HIGH-DIMENSIONAL REGRESSION AND NONLINEAR RANDOM MATRICES

Abstract. Regression is a foundational problem in statistics and we thought we understood all of it—until neural networks came along. In particular, the double-descent phenomenon seemingly contradicts the classical bias-variance trade-off. In this note, we showcase some recent results on double-descent for kernel ridge regression [HL22, Mis22], a class of flexible regression models reminiscent of neural networks. The analysis relies crucially on a certain Gaussian equivalence principle for inner-product (Gram) matrices with element-wise nonlinearity [LY25].

Contents

1. Ridge regression and the mystery of overparameterization

Consider the typical regression setting where one is given labelled data pairs $(x_i, y_i)_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, and one would like to obtain a best predictor $\hat{f}^n : \mathbb{R}^d \to \mathbb{R}$ in some sense. A typical setting is to assume the data is drawn from some distribution P_{x} and the output is some independently corrupted version of a true predictor $f_* : \mathbb{R}^d \to \mathbb{R}$, i.e.,

$$y_i = f_*(x_i) + \epsilon_i \qquad \text{where } \mathbb{E}[\epsilon_i] = 0 \text{ and } \mathbb{E}[\epsilon_i^2] = \sigma_\epsilon^2.$$

Then, one can talk about the best predictor with respect to some loss $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$; that is, one would like to minimize the statistical *risk* or *test error*

$$\mathsf{R}(f; f_*) = \mathbb{E}\left[\ell(f(x), y)\right]. \tag{1}$$

For example, a popular loss is the squared loss $\ell(y_1, y_2) = (y_1 - y_2)^2$. In this case, the *Bayes optimal predictor* is given by the conditional expectation

$$h(x) = \mathbb{E}[y|x] = \arg\min_{f:\mathbb{R}^d \to \mathbb{R}} \mathsf{R}(f; f_*).$$

Of course, we cannot compute the Bayes optimal with finitely many samples, and we need to restrict the class of functions f_* can be to have a shot at computing a good estimate. A convenient choice is if it is a linear function

$$f_*(x) = \langle \theta_*, x \rangle \quad \text{for some } \theta^* \in \mathbb{R}^d.$$

Under this *hypothesis class*, we can minimize the *empirical risk* (this is known as *empirical risk minimization*, or ERM) as a proxy for the true risk, i.e., replacing the expectation in (1) to be the expectation over the empirical measure:

$$\hat{\theta}^n = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \langle \theta, x_i \rangle)^2 + \lambda \|\theta\|_2^2 \right\} \tag{2}$$

for some regularizer $\lambda > 0$. The point of the regularizer is to prevent overfitting—we will come to this later. A Bayesian perspective gives a concrete interpretation: if we take $\epsilon_i \sim \mathsf{N}(0, \sigma_\epsilon^2)$, then the regularization can be thought of as the Bayesian prior $\theta \sim \mathsf{N}(0, \lambda^2)$. This is called the *ridge regression*.

Remark 1.1. In the case of ridge regression, minimizing the empirical risk is easy as 1) an analytic solution to (2) is available in terms of matrix inversion, and 2) the problem is (strongly) convex for a first-order optimizer is needed. For more complicated models—such as a neural network—empirical risks can be very difficult and the algorithm that one uses can influence the statistical risk greatly. We will not address this point in this note, but obtaining statistical lower bounds for these difficult learning problems is a very active field; see [AAM22, AAM23] or [CWPPS24, PPAP24] for two particular perspectives.

The class of linear functions seems restrictive at first sight. Taking $d = 1$, what if we want to fit a quadratic

$$f_*(x) = a_2 x^2 + a_2 x + a_0 \quad \text{for } a_1, a_2, a_3 \in \mathbb{R}.$$

One can be clever redefine the problem! We map the data x_i to include the polynomials and extend the coefficients to $\mathbb{R}^3$ to accommodate:

$$x_i \in \mathbb{R} \mapsto \tilde{x} = (1, x_i, x_i^2) \quad \text{and} \quad \theta \in \mathbb{R} \mapsto \tilde{\theta} = (\theta_1, \theta_2, \theta_3)^\intercal \in \mathbb{R}^3.$$

Then, we simply change our regression problem from (2) to the ones with the tilde. The problem is still linear, just in a higher-dimensional space.

1.1. **The "kernel trick": a machine learner's favorite.** We can take the lifting procedure to an extreme. Let the feature space $(\mathsf{F}, \langle \cdot, \cdot \rangle_\mathsf{F})$ be a Hilbert space and define a *feature map* $\Phi : \mathbb{R}^d \to \mathsf{F}$. Then, we can assume that $f_*(x) = \langle \theta_*, \Phi(x) \rangle_\mathsf{F}$ and the *kernel ridge regression* is the same ridge regression problem but on the feature space, that is, we want to solve the ERM problem

$$\hat{\theta}^n = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \langle \theta, \Phi(x_i) \rangle_\mathsf{F})^2 + \lambda \|\theta\|_\mathsf{F}^2 \right\}. \tag{3}$$

This is the so-called *kernel trick* and has been used by machine learners for a long time. It works in a much broader context than regression, e.g., most famously in support vector machines for building nonlinear classifiers. In spirit, whenever there is an inner product, one can lift it up to solve a linear problem in a higher-dimensional feature space F, which would seem like a nonlinear problem in the original space. But wait... where is the "kernel" in kernel trick? It turns out we need some functional analysis to explain where this kernel came from.

Definition 1.2. Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive, symmetric kernel. Then, a *reproducing kernel Hilbert space* (RKHS) with kernel K is a Hilbert space $(\mathsf{H}, \langle \cdot, \cdot \rangle_\mathsf{H})$ such that

$$K(x, \cdot) \in \mathsf{H} \text{ for all } x \in \mathbb{R}^d \quad \text{and} \quad \langle f, K(x, \cdot) \rangle = f(x) \text{ for all } f \in \mathsf{H}, x \in \mathbb{R}^d.$$

The space of regression functions on a feature space is an RKHS: define $f \in \mathsf{H}$ to be of the form

$$f(x) = \langle \theta, \Phi(x) \rangle_\mathsf{F} \quad \text{for some } \theta \in \mathsf{F}.$$

Then, the inner product

$$\langle f, f' \rangle_\mathsf{H} = \langle \theta, \theta' \rangle_\mathsf{F}$$

defines an RKHS with the kernel

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle_\mathsf{F}.$$

On the other hand, given a RKHS H, one can define the canonical feature map

$$\Phi : x \in \mathbb{R}^d \mapsto \Phi(x) \in \mathsf{H},$$

and functions $f \in \mathsf{H}$ can be thought of as functions from linear regression

$$f(x) = \langle \theta, \Phi(x) \rangle \quad \text{where } \theta = K(x, \cdot) \in \mathsf{F} = \mathsf{H}.$$

More generally, it turns out that the existence of a positive symmetric kernel is sufficient to define an RKHS.

Theorem 1.3 (Moore-Aronszajn)**.** *Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive symmetric kernel. Then, there exists a unique Hilbert space $\mathsf{H} \subset \{f : \mathbb{R}^d \to \mathbb{R}\}$ such that K is its reproducing kernel.*

Example 1.4 (Radial basis function)**.** The most canonical kernel choice is the *radial basis kernel*, which takes the form

$$K(x, x') = e^{-\|x-x'\|^2/\sigma}$$

for some $\sigma > 0$. The associated feature map is not obvious but we know that evaluating this kernel essentially performs an inner-product on some Hilbert space.

Example 1.5 (Inner-product kernels)**.** Let $h : \mathbb{R} \to \mathbb{R}$ be some (nice enough) nonlinear function. Then, *inner-product kernels* are kernels of the form

$$K(x, x') = h(\langle x, x' \rangle).$$

While these kernels are not particularly popular in practice, we will stick to this kernel for the remainder of the note because of the simplicity and resemblance with infinite-width neural networks. It has been argued that neural networks share many qualitative similarities as neural networks [BMM18], particularly so in the infinite-width limit known as the *neural tangent kernel* [JGH18].

Using the equivalence, we can now rewrite the ERM problem (3) using the kernel representation rather than the feature map representation:

$$\hat{f}^n = \arg\min_{f \in \mathsf{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathsf{H}}^2 \right\}. \tag{4}$$

Despite the problem now looking infinite-dimensional, using some functional analysis, one can show that the solution to the optimization problem (4) lives in the $\operatorname{span}\{K(x_i, \cdot)\}_{i \in [n]}$. Letting

$$k(x) = (K(x_i, x))_{i \in [n]} \quad \text{and} \quad K = (K(x_i, x_j))_{i,j \in [n]},$$

the ERM problem (4) can be rewritten and solved:

$$\hat{a}^n = \arg\min_{a \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} (y_i - a^\intercal k(x_i))^2 + \lambda a^\intercal K a \right\} = (K + \lambda \mathbb{I}_n)^{-1} y \tag{5}$$

with the solution being

$$\hat{f}^n(x) = \sum_{i=1}^{n} \hat{a}_i^n K(x_i, x).$$

Note that, for the particular case of the inner product kernels, the kernel matrix K (or written as H below) takes the form

$$H = (h(\langle x_i, x_j \rangle))_{i,j \in [n]}. \tag{6}$$

1.2. **You don't have to pay for the extra complexity.** Once we've obtained an estimator $\hat{f}^n$ from data, we would like to know how well it generalizes, i.e., given a new data point x, how far (on average) is the predicted point $f^n(x)$ from the true point $f_*(x) + \epsilon$. The error from this test is known as the *test error*, compactly expressed as $\mathsf{R}(\hat{f}^n; f_*)$. The featurization map now plays a huge role:

(1) if f_* is linear and we choose to include polynomial features, then our estimate will have lots of "wiggles" and perform terribly on a new test point; but
(2) if f_* is a high-degree polynomial but we only try to fit linearly, then our average test error will be high as all the polynomial features are lost.

This is standard in statistical theory known as the *bias-variance trade-off*. Take the squared loss, then the risk can be separated into

$$\mathsf{R}(\hat{f}^n, f_*) = \underbrace{\mathbb{E}[\hat{f}^n - f_*]^2}_{\text{bias}} + \underbrace{\mathbb{E}[(\hat{f}^n - \mathbb{E}[\hat{f}^n])^2]}_{\text{variance}}.$$

Case (1) corresponds to having a low variance but high bias, and Case (2) the other way. Classical statistics is all about finding the sweet spot that minimizes both or picking one over the other when necessary; one's model shouldn't be too simple nor too complicated.

Neural networks—which one can regard as a fancy regressor—came and destroyed this picture. People discovered that they can get away with making the neural network as big as they can (adding more degrees of

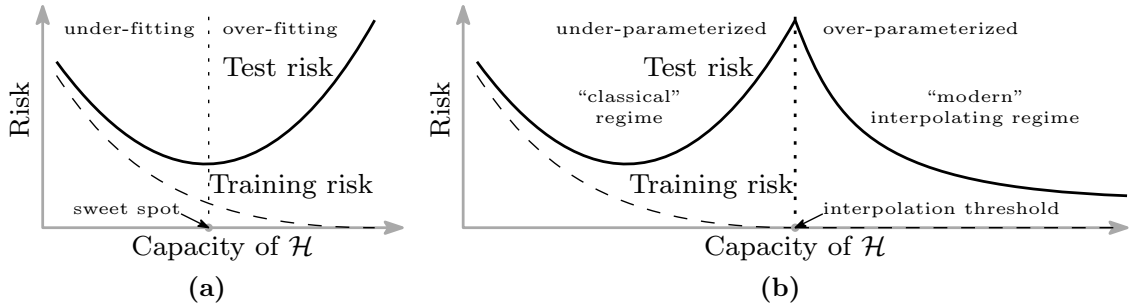

FIGURE 1. [Taken from [BHMM19]] (**a**) The classical *U-shaped risk curve* arising from the bias-variance trade-off. (**b**) The *double descent risk curve*, which incorporates the U-shaped risk curve (i.e., the "classical" regime) together with the observed behavior from using high capacity function classes (i.e., the "modern" interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

freedom than data points), training to zero empirical error (a sign of overfitting), and enjoy unprecedentedly good generalization error. This is known as the *double-descent phenomenon* [BHMM19] (cf. Figure 1) and it has been demonstrated empirically in many other related statistical models, including kernel ridge regression.

The double-descent phenomenon for kernel ridge regression has been noticed before by physicists via a hand-wavy central limit theorem (sometimes called the *Gaussian design model*) [CBP21, CLKZ21]. The observation is the following: suppose the kernel can be diagonalized by basis $(Y_{d,k})_{k\in\mathbb{N}}$, i.e.,

$$K_d(x_1, x_2) = \sum_{k=0}^{\infty} \mu_{d,k} Y_{d,k}(x_1) Y_{d,k}(x_2),$$

then, we can lift the problem to the feature space $\mathsf{F} = \ell^2(\mathbb{R})$ where

$$f_*(x) = \langle \theta_*, \Phi(x) \rangle_{\mathsf{F}} \quad \text{with } \Phi(x) = \left(\sqrt{\mu_{d,k}} Y_{d,k}(x)\right)_{k\in\mathbb{N}}.$$

Now, observe that the orthogonality of the basis gives that the features are uncorrelated

$$\mathbb{E}[\Phi(x)\Phi(x)^\intercal] = \mathbb{I}.$$

Moreover, without loss of generality, we can center the features so they are mean-zero. If we close our eyes and assert a CLT, i.e., claim that

$$\Phi(x) \sim \mathsf{N}(0, \operatorname{diag}(\mu_d)),$$

then one can hope to do calculation for this model. However, asserting Gaussian features is typically dangerous because they are certainly not independent and probably (as we will see) not sub-Gaussian. Nonetheless, the prediction using Gaussian design is incredibly accurate and we will show that, in fact, it is exact.

2. Spectrum of linear inner-product matrices

Before investigating the nonlinear counterpart, we review standard results in random matrix theory about the spectrum of inner-product matrices. Consider the data matrix

$$X = (x_1 \dots x_n) \in \mathbb{R}^{d\times n} \quad \text{where } x_i \overset{\text{i.i.d.}}{\sim} \mu^{\times d} \in \mathcal{P}(\mathbb{R}^d).$$

The protagonist of this note is the inner-product matrix

$$X^\intercal X = (x_i^\intercal x_j)_{i,j\in[n]} \in \mathbb{R}^{n\times n}$$

and its spectrum. We use σ to denote the empirical spectrum, i.e., let $(\lambda_i)_{i\in[n]}$ be the eigenvalues of some symmetric matrix $A \in \mathbb{R}^{n\times n}$, then

$$\sigma(A) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i}.$$

As is, the operator norm of $X^\intercal X$ explodes and the empirical spectrum probably does not have a well-defined limit. So, some scaling is needed. If we believe that the inner-product of independent vectors remains $\mathcal{O}(1)$ as $d \to \infty$, then rescaling each entry by $1/\sqrt{n}$ is sufficient to keep $\|X^\intercal X\| = \mathcal{O}(1)$. In summary, we consider

$$\sigma(\frac{1}{n} X^\intercal X) \to [\text{something}] \quad \text{as } n, d \to \infty,\ d/n \to \phi. \tag{7}$$

We call ϕ the *aspect ratio*.

Theorem 2.1 (Marchenko-Pastur [MP67] (see also [BS+10, Chapter 3])**.** *Suppose μ has zero mean and unit variance. Then, (7) holds with probability one with the limiting distribution being the Marchenko-Pastur (MP) law, defined by*

$$\rho_{\mathsf{MP},\phi}(dx) = \mathbf{1}_{\{[\lambda_-,\lambda_+]\}} \frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{2\pi x \phi} dx + (1-\phi^{-1})_+ \delta_0(dx) \quad \text{where } \lambda_\pm = (1 \pm \sqrt{\phi})^2. \tag{8}$$

A few words about the proof. A key tool in random matrix theory is the *Stieltjes transform* (sometimes also known as the *Cauchy transform*). For a real-valued function F of bounded variation, the Stieltjes transform of F is defined as

$$s_F(z) = \int_{\mathbb{R}} \frac{1}{\lambda - z} F(d\lambda)\ z \in \mathbb{C}^+ = \{w \in \mathbb{C} : \Im w > 0\}.$$

For us, the function F will be the empirical CDF of the spectrum. Moreover, we know that

(1) the Stieltjes transform converges if and only if the CDF converges weakly [BS+10, Theorem B.9], and
(2) given a Stieltjes transform, there is an inversion formula to get back F [BS+10, Theorem B.8].

So, let F^n be the empirical CDF of the spectrum of some random matrix A^n, we only need to show that

$$s_{F^n}(z) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\lambda_i - z} = \frac{1}{n} \operatorname{tr}\left[(A^n - z\mathbb{I}_n)^{-1}\right] \to s(z) \quad \text{for a.e. } z \in \mathbb{C}^+.$$

Lastly, we note that the random matrix that we will be considering is one whose diagonal is zero, and the resulting MP law is of a slightly different form.

Theorem 2.2 ([LY25, Appendix C])**.** *Suppose μ has zero mean and unit variance. Then, with probability one and aspect ratio ϕ,*

$$\sigma(\frac{t}{n}(X^\intercal X - d\,\mathbb{I}_n)) \to \rho_{\mathsf{sMP},\phi}$$

where the shifted-and-scaled Marchenko-Pastur law takes the form

$$\rho_{\mathsf{sMP},\phi}(dx) = \frac{(\sqrt{2\sqrt{\phi} + x/t - 1})(2\sqrt{\phi} - x/t + 1)_+}{2\pi \operatorname{sgn}(t)(x + t\phi)} dx + (1-\phi)_+ \delta_{x+t\phi}(dx). \tag{9}$$

3. A Gaussian equivalence principle

The spectrum of kernel inner-product matrices was original studied by [Kar10]. The linear scaling was initially done by Cheng and Singer [CS13], and they characterized the spectrum with the Stieltjes transform being the solution to some cubic equation. Later, Fan and Montanari [FM19] showed that the spectrum is actually a linear combination of (shifted) Wishart matrix and an independent GOE.

3.1. Gaussian equivalence on the sphere.

Throughout, we will assume $x_i \sim \mathsf{Unif}(\mathbb{S}^{d-1})$ independently and consider the inner product matrix

$$A = \left(\frac{1}{\sqrt{n}} h(\sqrt{d}\langle x_i, x_j\rangle)\right)_{i \neq j \in [n]}$$

for some $h : \mathbb{R} \to \mathbb{R}$ (to be qualified).

Remark 3.1. The diagonal entries of A are set to zero for convenience. In this case, the diagonal is identically $h(\sqrt{d})/\sqrt{n}$, which can drift away to infinity or diminish to zero for certain scaling of d and n. By enforcing the diagonal to be zero, we make sure that the spectrum stays $\mathcal{O}(1)$ in some sense.

Let $(Q_{d,k})_{k\in\mathbb{N}}$ denote the set of *Gegenbauer polynomials* or *ultraspherical polynomials*. For each d, $Q_{d,k}$ is a k-th degree polynomial and forms an orthonormal basis in the sense that

$$\mathbb{E}_{x,y\sim\mathsf{Unif}(\mathbb{S}^{d-1})}\left[Q_{d,k}(\sqrt{d}\langle x,y\rangle)Q_{d,\ell}(\sqrt{d}\langle x,y\rangle)\right]=\delta_{k\ell}.$$

Notably, the coefficients in front of the Gegenbauer polynomials depend on the dimension. Moreover, let

$$N_{d,k}=\frac{d^k}{k!}+o(d^k)$$

be the number of degree-k spherical harmonics $(Y_{k,s})_{s\in[N_k]}$. The Gegenbauer polynomials can be alternatively defined via

$$Q_{d,k}(\sqrt{d}x_i^\intercal x_j)=\frac{1}{\sqrt{N_{d,k}}}\sum_{s\in[N_{d,k}]}Y_{k,s}(x_i)Y_{k,s}(x_j).$$

For simplicity, suppose the nonlinearity h is a degree-L polynomial

$$h(x)=\sum_{k=0}^{L}\mu_k Q_{d,k}(x).$$

Then, the inner-product matrix H can be decomposed as

$$A=\sum_{k=0}^{L}\mu_k A_k\quad\text{where } A_k=\left(\frac{1}{\sqrt{n}}Q_{d,k}(\sqrt{d}\langle x_i,x_j\rangle)\right)_{i\neq j\in[n]}=\left(\frac{1}{\sqrt{nN_{d,k}}}\sum_{s\in[N_{d,k}]}Y_{k,s}(x_i)Y_{k,s}(x_j)\right)_{i\neq j\in[n]},$$

or rewritten further

$$A_k=\frac{1}{\sqrt{nN_{d,k}}}S_k^\intercal S_k-\sqrt{\frac{N_{d,k}}{n}}\mathbb{I}_n\quad\text{where } S_k=\left(Y_{k,s}(x_i)\right)_{s\in[N_{k,d}],i\in[n]}.$$

We will rewrite it one more time to match the (shifted-and-scaled) Marchenko-Pastur scaling (9):

$$A=\frac{1}{n}\sum_{k=0}^{L}\frac{\mu_k\sqrt{n}}{\sqrt{N_{d,k}}}\underbrace{\left(S_k^\intercal S_k-N_{d,k}\mathbb{I}_n\right)}_{A_k}.$$

Notice that the columns of S_k depend only on x_i and the data is assumed to be independent. By the orthonormality of the spherical harmonics, the entries themselves are uncorrelated, i.e.,

$$\mathbb{E}[Y_{k,s}(x_i)Y_{\ell,r}(x_i)]=\delta_{k\ell}\delta_{sr}. \tag{10}$$

The Gaussian equivalence asserts that replacing the asymptotic spectrum of S is the same while replacing the (uncorrelated, non-Gaussian) entries by i.i.d. Gaussians with matching first and second moments. Take this for granted temporarily, then for all k,

$$\sigma(A_k)\to\mathsf{MP}(\phi_k)\quad\text{for some aspect ratio }\phi_k=\lim_{n,d\to\infty}\frac{N_{d,k}}{n}.$$

Now, we take $n,d\to\infty$ such that $d^\ell/n\to\kappa>0$ for some $\ell\in\mathbb{N}$. Then, there are three regimes:

(1) for $k<\ell$, then $\phi_k=d^{k-\ell}/k!=o(d^{-1})$ and the spectrum degenerates to zero asymptotically. Moreover (and more crucially), they are low-rank (check dimension) in comparison to higher-order A_k's so they do not contribute to the global spectrum. This can be seen from the shifted MP law having only spike component.
(2) for $k=\ell$, the aspect ratio $\phi_k=\mathcal{O}(1)$ and the spectrum converges to a non-degenerate Marchenko-Pastur law.
(3) for $k>\ell$, the limiting spectrum is distributed like a shifted-and-scaled MP law with aspect ratio $\phi=\infty$ and $t=1/\sqrt{\phi}$. Ignoring dividing by zero, we have that

$$\rho_{\mathsf{sMP},\phi,t}(dx)=\frac{1}{2\pi}\sqrt{(4-x^2)_+}dx$$

and the Marchenko-Pastur degenerates into a semi-circle law.

The matrix of interest is actually a sum of the A_k's, so asymptotically we should expect

$$\sigma(A) = \sigma\left(\sum_{k=0}^{\ell-1} \mu_k A_k + \mu_\ell A_\ell + \sum_{k=\ell+1}^{L} A_k\right) \to [\text{nothing}] \boxplus \mathsf{MP} \boxplus \mathsf{SC} \quad \text{as } n, d \to \infty \text{ and } d^\ell/n = \kappa, \tag{11}$$

where $\boxplus$ is the additive free convolution.

Theorem 3.2 ([LY25, Theorem 1 and 2])**.** *Gaussian equivalence in the sense of (11) holds almost surely. Moreover, the nonlinearity h can be extended to functions with sub-exponential growth.*

3.2. **Universality of Gaussian equivalence.** It turns out the equivalence principle is universal!

Theorem 3.3 ([DLMY23])**.** *Suppose the data has i.i.d. entries sampled from some $\mu \in \mathcal{P}(\mathbb{R})$ that has finite moments, i.e.,*

$$\mathbb{E}_\mu[|x|^p] < c_p \quad \textit{for all } p \in \mathbb{N}.$$

Then, the Gaussian equivalence principle holds for all nonlinearity $h \in L^2(\gamma)$ where $\gamma = \mathsf{N}(0,1)$.

We want a similar proof strategy, but the spherical assumption informed our choice of the Gegenbauer polynomial. This is a crucial step because the orthonormality of the spherical harmonics allows us to conclude that the feature vectors are isotropic (cf. (10)).

Since we are showing a Gaussian equivalence, perhaps a "universal" choice of polynomials is the *Hermite polynomials* $(H_k)_{k\in\mathbb{N}}$, a collection of orthonormal polynomials in $L^2(\mathbb{R}, \gamma)$ where $\gamma = \mathsf{N}(0,1)$ is the standard Gaussian measure. In particular, for each $k \in \mathbb{N}$, H_k is a degree-k polynomial.

Again, let's start by assuming that the nonlinearity is a polynomial

$$h(x) = \sum_{k=0}^{L} \mu_k H_k(x).$$

We cannot immediately proceed as the spherical case as there is no general relationship between the Hermite polynomial and something like spherical harmonics. However, if we define the feature maps to be the polynomials:

$$\Phi_k : x \in \mathbb{R}^d \mapsto (x_{a_1} x_{a_2} \dots x_{a_k})_{1 \le a_1 < \dots < a_k \le d} \in \mathbb{R}^{N_{d,k}} \quad \text{where } N_{d,k} = \binom{d}{k},$$

then, we claim that

$$H_k(\sqrt{d} x_i^\intercal x_j) = \frac{1}{\sqrt{N_{d,k}}} \langle \Phi_k(x_i), \Phi_k(x_j) \rangle + \mathcal{O}_p(d^{-1/2}).$$

This is the so-called *Walsh chaos* and partly justifies the use of the Hermite polynomial as $\sqrt{d} x_i^\intercal x_j$ satisfies a CLT. The claim is easy to see for the case of $k = 2$.

Example 3.4 (Degree-2 polynomial on the hypercube)**.** Let $x_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}(\{\pm d^{-1/2}\}^d)$ and let $h(x) = H_2(x) = (x^2 - 1)/\sqrt{2}$. Then, from straightforward computation,

$$\begin{aligned} H_2(\sqrt{d} x_i^\intercal x_j) &= \left(\frac{x_i^\intercal x_j}{\sqrt{d}}\right)^2 - 1 = \frac{1}{d} \sum_{a,b=1}^{d} x_{ia} x_{ib} x_{ja} x_{jb} - 1 \\ &= \frac{2}{d} \sum_{a<b} (x_{ia} x_{ib})(x_{ja} x_{jb}) \\ &= \frac{1}{\sqrt{N_{d,2}}} \langle \Phi_2(x_i), \Phi_2(x_j) \rangle + \mathcal{O}(d^{-1}). \end{aligned}$$

Just like before, the features are centered and isotropic (by independence and non-overlapping entries):

$$\mathbb{E}[\Phi_k(x)] = 0 \text{ and } \mathbb{E}[\Phi_k(x_i)\Phi_k(x_i)] = \mathbb{I}_{N_{d,k}} \quad \text{for all } k \in \mathbb{N}.$$

So, we can proceed as before and write

$$A = \frac{1}{\sqrt{n}} \sum_{k=1}^{L} \mu^k H_k(\sqrt{d} X^T X) = \sum_{k=1}^{L} \mu_k A_k + [\text{error}] \quad \text{where } A_k = \left(\frac{1}{\sqrt{n N_{d,k}}} \langle \Phi_k(x_i), \Phi_k(x_j) \rangle\right)_{i \ne j},$$

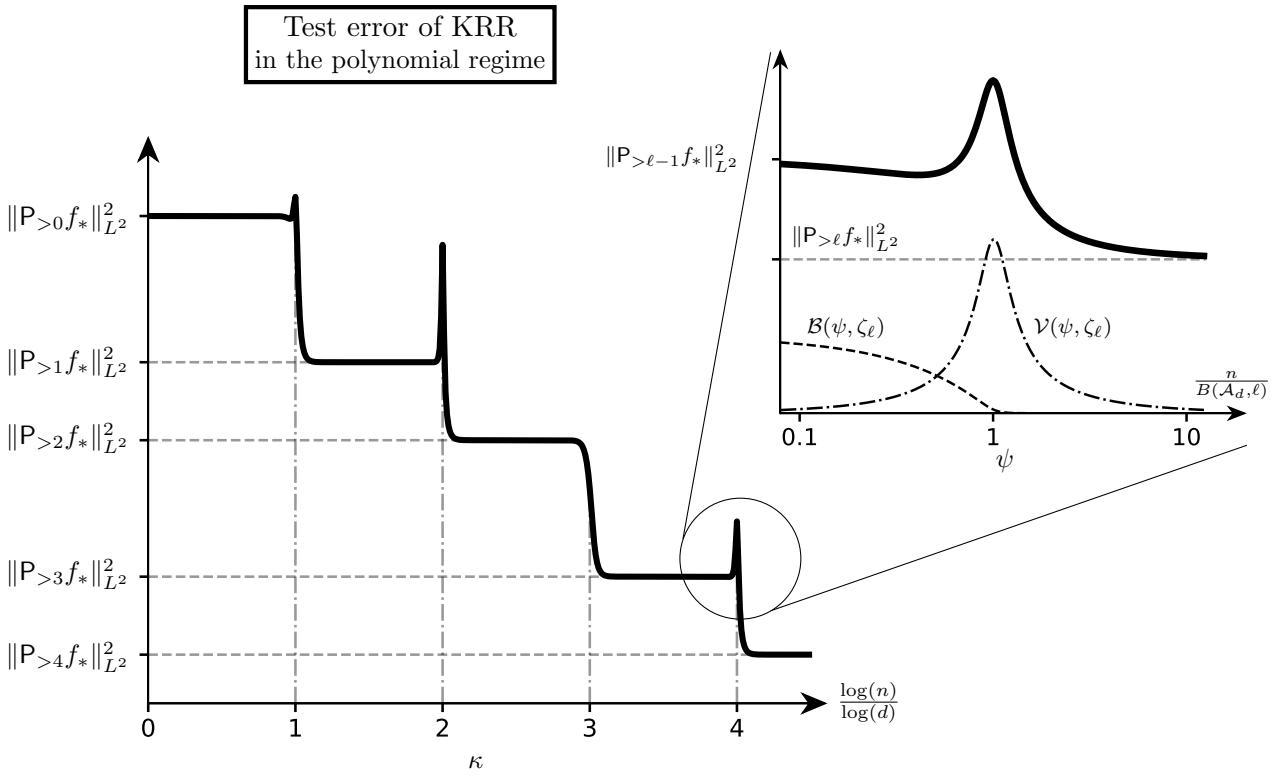

FIGURE 2. (Taken from [Mis22, Figure 1]) A cartoon illustration of the test error of KRR in the polynomial regime $n/d^\kappa \to \psi$, as $n, d \to \infty$, for any $\kappa, \psi \in \mathbb{R}_{>0}$. The test error follows a staircase, with peaks that can occur at each $\kappa = \ell \in \mathbb{N}$, depending on the effective regularization ζ_ℓ and effective signal-to-noise ratio SNR_ℓ at level ℓ (a peak occurs if ζ_ℓ and/or SNR_ℓ are small enough).

and now we are at the form which suggests the Gaussian equivalence. A key aspect of the proof is the show *approximate rotational invariance*, i.e., showing that the size of the feature vector (which is of size $N_{d,k} = \mathcal{O}(d^k)$) is of the same magnitude as that of a Gaussian.

3.3. **Application: double-descent for kernel ridge regression.** To conclude, we revisit kernel ridge regression and interpret the result in light of the Gaussian equivalence principle. Recall that the solution to the regression problem depends on the kernel inner-product matrix H (defined in (6)), i.e., we obtain the regression coefficients and estimate

$$\hat{a}^n = (H - \lambda \mathbb{I})^{-1} y \quad \text{and} \quad \hat{f}^n(x) = \sum_{i=1}^n \hat{a}_i^n h(\sqrt{d}\langle x_i, x\rangle).$$

We are interested in the test error

$$\mathsf{R}(\hat{f}^n; f_*) = \mathbb{E}[(\hat{f}^n(x) - f_*(x))^2].$$

If we suppose that the (class of) true function is a polynomial, represented by spherical harmonics

$$f_*(x) = \sum_{k=0}^{\ell-1} \sum_{s=1}^{N_{d,k}} \beta_{k,s}^* Y_{k,s}(x) + \sum_{k \geq \ell} \sum_{s=1}^{N_{d,k}} \tilde{\beta}_{k,s} Y_{k,s}(x),$$

and assume

$$\mathbb{E}[\tilde{\beta}_{k,s}^2] = \frac{1}{N_{d,k}} F_{d,k}^2 \text{ with } \lim_{d \to \infty} F_{d,\ell}^2 \to F_\ell^2.$$

Then, depending on the number of samples (aspect ratio), we can quantify the contribution from different subspaces of polynomials to the test error.

Theorem 3.5 ([Mis22, Theorem 2], see also [HL22]). *Suppose that $n, d \to \infty$ with $d^\ell/n = \kappa$ and let $\mathcal{V}_{<\ell} = span(Y_k)_{k<\ell}$, $\mathcal{V}_\ell = spanY_k$, $\mathcal{V}_{>\ell} = span(Y_k)_{k>\ell}$ respectively. Then,*

(1) on $\mathcal{V}_{<\ell}$, the function learns perfectly and does not contribute to the test error;
(2) on $\mathcal{V}_\ell$, classical picture re-enters for these high-order features and there are non-trivial bias and variance contributions;
(3) on $\mathcal{V}_{>\ell}$, the regressor does not learn any higher order features and contributes to a bias of F_ℓ^2.

The proof actually uses the spectral result only on the subspace $\mathcal{V}_\ell$, which asked for the resolvent of the Gegenbauer matrix $S_\ell^\intercal S_\ell$. The Gaussian equivalence discussed is stronger in the sense that it handles the whole hierarchy.

Remark 3.6. Notice that this means kernel methods need d^ℓ samples to learn features from ℓ-degree polynomials. This is believed not to be true for neural networks (e.g., for particular polynomials) and, hence, neural networks are not kernel methods.

References

[AAM22] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

[AAM23] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2552–2623. PMLR, 2023.

[BHMM19] 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, 2019.

[BMM18] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International conference on machine learning*, pages 541–549. PMLR, 2018.

[BS+10] Zhidong Bai, Jack William Silverstein, et al. *Spectral analysis of large dimensional random matrices*. Springer, 2010.

[CBP21] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

[CLKZ21] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

[CS13] Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

[CWPPS24] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ode for sgd learning dynamics on glms and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4):iaae028, 2024.

[DLMY23] Sofiia Dubova, Yue M Lu, Benjamin McKenna, and Horng-Tzer Yau. Universality for the global spectrum of random inner-product kernel matrices in the polynomial regime. *arXiv preprint arXiv:2310.18280*, 2023.

[FM19] Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1):27–85, 2019.

[HL22] Hong Hu and Yue M Lu. Sharp asymptotics of kernel ridge regression beyond the linear regime. *arXiv preprint arXiv:2205.06798*, 2022.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[Kar10] Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1 – 50, 2010.

[LY25] Yue M Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product kernel matrices with polynomial scalings. *The Annals of Applied Probability*, 35(4):2411–2470, 2025.

[Mis22] Theodor Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv preprint arXiv:2204.10425*, 2022.

[MP67] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[PPAP24] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of sgd in high-dimensions: Exact dynamics and generalization properties. *Mathematical Programming*, pages 1–90, 2024.