU function are linear, ReLU networks have a significantly higher chance than SELU or ELU networks to be initialized with "bad" parameters that are likely to generate "degenerate" feature matrices $\boldsymbol{Z}$ that do not have full rank at $n = p$ and only become full rank for some $n > p$. When inspecting the data underlying the plots, the estimate of $\mathcal{E}_{\text{Noise}}$ for ReLU networks in the underparameterized regime seems to be dominated by few outliers. It seems that the distribution of $\operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+)$ for ReLU networks is often so heavy-tailed that computing more Monte Carlo samples does not significantly reduce the estimation uncertainty.

Whitening For computing $\mathbb{E}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) = \mathbb{E}((\boldsymbol{Z}\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}^\top)^{-1})$ in the overparameterized case in Figure C.2, we regularize both matrix inversions on the right-hand side as above: For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{m \times m}$, we use a symmetric eigendecomposition $\boldsymbol{A} = \boldsymbol{U} \operatorname{diag}(s_1, \dots, s_m) \boldsymbol{U}^\top$ and approximate

$$\boldsymbol{A}^{-1} \approx \boldsymbol{U} \operatorname{diag}\left(\frac{s_1}{s_1^2+\lambda}, \dots, \frac{s_m}{s_m^2+\lambda}\right) \boldsymbol{U}^\top .$$

Optimization For our optimized feature maps in Figure 1 with $p = 30$, we use a neural network feature map with $d_0 = d = p = 30, d_1 = d_2 = 256, d_3 = p = 30$ and $\tanh$ activation function. We use NTK parameterization and zero-initialized biases, leading to

$$\phi_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sigma(\boldsymbol{b}^{(2)} + V_2^{-1/2} \boldsymbol{W}^{(2)} \sigma(\boldsymbol{b}^{(1)} + V_1^{-1/2} \boldsymbol{W}^{(1)} \sigma(\boldsymbol{b}^{(0)} + V_0^{-1/2} \boldsymbol{W}^{(0)} \boldsymbol{x})))$$

with independent initialization $W_{ij}^{(l)} \sim \mathcal{N}(0,1)$, $b_i^{(l)} = 0$. As the input distribution, we use $P_X = \mathcal{N}(0, \boldsymbol{I}_d)$. For given $\boldsymbol{\theta}$, let $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} \coloneqq \mathbb{E}_{\boldsymbol{x}} \phi_{\boldsymbol{\theta}}(\boldsymbol{x}) \phi_{\boldsymbol{\theta}}(\boldsymbol{x})^\top$, i.e. we define the second moment matrix $\boldsymbol{\Sigma}$ as depending on $\boldsymbol{\theta}$. We then optimize the loss function

$$L(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\boldsymbol{X}} \operatorname{tr}((\phi_{\boldsymbol{\theta}}(\boldsymbol{X})^+)^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \phi_{\boldsymbol{\theta}}(\boldsymbol{X})^+)$$

using AMSGrad (Reddi et al., 2018) with a learning rate that linearly decays from 10^{-3} to 0 over 1000 iterations. In order to approximate $L(\boldsymbol{\theta})$ in each iteration, we approximate $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ using 1000 Monte Carlo points and we draw 1024 different realizations of $\boldsymbol{X}$ (this can be considered as using batch size 1024). We also use a regularized version as in Eq. (3), but we omit the SVD for reasons of differentiability.

D A COUNTEREXAMPLE

Example D.1. Let $p \geq 2$. Consider the uniform distribution P_Z on an orthonormal basis $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_p\} \subseteq \mathbb{R}^p$. Then, $\boldsymbol{\Sigma} = \frac{1}{p}\sum_{i=1}^p \boldsymbol{e}_i \boldsymbol{e}_i^\top = \frac{1}{p}\boldsymbol{I}_p$ and hence (COV) is satisfied. However, from Proposition 6 it is easy to see that for any $n \geq 2$, (FRK) is not satisfied. Indeed, if the vector $\boldsymbol{e}_i$ occurs m_i times among the samples $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$, then $\boldsymbol{Z}^\top \boldsymbol{Z} = \operatorname{diag}(m_1, \dots, m_p)$. Assuming $\operatorname{Var}(y|\boldsymbol{z}) = \sigma^2 := 1$ for all $\boldsymbol{z}$, we then obtain from Theorem 3 with the convention $\frac{1}{0} := 0$:

$$\begin{aligned}\mathcal{E}_{\text{Noise}} &= \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) = \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}(\boldsymbol{Z}^+(\boldsymbol{Z}^+)^\top) = \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^\top \boldsymbol{Z})^+) \\ &= \frac{1}{p}\sum_{i=1}^p \mathbb{E}\frac{1}{m_i} = \mathbb{E}\frac{1}{m_1}\,,\end{aligned}$$

where m_1 follows a binomial distribution with parameters n and $\frac{1}{p}$. Using $\mathbb{E}\frac{1}{m_1} \leq \mathbb{E}1 = 1$, it is easy to see that the lower bound is violated for some values of n. For example, Figure D.1 shows that for $p = 30$, the lower bound is violated in a large region, especially in the overparameterized regime.[4] The underlying reason is that the function $x \mapsto \frac{1}{x}$ is convex on $(0, \infty)$, but not on $[0, \infty)$. If $\boldsymbol{Z}$ has singular values s_i, the pseudo-inverse $\boldsymbol{Z}^+$ has singular values $\frac{1}{s_i}$. If $s_i = 0$ is possible, we cannot use a convexity-based argument anymore. ◀

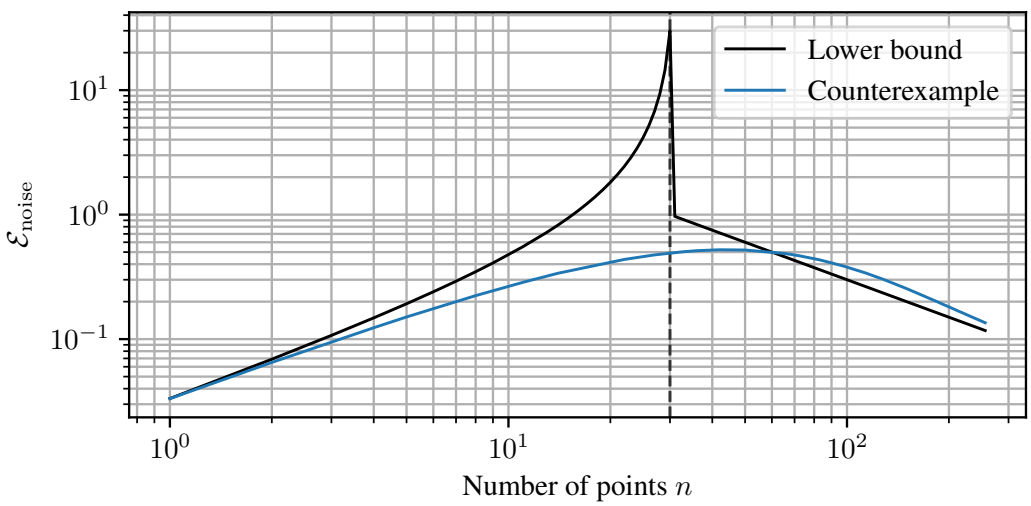

Figure D.1: The counterexample given in Example D.1 for $p = 30$ often has lower $\mathcal{E}_{\text{Noise}}$ than the lower bound from Theorem 3, but violates its assumption (FRK). For the counterexample, $\mathcal{E}_{\text{Noise}}$ was approximated as explained in Example D.1 using 10^6 Monte Carlo samples for each n. We assume $\operatorname{Var}(y|\boldsymbol{z}) = 1$ almost surely over $\boldsymbol{z}$.

E FULL-RANK RESULTS FOR RANDOM WEIGHT NEURAL NETWORKS

In the literature on neural networks with random weights (Schmidt et al., 1992) and the later virtually identical Extreme Learning Machine (ELM) (Huang et al., 2006), properties similar to (FRK) have been investigated for random neural network feature maps. In the following, we review the relevant approaches known to us and explain why they are flawed.

First, Sartori & Antsaklis (1991) state a result where the assumptions are not clearly specified, but which could look as follows:

[4] For $n = 1$, (FRK) holds and it is easy to see that the lower bound holds exactly in this case.

Claim 1 (Sartori & Antsaklis (1991), informally discussed after Lemma 1)**.** *For $n = p$, consider a single-layer random neural network with weights $\boldsymbol{W}^{(0)} \in \mathbb{R}^{(n-1)\times d}$ and signum activation σ that appends a 1 to its output:*

$$\phi_{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}^p, \boldsymbol{x} \mapsto (\sigma(\boldsymbol{W}^{(0)}\boldsymbol{x}), 1) \ .$$

If $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ are distinct, then for almost all $\boldsymbol{W}^{(0)}$, $\phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ is invertible.

This claim is evidently false for $n \geq 2$. For example, the following argument works for $n \geq 3$: For $\boldsymbol{\tau} \in \{-1, 0, 1\}^n$, let $\mathcal{U}_{\boldsymbol{\tau}} := \{\boldsymbol{w} \in \mathbb{R}^n \mid \sigma(\boldsymbol{w}^\top \boldsymbol{x}_j) = \tau_j \text{ for all } j\}$. Then, $\bigcup_{\boldsymbol{\tau}\in\{-1,0,1\}^n} \mathcal{U}_{\boldsymbol{\tau}} = \mathbb{R}^n$ and since $\{-1, 0, 1\}^n$ is finite, there exists $\boldsymbol{\tau}^*$ such that $\mathcal{U}_{\boldsymbol{\tau}^*}$ is not a Lebesgue null set. Hence the set $\mathcal{W} := \{\boldsymbol{W}^{(0)} \in \mathbb{R}^{(n-1)\times n} \mid \text{at least two rows of } \boldsymbol{W}^{(0)} \text{ are in } \mathcal{U}_{\boldsymbol{\tau}^*}\}$ is not a Lebesgue null set. But for all $\boldsymbol{W}^{(0)} \in \mathcal{W}$, the matrix $\phi_{\boldsymbol{\theta}}(\boldsymbol{X}) = (\sigma(\boldsymbol{X}(\boldsymbol{W}^{(0)})^\top) \mid \boldsymbol{1}_{n\times 1})$ has two columns equal to $\boldsymbol{\tau}^*$ and is therefore not invertible, contradicting Claim 1.

Second, Tamura & Tateishi (1997) state a result where the assumptions are not clearly formulated, but which could look as follows:

Claim 2 (Tamura & Tateishi (1997))**.** *For $n = p$, consider a single-layer random neural network with weights $\boldsymbol{W}^{(0)} \in \mathbb{R}^{(n-1)\times d}$, biases $\boldsymbol{b}^{(0)} \in \mathbb{R}^{n-1}$ and sigmoid activation σ that appends a 1 to its output:*

$$\phi_{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}^p, \boldsymbol{x} \mapsto (\sigma(\boldsymbol{W}^{(0)}\boldsymbol{x} + \boldsymbol{b}^{(0)}), 1) \ .$$

If $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ are distinct, then for almost all $\boldsymbol{W}^{(0)}$ and all $a < b$, there exists $\boldsymbol{b}^{(0)} \in [a, b]^{n-1}$ such that $\phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ is invertible.

Tamura & Tateishi (1997) attempt to prove this claim via showing that the curve $c_i : [a, b] \to \mathbb{R}^n, b_i \mapsto \left(\sigma((\boldsymbol{w}_i^{(0)})^\top \boldsymbol{x}_j + b_i)\right)_{j\in[n]}$ is not contained in any $(n-1)$-dimensional subspace of $\mathbb{R}^n$, which would allow them to pick a suitable bias b_i for each curve c_i such that the $c_i(b_i)$ and $(1, \dots, 1)^\top$ are linearly independent. Although this is part is formulated confusingly, it should work out. The major problem is that under the counterassumption that c_i lies in an $(n-1)$-dimensional subspace, they construct an infinite (overdetermined) system of equations involving derivatives of the sigmoid function which they claim has no solution. However, in general, overdetermined systems can have solutions and they do not prove why this would not be the case in their situation. Indeed, the only properties of the sigmoid function they use is that it is C^∞ and not a polynomial, and it is not hard to see that the exponential function has these properties as well and leads to a solvable overdetermined system since its derivatives are all identical.

This leads us to the next claim:

Claim 3 (Huang (2003), Theorem 2.1)**.** *Consider the setting of Claim 2 but with $\boldsymbol{W}^{(0)} \in \mathbb{R}^{n\times d}$ instead of appending a 1 to all feature vectors. If $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ are distinct and $\boldsymbol{\theta}$ has a Lebesgue density, then $\phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ is invertible almost surely over $\boldsymbol{\theta}$.*

Huang (2003) uses the "proof" by Tamura & Tateishi (1997) to show that from any nontrivial intervals one can choose $\boldsymbol{W}^{(0)}, \boldsymbol{b}^{(0)}$ such that $\phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ is invertible, setting all rows of $\boldsymbol{W}^{(0)}$ to be equal. He then concludes without further reasoning that $\phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ must then be invertible almost surely over $\boldsymbol{\theta}$. This major unexplained step cannot be fixed by a continuity-based argument since all $b_i^{(0)}$ were chosen from the same interval and similarly, all rows of $\boldsymbol{W}^{(0)}$ were chosen from the same interval. Since the sigmoid function is analytic, it would however be possible to prove this step using the multivariate identity theorem for analytic functions, Lemma I.3, in a fashion similar to the proof of (d) $\Rightarrow$ (a) in Proposition 6. The approach pursued in Huang (2003) thus introduces an additional problem on top of the problem of Tamura & Tateishi (1997) although this additional problem is in principle fixable.

A similar strategy is reused in the next claim issued in a particularly popular paper:

Claim 4 (Huang et al. (2006), Theorem 2.1)**.** *Claim 3 holds for all C^∞ activation functions σ.*

Not only does the "proof" in Huang et al. (2006) inherit the problems from the previous "proofs", but now the result is obviously false as well: For $\sigma \equiv 0$, the matrix $\phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ can never be invertible,

and for $\sigma = \mathrm{id}$, it is also easy to find counterexamples. Moreover, it is not sufficient to exclude (low-order) polynomials σ: For example, the well-known construction (e.g. Remark 3.4 (d) in Chapter V.3 in Amann & Escher, 2005)

$$\sigma(x) \coloneqq \begin{cases} 0 & , x \leq 0 \\ e^{-1/x} & , x > 0 \end{cases}$$

yields a C^∞ function σ that is zero on $(-\infty, 0]$ but not a polynomial. For this function, it is not hard to see that $\phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ would be singular with nonzero probability since there is a chance that $\boldsymbol{W}^{(0)}\boldsymbol{x} + \boldsymbol{b}^{(0)}$ contains only negative components.

Despite these evident problems and the paper's popularity, Claim 4 is restated years later as Theorem 1 in a survey paper by the same author (Huang et al., 2015). In his Theorem 1, Guo (2018) attempts to disprove Claim 4. However, this "disproof" is also invalid: For certain saturating activation functions like $\tanh$, Guo (2018) lets $\boldsymbol{\theta} \to \infty$ in a certain fashion, which causes $\phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ to converge to a singular matrix. The problem with this approach is that just because the limiting matrix is singular, the matrices for finite $\boldsymbol{\theta}$ do not need to be singular. However, this limiting behavior might at least explain the empirical results of Henriquez & Ruz (2017), which find in a certain setting with sigmoid activation that numerically, $\phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ is usually not a full-rank matrix. In contrast, Widrow et al. (2013) numerically reach the opposite conclusion. This is not a contradiction, considering that very small eigenvalues of $\phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ may be numerically rounded to zero, and the probability of having such small eigenvalues depends on the chosen distributions of $\boldsymbol{x}$ and $\boldsymbol{\theta}$.

In conclusion, there are multiple flawed attempts to prove or disprove the full-rank property for $\phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ for various single-layer random networks. Our Theorem 10 resolves these problems by providing an affirmative proof that is substantially different from the previous attempts. Moreover, it also applies to deeper networks. As argued above, it is not possible to replace the assumption that σ is analytic with $\sigma \in C^\infty$ and as shown in Lemma I.4, it is in general not possible to weaken the assumption that σ must not be a polynomial with less than p nonzero coefficients.

F Proofs for Section 4

In this section, we prove our main theorem and corollary from Section 4. Recall from Definition 2 that if $\mathbb{E}\|\boldsymbol{z}\|_2^2 < \infty$, we can define the second moment matrix

$$\boldsymbol{\Sigma} \coloneqq \mathbb{E}_{\boldsymbol{z} \sim P_Z}\left[\boldsymbol{z}\boldsymbol{z}^\top\right] \in \mathbb{R}^{p \times p} ,$$

and if $\boldsymbol{\Sigma}$ is invertible, we can also define the "whitened" data matrix $\boldsymbol{W} \coloneqq \boldsymbol{Z}\boldsymbol{\Sigma}^{-1/2}$, whose rows $\boldsymbol{w}_i = \boldsymbol{\Sigma}^{-1/2}$ satisfy $\mathbb{E}\boldsymbol{w}_i\boldsymbol{w}_i^\top = \boldsymbol{I}_p$.

Theorem 3 (Main result)**.** *Let $n, p \geq 1$. Assume that P and ϕ satisfy:*

(INT) $\mathbb{E}y^2 < \infty$ *and hence* $R_P(f_P^*) < \infty$*,*
(NOI) $\mathrm{Var}(y|\boldsymbol{z}) \geq \sigma^2$ *almost surely over* $\boldsymbol{z}$*,*
(MOM) $\mathbb{E}\|\boldsymbol{z}\|_2^2 < \infty$*, i.e.* $\boldsymbol{\Sigma}$ *exists and is finite,*
(COV) $\boldsymbol{\Sigma}$ *is invertible,*
(FRK) $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ *almost surely has full rank (i.e.* $\operatorname{rank} \boldsymbol{Z} = \min\{n, p\}$*).*

Then, for the ridgeless linear regression estimator $f_{\boldsymbol{Z},\boldsymbol{y}}(\boldsymbol{z}) = \boldsymbol{z}^\top \boldsymbol{Z}^+ \boldsymbol{y}$, the following holds:

$$\text{If } p \geq n, \qquad \mathcal{E}_{\text{Noise}} \overset{(*)}{\geq} \sigma^2 \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) \overset{(**)}{\geq} \sigma^2 \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) \geq \sigma^2 \frac{n}{p+1-n} .$$

$$\text{If } p \leq n, \qquad \mathcal{E}_{\text{Noise}} \overset{(*)}{\geq} \sigma^2 \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) \overset{(***)}{=} \sigma^2 \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{W}^\top\boldsymbol{W})^{-1}) \geq \sigma^2 \frac{p}{n} .$$

Here, the matrix inverses exist almost surely in the considered cases. Moreover, we have:

- *If (NOI) holds with equality, then $(*)$ holds with equality.*
- *If $n = p$ or $\boldsymbol{\Sigma} = \lambda \boldsymbol{I}_p$ for some $\lambda > 0$, then $(**)$ holds with equality.*

Proof. By (FRK), we only consider the case where $\boldsymbol{Z}$ has full rank. For further details on some matrix computations, we refer to Appendix B.

Step 1: Preparation. Since the pairs $(\boldsymbol{z}_1, y_1), \dots, (\boldsymbol{z}_n, y_n)$ are independent, we have the conditional covariance matrix

$$\operatorname{Cov}(\boldsymbol{y}|\boldsymbol{Z}) = \begin{pmatrix} \operatorname{Var}(y_1|\boldsymbol{z}_1) & & \\ & \ddots & \\ & & \operatorname{Var}(y_n|\boldsymbol{z}_n) \end{pmatrix} \overset{(*)}{\succeq} \sigma^2 \boldsymbol{I}_n$$

with equality in $(*)$ if (NOI) holds with equality. Therefore,

$$\begin{aligned}
\mathcal{E}_{\text{Noise}} &\overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{Z},\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{z}} \left(f_{\boldsymbol{Z},\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{z}) - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{Z}} f_{\boldsymbol{Z},\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{z}) \right)^2 \\
&= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{y},\boldsymbol{z}} \left(\boldsymbol{z}^\top \boldsymbol{Z}^+ \boldsymbol{y} - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{Z}} \boldsymbol{z}^\top \boldsymbol{Z}^+ \boldsymbol{y} \right)^2 \\
&= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{z}} \mathbb{E}_{\boldsymbol{y}|\boldsymbol{Z}} \boldsymbol{z}^\top \boldsymbol{Z}^+ (\boldsymbol{y} - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{Z}} \boldsymbol{y})(\boldsymbol{y} - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{Z}} \boldsymbol{y})^\top (\boldsymbol{Z}^+)^\top \boldsymbol{z} \\
&= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{z}} \boldsymbol{z}^\top \boldsymbol{Z}^+ \operatorname{Cov}(\boldsymbol{y}|\boldsymbol{Z}) (\boldsymbol{Z}^+)^\top \boldsymbol{z} \\
&\overset{(*)}{\geq} \sigma^2 \mathbb{E}_{\boldsymbol{Z},\boldsymbol{z}} \boldsymbol{z}^\top \boldsymbol{Z}^+ (\boldsymbol{Z}^+)^\top \boldsymbol{z} \\
&= \sigma^2 \mathbb{E}_{\boldsymbol{Z},\boldsymbol{z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{z} \boldsymbol{z}^\top \boldsymbol{Z}^+) \\
&= \sigma^2 \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} (\boldsymbol{Z}^+)^\top) \,.
\end{aligned}$$

Step 2.1: Reformulation, underparameterized case. In the case $p \leq n$, we have $\boldsymbol{Z}^+ = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top$ thanks to (FRK) and thus

$$\operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) = \operatorname{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{Z}^+ (\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma}^{1/2}) = \operatorname{tr}(\boldsymbol{\Sigma}^{1/2} (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{\Sigma}^{1/2}) = \operatorname{tr}((\boldsymbol{W}^\top \boldsymbol{W})^{-1}) \,.$$

Step 2.2: Reformulation, overparameterized case. In the case $p \geq n$, we have $\boldsymbol{Z}^+ = \boldsymbol{Z}^\top (\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1}$ thanks to (FRK). For $\boldsymbol{\Sigma} = \lambda \boldsymbol{I}_p$, we can show $\operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) = \operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1})$ similar to Step 2.1. For general $\boldsymbol{\Sigma}$, we can obtain a lower bound by "removing a projection": First, let

$$\begin{aligned}
\boldsymbol{S} &\coloneqq (\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+ = (\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1} \boldsymbol{Z} \boldsymbol{\Sigma} \boldsymbol{Z}^\top (\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1} \,, \\
\boldsymbol{A} &\coloneqq \boldsymbol{\Sigma}^{1/2} \boldsymbol{Z}^\top \,.
\end{aligned}$$

Now, since $\boldsymbol{Z}$ and $\boldsymbol{\Sigma}$ have full rank, we can compute

$$\boldsymbol{S}^{-1} = \boldsymbol{Z}\boldsymbol{Z}^\top (\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^\top)^{-1} \boldsymbol{Z}\boldsymbol{Z}^\top = \boldsymbol{W}\boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{W}^\top \,.$$

Since $\boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top$ is the orthogonal projection onto the column space of $\boldsymbol{A}$, we have $\boldsymbol{S}^{-1} \preceq \boldsymbol{W}\boldsymbol{W}^\top$ and hence $\lambda_i(\boldsymbol{S}^{-1}) \leq \lambda_i(\boldsymbol{W}\boldsymbol{W}^\top)$ by the Courant-Fischer-Weyl theorem. This yields

$$\begin{aligned}
\operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) &= \operatorname{tr}(\boldsymbol{S}) = \sum_{i=1}^n \lambda_{n+1-i}(\boldsymbol{S}) = \sum_{i=1}^n \frac{1}{\lambda_i(\boldsymbol{S}^{-1})} \\
&\geq \sum_{i=1}^n \frac{1}{\lambda_i(\boldsymbol{W}\boldsymbol{W}^\top)} = \sum_{i=1}^n \lambda_{n+1-i}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) = \operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) \,.
\end{aligned}$$

For $n = p$, $\boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top$ projects onto a p-dimensional space and is therefore the identity matrix, yielding equality.

Step 3.0: Random matrix bound, $p = 1$ or $n = 1$. If $n \geq p = 1$, $\boldsymbol{w}_i = w_i$ is a scalar and we have

$$\mathbb{E} \operatorname{tr}((\boldsymbol{W}^\top \boldsymbol{W})^{-1}) = \mathbb{E} \frac{1}{\boldsymbol{W}^\top \boldsymbol{W}} = \mathbb{E} \frac{1}{\sum_{i=1}^n w_i^2} \geq \frac{1}{\mathbb{E} \sum_{i=1}^n w_i^2} = \frac{1}{\sum_{i=1}^n \operatorname{tr}(\mathbb{E} w_i w_i^\top)} = \frac{1}{n} = \frac{p}{n}$$

by Jensen's inequality. Similarly, for $p \geq n = 1$, we obtain

$$\mathbb{E} \operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) = \mathbb{E} \frac{1}{\boldsymbol{w}_1^\top \boldsymbol{w}_1} \geq \frac{1}{\mathbb{E} \boldsymbol{w}_1^\top \boldsymbol{w}_1} = \frac{1}{\operatorname{tr}(\mathbb{E} \boldsymbol{w}_1 \boldsymbol{w}_1^\top)} = \frac{1}{p} = \frac{n}{p+1-n} \,.$$

Step 3.1: Random matrix bound, overparameterized case. We first consider the overparameterized case $p \geq n \geq 2$ and block-decompose

$$\boldsymbol{W} =: \begin{pmatrix} \boldsymbol{w}_1^\top \\ \boldsymbol{W}_2 \end{pmatrix} \in \mathbb{R}^{(1+(n-1)) \times p} \quad \Rightarrow \quad \boldsymbol{W}\boldsymbol{W}^\top = \begin{pmatrix} \boldsymbol{w}_1^\top \boldsymbol{w}_1 & \boldsymbol{w}_1^\top \boldsymbol{W}_2^\top \\ \boldsymbol{W}_2 \boldsymbol{w}_1 & \boldsymbol{W}_2 \boldsymbol{W}_2^\top \end{pmatrix} \,.$$

Since $\boldsymbol{Z}$ has full rank, $\boldsymbol{W}$ has full rank. Because of $n \leq p$, it follows that $\boldsymbol{W}\boldsymbol{W}^\top \succ 0$. Therefore,

$$\left((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}\right)_{11} = \left(\boldsymbol{w}_1^\top \boldsymbol{w}_1 - \boldsymbol{w}_1^\top \boldsymbol{W}_2^\top (\boldsymbol{W}_2 \boldsymbol{W}_2^\top)^{-1} \boldsymbol{W}_2 \boldsymbol{w}_1\right)^{-1} = \left(\boldsymbol{w}_1^\top (\boldsymbol{I}_p - \boldsymbol{P}_2) \boldsymbol{w}_1\right)^{-1} ,$$

where

$$\boldsymbol{P}_2 \coloneqq \boldsymbol{W}_2^\top \left(\boldsymbol{W}_2 \boldsymbol{W}_2^\top\right)^{-1} \boldsymbol{W}_2 \in \mathbb{R}^{p \times p}$$

is the orthogonal projection onto the column space of $\boldsymbol{W}_2^\top$. Thus, $\boldsymbol{P}_2$ has the eigenvalues 1 with multiplicity $n-1$ and 0 with multiplicity $p-(n-1)$, which yields $\operatorname{tr}(\boldsymbol{P}_2) = n-1$. Since the $\boldsymbol{z}_i$ are stochastically independent, $\boldsymbol{w}_1$ and $\boldsymbol{W}_2$ are also stochastically independent and we obtain

$$\mathbb{E}\boldsymbol{w}_1^\top (\boldsymbol{I}_p - \boldsymbol{P}_2) \boldsymbol{w}_1 = \mathbb{E} \operatorname{tr}((\boldsymbol{I}_p - \boldsymbol{P}_2)(\mathbb{E}_{\boldsymbol{w}_1} \boldsymbol{w}_1 \boldsymbol{w}_1^\top)) = \mathbb{E} \operatorname{tr}(\boldsymbol{I}_p - \boldsymbol{P}_2) = p + 1 - n .$$

Using Jensen's inequality with the convex function $(0,\infty) \to (0,\infty), x \mapsto 1/x$, we thus find that

$$\mathbb{E}\left((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}\right)_{11} = \mathbb{E}\left(\boldsymbol{w}_1^\top (\boldsymbol{I}_p - \boldsymbol{P}_2) \boldsymbol{w}_1\right)^{-1} \geq \frac{1}{\mathbb{E}\boldsymbol{w}_1^\top (\boldsymbol{I}_p - \boldsymbol{P}_2) \boldsymbol{w}_1} = \frac{1}{p+1-n} .$$

Since $\operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) = \sum_{i=1}^n ((\boldsymbol{W}\boldsymbol{W}^\top)^{-1})_{ii}$ and all diagonal entries can be treated in the same fashion (e.g. via permutation of the $\boldsymbol{w}_i$), it follows that

$$\mathbb{E} \operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) \geq \frac{n}{p+1-n} .$$

Step 3.2: Random matrix bound, underparameterized case. In the underparameterized case $n \geq p \geq 2$, we similarly block-decompose

$$\boldsymbol{W} = (\boldsymbol{W}_{:,1} \quad \boldsymbol{W}_{:,2:p}) \in \mathbb{R}^{n \times (1+(p-1))} \quad \Rightarrow \quad \boldsymbol{W}^\top \boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{:,1}^\top \boldsymbol{W}_{:,1} & \boldsymbol{W}_{:,1}^\top \boldsymbol{W}_{:,2:p} \\ \boldsymbol{W}_{:,2:p}^\top \boldsymbol{W}_{:,1} & \boldsymbol{W}_{:,2:p}^\top \boldsymbol{W}_{:,2:p} \end{pmatrix} .$$

Again using the same Schur complement argument, we obtain

$$\begin{aligned} ((\boldsymbol{W}^\top \boldsymbol{W})^{-1})_{11} &= \left(\boldsymbol{W}_{:,1}^\top \boldsymbol{W}_{:,1} - \boldsymbol{W}_{:,1}^\top \boldsymbol{W}_{:,2:p} \left(\boldsymbol{W}_{:,2:p}^\top \boldsymbol{W}_{:,2:p}\right)^{-1} \boldsymbol{W}_{:,2:p}^\top \boldsymbol{W}_{:,1}\right)^{-1} \\ &\geq \left(\boldsymbol{W}_{:,1}^\top \boldsymbol{W}_{:,1}\right)^{-1} . \end{aligned}$$

Moreover, since

$$\mathbb{E}\boldsymbol{W}_{:,1}^\top \boldsymbol{W}_{:,1} = \sum_{i=1}^n \mathbb{E}(W_{i,1})^2 = \sum_{i=1}^n (\Sigma_{11})^2 = n ,$$

we can again use Jensen (and again treat all diagonal entries in the same fashion) to obtain

$$\mathbb{E} \operatorname{tr}\left((\boldsymbol{W}^\top \boldsymbol{W})^{-1}\right) = \sum_{i=1}^p \mathbb{E}((\boldsymbol{W}^\top \boldsymbol{W})^{-1})_{ii} \geq \sum_{i=1}^p \frac{1}{\mathbb{E}\boldsymbol{W}_{:,i}^\top \boldsymbol{W}_{:,i}} = \frac{p}{n} . \qquad \square$$

Note that it is not possible to analyze Step 3.2 like Step 3.1 since $\boldsymbol{W}_{:,1}$ and $\boldsymbol{W}_{:,2:p}$ are not stochastically independent.

Corollary 4 (Random features)**.** *Let $\boldsymbol{\theta} \sim P_\Theta$ be a random variable such that $\phi_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^p$ is a random feature map. Consider the random features regression estimator $f_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{z}_{\boldsymbol{\theta}}^\top \boldsymbol{Z}_{\boldsymbol{\theta}}^+ \boldsymbol{y}$ with $\boldsymbol{z}_{\boldsymbol{\theta}} \coloneqq \phi_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $\boldsymbol{Z}_{\boldsymbol{\theta}} \coloneqq \phi_{\boldsymbol{\theta}}(\boldsymbol{X})$. If for P_Θ-almost all $\boldsymbol{\theta}_{\text{const}}$, the assumptions of Theorem 3 are satisfied for $\boldsymbol{z} = \boldsymbol{z}_{\boldsymbol{\theta}_{\text{const}}}$ and $\boldsymbol{Z} = \boldsymbol{Z}_{\boldsymbol{\theta}_{\text{const}}}$ (with the corresponding matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\text{const}}}$), then*

$$\mathcal{E}_{\text{Noise}} \geq \begin{cases} \sigma^2 \frac{n}{p+1-n} & \text{if } p \geq n, \\ \sigma^2 \frac{p}{n} & \text{if } p \leq n. \end{cases}$$

Proof. First, let $p \leq n$. Since $\boldsymbol{\theta}$ is independent from $\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{y}$,

$$\begin{aligned} \mathcal{E}_{\text{Noise}} &= \mathbb{E}_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{x}} \left(\boldsymbol{z}_{\boldsymbol{\theta}}^\top \boldsymbol{Z}_{\boldsymbol{\theta}}^+ \boldsymbol{y} - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}} \boldsymbol{z}_{\boldsymbol{\theta}}^\top \boldsymbol{Z}_{\boldsymbol{\theta}}^+ \boldsymbol{y}\right)^2 \\ &= \mathbb{E}_{\boldsymbol{\theta}} \left[\mathbb{E}_{\boldsymbol{X},\boldsymbol{y},\boldsymbol{x}} \left(\boldsymbol{z}_{\boldsymbol{\theta}}^\top \boldsymbol{Z}_{\boldsymbol{\theta}}^+ \boldsymbol{y} - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}} \boldsymbol{z}_{\boldsymbol{\theta}}^\top \boldsymbol{Z}_{\boldsymbol{\theta}}^+ \boldsymbol{y}\right)^2\right] \overset{\text{Theorem 3}}{\geq} \mathbb{E}_{\boldsymbol{\theta}} \sigma^2 \frac{p}{n} = \sigma^2 \frac{p}{n} . \end{aligned}$$

The case $p \geq n$ can be treated analogously. $\square$

G Discussion of the Main Theorem

Remark G.1 (Dependence on $\boldsymbol{\Sigma}$)**.** Hastie et al. (2019) discuss that $\boldsymbol{\Sigma}$ only influences the expected excess risk in the overparameterized regime. In the following, we will illustrate that Theorem 3 even implies that in the overparameterized case, $\boldsymbol{\Sigma} = \lambda \boldsymbol{I}_d$ yields the lowest $\mathcal{E}_{\text{Noise}}$. This fact is also discussed in Muthukumar et al. (2020) in a slightly different setting. Note that since P_X is unknown in general, $\boldsymbol{\Sigma}$ is also unknown in general.

Assume that (NOI) holds with equality such that we have $\mathcal{E}_{\text{Noise}} = \sigma^2 \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+)$ by Theorem 3. Suppose that we perform linear regression on the whitened data $\tilde{\boldsymbol{z}} \coloneqq \boldsymbol{\Sigma}^{-1/2}\boldsymbol{z}$. Then,

$$\begin{aligned}
\tilde{\boldsymbol{Z}} &= \boldsymbol{Z}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{W} \,, \\
\tilde{\boldsymbol{\Sigma}} &= \mathbb{E}_{\tilde{\boldsymbol{z}}} \tilde{\boldsymbol{z}}\tilde{\boldsymbol{z}}^\top = \boldsymbol{\Sigma}^{-1/2} \left[\mathbb{E}_{\boldsymbol{z}} \boldsymbol{z}\boldsymbol{z}^\top\right] \boldsymbol{\Sigma}^{-1/2} = \boldsymbol{I}_p \,, \\
\tilde{\boldsymbol{W}} &= \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\Sigma}}^{-1/2} = \tilde{\boldsymbol{Z}} = \boldsymbol{W} \,.
\end{aligned}$$

In the underparameterized case $p \leq n$, we then obtain

$$\mathbb{E}_{\tilde{\boldsymbol{Z}}} \operatorname{tr}((\tilde{\boldsymbol{Z}}^+)^\top \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{Z}}^+) = \mathbb{E}_{\tilde{\boldsymbol{Z}}} \operatorname{tr}((\tilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{W}})^{-1}) = \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{W}^\top \boldsymbol{W})^{-1}) = \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) \,.$$

Therefore, whitening the data does not make a difference if $p \leq n$. In contrast, for $p > n$, we only know

$$\mathbb{E}_{\tilde{\boldsymbol{Z}}} \operatorname{tr}((\tilde{\boldsymbol{Z}}^+)^\top \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{Z}}^+) \overset{(**)}{=} \mathbb{E}_{\tilde{\boldsymbol{Z}}} \operatorname{tr}((\tilde{\boldsymbol{W}} \tilde{\boldsymbol{W}}^\top)^{-1}) = \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) \overset{(**)}{\leq} \mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+)$$

since $(**)$ holds with equality for whitened features. From Step 2.1 in the proof of Theorem 3, it is obvious that $(**)$ in general does not hold with equality. Hence, in the overparameterized case $p > n$, whitening the features often reduces and never increases $\mathcal{E}_{\text{Noise}}$. Since $\mathcal{E}_{\text{Noise}}$ is just a lower bound for the expected excess risk, whitening does not necessarily reduce the expected excess risk.

This phenomenon also has a different kernel-based interpretation: Under the assumptions (MOM) and (COV), we can choose an ONB $\boldsymbol{u}_1, \dots, \boldsymbol{u}_p$ of eigenvectors of $\boldsymbol{\Sigma}$ with corresponding eigenvalues $\lambda_1, \dots, \lambda_p > 0$. If $\boldsymbol{z} = \phi(\boldsymbol{x})$ and $k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$, we can write

$$k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\top \left(\sum_{i=1}^n \boldsymbol{u}_i \boldsymbol{u}_i^\top \right) \phi(\boldsymbol{x}') = \sum_{i=1}^p \lambda_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{x}') \,, \tag{5}$$

where the functions $\psi_i : \mathbb{R}^d \to \mathbb{R}, \boldsymbol{x} \mapsto \frac{1}{\sqrt{\lambda_i}} \boldsymbol{u}_i^\top \phi(\boldsymbol{x})$ form an orthonormal system in $L_2(P_X)$:

$$\mathbb{E}_{\boldsymbol{x}} \psi_i(\boldsymbol{x}) \psi_j(\boldsymbol{x}) = \frac{1}{\sqrt{\lambda_i \lambda_j}} \boldsymbol{u}_i^\top \left[\mathbb{E}_{\boldsymbol{x}} \phi(\boldsymbol{x}) \phi(\boldsymbol{x})^\top\right] \boldsymbol{u}_j = \frac{1}{\sqrt{\lambda_i \lambda_j}} \boldsymbol{u}_i^\top \boldsymbol{\Sigma} \boldsymbol{u}_j = \frac{\lambda_j}{\sqrt{\lambda_i \lambda_j}} \boldsymbol{u}_i^\top \boldsymbol{u}_j = \delta_{ij}. \tag{6}$$

Therefore, Eq. (5) is a Mercer representation of k and the eigenvalues λ_i of $\boldsymbol{\Sigma}$ are also the eigenvalues of the integral operator $T_k : L_2(P_X) \to L_2(P_X)$ associated with k:

$$(T_k f)(\boldsymbol{x}) \coloneqq \int k(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \, \mathrm{d}P_X(\boldsymbol{x}') = \sum_{i=1}^n \lambda_i \langle \psi_i, f \rangle_{L_2(P_X)} \psi_i(\boldsymbol{x}) \,.$$

We can define a kernel $\tilde{k}$ with flattened eigenspectrum via

$$\tilde{k}(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \sum_{i=1}^p \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{x}') \,.$$

Its feature map $\boldsymbol{x} \mapsto (\psi_i(\boldsymbol{x}))_{i \in [n]}$ by Eq. (6) satisfies $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. By the above discussion, $\mathcal{E}_{\text{Noise}}(\tilde{k}) \leq \mathcal{E}_{\text{Noise}}(k)$. However, this needs to be taken with caution: Assume for simplicity that for independent $\boldsymbol{x}, \tilde{\boldsymbol{x}}$, the feature map ϕ yields $\phi(\boldsymbol{x}), \phi(\tilde{\boldsymbol{x}}) \sim \mathcal{N}(0, \frac{1}{p}\boldsymbol{I}_p)$. Then, $k(\boldsymbol{x}, \boldsymbol{x}) = 1$ but $\mathbb{E} k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = 0$ and $\operatorname{Var} k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{1}{p^2} \sum_{i=1}^p \mathbb{E}_{u,v \sim \mathcal{N}(0,1)} u^2 v^2 = \frac{1}{p}$. In this sense, for $p \to \infty$, k converges to the Dirac kernel $k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \delta_{\boldsymbol{x}, \tilde{\boldsymbol{x}}}$, which satisfies $\mathcal{E}_{\text{Noise}} = 0$ but provides bad interpolation performance if $f_P^* \not\equiv 0$. ◀

The next lemma and its proof are an adaptation of Lemma 4.14 in Bordenave & Chafaï (2012).

Lemma G.2. *For $i \in [n]$, let $\mathcal{W}_{-i} \coloneqq \operatorname{Span}\{\boldsymbol{w}_j \mid j \in [n] \setminus \{i\}\}$. Then, under the assumptions of Theorem 3, in the overparameterized case $p \geq n$,*

$$\operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) = \sum_{i=1}^{n} \operatorname{dist}(\boldsymbol{w}_i, \mathcal{W}_{-i})^{-2} \,. \tag{7}$$

Proof. The case $n = 1$ is trivial, hence let $n \geq 2$. In Step 3.1 in the proof of Theorem 3, since $\boldsymbol{P}_2$ is the orthogonal projection onto $\mathcal{W}_{-1}$, we have

$$\boldsymbol{w}_1^\top (\boldsymbol{I}_p - \boldsymbol{P}_2)\boldsymbol{w}_1 = \|\boldsymbol{w}_1\|_2^2 - \|\boldsymbol{P}_2\boldsymbol{w}_1\|_2^2 \overset{\text{Pythagoras}}{=} \operatorname{dist}(\boldsymbol{w}_1, \mathcal{W}_{-1})^2 \,,$$

where $\operatorname{dist}(\boldsymbol{w}_1, \mathcal{W}_{-1})$ is the Euclidean distance between $\boldsymbol{w}_1$ and $\mathcal{W}_{-1}$. □

Remark G.3 (Is $\mathcal{U}(\mathbb{S}^{p-1})$ optimal?)**.** From $(*)$ in Theorem 3 it is clear that the best possible lower bound for $\mathcal{E}_{\text{Noise}}$ under the assumptions of Theorem 3 given $n, p \geq 1$ is

$$\mathcal{E}_{\text{Noise}} \geq \sigma^2 \inf_{\text{Distribution } P_Z \text{ on } \mathbb{R}^p \text{ satisfying (MOM), (COV), (FRK)}} \mathbb{E}_{\boldsymbol{Z}\sim P_Z^n} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) \,. \tag{8}$$

Here, we want to discuss the hypothesis that the infimum in Eq. (8) is achieved (for example) by $P_Z = \mathcal{U}(\mathbb{S}^{p-1})$. Figure 1 shows that we were not able to obtain a lower $\mathcal{E}_{\text{Noise}}$ by optimizing a neural network feature map to minimize $\mathcal{E}_{\text{Noise}}$. In the following, we want to discuss some theoretical evidence as to why this is plausible in the overparameterized case $p \geq n$. Lemma G.2 shows that Step 3.1 of the proof of Theorem 3 has a distance-based interpretation. In this interpretation, Step 3.1 then applies Jensen's inequality to the convex function $(0, \infty) \to (0, \infty), x \mapsto 1/x$ using

$$\mathbb{E} \operatorname{dist}(\boldsymbol{w}_i, \mathcal{W}_{-i})^2 = p + 1 - n \,. \tag{9}$$

We can use this perspective to gain insights on how distributions P_Z with small $\mathcal{E}_{\text{Noise}}$ in the overparameterized case $p \geq n$ look like. First of all, Remark G.1 suggests that for minimizing $\mathcal{E}_{\text{Noise}}$ in the overparameterized case, $\boldsymbol{\Sigma}$ should be a multiple of $\boldsymbol{I}_p$, which is clearly satisfied for $\mathcal{U}(\mathbb{S}^{p-1})$ by Theorem 11. Since the lower bound obtained from (9) is independent of the distribution of $\boldsymbol{W}$, minimizing $\mathbb{E} \operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1})$ amounts to minimizing the error made by Jensen's inequality, which essentially amounts to reducing the variance of the random variables $\operatorname{dist}(\boldsymbol{w}_i, \mathcal{W}_{-i})$. We can decompose $\operatorname{dist}(\boldsymbol{w}_i, \mathcal{W}_{-i}) = \|\boldsymbol{w}_i\|_2 \cdot \operatorname{dist}(\boldsymbol{w}_i/\|\boldsymbol{w}_i\|_2, \mathcal{W}_{-i})$, where $\operatorname{dist}(\boldsymbol{w}_i/\|\boldsymbol{w}_i\|_2, \mathcal{W}_{-i})$ only depends on the angular components $\boldsymbol{w}_j/\|\boldsymbol{w}_j\|_2$ for $j \in [n]$. This suggests that for a "good" distribution P_W,

- the radial component $\|\boldsymbol{w}_i\|_2$ should have low variance, and
- the distribution of the angular component $\boldsymbol{w}_i/\|\boldsymbol{w}_i\|_2$ should not contain "clusters", since clusters would increase the probability of $\operatorname{dist}(\boldsymbol{w}_i, \mathcal{W}_{-i})$ being very small.

Clearly, both points are perfectly achieved for $P_Z = \mathcal{U}(\mathbb{S}^{p-1})$. ◀

The next proposition shows that a small input dimension d does not necessarily provide a limitation:

Proposition G.4. *Let $p, d \geq 1$. Then, there exists a probability distribution P_X on $\mathbb{R}^d$ (with bounded support) and a continuous feature map $\phi : \mathbb{R}^d \to \mathbb{R}^p$ such that for $\boldsymbol{x} \sim P_X$, $\phi(\boldsymbol{x}) \sim \mathcal{U}(\mathbb{S}^{p-1})$.*

Proof. For $p = 1$, the result is trivial, we will therefore assume $p \geq 2$. We will prove the result for any d by a reduction to the case $d = 1$, although substantially simpler constructions are possible for $d \geq p - 1$. First, introduce the spaces

$$\begin{aligned} \mathbb{S}_+^{p-1} &\coloneqq \{\boldsymbol{z} \in \mathbb{S}^{p-1} \mid z_p \geq 0\} \subseteq \mathbb{R}^p \\ \mathbb{B}^{p-1} &\coloneqq \{\boldsymbol{z} \in \mathbb{R}^{p-1} \mid \|\boldsymbol{z}\|_2 \leq 1\} \\ \mathcal{X} &\coloneqq [0, 3] \times \{0\}^{p-1} \subseteq \mathbb{R}^p \,. \end{aligned}$$

Step 1: Space-filling curve on the sphere. In this step, we show that there exists a continuous surjective map $\phi : \mathcal{X} \to \mathbb{S}^{p-1}$. First of all, let $f_1 : [0, 1] \to [0, 1]^{p-1}$ be continuous and surjective, e.g. a Hilbert or Peano curve (see e.g. Sagan, 2012). We define the following maps:

$$f_2 : [0, 1]^{p-1} \to \mathbb{B}^{p-1}, \boldsymbol{u} \mapsto \begin{cases} \boldsymbol{0} & \text{if } \boldsymbol{u} = \boldsymbol{0} \\ \frac{\|\boldsymbol{u}\|_\infty}{\|\boldsymbol{u}\|_2} \boldsymbol{u} & \text{if } \boldsymbol{u} \neq \boldsymbol{0} \,, \end{cases}$$

$$f_3 : \mathbb{B}^{p-1} \to \mathbb{S}_+^{p-1}, \boldsymbol{v} \mapsto \left(\boldsymbol{v}, \sqrt{1 - \|\boldsymbol{v}\|_2^2}\right) .$$

It is not hard to verify that f_2 and f_3 are continuous and surjective as well. For example, f_2 is continuous in $\mathbf{0}$ since $\|\boldsymbol{u}\|_\infty \leq \|\boldsymbol{u}\|_2$ for all $\boldsymbol{u} \in \mathbb{R}^{p-1}$. Thus, the map $f := f_3 \circ f_2 \circ f_1 : [0,1] \to \mathbb{S}_+^{p-1}$ is continuous and surjective. Define the map

$$\tau : \mathbb{R}^p \to \mathbb{R}^p, \boldsymbol{z} \mapsto (z_1, \ldots, z_{p-1}, -z_p) .$$

By the previous considerations, it is not hard to verify that the map

$$g : [0,3] \to \mathbb{S}^{p-1}, x \mapsto \begin{cases} f(x) & \text{if } x \in [0,1] \\ \frac{x-1}{2-1} f(1) + \left(1 - \frac{x-1}{2-1}\right) \tau(f(1)) & \text{if } x \in [1,2] \\ \tau(f(3-x)) & \text{if } x \in [2,3] \end{cases}$$

is continuous and surjective as well. We can therefore define the continuous and surjective map $\phi : \mathcal{X} \to \mathbb{S}^{p-1}, \boldsymbol{x} \mapsto g(x_1)$.

Step 2: Existence of a pull-back measure. We consider the Borel σ-algebras $\mathcal{B}(\mathcal{X}), \mathcal{B}(\mathbb{S}^{p-1})$ on $\mathcal{X}$ and $\mathbb{S}^{p-1}$. The uniform distribution $P_Z = \mathcal{U}(\mathbb{S}^{p-1})$ on the sphere is defined with respect to $\mathcal{B}(\mathbb{S}^{p-1})$ and is therefore a Borel measure. Since ϕ is continuous, it is Borel measurable. Moreover, since $\mathcal{X}$ and $\mathbb{S}^{p-1}$ are complete separable metric spaces, they are also Souslin spaces, cf. Section 6.6 in Bogachev (2007). Since ϕ is surjective, Theorem 9.1.5 in Bogachev (2007) guarantees the existence of a measure P_X such that if $\boldsymbol{x} \sim P_X$, then $\phi(\boldsymbol{x}) \sim \mathcal{U}(\mathbb{S}^{p-1})$. Since $P_X(\mathcal{X}) = P_Z(\mathbb{S}^{p-1}) = 1$, P_X is a probability measure.

Step 3: Continuation. We can arbitrarily extend the mapping $\phi : \mathcal{X} \to \mathbb{S}^{p-1}$ to a continuous mapping $\phi : \mathbb{R}^d \to \mathbb{R}^p$. Moreover, the domain $\mathcal{X}$ of P_X can be extended to $\mathbb{R}^d$ via $P_X(A) := P_X(A \cap \mathcal{X})$, the support of P_X is still bounded, and we still have $\phi(\boldsymbol{x}) \sim \mathcal{U}(\mathbb{S}^{p-1})$ if $\boldsymbol{x} \sim P_X$. □

Remark G.5. The proof of Proposition G.4 could be slightly shorter if we required $\phi(\boldsymbol{x}) \sim \mathcal{U}(\mathbb{S}_+^{p-1})$ instead of $\phi(\boldsymbol{x}) \sim \mathcal{U}(\mathbb{S}^{p-1})$. This would be of similar interest since the uniform distribution $\mathcal{U}(\mathbb{S}_+^{p-1})$ on the "half-sphere" leads to the same $\mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma}(\boldsymbol{Z}^+))$ as the uniform distribution $\mathcal{U}(\mathbb{S}^{p-1})$ on the full sphere: If $\boldsymbol{z}_i \sim \mathcal{U}(\mathbb{S}_+^{p-1})$ and $\varepsilon_i \sim \mathcal{U}(\{-1,1\})$ are stochastically independent, then $\tilde{\boldsymbol{z}}_i := \varepsilon_i \boldsymbol{z}_i \sim \mathcal{U}(\mathbb{S}^{p-1})$. Therefore, $\boldsymbol{\Sigma} = \tilde{\boldsymbol{\Sigma}}$, $\tilde{\boldsymbol{Z}} = \operatorname{diag}(\varepsilon_1, \ldots, \varepsilon_n)\boldsymbol{Z}$, and if $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ is a SVD of $\boldsymbol{Z}$, then $(\operatorname{diag}(\varepsilon_1, \ldots, \varepsilon_n)\boldsymbol{U})\boldsymbol{D}\boldsymbol{V}^\top$ is a SVD of $\tilde{\boldsymbol{Z}}$. Therefore, $\boldsymbol{Z}$ and $\tilde{\boldsymbol{Z}}$ have the same singular values, hence $\boldsymbol{W}$ and $\tilde{\boldsymbol{W}}$ have the same singular values, hence $\operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) = \operatorname{tr}((\tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^\top)^{-1})$ for $p \geq n$ and $\operatorname{tr}((\boldsymbol{W}^\top\boldsymbol{W})^{-1}) = \operatorname{tr}((\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}})^{-1})$ for $p \leq n$. ◀

Remark G.6. One might ask whether it is possible in Proposition G.4 to choose P_X as a "nice" distribution, like a uniform distribution on a cube or a Gaussian distribution. The answer to this question is affirmative if there exists an area-preserving space-filling curve $\phi : [0, \operatorname{volume}(\mathbb{S}^{p-1})] \to \mathbb{S}^{p-1}$. For $p = 3$, such a construction is informally described by Purser et al. (2009) and it seems plausible that such a construction is possible for all p. ◀

H Exact Computations

Theorem 11. *Let $P_Z = \mathcal{U}(\mathbb{S}^{p-1})$. Then, P_Z satisfies the assumptions (MOM), (COV) and (FRK) for all n with $\boldsymbol{\Sigma} = \frac{1}{p}\boldsymbol{I}_p$. Moreover, for $n \geq p = 1$ or $p \geq n \geq 1$, we can compute*

$$\mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) = \begin{cases} \frac{1}{n} & \text{if } n \geq p = 1, \\ \frac{1}{p} & \text{if } p \geq n = 1, \\ \infty & \text{if } 2 \leq n \leq p \leq n+1, \\ \frac{n}{p-1-n} \cdot \frac{p-2}{p} & \text{if } 2 \leq n \leq n+2 \leq p. \end{cases}$$

Proof. **Step 1: Verify (MOM).** Let $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_p)$ for $i \in [n]$ be independent. Then, $\boldsymbol{z}_i := \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_2} \sim \mathcal{U}(\mathbb{S}^{p-1})$. Since $\mathbb{E}\|\boldsymbol{z}_i\|_2^2 = \mathbb{E}1 = 1$, (MOM) is satisfied and thus, $\boldsymbol{\Sigma}$ is well-defined.

Step 2: Compute $\boldsymbol{\Sigma}$. We can use rotational invariance as follows: Let $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ be an arbitrary fixed orthogonal matrix. Then, $\boldsymbol{V}\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{V}\boldsymbol{V}^\top) = \mathcal{N}(0, \boldsymbol{I}_p)$ and hence $\boldsymbol{V}\boldsymbol{z}_i = \frac{\boldsymbol{V}\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_2} = \frac{\boldsymbol{V}\boldsymbol{x}_i}{\|\boldsymbol{V}\boldsymbol{x}_i\|_2} \sim \mathcal{U}(\mathbb{S}^{p-1})$. Therefore,

$$\boldsymbol{\Sigma} = \mathbb{E}\boldsymbol{z}_i\boldsymbol{z}_i^\top = \mathbb{E}\boldsymbol{V}\boldsymbol{z}_i\boldsymbol{z}_i^\top\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top \,. \tag{10}$$

If $0 \neq \boldsymbol{v} \in \mathbb{R}^p$ is an eigenvector of $\boldsymbol{\Sigma}$ with eigenvalue λ, then $\boldsymbol{V}\boldsymbol{v}$ must by Eq. (10) also be an eigenvector of $\boldsymbol{\Sigma}$ with eigenvalue λ. But since $\boldsymbol{V}$ is an arbitrary orthogonal matrix, this means that $\boldsymbol{V}\boldsymbol{v}$ is an arbitrary rotation of $\boldsymbol{v}$. From this it is easy to conclude that $\boldsymbol{\Sigma} = \lambda\boldsymbol{I}_p$, and from

$$p\lambda = \operatorname{tr}(\boldsymbol{\Sigma}) = \mathbb{E}\operatorname{tr}(\boldsymbol{z}_i\boldsymbol{z}_i^\top) = \mathbb{E}\boldsymbol{z}_i^\top\boldsymbol{z}_i = \mathbb{E}1 = 1 \,,$$

it follows that $\boldsymbol{\Sigma} = \frac{1}{p}\boldsymbol{I}_p$. Hence, (COV) is satisfied and $\boldsymbol{w}_i = \sqrt{p}\boldsymbol{z}_i$.

Step 3: Verify (FRK) for all n. By Proposition 6, it is sufficient to verify (FRK) for $n = p$. Therefore, let $n = p$. It is obvious from Proposition 6 with $\phi = \mathrm{id}$ that $\mathcal{N}(0, \boldsymbol{I}_p)$ satisfies (FRK). Hence, $\boldsymbol{X}$ almost surely has full rank. But then, since $\|\boldsymbol{x}_i\|_2 > 0$ almost surely,

$$\boldsymbol{Z} = \operatorname{diag}\left(\frac{1}{\|\boldsymbol{x}_1\|}, \ldots, \frac{1}{\|\boldsymbol{x}_n\|}\right)\boldsymbol{X}$$

almost surely has full rank as well, which proves (FRK).

Step 4.1: Computation for $n \geq p = 1$. In the underparameterized case $n \geq p = 1$, we can compute

$$\mathbb{E}_{\boldsymbol{Z}}\operatorname{tr}((\boldsymbol{Z}^+)^\top\boldsymbol{\Sigma}\boldsymbol{Z}^+) = \mathbb{E}_{\boldsymbol{Z}}\operatorname{tr}((\boldsymbol{W}^\top\boldsymbol{W})^{-1}) = \mathbb{E}_{\boldsymbol{Z}}\frac{1}{\sum_{i=1}^n w_i^2} = \mathbb{E}_{\boldsymbol{Z}}\frac{1}{n} = \frac{1}{n} \,.$$

Step 4.2: Computation for $p \geq n = 1$. In the overparameterized case $p \geq n = 1$, we can compute

$$\mathbb{E}_{\boldsymbol{Z}}\operatorname{tr}((\boldsymbol{Z}^+)^\top\boldsymbol{\Sigma}\boldsymbol{Z}^+) = \mathbb{E}_{\boldsymbol{Z}}\operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) = \mathbb{E}_{\boldsymbol{Z}}\frac{1}{\boldsymbol{w}_1^\top\boldsymbol{w}_1} = \mathbb{E}_{\boldsymbol{Z}}\frac{1}{p} = \frac{1}{p} \,,$$

where we used that since $\boldsymbol{\Sigma} = \frac{1}{p}\boldsymbol{I}_p$, $\boldsymbol{w}_1^\top\boldsymbol{w}_1 = \|\boldsymbol{w}_1\|_2^2 = \|\sqrt{p}\boldsymbol{z}_1\|_2^2 = p$.

Step 4.3: Computation for $p \geq n \geq 2$. Now, let $p \geq n \geq 2$. Since $\boldsymbol{\Sigma} = \frac{1}{p}\boldsymbol{I}_p$, we have

$$\mathbb{E}_{\boldsymbol{Z}}\operatorname{tr}((\boldsymbol{Z}^+)^\top\boldsymbol{\Sigma}\boldsymbol{Z}^+) = \mathbb{E}_{\boldsymbol{Z}}\operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1})$$

by Theorem 3. Using that the $\boldsymbol{w}_i$ are i.i.d., we obtain from Lemma G.2 that $\mathbb{E}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) = n\mathbb{E}\operatorname{dist}(\boldsymbol{w}_1, \mathcal{W}_{-1})^{-2}$, where $\mathcal{W}_{-1}$ is the space spanned by $\boldsymbol{w}_2, \ldots, \boldsymbol{w}_n$. Define the subspace $\mathcal{U}_n \coloneqq \{\boldsymbol{z} \in \mathbb{R}^p \mid z_n = z_{n+1} = \ldots = z_p = 0\}$. By (FRK), we almost surely have $\dim(\mathcal{W}_{-1}) = n-1$. Thus, there is an orthogonal matrix $\boldsymbol{U}_{-1}$ depending only on $\mathcal{W}_{-1}$ that rotates $\mathcal{W}_{-1}$ to $\mathcal{U}_n$:

$$\mathcal{U}_n = \boldsymbol{U}_{-1}\mathcal{W}_{-1} \,.$$

Because $\boldsymbol{w}_1$ is stochastically independent from $\mathcal{W}_{-1}$ and $\boldsymbol{U}_{-1}$ and its distribution is rotationally symmetric, we have the distributional equivalence (using the $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ from Step 1)

$$\begin{aligned}\operatorname{dist}(\boldsymbol{w}_1, \mathcal{W}_{-1})^2 &= \operatorname{dist}(\boldsymbol{U}_{-1}\boldsymbol{w}_1, \boldsymbol{U}_{-1}\mathcal{W}_{-1})^2 \overset{\text{distrib.}}{=} \operatorname{dist}(\boldsymbol{w}_1, \mathcal{U}_n)^2 = p(z_{1,n}^2 + \ldots + z_{1,p}^2) \\ &= p\frac{x_{1,n}^2 + \ldots + x_{1,p}^2}{\|\boldsymbol{x}_1\|_2^2} = p\frac{A}{A+B} \,,\end{aligned}$$

where $A \coloneqq x_{1,n}^2 + x_{1,n+1}^2 + \ldots + x_{1,p}^2$ has a χ^2_{p+1-n} distribution and $B \coloneqq x_{1,1}^2 + \ldots + x_{1,n-1}^2$ has a χ^2_{n-1} distribution. Hence,

$$\mathbb{E}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) = n\mathbb{E}\operatorname{dist}(\boldsymbol{w}_1, \mathcal{W}_{-1})^{-2} = \frac{n}{p}\left(1 + \mathbb{E}\frac{B}{A}\right) \,.$$

Since $p \geq n \geq 2$, $n-1$ and $p+1-n$ are positive. Since A and B are independent, $\frac{B/(n-1)}{A/(p+1-n)}$ follows a Variance-Ratio F-distribution with parameters $n-1$ and $p+1-n$, whose mean is known (see e.g. Chapter 20 in Forbes et al., 2011):

$$\mathbb{E}\frac{B}{A} = \frac{n-1}{p+1-n}\mathbb{E}\frac{B/(n-1)}{A/(p+1-n)} = \begin{cases}\infty & \text{, if } p+1-n \leq 2, \\ \frac{n-1}{p+1-n}\frac{p+1-n}{(p+1-n)-2} = \frac{n-1}{p-1-n} & \text{, if } p+1-n > 2.\end{cases} \tag{11}$$

The infinite expectation for $p+1-n \leq 2$ is not explicitly specified in Forbes et al. (2011), but it is easy to obtain from the p.d.f. of the F-distribution: The p.d.f. $f(x)$ of the $F(a,b)$-distribution for $x \geq 0$ is

$$f(x) = C_{a,b} \frac{x^{(a-2)/2}}{(1+(a/b)x)^{(a+b)/2}} = \Theta(x^{-b/2-1}) \qquad (x \to \infty)$$

for some constant $C_{a,b}$ (cf. Chapter 20 in Forbes et al., 2011), and the expected value is therefore

$$\int_0^\infty x f(x)\,\mathrm{d}x = \int_0^\infty \Theta(x^{-b/2})\,\mathrm{d}x\ ,$$

which is infinite for $p+1-n = b \leq 2$. For $n \in \{p, p-1\}$, we therefore obtain $\mathbb{E}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) = \infty$. For $n \leq p-2$, we compute

$$\mathbb{E}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) = \frac{n}{p}\left(1 + \frac{n-1}{p-1-n}\right) = \frac{n}{p} \cdot \frac{p-2}{p-1-n}\ . \qquad \square$$

In the following, we will prove Theorem 12. The formulas in this theorem have in principle already been computed by Belkin et al. (2019b). However, our proof circumvents a technical problem in their proof: Consider for example the case $p \leq n$. Belkin et al. (2019b) mention that $(\boldsymbol{W}^\top\boldsymbol{W})^{-1}$ has an inverse Wishart distribution, which for $p \leq n-2$ has expectation $\frac{1}{n-1-p}\boldsymbol{I}_p$, and then use $\mathbb{E}\operatorname{tr}((\boldsymbol{W}^\top\boldsymbol{W})^{-1}) = \operatorname{tr}(\mathbb{E}(\boldsymbol{W}^\top\boldsymbol{W})^{-1})$. However, for $p \geq n-1$, the latter expectation is not specified in common literature[5] on the inverse Wishart distribution (Mardia et al., 1979; Press, 2005; von Rosen, 1988), presumably because it is ∞ for diagonal elements but is not well-defined for off-diagonal matrix elements.

Theorem 12. *Let $P_Z = \mathcal{N}(0, \boldsymbol{I}_p)$. Then, P_Z satisfies the assumptions (MOM), (COV) and (FRK) for all n with $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Moreover, for $n, p \geq 1$,*

$$\mathbb{E}_{\boldsymbol{Z}} \operatorname{tr}((\boldsymbol{Z}^+)^\top \boldsymbol{\Sigma} \boldsymbol{Z}^+) = \begin{cases} \frac{n}{p-1-n} & \text{if } p \geq n+2, \\ \infty & \text{if } p \in \{n-1, n, n+1\}, \\ \frac{p}{n-1-p} & \text{if } p \leq n-2. \end{cases}$$

Proof. **Step 1: Assumptions.** Verifying (MOM), (COV) and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ is trivial and (FRK) for all n follows from Proposition 6 with $\boldsymbol{x} = \boldsymbol{z}$ and $\phi = \mathrm{id}$.

Step 2: Overparameterized case. For the expectation, we first follow Step 4.3 in the proof of Theorem 11 in the overparameterized case $p \geq n \geq 1$, the main difference being that instead of $\boldsymbol{w}_i = \sqrt{p}\frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_2}$, we now have $\boldsymbol{w}_i = \boldsymbol{x}_i$, which translates to the simpler equation

$$\operatorname{dist}(\boldsymbol{w}_1, \mathcal{W}_{-1})^2 \overset{\text{distrib.}}{=} A$$

with $A \sim \chi^2_{p+1-n}$. Let $B \sim \chi^2_1$ be independent of A, then we can compute similar to Eq. (11)

$$\begin{aligned} \mathbb{E}\operatorname{tr}((\boldsymbol{W}\boldsymbol{W}^\top)^{-1}) &= n\mathbb{E}\operatorname{dist}(\boldsymbol{w}_1, \mathcal{W}_{-1})^{-2} = n\mathbb{E}\frac{1}{A} = n\,(\mathbb{E}B)\left(\mathbb{E}\frac{1}{A}\right) = n\mathbb{E}\frac{B}{A} \\ &= \frac{n}{p+1-n}\mathbb{E}\frac{B/1}{A/(p+1-n)} \\ &= \begin{cases} \infty & \text{if } p+1-n \leq 2, \\ \frac{n}{p+1-n}\frac{p+1-n}{(p+1-n)-2} = \frac{n}{p-1-n} & \text{if } p+1-n > 2. \end{cases} \end{aligned}$$

This proves the over-parameterized case.

Step 3: Underparameterized case. Since the rows $\boldsymbol{w}_i$ of $\boldsymbol{W} \in \mathbb{R}^{n\times p}$ are independent and follow a $\mathcal{N}(0, \boldsymbol{I}_p)$ distribution, the rows of $\boldsymbol{W}^\top \in \mathbb{R}^{p\times n}$ are independent and follow a $\mathcal{N}(0, \boldsymbol{I}_n)$ distribution. Therefore, the underparameterized case $p \leq n$ follows from the overparameterized case $n \leq p$ by switching the roles of n and p. $\square$

Remark H.1. An alternative (and presumably similar) way to prove Theorem 12 is to use that the diagonal elements of a matrix with an inverse Wishart distribution follow an inverse Gamma distribution as specified in Example 5.2.2 in Press (2005). ◀

[5] Belkin et al. (2019b) do not cite any source on the inverse Wishart distribution.

I PROOFS FOR SECTION 5

First, we prove a statement about conditional variances.

Lemma I.1. *In the setting of Theorem 3, we have*

$$\mathrm{Var}(y|\boldsymbol{z}) = \mathbb{E}(\mathrm{Var}(y|\boldsymbol{x})|\boldsymbol{z}) + \mathrm{Var}(\mathbb{E}(y|\boldsymbol{x})|\boldsymbol{z}) \ .$$

Hence, if $\mathrm{Var}(y|\boldsymbol{x}) \geq \sigma^2$ *almost surely over* $\boldsymbol{x}$*, then* $\mathrm{Var}(y|\boldsymbol{z}) \geq \sigma^2$ *almost surely over* $\boldsymbol{z}$*. The converse holds, for example, if* ϕ *is injective.*

Proof. For properties of conditional expectations, we refer to the literature, e.g. Chapter 4.1 in Durrett (2019). Since $\boldsymbol{z} = \phi(\boldsymbol{x})$ is a function of $\boldsymbol{x}$, we have $\mathbb{E}[\cdot|\boldsymbol{z}] = \mathbb{E}[\mathbb{E}(\cdot|\boldsymbol{x})|\boldsymbol{z}]$. Thus,

$$\begin{aligned}
\mathrm{Var}(y|\boldsymbol{z}) &= \mathbb{E}\left[(y - \mathbb{E}(y|\boldsymbol{z}))^2 \middle| \boldsymbol{z}\right] \\
&= \mathbb{E}\left[\left((y - \mathbb{E}(y|\boldsymbol{x})) + (\mathbb{E}(y|\boldsymbol{x}) - \mathbb{E}(\mathbb{E}(y|\boldsymbol{x})|\boldsymbol{z}))\right)^2 \middle| \boldsymbol{z}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left((y - \mathbb{E}(y|\boldsymbol{x}))^2|\boldsymbol{x}\right)\middle|\boldsymbol{z}\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left((y - \mathbb{E}(y|\boldsymbol{x}))(\mathbb{E}(y|\boldsymbol{x}) - \mathbb{E}(\mathbb{E}(y|\boldsymbol{x})|\boldsymbol{z}))|\boldsymbol{x}\right)|\boldsymbol{z}\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left((\mathbb{E}(y|\boldsymbol{x}) - \mathbb{E}(\mathbb{E}(y|\boldsymbol{x})|\boldsymbol{z}))^2|\boldsymbol{x}\right)\middle|\boldsymbol{z}\right] \ .
\end{aligned}$$

For the first term, we have $\mathbb{E}\left((y - \mathbb{E}(y|\boldsymbol{x}))^2|\boldsymbol{x}\right) = \mathrm{Var}(y|\boldsymbol{x})$ by definition. The second term is zero: Since $\mathbb{E}(y|\boldsymbol{x}) - \mathbb{E}(\mathbb{E}(y|\boldsymbol{x})|\boldsymbol{z})$ is already a function of $\boldsymbol{x}$, we have

$$\begin{aligned}
\mathbb{E}\left((y - \mathbb{E}(y|\boldsymbol{x}))(\mathbb{E}(y|\boldsymbol{x}) - \mathbb{E}(\mathbb{E}(y|\boldsymbol{x})|\boldsymbol{z}))|\boldsymbol{x}\right) &= \mathbb{E}\left((y - \mathbb{E}(y|\boldsymbol{x}))|\boldsymbol{x}\right)(\mathbb{E}(y|\boldsymbol{x}) - \mathbb{E}(\mathbb{E}(y|\boldsymbol{x})|\boldsymbol{z})) \\
&= (\mathbb{E}(y|\boldsymbol{x}) - \mathbb{E}(y|\boldsymbol{x}))(\mathbb{E}(y|\boldsymbol{x}) - \mathbb{E}(\mathbb{E}(y|\boldsymbol{x})|\boldsymbol{z})) \\
&= 0 \ .
\end{aligned}$$

Finally, the third term is by definition equal to $\mathrm{Var}(\mathbb{E}(y|\boldsymbol{x})|\boldsymbol{z})$. Therefore,

$$\mathrm{Var}(y|\boldsymbol{z}) = \mathbb{E}(\mathrm{Var}(y|\boldsymbol{x})|\boldsymbol{z}) + \mathrm{Var}(\mathbb{E}(y|\boldsymbol{x})|\boldsymbol{z}) \ .$$

If ϕ is injective, then $\boldsymbol{x}$ is also a function of $\boldsymbol{z}$ and we obtain $\mathrm{Var}(y|\boldsymbol{z}) = \mathrm{Var}(y|\boldsymbol{x})$. □

Our main ingredient for analyzing analytic activation functions will be the univariate and multivariate identity theorems.

Theorem I.2 (Identity theorem for univariate real analytic functions)**.** *Let* $f : \mathbb{R} \to \mathbb{R}$ *be analytic. If the set* $\mathcal{N}(f) := \{z \in \mathbb{R} \mid f(z) = 0\}$ *has an accumulation point, i.e. a point* z *such that* $\mathcal{N}(f) \cap ((z - \varepsilon, z + \varepsilon) \setminus \{z\})$ *is nonempty for all* $\varepsilon > 0$*, then* $f \equiv 0$*.*

Proof. See e.g. Corollary 1.2.7 in Krantz & Parks (2002). □

The following multivariate version of Theorem I.2 is briefly mentioned at the beginning of Section 4.1 in Krantz & Parks (2002).

Lemma I.3 (Multivariate version of the identity theorem)**.** *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be analytic. If* $\mathcal{N}(f) := \{\boldsymbol{x} \in \mathbb{R}^d \mid f(\boldsymbol{x}) = 0\}$ *is not a (Lebesgue-) null set, then* $f \equiv 0$*.*

Proof. Let λ^d be the Lebesgue measure on $\mathbb{R}^d$ and let $\lambda := \lambda^1$. We prove the statement by induction on $d \geq 1$. For $d = 1$, if $\lambda(\mathcal{N}(f)) > 0$, then there exists a finite interval $[a, b]$ with $\lambda(\mathcal{N}(f) \cap [a, b]) > 0$. Then, $\mathcal{N}(f) \cap [a, b]$ is uncountable and hence $\mathcal{N}(f)$ has an accumulation point in $[a, b]$. By Theorem I.2, $f \equiv 0$.

Now, let the statement hold for $d - 1 \geq 1$ and assume $\lambda(\mathcal{N}(f)) > 0$. For $a \in \mathbb{R}$, define the functions $f_a : \mathbb{R}^{d-1} \to \mathbb{R}, f_a(\boldsymbol{x}) = f(a, \boldsymbol{x})$. Then,

$$\begin{aligned}
0 < \lambda(\mathcal{N}(f)) &= \int_{\mathbb{R}^p} \mathbb{1}_{\mathcal{N}(f)} \,\mathrm{d}\lambda^p = \int_{\mathbb{R}} \int_{\mathbb{R}^{p-1}} \mathbb{1}_{\mathcal{N}(f)}(a, \boldsymbol{x}) \,\mathrm{d}\lambda^{p-1}(\boldsymbol{x}) \,\mathrm{d}\lambda(a) \\
&= \int_{\mathbb{R}} \lambda^{p-1}(\mathcal{N}(f_a)) \,\mathrm{d}a \ .
\end{aligned}$$

It follows that the set $U := \{x \in \mathbb{R} \mid \lambda^{p-1}(\mathcal{N}(f_x)) > 0\}$ satisfies $\lambda(U) > 0$. By induction, for all $x \in U$, we have $f_x \equiv 0$. Then, for all $\boldsymbol{x} \in \mathbb{R}^{d-1}$, we can conclude that the function $f_{\boldsymbol{x}} : \mathbb{R} \to \mathbb{R}, a \mapsto f(a, \boldsymbol{x})$ satisfies $\mathcal{N}(f_{\boldsymbol{x}}) \supseteq U$ and therefore $\lambda(\mathcal{N}(f_{\boldsymbol{x}})) \geq \lambda(U) > 0$. Using the case $d = 1$ again, it follows that $f_{\boldsymbol{x}} \equiv 0$ and therefore $f(a, \boldsymbol{x}) = 0$ for all $a \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^{d-1}$. □

Proposition 6 (Characterization of (COV) and (FRK))**.** *Consider the setting of Theorem 3 and let* $\mathrm{FRK}(n)$ *be the statement that (FRK) holds for* n*. Then,*

(i) Let $n \geq 1$*. Then,* $\mathrm{FRK}(n)$ *iff* $P(\boldsymbol{z} \in U) = 0$ *for all linear subspaces* $U \subseteq \mathbb{R}^p$ *of dimension* $\min\{n, p\} - 1$*.*

(ii) Let (MOM) hold. Then, (COV) holds iff $P(\boldsymbol{z} \in U) < 1$ *for all linear subspaces* $U \subseteq \mathbb{R}^p$ *of dimension* $p - 1$*.*

Assuming that (MOM) holds such that (COV) is well-defined, consider the following statements:

(a) $\mathrm{FRK}(p)$ *holds.*

(b) $\mathrm{FRK}(n)$ *holds for all* $n \geq 1$*.*

(c) (COV) holds.

(d) There exists a fixed deterministic matrix $\boldsymbol{X}_{\mathrm{const}} \in \mathbb{R}^{p \times d}$ *such that* $\det(\phi(\boldsymbol{X}_{\mathrm{const}})) \neq 0$*.*

We have (a) $\Leftrightarrow$ *(b)* $\Rightarrow$ *(c)* $\Rightarrow$ *(d). Furthermore, if* $\boldsymbol{x} \in \mathbb{R}^d$ *has a Lebesgue density and* ϕ *is analytic, then (a) – (d) are equivalent.*

Proof. **Step 1: Prove (i) and (ii).**

(i) Denote the n (stochastically independent) rows of $\boldsymbol{Z}$ by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. First, assume $P(\boldsymbol{z} \in U) > 0$ for some subspace U of dimension $\min\{n, p\} - 1$. Then,

$$P\Big(\operatorname{rank}(\boldsymbol{Z}) \leq \min\{n, p\} - 1\Big) \geq P(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in U) = (P(\boldsymbol{z} \in U))^n > 0 \,.$$

For the converse, it suffices to consider the case $n \leq p$ since if $n > p$ and if an arbitrary $p \times p$ submatrix of $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ almost surely has full rank, then $\boldsymbol{Z}$ also almost surely has full rank. We prove the statement for $n \leq p$ by induction on n. For $n = 1$, the claim is trivial. Thus, let $n > 1$ and let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ be the (stochastically independent) rows of $\boldsymbol{Z}$. Assume $P(\boldsymbol{z} \in U) = 0$ for all linear subspaces $U \subseteq \mathbb{R}^p$ of dimension $n - 1$. Then, we also have $P(\boldsymbol{z} \in U) = 0$ for all linear subspaces $U \subseteq \mathbb{R}^p$ of dimension $n - 2$ and by the induction hypothesis, we obtain that $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n-1}$ are almost surely linearly independent. Hence, almost surely, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are linearly independent iff $\boldsymbol{z}_n \notin U_{n-1} := \operatorname{Span}\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n-1}\}$. Then, since $\boldsymbol{z}_n$ and U_{n-1} are stochastically independent,

$$\begin{aligned} P(\boldsymbol{Z} \in \mathbb{R}^{n \times p} \text{ has full rank}) &= P(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \text{ are linearly independent}) \\ &= P(\boldsymbol{z}_n \notin U_{n-1}) = \iint \mathbb{1}_{U_{n-1}^c}(\boldsymbol{z}_n) \,\mathrm{d}P_Z(\boldsymbol{z}_n) \,\mathrm{d}P(U_{n-1}) \\ &= \int P_Z(\boldsymbol{z}_n \notin U_{n-1}) \,\mathrm{d}P(U_{n-1}) = \int 1 \,\mathrm{d}P(U_{n-1}) = 1 \,. \end{aligned}$$

(ii) If (COV) does not hold, there exists a vector $0 \neq \boldsymbol{v} \in \mathbb{R}^p$ with $\boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v} = 0$, hence

$$0 = \boldsymbol{v}^\top \left(\mathbb{E} \boldsymbol{z} \boldsymbol{z}^\top\right) \boldsymbol{v} = \mathbb{E}(\boldsymbol{v}^\top \boldsymbol{z})^2 \,,$$

and therefore $\boldsymbol{z}$ is almost surely orthogonal to $\boldsymbol{v}$. If U is the orthogonal complement of $\operatorname{Span}\{\boldsymbol{v}\}$ in $\mathbb{R}^p$, then $P(\boldsymbol{z} \in U) = 1$.
For the converse, if there exists a $(p-1)$-dimensional subspace U with $P(\boldsymbol{z} \in U) = 1$, then we can again take a vector $0 \neq \boldsymbol{v} \in \mathbb{R}^p$ that is orthogonal to U, reverse the above computation and obtain $\boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v} = 0$, hence (COV) does not hold.

Step 2: (a) $\Leftrightarrow$ (b). The implication (b) $\Rightarrow$ (a) is trivial and the implication (a) $\Rightarrow$ (b) follows immediately from (i).

Step 3: (b) $\Rightarrow$ (c). This also follows from (i) and (ii).

Step 4: (c) $\Rightarrow$ (d). We will prove by induction on n that for all $n \in \{0, \ldots, p\}$, there exist $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ such that $\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_n)$ are linearly independent. For $n = 0$, the statement is trivial. Now assume that the statement holds for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}$, where $0 \leq n - 1 \leq p - 1$. Since (COV) holds, by (ii) the subspace $U := \operatorname{Span}\{\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_{n-1})\}$ satisfies $P(\phi(\boldsymbol{x}) \in U) < 1$, hence there is $\boldsymbol{x}_n$ such that $\boldsymbol{x}_n \notin U$ and for this choice, $\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_n)$ are linearly independent.

Finally, the statement for $n = p$ yields the existence of $\boldsymbol{X} \in \mathbb{R}^{p \times d}$ such that $\phi(\boldsymbol{X})$ has linearly independent rows, which implies $\det(\phi(\boldsymbol{X})) \neq 0$.

Step 5: Analytic feature map. Assume that ϕ is analytic, $\boldsymbol{x}$ has a Lebesgue density and (d) holds. Let $n = p$. For $\boldsymbol{X}_{\text{const}} \in \mathbb{R}^{p\times d}$, consider the analytic function $f(\boldsymbol{X}_{\text{const}}) \coloneqq \det(\phi(\boldsymbol{X}_{\text{const}}))$. (The determinant is analytic since it is a polynomial in the matrix entries.) Since (d) holds, Lemma I.3 shows that $f(\boldsymbol{X}_{\text{const}}) \neq 0$ for (Lebesgue-) almost all $\boldsymbol{X}_{\text{const}}$. Since $\boldsymbol{x}$ has a Lebesgue density and $\boldsymbol{X}$ has independent $\boldsymbol{x}$-distributed rows, $\boldsymbol{X}$ has a Lebesgue density. Therefore, $f(\boldsymbol{X}) \neq 0$ almost surely over $\boldsymbol{X}$, hence (a) holds. □

Proposition 8 (Polynomial kernel)**.** *Let $m, d \geq 1$ and $c > 0$. For $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{R}^d$, define the polynomial kernel $k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \coloneqq (\boldsymbol{x}^\top \tilde{\boldsymbol{x}} + c)^m$. Then, there exists a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^p$, $p \coloneqq \binom{m+d}{m}$, such that*

(a) $k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \phi(\boldsymbol{x})^\top \phi(\tilde{\boldsymbol{x}})$ for all $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{R}^d$, and
(b) if $\boldsymbol{x} \in \mathbb{R}^d$ has a Lebesgue density and we use $\boldsymbol{z} = \phi(\boldsymbol{x})$, then (FRK) is satisfied for all n.

Proof. Let $\mathcal{M} \coloneqq \{(m_1, \ldots, m_{d+1}) \in \mathbb{N}_0^{d+1} \mid m_1 + \ldots + m_{d+1} = m\}$ and for $\boldsymbol{m} = (m_1, \ldots, m_{d+1}) \in \mathcal{M}$, let

$$C(\boldsymbol{m}) \coloneqq \binom{m}{m_1 \quad \ldots \quad m_{d+1}}$$

be the corresponding multinomial coefficient. Then, $|\mathcal{M}| = \binom{m+d}{d} = p$. Define the feature map $\phi : \mathbb{R}^d \to \mathbb{R}^p$ by

$$\phi(\boldsymbol{x}) \coloneqq \left(\sqrt{C(\boldsymbol{m})} z_1^{m_1} \cdots z_d^{m_d} \cdot (\sqrt{c})^{m_{d+1}}\right)_{\boldsymbol{m}\in\mathcal{M}} .$$

(a) We have

$$\begin{aligned}\phi(\boldsymbol{x})^\top \phi(\tilde{\boldsymbol{x}}) &= \sum_{\boldsymbol{m}\in\mathcal{M}} C(\boldsymbol{m})(x_1\tilde{x}_1)^{m_1} \cdots (x_d\tilde{x}_d)^{m_d} \cdot c^{m_{d+1}} \\ &= (x_1\tilde{x}_1 + \ldots + x_d\tilde{x}_d + c)^m = k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) .\end{aligned}$$

(b) Assume that $\boldsymbol{x}$ has a Lebesgue density. Let U be an arbitrary $(p-1)$-dimensional linear subspace of $\mathbb{R}^p$. Then, there exists $0 \neq \boldsymbol{v} \in \mathbb{R}^p$ such that $U = (\operatorname{Span}\{\boldsymbol{v}\})^\perp$. Since the monomials $x_1^{m_1} \cdots x_d^{m_d}$ for $\boldsymbol{m} \in \mathcal{M}$ are all distinct, the polynomial

$$\boldsymbol{v}^\top \phi(\boldsymbol{x}) = \sum_{\boldsymbol{m}\in\mathcal{M}} \left(v_{\boldsymbol{m}} \sqrt{C(\boldsymbol{m}) c^{m_{d+1}}} x_1^{m_1} \cdots x_d^{m_d}\right)$$

is not the zero polynomial. By the identity theorem (Lemma I.3), since $\boldsymbol{x}$ has a Lebesgue density and since polynomials are analytic,

$$P(\phi(\boldsymbol{x}) \in U) = P(\boldsymbol{v}^\top \phi(\boldsymbol{x}) = 0) = 0 .$$

Hence, Proposition 6 shows that (FRK) is satisfied for $n = p$ and hence for all n. □

We want to remark at this point that the proof strategy of Proposition 8, where the identity theorem is applied to the functions $\boldsymbol{v}^\top \phi(\boldsymbol{x})$ for all $0 \neq \boldsymbol{v} \in \mathbb{R}^p$, does not work for random feature maps: The statements

- For all $0 \neq \boldsymbol{v} \in \mathbb{R}^p$ for almost all $\boldsymbol{x}$ for almost all $\boldsymbol{\theta}$, $\boldsymbol{v}^\top \phi_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq 0$
- For almost all $\boldsymbol{\theta}$ for all $0 \neq \boldsymbol{v} \in \mathbb{R}^p$ for almost all $\boldsymbol{x}$, $\boldsymbol{v}^\top \phi_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq 0$

are not equivalent since in the first statement, the null set for $\boldsymbol{\theta}$ may depend on $\boldsymbol{v}$, and the union of the null sets for all $\boldsymbol{v}$ is an uncountable union. Perhaps the simplest counterexample is $n = p = 2$, $\boldsymbol{\theta} \sim \mathcal{N}(0, \boldsymbol{I}_2)$ and $\phi_{\boldsymbol{\theta}}(\boldsymbol{x}) \coloneqq \boldsymbol{\theta}$, which satisfies the first but not the second statement. For our rescue, we can replace the uncountable analytic function family $(\boldsymbol{x} \mapsto \boldsymbol{v}^\top \phi_{\boldsymbol{\theta}}(\boldsymbol{x}))_{\boldsymbol{v}\in\mathbb{R}^p\setminus\{0\}}$ by the single analytic function $(\boldsymbol{\theta}, \boldsymbol{X}) \mapsto \det(\phi_{\boldsymbol{\theta}}(\boldsymbol{X}))$:

Proposition 9 (Random feature maps)**.** *Consider feature maps $\phi_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^p$ with (random) parameter $\boldsymbol{\theta} \in \mathbb{R}^q$. Suppose the map $(\boldsymbol{\theta}, \boldsymbol{x}) \mapsto \phi_{\boldsymbol{\theta}}(\boldsymbol{x})$ is analytic and that $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are independent and have Lebesgue densities. If there exist fixed $\boldsymbol{\theta}_{\text{const}} \in \mathbb{R}^q, \boldsymbol{X}_{\text{const}} \in \mathbb{R}^{p\times d}$ with $\det(\phi_{\boldsymbol{\theta}_{\text{const}}}(\boldsymbol{X}_{\text{const}})) \neq 0$, then almost surely over $\boldsymbol{\theta}$, (FRK) holds for all n for $\boldsymbol{z} = \phi_{\boldsymbol{\theta}}(\boldsymbol{x})$.*

Proof. Consider the analytic map $(\boldsymbol{\theta}, \boldsymbol{X}) \mapsto \det(\phi_{\boldsymbol{\theta}}(\boldsymbol{X}))$. Suppose there exist $\boldsymbol{\theta}_{\text{const}} \in \mathbb{R}^q$ and $\boldsymbol{X}_{\text{const}} \in \mathbb{R}^{p\times d}$ with $\det(\phi_{\boldsymbol{\theta}_{\text{const}}}(\boldsymbol{X}_{\text{const}})) \neq 0$. Then, Lemma I.3 tells us that $\det(\tilde{\phi}(\boldsymbol{\theta}, \boldsymbol{X})) \neq 0$ for almost all $(\boldsymbol{\theta}, \boldsymbol{X})$. This implies that for almost all $\boldsymbol{\theta}$, we have for almost all $\boldsymbol{X}$ that $\det(\phi_{\boldsymbol{\theta}}(\boldsymbol{X})) \neq 0$. Since by assumption, $\boldsymbol{\theta}$ has a density, this implies that almost surely over $\boldsymbol{\theta}$, there exists $\boldsymbol{X}$ such that $\det(\phi_{\boldsymbol{\theta}}(\boldsymbol{X})) \neq 0$. Since all $\phi_{\boldsymbol{\theta}}$ are analytic, the claim now follows using (d) $\Rightarrow$ (b) from Proposition 6. (The proof of (d) $\Rightarrow$ (a) $\Leftrightarrow$ (b) in Proposition 6 does not require (MOM).) □

The next proposition will help to analyze feature maps given by random neural networks.

Lemma I.4 (Shallow network with $d = 1$)**.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be analytic. Then, $\det(\sigma(\boldsymbol{A})) = 0$ for all rank-1 matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ if and only if σ is a polynomial with at most $n-1$ nonzero coefficients. Here, σ is applied element-wise to the entries of $\boldsymbol{A}$*

Proof. Let $\sigma(x) = \sum_{k\in I} a_k x^k$ for some index set $I \subseteq \mathbb{N}_0$ with $|I| \leq n-1$. Then, for some arbitrary rank-1 matrix $\boldsymbol{A} = \boldsymbol{y}\boldsymbol{x}^\top$, $\boldsymbol{y}, \boldsymbol{x} \in \mathbb{R}^n$, the matrix

$$\sigma(\boldsymbol{A}) = \sum_{k\in I} a_k (y_i^k x_j^k)_{i,j\in[n]}$$

is a sum of $|I|$ rank-1 matrices and therefore has rank $\leq |I| \leq n-1$, which implies $\det(\sigma(\boldsymbol{A})) = 0$.

For the converse, write $\sigma(x) = \sum_{k=0}^\infty a_k x^k$ in a neighborhood of 0. For two arbitrary vectors $\boldsymbol{y}, \boldsymbol{x} \in \mathbb{R}^p$ in a suitable neighborhood of 0, we then obtain (using the Leibniz formula involving the permutation group S_n on $[n]$)

$$\begin{aligned}
\det(\sigma(\boldsymbol{y}\boldsymbol{x}^\top)) &= \sum_{\pi\in S_n} \operatorname{sgn}(\pi) \prod_{i=1}^n \sum_{k=0}^\infty a_k (y_i x_{\pi(i)})^k \\
&= \sum_{k_1,\ldots,k_n\geq 0} \sum_{\pi\in S_n} \left(\operatorname{sgn}(\pi) \prod_{i=1}^n a_{k_i}\right) \left(\prod_{i=1}^n y_i^{k_i} x_{\pi(i)}^{k_i}\right) \\
&= \sum_{k_1,\ldots,k_n\geq 0} \sum_{l_1,\ldots,l_n\geq 0} c_{k_1,\ldots,k_n,l_1,\ldots,l_n} y_1^{k_1}\cdots y_n^{k_n} x_1^{l_1}\cdots x_n^{l_n}
\end{aligned} \tag{12}$$

for certain coefficients $c_{k_1,\ldots,k_n,l_1,\ldots,l_n}$. Assuming $\det(\sigma(\boldsymbol{y}\boldsymbol{x}^\top)) = 0$ for all such $\boldsymbol{y}, \boldsymbol{x}$, we obtain that all of these coefficients must be zero, since they are determined by the derivatives of $\det(\sigma(\boldsymbol{y}\boldsymbol{x}^\top))$. Now, consider n *distinct* indices $k_1, \ldots, k_n \geq 0$. Then, the monomial

$$y_1^{k_1}\cdots y_n^{k_n} x_1^{k_1}\cdots x_n^{k_n}$$

only occurs in (12) for these specific values of $k_1, \ldots, k_n$ and, because all k_i are distinct, only for $\pi = \text{id}$. Thus,

$$0 = c_{k_1,\ldots,k_n,k_1,\ldots,k_n} = a_{k_1}\cdots a_{k_n} \ .$$

Therefore, each product of n distinct coefficients $a_{k_1}, \ldots, a_{k_n}$ is zero, hence at most $n-1$ coefficients can be nonzero. Let I be the set of nonzero coefficients. Since $|I| \leq n-1$, I is finite and the polynomial $f(x) := \sum_{k\in I} a_k x^k$ can be defined on all of $\mathbb{R}$. Since polynomials are analytic, the difference $g = \sigma - f$ is also analytic. For all x in the neighborhood of 0 where the expansion $\sigma(x) = \sum_{k=0}^\infty a_k x^k$ holds, we have $g(x) = 0$ by definition. But then, by the identity theorem (Theorem I.2), g is the zero function and hence $\sigma(x) = f(x) = \sum_{k\in I} a_k x^k$ for all $x \in \mathbb{R}$. □

Theorem 10 (Random neural networks)**.** *Let $d, p, L \geq 1$ and let $\sigma : \mathbb{R} \to \mathbb{R}$ be analytic. Consider layer sizes $d_0 = d, d_1, \ldots, d_{L-1} \geq p$ and $d_L = p$. For weights $\boldsymbol{\theta} := (\boldsymbol{W}^{(0)}, \ldots, \boldsymbol{W}^{(L-1)})$ with $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d_{l+1}\times d_l}$, we define the feature map*

$$\phi_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^p, \boldsymbol{x} \mapsto \sigma(\boldsymbol{W}^{(L-1)}\sigma(\ldots\sigma(\boldsymbol{W}^{(0)}\boldsymbol{x}))) \ ,$$

where σ is applied element-wise.

If $\boldsymbol{\theta}$ and $\boldsymbol{x}$ have Lebesgue densities and σ is not a polynomial with less than p nonzero coefficients, (FRK) holds for all n and almost surely over $\boldsymbol{\theta}$. This statement also holds if $\phi_{\boldsymbol{\theta}}$ includes biases that are initialized from a probability distribution with a density.

Proof. By Proposition 9, we only need to show that there exist fixed deterministic $(\boldsymbol{\theta}, \boldsymbol{X})$ with $\det(\phi_{\boldsymbol{\theta}}(\boldsymbol{X})) \neq 0$. It suffices to do this for the case without biases, since we can otherwise set all biases to zero. Moreover, we can without loss of generality assume that $d_1, \ldots, d_{L-1} = p$, since we can otherwise set the incoming and outgoing weights of the additional neurons to zero. We recursively define $\boldsymbol{X}^{(0)} := \boldsymbol{X}$ and

$$\left(\boldsymbol{X}^{(l+1)}\right)^\top := \sigma\left(\boldsymbol{W}^{(l)}\left(\boldsymbol{X}^{(l)}\right)^\top\right)$$

such that $\phi_{\boldsymbol{\theta}}(\boldsymbol{X}) = \boldsymbol{X}^{(L)}$. By Lemma I.4, there exist $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$ such that $\sigma(\boldsymbol{u}\boldsymbol{v}^\top)$ is invertible. We can obviously choose $\boldsymbol{W}^{(0)}, \boldsymbol{X}^{(0)}$ such that

$$\boldsymbol{W}^{(0)}\left(\boldsymbol{X}^{(0)}\right)^\top = \boldsymbol{u}\boldsymbol{v}^\top ,$$

which makes $\boldsymbol{X}^{(1)}$ invertible. Moreover, if $\boldsymbol{X}^{(l)}$ is invertible for $l \geq 1$, we can choose

$$\boldsymbol{W}^{(l)} := \boldsymbol{u}\boldsymbol{v}^\top\left(\left(\boldsymbol{X}^{(l)}\right)^\top\right)^{-1} .$$

With this choice, $\boldsymbol{X}^{(l+1)}$ is also invertible. By induction, this shows that there exist $\boldsymbol{\theta}, \boldsymbol{X}$ such that $\phi_{\boldsymbol{\theta}}(\boldsymbol{X}) = \boldsymbol{X}^{(L)}$ is invertible. □

J Random Fourier Features

In a celebrated paper, Rahimi & Recht (2008) propose to approximate a shift-invariant positive definite kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \overline{k}(\boldsymbol{x} - \boldsymbol{x}')$ with a potentially infinite-dimensional feature map by a random finite-dimensional feature map, yielding so-called *random Fourier features*. If $\overline{k}$ is (up to scaling) the Fourier transform of a probability distribution P_k on $\mathbb{R}^d$, two versions of random Fourier features are proposed:

(1) One version uses $\phi_{\boldsymbol{W},\boldsymbol{b}}(\boldsymbol{x}) = \sqrt{2}\cos(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$, where the rows of $\boldsymbol{W} \in \mathbb{R}^{p \times d}$ are independently sampled from P_k and the entries of $\boldsymbol{b} \in \mathbb{R}^p$ are independently sampled from the uniform distribution on $[0, 2\pi]$. This feature map is covered by Theorem 10 and hence, if P_k and $\boldsymbol{x}$ have Lebesgue densities, (FRK) is satisfied for all n. For example, if k is a Gaussian kernel, P_k is a Gaussian distribution and therefore has a Lebesgue density.

(2) The other version uses

$$\phi_{\boldsymbol{W}}(\boldsymbol{x}) = \begin{pmatrix} \sin(\boldsymbol{W}\boldsymbol{x}) \\ \cos(\boldsymbol{W}\boldsymbol{x}) \end{pmatrix}$$

with the same distribution over $\boldsymbol{W}$. It is not covered by Theorem 10 because of the different "activation functions" and the "weight sharing" between these activation functions. In the following proposition, we show that the proof of Theorem 10 can be adjusted to this setting and the conclusions still hold.

Proposition J.1. *For $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{W} \in \mathbb{R}^{q \times d}$ and $p := 2q$, define*

$$\phi_{\boldsymbol{W}}(\boldsymbol{x}) := \begin{pmatrix} \sin(\boldsymbol{W}\boldsymbol{x}) \\ \cos(\boldsymbol{W}\boldsymbol{x}) \end{pmatrix} \in \mathbb{R}^p .$$

If the distributions over $\boldsymbol{x}$ and $\boldsymbol{W}$ have a Lebesgue density, then for $\boldsymbol{z} = \phi(\boldsymbol{x})$, (FRK) holds for all n.

Proof. **Step 1: Reduction to rank-1 case.** We will proceed similarly to the proofs of Lemma I.4 and Theorem 10 here. By Proposition 9, we only need to show that there exist $\boldsymbol{W} \in \mathbb{R}^{q \times d}$ and $\boldsymbol{X} \in \mathbb{R}^{d \times 2q}$ with $\det(\phi_{\boldsymbol{W}}(\boldsymbol{X})) \neq 0$. We will show that there exist vectors $\boldsymbol{w} \in \mathbb{R}^q, \boldsymbol{x} \in \mathbb{R}^{2q}$ such that

$$\begin{pmatrix} \sin(\boldsymbol{w}\boldsymbol{x}^\top) \\ \cos(\boldsymbol{w}\boldsymbol{x}^\top) \end{pmatrix} \in \mathbb{R}^{p \times p}$$

is invertible. Once this is proven, it is clear that we can choose $\boldsymbol{W}$ and $\boldsymbol{X}$ such that $\boldsymbol{W}\boldsymbol{X}^\top = \boldsymbol{w}\boldsymbol{x}^\top$.

Step 2: Proof of rank-1 case. Take the power series of $\sin$ and $\cos$ as

$$\sin(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} =: \sum_{k=0}^{\infty} a_k x^k$$
$$\cos(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} =: \sum_{k=0}^{\infty} b_k x^k \,.$$

Then, similar to the proof of Lemma I.4, we can compute

$$\det \begin{pmatrix} \sin(\boldsymbol{w}\boldsymbol{x}^\top) \\ \cos(\boldsymbol{w}\boldsymbol{x}^\top) \end{pmatrix} = \sum_{k_1,\ldots,k_{2q}\geq 0} \sum_{\pi \in S_{2q}} \operatorname{sgn}(\pi) a_{k_1} \cdots a_{k_q} b_{k_{q+1}} \cdots b_{k_{2q}} w_1^{k_1+k_{q+1}} \cdots w_q^{k_q+k_{2q}} x_{\pi(1)}^{k_1} \cdots x_{\pi(2q)}^{k_{2q}} \,. \tag{13}$$

Now, choose $k_1 = 3^1, \ldots, k_q = 3^q, k_{q+1} = 3^{q+1}+1, \ldots, k_{2q} = 3^{2q}+1$. We want to compute the coefficient of the monomial

$$w_1^{k_1+k_{q+1}} \cdots w_q^{k_q+k_{2q}} x_1^{k_1} \cdots x_{2q}^{k_{2q}} \tag{14}$$

in the expansion in Eq. (13). For $(k'_1, \ldots, k'_{2q}) \in \mathbb{N}_0^{2q}$ and $\pi \in S_{2q}$, the corresponding term in (13) is a multiple of the monomial

$$w_1^{k'_1+k'_{q+1}} \cdots w_q^{k'_q+k'_{2q}} x_{\pi(1)}^{k'_1} \cdots x_{\pi(2q)}^{k'_{2q}} \,. \tag{15}$$

The monomials in (14) and (15) are equal if and only if

$$\begin{aligned} k_1 + k_{q+1} &= k'_1 + k'_{q+1} \\ &\vdots \\ k_q + k_{2q} &= k'_q + k'_{2q} \\ k_{\pi(1)} &= k'_1 \\ &\vdots \\ k_{\pi(2q)} &= k'_{2q} \,. \end{aligned}$$

Assume that these identities hold. Then, from the latter identities, we can conclude $\{k'_1, \ldots, k'_{2q}\} = \{k_1, \ldots, k_{2q}\}$. By construction of the set $\{k_1, \ldots, k_{2q}\}$, $k_i + k_j = k_{i'} + k_{j'}$ implies $\{i, j\} = \{i', j'\}$.[6] Hence, we must have $\{k_i, k_{q+i}\} = \{k'_i, k'_{q+i}\}$ for all $i \in [q]$. It follows that there are 2^q choices for $(k'_1, \ldots, k'_{2q}) \in \mathbb{N}_0^{2q}$ leading to the monomial (14), by choosing either

(a) $k'_i = k_i$ and $k'_{q+i} = k_{q+i}$, or
(b) $k'_i = k_{q+i}$ and $k'_{q+i} = k_i$

for all $i \in [q]$. We know that k_i is odd iff $i \leq q$, $a_k \neq 0$ iff k is odd and $b_k \neq 0$ iff k is even. Hence, choosing option (a) for an index $i \in [q]$ yields $a_{k'_i}, b_{k'_{q+i}} \neq 0$, while choosing option (b) yields $a_{k'_i} = b_{k'_{q+i}} = 0$. Therefore, the coefficient

$$\operatorname{sgn}(\pi) a_{k'_1} \cdots a_{k'_q} b_{k'_{q+1}} \cdots b_{k'_{2q}}$$

is nonzero if and only if $k_1 = k'_1, \ldots, k_{2q} = k'_{2q}$ (and hence $\pi = \mathrm{id}$). Consequently, the coefficient of the monomial (14) in Eq. (13) is

$$a_{k_1} \cdots a_{k_q} b_{k_{q+1}} \cdots b_{k_{2q}} \neq 0,$$

and hence there must be $\boldsymbol{w}, \boldsymbol{x}$ with

$$\det \begin{pmatrix} \sin(\boldsymbol{w}\boldsymbol{x}^\top) \\ \cos(\boldsymbol{w}\boldsymbol{x}^\top) \end{pmatrix} \neq 0 \,. \qquad \square$$

[6] This can be seen, for example, by writing the k_i in base-3 notation.