1. Background on Kernel Machines

1.1. Reproducing Kernel Hilbert Space

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $p(\mathbf{x})$ be a probability distribution over $\mathcal{X}$. Let $\mathcal{H}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. A kernel $K(\cdot, \cdot)$ is said to be reproducing for $\mathcal{H}$ if function evaluation at any $\mathbf{x} \in \mathcal{X}$ is the equivalent to the Hilbert inner product with $K(\cdot, \mathbf{x})$: K is reproducing for $\mathcal{H}$ if for all $g \in \mathcal{H}$ and all $\mathbf{x} \in \mathcal{X}$

$$\langle K(\cdot, \mathbf{x}), g \rangle_{\mathcal{H}} = g(\mathbf{x}). \tag{SI.1}$$

If such a kernel exists for a Hilbert space, then it is unique and defined as the reproducing kernel for the RKHS (Evgeniou et al., 1999; Schölkopf & Smola, 2001).

1.2. Mercer's Theorem

Let $\mathcal{H}$ be a RKHS with kernel K. Mercer's theorem (Mercer, 1909; Rasmussen & Williams, 2005) allows the eigendecomposition of K

$$K(\mathbf{x}, \mathbf{x}') = \sum_{\rho} \lambda_\rho \phi_\rho(\mathbf{x}) \phi_\rho(\mathbf{x}), \tag{SI.2}$$

where the eigenvalue statement is

$$\int d\mathbf{x}' p(\mathbf{x}') K(\mathbf{x}, \mathbf{x}') \phi_\rho(\mathbf{x}') = \lambda_\rho \phi_\rho(\mathbf{x}). \tag{SI.3}$$

1.3. Representer Theorem

Let $\mathcal{H}$ be a RKHS with inner product $\langle ., . \rangle_{\mathcal{H}}$. Consider the regularized learning problem

$$\min_{f \in \mathcal{H}} \hat{\mathcal{L}}[f] + \lambda ||f||^2_{\mathcal{H}}, \tag{SI.4}$$

where $\hat{\mathcal{L}}[f]$ is an empirical cost defined on the discrete support of the dataset and $\lambda > 0$. The optimal solution to the optimization problem above can always be written as (Schölkopf & Smola, 2001)

$$f(x) = \sum_{i=1}^{p} \alpha_i K(x_i, x). \tag{SI.5}$$

1.4. Solution to Least Squares

Specializing to the case of least squares regression, let

$$\hat{\mathcal{L}}[f] = \sum_{i=1}^{p} (f(\mathbf{x}_i) - y_i)^2. \tag{SI.6}$$

Using the representer theorem, we may reformulate the entire objective in terms of the p coefficients α_i

$$\begin{aligned}
\mathcal{L}[f] &= \sum_{i=1}^{p} (f(\mathbf{x}_i) - y_i)^2 + \lambda ||f||^2_{\mathcal{H}} \\
&= \sum_{i=1}^{p} (\sum_{j=1}^{p} \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) - y_i)^2 \\
&\quad + \lambda \sum_{ij} \alpha_i \alpha_j \Big\langle K(\mathbf{x}_i, \cdot), K(\mathbf{x}_j, \cdot) \Big\rangle_{\mathcal{H}} \\
&= \boldsymbol{\alpha}^\top \mathbf{K}^2 \boldsymbol{\alpha} - 2\mathbf{y}^\top \mathbf{K} \boldsymbol{\alpha} + \mathbf{y}^\top \mathbf{y} + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}.
\end{aligned} \tag{SI.7}$$

Optimizing this loss with respect to $\boldsymbol{\alpha}$ gives

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y}. \tag{SI.8}$$

Therefore the optimal function evaluated at a test point is

$$f(\mathbf{x}) = \boldsymbol{\alpha}^\top \mathbf{k}(\mathbf{x}) = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}). \tag{SI.9}$$

2. Derivation of the Generalization Error

Let the RKHS $\mathcal{H}$ have eigenvalues λ_ρ for $\rho \in \mathbb{Z}^+$. Define $\psi_\rho(\mathbf{x}) = \sqrt{\lambda_\rho} \phi_\rho(\mathbf{x})$, where ϕ_ρ are the eigenfunctions of the reproducing kernel for $\mathcal{H}$. Let the target function have the following expansion in terms of the kernel eigenfunctions $f^*(\mathbf{x}) = \sum_\rho \overline{w}_\rho \psi_\rho(\mathbf{x})$. Define the design matrices $\boldsymbol{\Phi}_{\rho,i} = \phi_\rho(\mathbf{x}_i)$ and $\boldsymbol{\Lambda}_{\rho\gamma} = \lambda_\rho \delta_{\rho\gamma}$. Then the average generalization error for kernel regression is

$$E_g = \mathrm{Tr} \left(\mathbf{D} \left\langle \mathbf{G}^2 \right\rangle_{\{\mathbf{x}_i\}} \right) \tag{SI.10}$$

where

$$\mathbf{G} = \left(\frac{1}{\lambda} \boldsymbol{\Phi} \boldsymbol{\Phi}^\top + \boldsymbol{\Lambda}^{-1} \right)^{-1}, \quad \boldsymbol{\Phi} = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\Psi}. \tag{SI.11}$$

and

$$\mathbf{D} = \boldsymbol{\Lambda}^{-1/2} \left\langle \overline{\mathbf{w}} \overline{\mathbf{w}}^\top \right\rangle_{\overline{\mathbf{w}}} \boldsymbol{\Lambda}^{-1/2}. \tag{SI.12}$$

Proof. Define the student's eigenfunction expansion $f(\mathbf{x}) = \sum_\rho w_\rho \psi_\rho(\mathbf{x})$ and decompose the risk in the basis of eigenfunctions:

$$\begin{aligned}
E_g(\{\mathbf{x}_i\}, f^*) &= \left\langle (f(\mathbf{x}) - y(\mathbf{x}))^2 \right\rangle_{\mathbf{x}} \\
&= \sum_{\rho,\gamma} (w_\rho - \overline{w}_\rho)(w_\gamma - \overline{w}_\gamma) \left\langle \psi_\rho(\mathbf{x}) \psi_\gamma(\mathbf{x}) \right\rangle_{\mathbf{x}} \\
&= \sum_{\rho} \lambda_\rho (w_\rho - \overline{w}_\rho)^2 \\
&= (\mathbf{w} - \overline{\mathbf{w}})^\top \boldsymbol{\Lambda} (\mathbf{w} - \overline{\mathbf{w}}).
\end{aligned} \tag{SI.13}$$

Next, it suffices to calculate the weights $\mathbf{w}$ learned through kernel regression. Define a matrix with elements $\boldsymbol{\Psi}_{\rho,i} = \psi_\rho(\mathbf{x}_i)$. The training error for kernel regression is

$$E_{tr} = ||\boldsymbol{\Psi}^\top \mathbf{w} - \mathbf{y}||^2 + \lambda ||\mathbf{w}||_2^2 \tag{SI.14}$$

The ℓ_2 norm on $\mathbf{w}$ is equivalent to the Hilbert norm on the student function. If $f(\mathbf{x}) = \sum_\rho w_\rho \psi_\rho(\mathbf{x})$ then

$$||f||^2_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}} = \sum_{\rho\gamma} w_\rho w_\gamma \langle \psi_\rho(\cdot), \psi_\gamma(\cdot) \rangle_{\mathcal{H}} = \sum_\rho w_\rho^2, \quad \text{(SI.15)}$$

since $\langle \psi_\rho(\cdot), \psi_\gamma(\cdot) \rangle_{\mathcal{H}} = \delta_{\rho,\gamma}$ (Bietti & Mairal, 2019). This fact can be verified by invoking the reproducing property of the kernel and it's Mercer decomposition. Let $g(\cdot) = \sum_\rho a_\rho \psi_\rho(\cdot)$. By the reproducing property

$$\langle K(\cdot, \mathbf{x}), g(\cdot) \rangle_{\mathcal{H}} = \sum_{\rho,\gamma} a_\gamma \psi_\rho(\mathbf{x}) \langle \psi_\rho(\cdot), \psi_\gamma(\cdot) \rangle_{\mathcal{H}} = g(\mathbf{x}) = \sum_\rho a_\rho \psi_\rho(\mathbf{x}) \quad \text{(SI.16)}$$

Demanding equality of each term, we find

$$\sum_\gamma a_\gamma \langle \psi_\rho(\cdot), \psi_\gamma(\cdot) \rangle_{\mathcal{H}} = a_\rho \quad \text{(SI.17)}$$

Due to the arbitrariness of a_ρ, we must have $\langle \psi_\rho(\cdot), \psi_\gamma(\cdot) \rangle_{\mathcal{H}} = \delta_{\rho,\gamma}$. We stress the difference between the action of the Hilbert inner product and averaging feature functions over a dataset $\langle \psi_\rho(\mathbf{x}) \psi_\gamma(\mathbf{x}) \rangle_{\mathbf{x}} = \lambda_\rho \delta_{\rho,\gamma}$ which produce different results. We will always decorate angular brackets with $\mathcal{H}$ to denote Hilbert inner product.

The training error has a unique minimum

$$\mathbf{w} = (\mathbf{\Psi}\mathbf{\Psi}^\top + \lambda \mathbf{I})^{-1} \mathbf{\Psi} \mathbf{y} = (\mathbf{\Psi}\mathbf{\Psi}^\top + \lambda \mathbf{I})^{-1} \mathbf{\Psi}\mathbf{\Psi}^\top \overline{\mathbf{w}} = \overline{\mathbf{w}} - \lambda (\mathbf{\Psi}\mathbf{\Psi}^\top + \lambda \mathbf{I})^{-1} \overline{\mathbf{w}}, \quad \text{(SI.18)}$$

where the target function is produced according to $\mathbf{y} = \mathbf{\Psi}^\top \overline{\mathbf{w}}$.

Plugging in the $\mathbf{w}$ that minimizes the training error into the formula for the generalization error, we find

$$E_g(\{\mathbf{x}_i\}, \overline{\mathbf{w}}) = \lambda^2 \left\langle \overline{\mathbf{w}} (\mathbf{\Psi}\mathbf{\Psi}^\top + \lambda \mathbf{I})^{-1} \mathbf{\Lambda} (\mathbf{\Psi}\mathbf{\Psi}^\top + \lambda \mathbf{I})^{-1} \overline{\mathbf{w}} \right\rangle. \quad \text{(SI.19)}$$

Defining

$$\mathbf{G} = \lambda \mathbf{\Lambda}^{1/2} (\mathbf{\Psi}\mathbf{\Psi}^\top + \lambda \mathbf{I})^{-1} \mathbf{\Lambda}^{1/2} = \left(\frac{1}{\lambda} \mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}^{-1} \right)^{-1}, \quad \text{(SI.20)}$$

and

$$\mathbf{D} = \mathbf{\Lambda}^{-1/2} \left\langle \overline{\mathbf{w}} \overline{\mathbf{w}}^\top \right\rangle \mathbf{\Lambda}^{-1/2}, \quad \text{(SI.21)}$$

and identifying the terms in (SI.19) with these definitions, we obtain the desired result. Then each component of the mode error is given by:

$$E_\rho = \sum_\gamma \mathbf{D}_{\rho,\gamma} \langle \mathbf{G}^2_{\gamma,\rho} \rangle \quad \text{(SI.22)}$$

□

3. Solution of the PDE Using Method of Characteristics

Here we derive the solution to the PDE in equation 17 of the main text by adapting the method used by (Sollich, 1999). We will prove both Propositions 2 and 3.

Let

$$g_\rho(p, v) \equiv \left\langle \tilde{\mathbf{G}}(p, v)_{\rho\rho} \right\rangle, \quad \text{(SI.23)}$$

and

$$t(p, v) \equiv \mathrm{Tr} \langle \mathbf{G}(p, v) \rangle = \sum_\rho g_\rho(p, v). \quad \text{(SI.24)}$$

It follows from equation 17 that t obeys the PDE

$$\frac{\partial t(p, v)}{\partial p} = \frac{1}{\lambda + t} \frac{\partial t(p, v)}{\partial v}, \quad \text{(SI.25)}$$

with an initial condition $t(0, v) = \mathrm{Tr}(\mathbf{\Lambda}^{-1} + v\mathbf{I})^{-1}$. The solution to first order PDEs of the form is given by the method of characteristics (Arfken, 1985), which we describe below, and prove Proposition 2.

Proof of Proposition 2. The solution to (SI.25) is a surface $(t, p, v) \subset \mathbb{R}^3$ that passes through the line $(\mathrm{Tr}(\mathbf{\Lambda}^{-1} + v\mathbf{I})^{-1}, 0, v)$ and satisfies the PDE at all points. The tangent plane to the solution surface at a point (t, p, v) is $\mathrm{span}\{(\frac{\partial t}{\partial p}, 1, 0), (\frac{\partial t}{\partial v}, 0, 1)\}$. Therefore a vector $\mathbf{a} = (a_t, a_p, a_v) \in \mathbb{R}^3$ normal to the solution surface must satisfy

$$a_t \frac{\partial t}{\partial p} + a_p = 0, \quad a_t \frac{\partial t}{\partial v} + a_v = 0.$$

One such normal vector is $(-1, \frac{\partial t}{\partial p}, \frac{\partial t}{\partial v})$.

The PDE can be written as a dot product involving this normal vector,

$$\left(-1, \frac{\partial t}{\partial p}, \frac{\partial t}{\partial v} \right) \cdot \left(0, 1, -\frac{1}{\lambda + t} \right) = 0, \quad \text{(SI.26)}$$

demonstrating that $(0, 1, -\frac{1}{\lambda+t})$ is tangent to the solution surface. This allows us to parameterize one dimensional curves along the solution in these tangent directions. Such curves are known as characteristics. Introducing a parameter $s \in \mathbb{R}$ that varies along the one dimensional characteristic curves, we get

$$\frac{dt}{ds} = 0, \quad \frac{dp}{ds} = 1, \quad \frac{dv}{ds} = -\frac{1}{\lambda + t}. \quad \text{(SI.27)}$$

The first of these equations indicate that t is constant along each characteristic curve. Integrating along the parameter, $p = s + p_0$ and $v = -\frac{s}{\lambda+t} + v_0$ where p_0 is the value of p when $s = 0$ and v_0 is the value of v at $s = 0$. Without loss

of generality, take $p_0 = 0$ so that $s = p$. At $s = 0$, we have our initial condition

$$t(0, v) = \mathrm{Tr}\left(\mathbf{\Lambda}^{-1} + v_0\mathbf{I}\right)^{-1}. \tag{SI.28}$$

Since t takes on the same value for each characteristic

$$t(p, v) = \mathrm{Tr}\left(\mathbf{\Lambda}^{-1} + \left(v + \frac{p}{\lambda + t(p, v)}\right)\mathbf{I}\right)^{-1}, \tag{SI.29}$$

which gives an implicit solution for $t(p, v)$. Now that we have solved for $t(p, v)$, remembering (SI.24), we may write

$$g_\rho(p, v) = \left(\frac{1}{\lambda_\rho} + v + \frac{p}{\lambda + t(p, v)}\right)^{-1}. \tag{SI.30}$$

This equation proves Proposition 2 of the main text. □

Next, we compute the modal generalization errors E_ρ and prove Proposition 3.

Proof of Proposition 3. Computing generalization error of kernel regression requires the differentiation with respect to v at $v = 0$ (eq.s (11) and (16) of main text). Since $\left\langle \mathbf{G}^2 \right\rangle$ is diagonal, the mode errors only depend on the diagonals of $\mathbf{D}$ and on $\left\langle \mathbf{G}^2_{\rho,\rho} \right\rangle = -\frac{\partial g_\rho}{\partial v}|_{v=0}$:

$$E_\rho = \sum_\gamma \mathbf{D}_{\rho,\gamma} \left\langle \mathbf{G}^2_{\gamma,\rho} \right\rangle = -\frac{\langle \overline{w}^2_\rho \rangle}{\lambda_\rho} \left.\frac{\partial g_\rho}{\partial v}\right|_{v=0}. \tag{SI.31}$$

We proceed with calculating the derivative in the above equation.

$$\frac{\partial g_\rho(p, 0)}{\partial v} = -\left(\frac{1}{\lambda_\rho} + \frac{p}{\lambda + t(p, 0)}\right)^{-2} \times \left(1 - \frac{p}{(\lambda + t)^2}\frac{\partial t(p, 0)}{\partial v}\right). \tag{SI.32}$$

We need to calculate $\frac{\partial t(p,v)}{\partial v}|_{v=0}$

$$\frac{\partial t(p, 0)}{\partial v} = -\gamma\left(1 - \frac{p}{(\lambda + t)^2}\frac{\partial t(p, 0)}{\partial v}\right), \tag{SI.33}$$

where

$$\gamma \equiv \sum_\rho \left(\frac{1}{\lambda_\rho} + \frac{p}{\lambda + t(p, 0)}\right)^{-2}. \tag{SI.34}$$

Solving for the derivative, we get

$$\frac{\partial t(p, 0)}{\partial v} = -\frac{\gamma}{1 - \gamma\frac{p}{(\lambda+t)^2}}, \tag{SI.35}$$

and

$$\frac{\partial g_\rho(p, 0)}{\partial v} = -\left(\frac{1}{\lambda_\rho} + \frac{p}{\lambda + t}\right)^{-2}\left(1 - \frac{\gamma p}{(\lambda + t)^2}\right)^{-1}. \tag{SI.36}$$

The error in mode ρ is therefore

$$E_\rho = \frac{\langle \overline{w}^2_\rho \rangle}{\lambda_\rho}\left(\frac{1}{\lambda_\rho} + \frac{p}{\lambda + t(p)}\right)^{-2}\left(1 - \frac{p\gamma(p)}{(\lambda + t(p))^2}\right)^{-1}, \tag{SI.37}$$

so it suffices to numerically solve for $t(p, 0)$ to recover predictions of the mode errors. Equations (SI.29) (evaluated at $v = 0$), (SI.34) and (SI.37) collectively prove Proposition 3. □

4. Learning Curve for Power Law Spectra

For $\lambda > 0$, the mode errors asymptotically satisfy $E_\rho \sim \mathcal{O}(p^{-2})$ since $\frac{p}{\lambda+t} \sim \frac{p}{\lambda}$ and $\frac{(\lambda+t)^2}{(\lambda+t)^2-\gamma p} \sim \mathcal{O}_p(1)$ (see below). Although each mode error decays asymptotically like p^{-2}, the total generalization error can have nontrivial scaling with p that depends on both the kernel and the target function.

To illustrate the dependence of the learning curves on the choice of kernel and target function, we consider a case where both have power law spectra. Specifically, we assume that $\lambda_\rho = \rho^{-b}$ and $a^2_\rho \equiv \overline{w}^2_\rho\lambda_\rho = \rho^{-a}$ for $\rho = 1, 2, ...$ We introduce the variable $z = t+\lambda$ to simplify the computations below. We further approximate the sums over modes with integrals

$$E_g \approx \frac{z^2}{z^2 - p\gamma}\int_1^\infty \frac{d\rho\, \rho^{-a}}{\left(\frac{p}{z}\rho^{-b} + 1\right)^2}. \tag{SI.38}$$

We use the same approximation technique to study the behavior of $z(p)$

$$z = \lambda + \frac{z}{p}\int_1^\infty \frac{d\rho}{1 + \frac{z}{p}\rho^b} = \lambda + \left(\frac{z}{p}\right)^{1-\frac{1}{b}}\int_{(z/p)^{1/b}}^\infty \frac{du}{1 + u^b}$$
$$= \lambda + \left(\frac{z}{p}\right)^{1-\frac{1}{b}} F(b, p, z), \tag{SI.39}$$

where $F(b, p, z) = \int_{(z/p)^{1/b}}^\infty \frac{du}{1+u^b}$. If $p \gg \lambda^{-1/(b-1)}$ then $z \approx \lambda$, otherwise $z \approx p^{1-b}F(b, p, z)^b$. Further, the scaling $z \sim \mathcal{O}(p^{1-b})$ is self-consistent since the lower endpoint of integration $(z/p)^{1/b} \sim p^{-1} \to 0$ so $F(b, z, p)$ approaches a constant $F(b)$ for $p \to \infty$

$$z \sim p^{1-b}F(b)^b\ ,\ F(b, z, p) \sim F(b) = \int_0^\infty \frac{du}{1 + u^b}. \tag{SI.40}$$

We similarly find that $p\gamma(p) \sim \mathcal{O}(p^{2-2b})$ if $p \ll \lambda^{-1/(b-1)}$. The mode-independent prefactor is approximately constant $\frac{z^2}{z^2-\gamma p} \sim \mathcal{O}_p(1)$.

We can use all of these facts to identify scalings of E_g. We

will first consider the case where $p \ll \lambda^{-1/(b-1)}$:

$$\begin{aligned} E_g &\sim \int_1^\infty \frac{d\rho \rho^{-a}}{(p^b \rho^{-b}+1)^2} \\ &\approx p^{-2b}\int_1^p d\rho \rho^{-a+2b} + \int_p^\infty d\rho \rho^{-a} \\ &= \frac{1}{a-1-2b}p^{-2b} + \frac{2b}{(a-1)(2b+1-a)}p^{-(a-1)}. \end{aligned} \tag{SI.41}$$

If $2b > a-1$ then the second term dominates, indicating that higher frequency modes $k > p$ provide a greater contribution to the error due to the slow decay in the target power. In this case $E_g \sim p^{-(a-1)}$. If, on the other hand, $2b < a-1$ then lower frequency modes $k < p$ dominate the error and $E_g \sim p^{-2b}$.

Now, suppose that $p > \lambda^{-1/(b-1)}$. In this regime

$$\begin{aligned} E_g &\sim \int_1^\infty \frac{d\rho \rho^{-a}}{(\frac{p}{\lambda}\rho^{-b}+1)^2} \\ &\approx \frac{\lambda^2}{p^2}\int_1^{(p/\lambda)^{1/b}} d\rho\, \rho^{2b-a} + \int_{(p/\lambda)^{1/b}}^\infty d\rho \rho^{-a} \\ &= \frac{\lambda^2}{p^2}\frac{1}{2b-a+1}\left[\left(\frac{p}{\lambda}\right)^{(2b-a+1)/b} - 1\right] \\ &\quad + \frac{1}{a-1}\left(\frac{p}{\lambda}\right)^{(1-a)/b}. \end{aligned} \tag{SI.42}$$

Here there are two possible scalings. If $2b > a-1$ then $E_g \sim p^{-(a-1)/b}$ while $2b < a-1$ implies $E_g \sim p^{-2}$.

So the total error scales like

$$\begin{aligned} E_g &\sim p^{-\min\{a-1,2b\}}, \qquad p < \lambda^{-1/(b-1)} \\ E_g &\sim p^{-\min\{a-1,2b\}/b}, \qquad p > \lambda^{-1/(b-1)}. \end{aligned} \tag{SI.43}$$

A verification of this scaling is provided in Figure SI.1, which shows the behavior of z and E_g in these two regimes. When the explicit regularization is low (or zero) ($p < \lambda^{-1/(b-1)}$), our equations reproduce the power law scalings derived with Fourier analysis in (Spigler et al., 2019)[2].

The slower asymptotic decays in generalization error when explicit regularization λ is large relative to the sample size indicates that explicit regularization hurts performance. The decay exponents also indicate that the RKHS eigenspectrum should decay with exponent at least as large as $b^* > \frac{a-1}{2}$ for optimal asymptotics. Kernels with slow decays in their RKHS spectra induce larger errors.

[2]We note that in a recent version of their paper, Spigler et al. (2019) used our formalism to independently derive the scalings in (SI.43) for the ridgeless ($\lambda = 0$) case. Our calculation in an earlier preprint had missed the possible $\sim p^{-2b}$ and $\sim p^{-2}$ scalings, which we corrected after their paper.

5. Replica Calculation

In this section, we present the replica trick and the saddle-point approximation summarized in main text Section 2.3. Our goal is to show that the continuous approximation of the main paper and previous section can be interpreted as a finite size saddle-point approximation to the replicated system under a replica symmetry ansatz. We will present a detailed treatment of the thermodynamic limit and the replica symmetric ansatz in a different paper.

Let $\tilde{\mathbf{G}}(p,v) = \left(\frac{1}{\lambda}\mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}^{-1} + v\mathbf{I}\right)^{-1}$. To obtain the average elements $\left\langle \tilde{\mathbf{G}}(p,v)_{\rho,\gamma} \right\rangle$ we will use a Gaussian integral representation of the matrix inverse

$$\begin{aligned} &\left\langle \tilde{\mathbf{G}}(p,v)_{\rho,\gamma} \right\rangle \\ &= \frac{\partial^2}{\partial h_\rho \partial h_\gamma}\left\langle \frac{1}{Z}\int d\mathbf{u} e^{-\frac{1}{2}\mathbf{u}(\frac{1}{\lambda}\mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}^{-1} + v\mathbf{I})\mathbf{u} + \mathbf{h}\cdot\mathbf{u}} \right\rangle_{\mathbf{\Phi}}, \end{aligned} \tag{SI.44}$$

where

$$Z = \int d\mathbf{u}\, e^{-\frac{1}{2}\mathbf{u}(\frac{1}{\lambda}\mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}^{-1} + v\mathbf{I})\mathbf{u}}, \tag{SI.45}$$

and make use of the identity $Z^{-1} = \lim_{n\to 0} Z^{n-1}$ to rewrite the entire average in the form

$$R(\mathbf{h}) = \int \prod_{a=1}^n d\mathbf{u}^a \left\langle e^{-\frac{1}{2}\sum_a \mathbf{u}^a(\frac{1}{\lambda}\mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}^{-1} + v\mathbf{I})\mathbf{u}^a + \mathbf{h}\cdot\mathbf{u}^{(1)}} \right\rangle \tag{SI.46}$$

with the identification that

$$\left\langle \tilde{\mathbf{G}}(p,v)_{\rho,\gamma} \right\rangle = \frac{\partial^2}{\partial h_\rho \partial h_\gamma}\lim_{n\to 0} R(\mathbf{h})|_{\mathbf{h}=0}. \tag{SI.47}$$

Following the replica method from the physics of disordered systems, we will first restrict ourselves to integer n and then analytically continue the resulting expressions to take the limit of $n \to 0$.

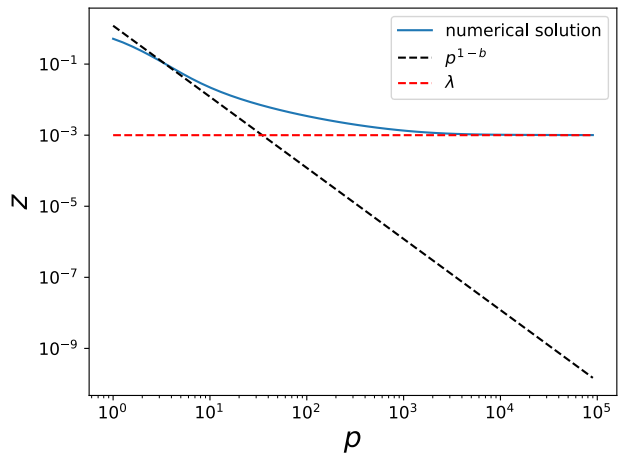
Averaging over the quenched disorder (dataset) with the assumption that the residual error $(\mathbf{w} - \overline{\mathbf{w}}) \cdot \mathbf{\Psi}(\mathbf{x}_i)$ is a Gaussian process, we find

$$\left\langle e^{-\frac{1}{2\lambda}\sum_a \mathbf{u}^a \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{u}^a} \right\rangle = e^{-\frac{p}{2}\log\det(\mathbf{I} + \frac{1}{\lambda}\mathbf{Q})}, \tag{SI.48}$$

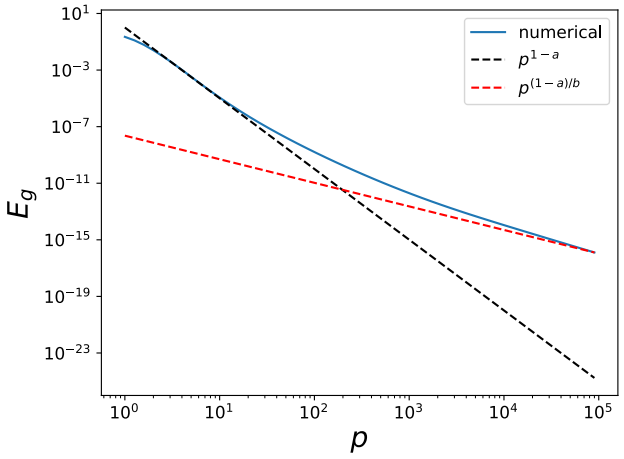
where order parameters $Q_{ab} = \mathbf{u}^a \cdot \mathbf{u}^b$ have been introduced.

To enforce the definition of these order parameters, Dirac delta functions are inserted into the expression for R. We then represent each delta function as a Fourier integral so that integrals over $\mathbf{u}^a$ can be computed

$$\delta(Q_{ab} - \mathbf{u}^a \cdot \mathbf{u}^b) = \int d\hat{Q}_{ab} e^{iQ_{ab}\hat{Q}_{ab} - i\hat{Q}_{ab}\mathbf{u}^a\cdot\mathbf{u}^b}. \tag{SI.49}$$



(a) $z = t + \lambda$



(b) $E_g(p)$

Figure SI.1. Approximate scaling of learning curve for spectra that decay as power laws $\lambda_k \sim k^{-b}$ and $a_k^2 \equiv \overline{w}_k^2 \lambda_k = k^{-a}$. Figure (a) shows a comparison of the numerical solution to the implicit equation for $t + \lambda$ as a function of p and its comparison to approximate scalings. There are two regimes which are separated by $p \approx \lambda^{-1/(b-1)}$. For small p, $z \sim p^{1-b}$ but for large p, $z \sim \lambda$. The total generalization error is shown in (b) which scales like p^{1-a} for small p and $p^{(1-a)/b}$ for large p.

After inserting delta functions to enforce order parameter definitions, we are left with integrals over the thermal degrees of freedom

$$\int \prod_{a=1}^{n} d\mathbf{u}^a e^{-\frac{1}{2}\sum_a \mathbf{u}^a \mathbf{\Lambda}^{-1} \mathbf{u}^a - i \sum_{ab} \hat{Q}_{ab} \mathbf{u}^a \mathbf{u}^b + \mathbf{u}^{(1)} \mathbf{h}} = e^{-\frac{1}{2}\sum_\rho \log\det(\frac{1}{\lambda_\rho}\mathbf{I} + 2i\hat{\mathbf{Q}}) + \frac{1}{2}\sum_\rho h_\rho^2 (\frac{1}{\lambda_\rho}\mathbf{I} + 2i\hat{\mathbf{Q}})^{-1}_{11}}. \quad \text{(SI.50)}$$

We now make a replica symmetric ansatz $Q_{ab} = q\delta_{ab} + q_0$ and $2i\hat{Q}_{ab} = \hat{q}\delta_{ab} + \hat{q}_0$. Under this ansatz $R(\mathbf{h})$ can be rewritten as

$$R(\mathbf{h}) = \int dq d\hat{q} dd\hat{q} d\hat{q}_0 e^{-pn\mathcal{F}(q,q_0,\hat{q},\hat{q}_0)} e^{\frac{1}{2}\sum_\rho h_\rho^2 (\frac{1}{\lambda_\rho}\mathbf{I} + 2i\hat{\mathbf{Q}})^{-1}_{11}}, \quad \text{(SI.51)}$$

where the free energy is

$$\begin{aligned} 2p\mathcal{F}(q, q_0, \hat{q}, \hat{q}_0) =& p \log\left(1 + \frac{q}{\lambda}\right) + p\frac{q_0}{\lambda + q} + v(q + q_0) \\ & - (q + q_0)(\hat{q} + \hat{q}_0) + q_0\hat{q}_0 \\ & + \sum_\rho \left[\log\left(\frac{1}{\lambda_\rho} + \hat{q}\right) + \frac{\hat{q}_0}{\frac{1}{\lambda_\rho} + \hat{q}}\right]. \end{aligned} \quad \text{(SI.52)}$$

In the limit $p \to \infty$, $R(\mathbf{h})$ is dominated by the saddle point of the free energy where $\nabla\mathcal{F}(q, \hat{q}, q_0, \hat{q}_0) = 0$. The saddle point equations are

$$\begin{aligned} \hat{q}^* &= \frac{p}{q^* + \lambda} + v, \\ q^* &= \sum_\rho \frac{1}{\frac{1}{\lambda_\rho} + \hat{q}^*} = \sum_\rho \frac{1}{\frac{1}{\lambda_\rho} + v + \frac{p}{q^* + \lambda}}, \\ q_0^* &= \hat{q}_0^* = 0. \end{aligned} \quad \text{(SI.53)}$$

We see that q^* is exactly equivalent to $t(p, v)$ defined in SI.29 for the continuous approximation. Under the saddle point approximation we find

$$R(\mathbf{h}) \approx e^{-np\mathcal{F}(q^*, q_0^*, \hat{q}^*, \hat{q}_0^*)} e^{\frac{1}{2}\sum_\rho h_\rho^2 \frac{1}{\frac{1}{\lambda_\rho} + \hat{q}^*}}. \quad \text{(SI.54)}$$

Taking the $n \to 0$ limit as promised, we obtain the normalized average

$$\tilde{R}(\mathbf{h}) \equiv \lim_{n \to 0} R(\mathbf{h}) = e^{\frac{1}{2}\sum_\rho h_\rho^2 \frac{1}{\frac{1}{\lambda_\rho} + \hat{q}^*}}, \quad \text{(SI.55)}$$

so that the matrix elements are

$$\begin{aligned} \left\langle \tilde{\mathbf{G}}(p, v)_{\rho,\gamma} \right\rangle &= \frac{\partial^2}{\partial h_\rho \partial h_\gamma} \tilde{R}(\mathbf{h})|_{\mathbf{h}=0} = \frac{\delta_{\rho,\gamma}}{\frac{1}{\lambda_\rho} + v + \frac{p}{\lambda + q^*}}, \\ q^* &= \sum_\rho \frac{1}{\frac{1}{\lambda_\rho} + v + \frac{p}{\lambda + q^*}}. \end{aligned} \quad \text{(SI.56)}$$

Using our formula for the mode errors, we find

$$\begin{aligned} E_\rho &= \sum_\gamma \mathbf{D}_{\rho,\gamma} \left\langle \tilde{\mathbf{G}}(p,v)^2_{\gamma,\rho} \right\rangle \\ &= -\mathbf{D}_{\rho,\rho} \frac{\partial}{\partial v} \left\langle \tilde{\mathbf{G}}(p,v)_{\rho,\rho} \right\rangle |_{v=0} \\ &= \frac{\langle \overline{w}^2_\rho \rangle}{\lambda_\rho} \frac{(\lambda+q^*)^2}{(\lambda+q^*)^2-\gamma p} \left(\frac{1}{\lambda_\rho} + \frac{p}{\lambda+q^*} \right)^{-2}, \end{aligned} \tag{SI.57}$$

consistent with our result from the continuous approximation.

6. Spectral Dependence of Learning Curves

We want to calculate how different mode errors change as we add one more sample. We study:

$$\frac{1}{2}\frac{d}{dp} \log \frac{E_\rho}{E_\gamma}, \tag{SI.58}$$

where E_ρ is given by eq. (21). Evaluating the derivative, we find:

$$\begin{aligned} &\frac{1}{2}\frac{d}{dp} \log \left(\frac{E_\rho}{E_\gamma} \right) \\ &= -\left(\frac{1}{\frac{1}{\lambda_\rho} + \frac{p}{\lambda+t}} - \frac{1}{\frac{1}{\lambda_\gamma} + \frac{p}{\lambda+t}} \right) \frac{\partial}{\partial p} \left(\frac{p}{\lambda+t} \right). \end{aligned} \tag{SI.59}$$

Using eq. (22),

$$\begin{aligned} \frac{\partial t}{\partial p} &= -\frac{\partial}{\partial p} \left(\frac{p}{\lambda+t} \right) \sum_\rho \left(\frac{1}{\lambda_\gamma} + \frac{p}{\lambda+t} \right)^{-2} \\ &= -\gamma \frac{\partial}{\partial p} \left(\frac{p}{\lambda+t} \right), \end{aligned} \tag{SI.60}$$

where we identified the sum with γ. Inserting this, we obtain:

$$\frac{1}{2}\frac{d}{dp} \log \left(\frac{E_\rho}{E_\gamma} \right) = \left[\frac{1}{\frac{1}{\lambda_\rho} + \frac{p}{\lambda+t}} - \frac{1}{\frac{1}{\lambda_\gamma} + \frac{p}{\lambda+t}} \right] \frac{1}{\gamma} \frac{\partial t}{\partial p}. \tag{SI.61}$$

Finally, solving for $\partial t/\partial p$ from (SI.60), we get:

$$\frac{\partial t}{\partial p} = -\frac{1}{\lambda+t} \frac{(\lambda+t)^2\gamma}{(\lambda+t)^2-p\gamma} = -\frac{1}{\lambda+t} \mathrm{Tr}\big(\mathbf{G}^2\big), \tag{SI.62}$$

proving that $\partial t/\partial p < 0$. Taking $\lambda_\gamma > \lambda_\rho$ without loss of generality, it follows that

$$\frac{d}{dp} \log \left(\frac{E_\rho}{E_\gamma} \right) > 0 \;\Rightarrow\; \frac{d}{dp} \log E_\rho > \frac{d}{dp} \log E_\gamma. \tag{SI.63}$$

7. Spherical Harmonics

Let $-\Delta$ represent the Laplace-Beltrami operator in $\mathbb{R}^d$. Spherical harmonics $\{Y_{km}\}$ in dimension d are harmonic ($-\Delta Y_{km}(\mathbf{x}) = 0$), homogeneous ($Y_{km}(t\mathbf{x}) = t^k Y_{km}(\mathbf{x})$) polynomials that are orthonormal with respect to the uniform measure on $\mathbb{S}^{d-1}$ (Efthimiou & Frye, 2014; Dai & Xu, 2013). The number of spherical harmonics of degree k in dimension d denoted by $N(d,k)$ is

$$N(d,k) = \frac{2k+d-2}{k} \binom{k+d-3}{k-1}. \tag{SI.64}$$

The Laplace Beltrami Operator can be decomposed into the radial and angular parts, allowing

$$-\Delta = -\Delta_r - \Delta_{\mathbb{S}^{d-1}} \tag{SI.65}$$

Using this decomposition, the spherical harmonics are eigenfunctions of the surface Laplacian

$$-\Delta_{\mathbb{S}^{d-1}} Y_{km}(\mathbf{x}) = k(k+d-2) Y_{km}(\mathbf{x}). \tag{SI.66}$$

The spherical harmonics are related to the Gegenbauer polynomials $\{Q_k\}$, which are orthogonal with respect to the measure $d\tau(z) = (1-z^2)^{(d-3)/2} dz$ of inner products $z = \mathbf{x}^\top \mathbf{x}'$ of uniformly sampled pairs on the sphere $\mathbf{x}, \mathbf{x}' \sim \mathbb{S}^{d-1}$. The Gegenbauer polynomials can be constructed with the Gram-Schmidt procedure and have the following properties

$$\begin{aligned} Q_k(\mathbf{x}^\top \mathbf{x}') &= \frac{1}{N(d,k)} \sum_{m=1}^{N(d,k)} Y_{km}(\mathbf{x}) Y_{km}(\mathbf{x}'), \\ \int_{-1}^{1} Q_k(z) Q_\ell(z) d\tau(z) &= \frac{\omega_{d-1}}{\omega_{d-2}} \frac{\delta_{k,\ell}}{N(d,k)}, \end{aligned} \tag{SI.67}$$

where $\omega_{d-1} = \frac{\pi^{d/2}}{\Gamma(d/2)}$ is the surface area of $\mathbb{S}^{d-1}$.

8. Decomposition of Dot Product Kernels on $\mathbb{S}^{d-1}$

For inputs sampled from the uniform measure on $\mathbb{S}^{d-1}$, dot product kernels can be decomposed into Gegenbauer polynomials introduced in SI Section 7.

Let $K(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x}^\top \mathbf{x}')$. The kernel's orthogonal decomposition is

$$\begin{aligned} \kappa(z) &= \sum_{k=0}^{\infty} \lambda_k N(d,k) Q_k(z), \\ \lambda_k &= \frac{\omega_{d-2}}{\omega_{d-1}} \int_{-1}^{1} \kappa(z) Q_k(z) d\tau(z). \end{aligned} \tag{SI.68}$$

To numerically calculate the kernel eigenvalues of κ, we use Gauss-Gegenbauer quadrature (Abramowitz & Stegun,

1972) for the measure $d\tau(z)$ so that for a quadrature scheme of order r

$$\int_{-1}^{1} \kappa(z) Q_k(z) d\tau(z) \approx \sum_{i=1}^{r} w_i Q_k(z_i) \kappa(z_i), \quad \text{(SI.69)}$$

where z_i are the r roots of $Q_r(z)$ and the weights w_i are chosen with

$$w_i = \frac{\Gamma(r+\alpha+1)^2}{\Gamma(r+2\alpha+1)} \frac{2^{2r+2\alpha+1} r!}{V_r'(z_i) V_{r+1}(z_i)}, \quad \text{(SI.70)}$$

where

$$V_r(z) = 2^r r! (-1)^r Q_r(z) \quad \text{(SI.71)}$$

For our calculations we take $r = 1000$.

9. Frequency Dependence of Learning Curves in $d \to \infty$ Limit

Here, we consider an informative limit where the number of input data dimension, d, goes to infinity.

Denoting the index $\rho = (k, m)$, we can write mode error (SI.37), after some rearranging, as:

$$E_{km} = \frac{(\lambda+t)^2}{1 - \frac{p\gamma}{(\lambda+t)^2}} \frac{\lambda_k \langle \overline{w}_{km}^2 \rangle}{(\lambda + t + p\lambda_k)^2}, \quad \text{(SI.72)}$$

where t and γ, after performing the sum over degenerate indices, are:

$$t = \sum_m \frac{N(d,m)(\lambda+t)\lambda_m}{\lambda + t + p\lambda_m},$$
$$\gamma = \sum_m \frac{N(d,m)(\lambda+t)^2 \lambda_m^2}{(\lambda + t + p\lambda_m)^2}. \quad \text{(SI.73)}$$

In the limit $d \to \infty$, the degeneracy factor (SI.64) approaches to $N(d,k) \sim \mathcal{O}(d^k)$. We note that for dot-product kernels λ_k scales with d as $\lambda_k \sim d^{-k}$ (Smola et al., 2001) (Figure 1), which leads us to define the $\mathcal{O}(1)$ parameter $\bar{\lambda}_k = d^k \lambda_k$. Plugging these in, we get:

$$E_{km}(g_k) = \frac{d^{-k}(t+\lambda)^2}{1-\tilde{\gamma}} \frac{\bar{\lambda}_k \langle \bar{w}_{km}^2 \rangle}{\left(t + \lambda + g_k \bar{\lambda}_k\right)^2}$$
$$t = \sum_m \frac{(t+\lambda)\bar{\lambda}_m}{t + \lambda + g_m \bar{\lambda}_m},$$
$$\tilde{\gamma} = \sum_m \frac{g_m \bar{\lambda}_m^2}{\left(t + \lambda + g_m \bar{\lambda}_m\right)^2}, \quad \text{(SI.74)}$$

where $g_k = p/d^k$ is the ratio of sample size to the degeneracy. Furthermore, we want to calculate the ratio $E_{km}(p)/E_{km}(0)$ to probe how much the mode errors move from their initial value:

$$\frac{E_{km}(p)}{E_{km}(0)} = \frac{1}{1-\tilde{\gamma}} \frac{1}{\left(1 + \frac{g_k \bar{\lambda}_k}{t+\lambda}\right)^2} \quad \text{(SI.75)}$$

Let us consider an integer l such that the scaling $P = \alpha d^l$ holds. This leads to three different asymptotic behavior of g_ks:

$$\begin{aligned} g_k &\sim \mathcal{O}(d^{l-k}) \gg \mathcal{O}(1), && k < l \\ g_k &= \alpha \sim \mathcal{O}(1), && k = l \\ g_k &\sim \mathcal{O}(d^{l-k}) \ll \mathcal{O}(1), && k > l \end{aligned} \quad \text{(SI.76)}$$

If we assume $t \sim \mathcal{O}(1)$, we get an asymptotically consistent set of equations:

$$\begin{aligned} t &\approx \sum_{m>l} \bar{\lambda}_m + a(\alpha, t, \lambda, \bar{\lambda}_l) \sim \mathcal{O}(1), \\ \tilde{\gamma} &\approx b(\alpha, t, \lambda, \bar{\lambda}_l) \sim \mathcal{O}(1), \end{aligned} \quad \text{(SI.77)}$$

where a and b are the l^{th} terms in the sums in t and $\tilde{\gamma}$, respectively, and are given by:

$$a(\alpha, t, \lambda, \bar{\lambda}_l) = \frac{(t+\lambda)\bar{\lambda}_l}{t + \lambda + \alpha\bar{\lambda}_l},$$
$$b(\alpha, t, \lambda, \bar{\lambda}_l) = \frac{\alpha \bar{\lambda}_l^2}{\left(t + \lambda + \alpha \bar{\lambda}_l\right)^2} \quad \text{(SI.78)}$$

Then using (SI.75), (SI.76) and (SI.77), we find the errors associated to different modes as:

$$\begin{aligned} k < l, &\quad \frac{E_{km}(\alpha)}{E_{km}(0)} \sim \mathcal{O}(d^{2(k-l)}) \approx 0, \\ k > l, &\quad \frac{E_{km}(\alpha)}{E_{km}(0)} \approx \frac{1}{1-\tilde{\gamma}(\alpha)}, \\ k = l, &\quad \frac{E_{km}(\alpha)}{E_{km}(0)} = s(\alpha) \sim \mathcal{O}(1), \end{aligned} \quad \text{(SI.79)}$$

where $s(\alpha)$ is given by:

$$s(\alpha) = \frac{1}{1-\tilde{\gamma}(\alpha)} \frac{1}{\left(1 + \alpha \frac{\bar{\lambda}_l}{t+\lambda}\right)^2}. \quad \text{(SI.80)}$$

Note that $\lim_{\alpha\to 0} \tilde{\gamma}(\alpha) = \lim_{\alpha\to\infty} \tilde{\gamma}(\alpha) = 0$ and non-zero in between. Then, for large α, in the limit we are considering

$$\begin{aligned} k < l, &\quad \frac{E_{km}(\alpha)}{E_{km}(0)} \approx 0, \\ k > l, &\quad \frac{E_{km}(\alpha)}{E_{km}(0)} \approx 1, \\ k = l, &\quad \frac{E_{km}(\alpha)}{E_{km}(0)} \approx \frac{(\lambda + \sum_{m>l} \bar{\lambda}_m)^2}{\bar{\lambda}_l^2} \frac{1}{\alpha^2}. \end{aligned} \quad \text{(SI.81)}$$

10. Neural Tangent Kernel

The neural tangent kernel is

$$K_{\text{NTK}}(\mathbf{x}, \mathbf{x}') = \sum_i \left\langle \frac{\partial f_\theta(\mathbf{x})}{\partial \theta_i} \frac{\partial f_\theta(\mathbf{x}')}{\partial \theta_i} \right\rangle_\theta. \quad \text{(SI.82)}$$

For a neural network, it is convenient to compute this recursively in terms of the Neural Network Gaussian Process (NNGP) kernel which corresponds to only training the readout weights from the final layer (Jacot et al., 2018; Arora et al., 2019). We will restrict our attention to networks with zero bias and nonlinear activation function σ. Then

$$\begin{aligned} &K^{(1)}_{NTK}(\mathbf{x},\mathbf{x}') \\ &\quad = K^{(1)}_{NNGP}(\mathbf{x},\mathbf{x}') \\ &K^{(2)}_{NTK}(\mathbf{x},\mathbf{x}') \\ &\quad = K^{(2)}_{NNGP}(\mathbf{x},\mathbf{x}') + K^{(1)}_{NTK}(\mathbf{x},\mathbf{x}')\dot{K}^{(2)}(\mathbf{x},\mathbf{x}') \\ &\quad \dots \\ &K^{(L)}_{NTK}(\mathbf{x},\mathbf{x}') \\ &\quad = K^{(L)}_{NNGP}(\mathbf{x},\mathbf{x}') + K^{(L-1)}_{NTK}(\mathbf{x},\mathbf{x}')\dot{K}^{(L)}(\mathbf{x},\mathbf{x}'), \end{aligned} \tag{SI.83}$$

where

$$\begin{aligned} &K^{(L)}_{NNGP}(\mathbf{x},\mathbf{x}') = \mathbb{E}_{(\alpha,\beta)\sim p^{(L-1)}_{\mathbf{x},\mathbf{x}'}}\sigma(\alpha)\sigma(\beta), \\ &\dot{K}^{(L)}(\mathbf{x},\mathbf{x}') = \mathbb{E}_{(\alpha,\beta)\sim p^{(L-1)}_{\mathbf{x},\mathbf{x}'}}\dot{\sigma}(\alpha)\dot{\sigma}(\beta), \\ &p^{(L-1)}_{\mathbf{x},\mathbf{x}'} = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} K^{(L-1)}(\mathbf{x},\mathbf{x}) & K^{(L-1)}(\mathbf{x},\mathbf{x}') \\ K^{(L-1)}(\mathbf{x},\mathbf{x}') & K^{(L-1)}(\mathbf{x}',\mathbf{x}')\end{pmatrix}\right), \\ &K^{(1)}_{NNGP}(\mathbf{x},\mathbf{x}') = \mathbf{x}^\top\mathbf{x}'. \end{aligned} \tag{SI.84}$$

If σ is chosen to be the ReLU activation, then we can analytically simplify the expression. Defining the following function

$$f(z) = \arccos\left(\frac{1}{\pi}\sqrt{1-z^2} + \left(1-\frac{1}{\pi}\arccos(z)\right)z\right), \tag{SI.85}$$

we obtain

$$\begin{aligned} K^{(L)}_{NNGP}(\mathbf{x},\mathbf{x}') &= \cos\left(f^{\circ(L-1)}(\mathbf{x}^\top\mathbf{x}')\right) \\ \dot{K}_L(\mathbf{x},\mathbf{x}') &= \left(1-\frac{1}{\pi}f^{\circ(L-2)}(\mathbf{x}^\top\mathbf{x}')\right), \end{aligned} \tag{SI.86}$$

where $f^{\circ(L-1)}(z)$ is the function f composed into itself $L-1$ times.

This simplification gives an exact recursive formula to compute the kernel as a function of $z = \mathbf{x}^\top\mathbf{x}'$, which is what we use to compute the eigenspectrum with the quadrature scheme described in the previous section.

11. Spectra of Fully Connected ReLU NTK

A plot of the RKHS spectra of fully connected ReLU NTK's of varying depth is shown in Figure SI.2. As the depth increases, the spectrum becomes more white, eventually, the kernel's trace $\langle K(\mathbf{x},\mathbf{x})\rangle_{\mathbf{x}} = \sum_k \lambda_k N(d,k)$ begins to diverge. Inference with such a kernel is equivalent to learning a function with infinite variance. Constraints on the variance of derivatives $\left\langle ||\nabla^n_{\mathbb{S}^{d-1}} f(\mathbf{x})||^2\right\rangle$ correspond to more restrictive constraints on the eigenspectrum of the RKHS. Specifically, $\lambda_k N(d,k) \sim \mathcal{O}(k^{-n-1/2})$ implies that the n-th gradient has finite variance $\left\langle ||\nabla^n_{\mathbb{S}^{d-1}} f(\mathbf{x})||^2\right\rangle < \infty$.

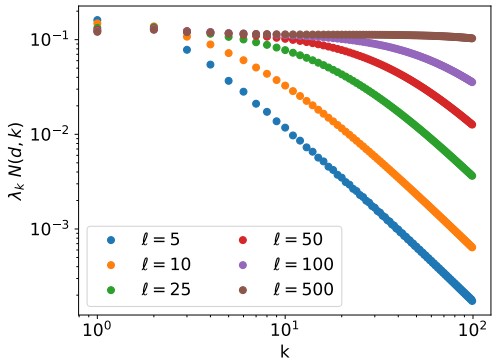


Figure SI.2. Spectrum of fully connected ReLU NTK without bias for varying depth ℓ. As the depth increases, the spectrum whitens, causing derivatives of lower order to have infinite variance. As $\ell \to \infty$, $\lambda_k N(d,k) \sim 1$ implying that the kernel becomes non-analytic at the origin.

Proof. By the representer theorem, let $f(\mathbf{x}) = \sum_{i=1}^p \alpha_i K(\mathbf{x},\mathbf{x}_i)$. By Green's theorem, the variance of the n-th derivative can be rewritten as

$$\begin{aligned} \left\langle ||\nabla^n_{\mathbb{S}^{d-1}} f(\mathbf{x})||^2\right\rangle &= \langle f(\mathbf{x})(-\Delta_{\mathbb{S}^{d-1}})^n f(\mathbf{x})\rangle \\ &= \sum_{kk'mm'ij} \alpha_i\alpha_j\lambda_k\lambda_{k'} Y_{km}(\mathbf{x}_i)Y_{k'm'}(\mathbf{x}_j) \\ &\quad \times \langle Y_{km}(\mathbf{x})(-\Delta_{\mathbb{S}^{d-1}})^n Y_{k'm'}(\mathbf{x})\rangle \\ &= \sum_{kij} \lambda_k^2 k^n (k+d-2)^n N(d,k)\alpha_i\alpha_j Q_k(\mathbf{x}_i^\top\mathbf{x}_j) \\ &\le Cp^2(\alpha^*)^2 \sum_k \lambda_k^2 k^n (k+d-2)^n N(d,k)^2, \end{aligned} \tag{SI.87}$$

where $\alpha^* = \max_j |\alpha_j|$ and $|Q_k(z)| \le CN(d,k)$ for a universal constant C. A sufficient condition for this sum to converge is that $\lambda_k^2 k^n (k+d-2)^n N(d,k)^2 \sim \mathcal{O}(k^{-1})$ which is equivalent to demanding $\lambda_k N(d,k) \sim \mathcal{O}(k^{-n-1/2})$ since $(k+d-2)^n \sim k^n$ as $k \to \infty$. □

12. Decomposition of Risk for Numerical Experiments

As we describe in Section 4.1 of the main text, the teacher functions for the kernel regression experiments are chosen

as

$$f^*(\mathbf{x}) = \sum_{i=1}^{p'} \overline{\alpha_i} K(\mathbf{x}, \overline{\mathbf{x}}_i), \tag{SI.88}$$

where the coefficients $\overline{\alpha_i} \sim \mathcal{B}(1/2)$ are randomly sampled from a centered Bernoulli distribution on $\{\pm 1\}$ and the points $\overline{\mathbf{x}}_i \sim p(\mathbf{x})$ are drawn from the same distribution as the training data. In general p' is not the same as the number of samples p. Choosing a function of this form is very convenient for producing theoretical predictions of mode errors as we discuss below.

12.1. Theoretical Mode Errors

Since the matrix elements $\left\langle \mathbf{G}^2_{\rho\rho} \right\rangle$ are determined completely by the kernel eigenvalues $\{\lambda_\rho\}$, it suffices to calculate the diagonal elements of $\mathbf{D}$ to find the generalization error. For the teacher function sampled in the way described above, there is a convenient expression for $\mathbf{D}_{\rho\rho}$.

The teacher function admits an expansion in the basis of kernel eigenfunctions

$$f^*(\mathbf{x}) = \sum_\rho \overline{w}_\rho \psi_\rho(\mathbf{x}). \tag{SI.89}$$

Using the Mercer decomposition of the kernel we can identify the coefficients

$$f^*(\mathbf{x}) = \sum_{i=1}^{p'} \overline{\alpha}_i K(\mathbf{x}, \overline{\mathbf{x}}_i) = \sum_\rho \Big(\sum_i \overline{\alpha_i} \psi_\rho(\overline{\mathbf{x}}_i) \Big) \psi_\rho(\mathbf{x}) \tag{SI.90}$$

Comparing each term in these two expressions, we identify the coefficient of the ρ-th eigenfunction

$$\overline{w_\rho} = \sum_i \overline{\alpha_i} \psi_\rho(\overline{\mathbf{x}}_i). \tag{SI.91}$$

We now need to compute the $D_{\rho\rho}$, by averaging $\overline{w}^2_\rho$ over all possible teachers

$$\begin{aligned} D_{\rho\rho} &= \frac{1}{\lambda_\rho} \left\langle \overline{w}^2_\rho \right\rangle = \frac{1}{\lambda_\rho} \sum_{ij} \left\langle \alpha_i \alpha_j \right\rangle \left\langle \psi_\rho(\overline{\mathbf{x}}_i) \psi_\rho(\overline{\mathbf{x}}_j) \right\rangle \\ &= \frac{1}{\lambda_\rho} \sum_i \left\langle \psi_\rho(\overline{\mathbf{x}}_i) \psi_\rho(\overline{\mathbf{x}}_i) \right\rangle = \frac{p' \lambda_\rho}{\lambda_\rho} = p', \end{aligned} \tag{SI.92}$$

since $\left\langle \psi_\rho(\mathbf{x}) \psi_\rho(\mathbf{x}) \right\rangle = \lambda_\rho$. Thus it suffices to calculate $\frac{\partial}{\partial v} g_\rho(p, v)$ for each mode and then compute mode errors with

$$E_\rho = -d_\rho \frac{\partial g_\rho(p, v)}{\partial v}|_{v=0}, \tag{SI.93}$$

where $\frac{\partial g_\rho}{\partial v}|_{v=0}$ is evaluated in terms of the numerical solution for $t(p, 0)$.

12.2. Empirical Mode Errors

By the representer theorem, we may represent the student function as $f(\mathbf{x}) = \sum_{i=1}^P \alpha_i K(\mathbf{x}, \mathbf{x}_i)$. Then, the generalization error is given by

$$\begin{aligned} E_g &= \left\langle (f(x) - f^*(x))^2 \right\rangle \\ &= \sum_{\rho\gamma} \lambda_\rho \lambda_\gamma \left(\sum_{j=1}^P \alpha_j \phi_\rho(x_j) - \sum_{i=1}^{P'} \overline{\alpha}_i \phi_\rho(\overline{x}_i) \right) \\ &\quad \left(\sum_{j=1}^P \alpha_j \phi_\gamma(x_j) - \sum_{i=1}^{P'} \overline{\alpha}_i \phi_\gamma(\overline{x}_i) \right) \left\langle \phi_\rho(x) \phi_\gamma(x) \right\rangle \\ &= \sum_\rho \lambda^2_\rho \left(\sum_{j,j'} \alpha_j \alpha_{j'} \phi_\rho(x_j) \phi_\rho(x_j) \right. \\ &\quad \left. -2 \sum_{i,j} \alpha_j \overline{\alpha}_i \phi_\rho(x_j) \phi_\rho(\overline{x}_i) + \sum_{i,i'} \overline{\alpha}_i \overline{\alpha}_{i'} \phi(\overline{x}_i) \phi(\overline{x}_{i'}) \right). \end{aligned} \tag{SI.94}$$

On the d-sphere, by defining $E_k = \sum_{m=1}^{N(d,k)} E_{km}$ we arrive at the formula

$$\begin{aligned} E_k = \lambda^2_k N(d,k) \left(\boldsymbol{\alpha}^\top Q_k(\mathbf{X}^T \mathbf{X}) \boldsymbol{\alpha} - 2 \boldsymbol{\alpha}^\top Q_k(\mathbf{X}^T \overline{\mathbf{X}}) \overline{\boldsymbol{\alpha}} \right. \\ \left. + \overline{\boldsymbol{\alpha}}^\top Q_k(\overline{\mathbf{X}}^T \overline{\mathbf{X}}) \overline{\boldsymbol{\alpha}} \right). \end{aligned} \tag{SI.95}$$

We randomly sample the $\overline{\alpha}$ variables for the teacher and fit $\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y}$ to the training data. Once these coefficients are known, we can obtain empirical mode errors.

13. Neural Network Experiments

For the "pure mode" experiments with neural networks, the target function was

$$\begin{aligned} f^*(\mathbf{x}) &= \sum_{i=1}^{P'} \overline{\alpha}_i Q_k(\mathbf{x}^\top \overline{\mathbf{x}}_i) \\ &= \sum_{m=1}^{N(d,k)} \left(\sum_{i=1}^{P'} \overline{\alpha}_i Y_{km}(\overline{\mathbf{x}}_i) \right) Y_{km}(\mathbf{x}), \end{aligned} \tag{SI.96}$$

whereas, for the composite experiment, the target function was a randomly sampled two layer neural network with ReLU activations

$$f^*(\mathbf{x}) = \overline{\mathbf{r}}^\top \sigma(\overline{\boldsymbol{\Theta}} \mathbf{x}). \tag{SI.97}$$

This target model is a special case of eq. (SI.90) so the same technology can be used to compute the theoretical learning curves. We can use a similar trick as that shown in equation (SI.92) to determine $\overline{w}_\rho$ for the NN teacher experiment. Let the Gegenbauer polynomial expansion of

$\sigma(z)$ be $\sigma(z) = \sum_{k=0}^{\infty} a_k N(d,k) Q_k(z)$. Then the mode error for mode k is $E_k = \frac{a_k^2}{\lambda_k^2} \langle g_k^2 \rangle$ where $\langle g_k^2 \rangle$ is computed with equation (SI.37).

A sample of some training error and generalization errors from pure mode experiments are provided below in Figures SI.3 and SI.4.

13.1. Hyperparameters

The choice of the number of hidden units N was based primarily on computational considerations. For two layer neural networks, the total number of parameters scales linearly with N, so to approach the overparameterized regime, we aimed to have $N \approx 10 p_{max}$ where p_{max} is the largest sample size used in our experiment. For $p_{max} = 500$, we chose $N = 4000, 10000$.

For the three and four layer networks, the number of parameters scales quadratically with N, making simulations with $N > 10^3$ computationally expensive. We chose N to give comparable training time for the 2 layer case which corresponded to $N = 500$ after experimenting with $\{100, 250, 500, 1000, 5000\}$.

We found that the learning rate needed to be quite large for the training loss to be reduced by a factor of $\approx 10^6$. For the 2 layer networks, we tried learning rates $\{10^{-3}, 10^{-2}, 1, 10, 32\}$ and found that a learning rate of 32 gave the lowest training error. For the three and four layer networks, we found that lower learning rates worked better and used learning rates in the range from $[0.5, 3]$.

14. Discrete Measure and Kernel PCA

We consider a special case of a discrete probability measure with equal mass on each point in a dataset of size $\tilde{p}$

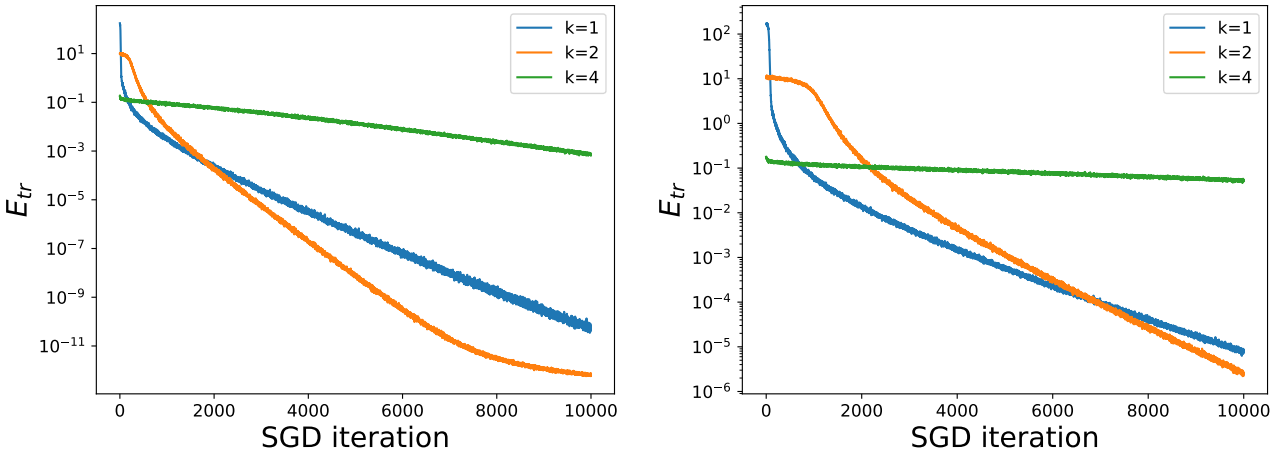
$$p(\mathbf{x}) = \frac{1}{\tilde{p}} \sum_{i=1}^{\tilde{p}} \delta(\mathbf{x} - \mathbf{x}_i). \tag{SI.98}$$

For this measure, the integral eigenvalue equation becomes

$$\begin{aligned} &\int d\mathbf{x}\, p(\mathbf{x}) K(\mathbf{x}, \mathbf{x}') \phi_\rho(\mathbf{x}) \\ &\quad = \frac{1}{\tilde{p}} \sum_{i=1}^{\tilde{p}} \int d\mathbf{x}\, \delta(\mathbf{x} - \mathbf{x}_i) K(\mathbf{x}, \mathbf{x}') \phi_\rho(\mathbf{x}) \\ &\quad = \frac{1}{\tilde{p}} \sum_{i=1}^{\tilde{p}} K(\mathbf{x}_i, \mathbf{x}') \phi_\rho(\mathbf{x}_i) = \lambda_\rho \phi_\rho(\mathbf{x}'). \end{aligned} \tag{SI.99}$$

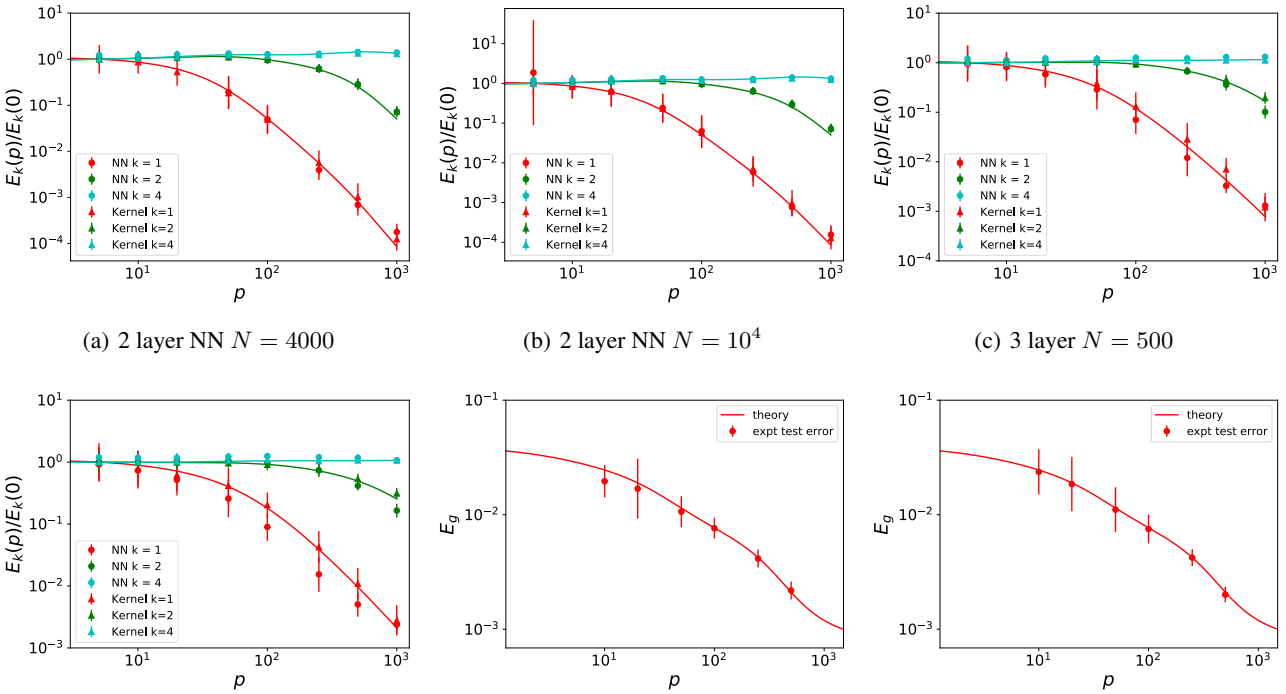
Evaluating $\mathbf{x}'$ at each of the points $\mathbf{x}_i$ in the dataset yields a matrix equation. Let $\mathbf{\Phi}_{\rho,i} = \phi_\rho(\mathbf{x}_i)$ and $\mathbf{\Lambda}_{\rho,\gamma} = \delta_{\rho,\gamma} \lambda_\rho$

$$\mathbf{K} \mathbf{\Phi}^\top = \tilde{p} \mathbf{\Phi}^\top \mathbf{\Lambda}. \tag{SI.100}$$



(a) 3 Layer Training Loss; lr =2 (b) 4 Layer Training Loss; lr = 0.5

Figure SI.3. Training error for different pure mode target functions on neural networks with 500 hidden units per hidden layer on a sample of size $p = 500$. Generally, we find that the low frequency modes have an initial rapid reduction in the training error but the higher frequencies $k \geq 4$ are trained at a slower rate.



(a) 2 layer NN $N = 4000$ (b) 2 layer NN $N = 10^4$ (c) 3 layer $N = 500$

(d) 4 layer $N = 500$ (e) 2 Layer NN Student-Teacher; $N = 2000$ (f) 2 Layer NN Student-Teacher; $N = 8000$

Figure SI.4. Learning curves for neural networks on "pure modes" and on student teacher experiments. The theory curves shown as solid lines. For the pure mode experiments, the test error for the finite width neural networks and NTK are shown with dots and triangles respectively. Logarithms are evaluated with base 10.