Failure and success of the spectral bias prediction for Laplace Kernel Ridge Regression: the case of low-dimensional data

Umberto M. Tomasini [1] **Antonio Sclocchi** [1] **Matthieu Wyart** [1]

Abstract

Recently, several theories including the replica method made predictions for the generalization error of Kernel Ridge Regression. In some regimes, they predict that the method has a 'spectral bias': decomposing the true function f^* on the eigenbasis of the kernel, it fits well the coefficients associated with the O(P) largest eigenvalues, where P is the size of the training set. This prediction works very well on benchmark data sets such as images, yet the assumptions these approaches make on the data are never satisfied in practice. To clarify when the spectral bias prediction holds, we first focus on a one-dimensional model where rigorous results are obtained and then use scaling arguments to generalize and test our findings in higher dimensions. Our predictions include the classification case $f(x)$ =sign(x_1) with a data distribution that vanishes at the decision boundary $p(x) \sim x_1^\chi$. For $\chi > 0$ and a Laplace kernel, we find that (i) there exists a cross-over ridge $\lambda^*_{d,\chi}(P) \sim P^{-\frac{1}{d+\chi}}$ such that for $\lambda \gg \lambda^*_{d,\chi}(P)$, the replica method applies, but not for $\lambda \ll \lambda^*_{d,\chi}(P)$, (ii) in the ridgeless case, spectral bias predicts the correct training curve exponent only in the limit $d \to \infty$.

1. Introduction and Motivations

Given the task of learning an unknown function f^*, a widely used algorithm is Kernel Ridge Regression (KRR) (Smola & Scholkopf, 2002). Given a set of P training points $\{x_i, f^*(x_i)\}_{i=1,...,P}$, KRR builds a predictor function f_P that is linear in a given kernel K, such that it minimizes the following training loss:

$$\sum_{i=1}^{P} |f^*(x_i) - f_P(x_i)|^2 + \lambda ||f_P||_K^2, \tag{1}$$

where λ is the ridge parameter which controls the regularisation of the kernel norm $||.||_K$ of f_P.

Minimising (1) is a convex problem, which yields the following explicit solution:

$$f_P(x) = \vec{k}(x)(K + \lambda \mathbb{1})^{-1}\vec{y}, \tag{2}$$

where $k(x)_i = K(x, x_i)$, $K_{ij} = K(x_i, x_j)$ is the $P \times P$ Gram matrix, in the noiseless setting we consider $y_i = f^*(x_i)$, λ is the ridge regularization parameter and $\mathbb{1}$ is the $P \times P$ identity matrix.

The generalization properties of KRR are an active field of research. In recent years, interest in the subject has been further increased by the discovery that for certain initializations, deep-learning behaves as a kernel method used in the ridge-less case (Jacot et al., 2018). The key quantity of interest is the generalization error ε_t, namely how much error the predictor function $f_P(x)$ does on average on the data distribution $p(x)$, with x in some space $\mathcal{D}$. Using the mean square loss, ε_t is given by:

$$\varepsilon_t = \int_{\mathcal{D}} p(d^d x)(f_P(x) - f^*(x))^2. \tag{3}$$

It is crucial to characterize ε_t with respect to the number P of training points since it allows quantification of how many samples are needed to achieve a given test error. It is empirically observed that, asymptotically for large P, ε_t often behaves as a power law in P, with a certain exponent β: $\varepsilon_t(P) \sim P^{-\beta}$ (Hestness et al., 2017; Spigler et al., 2020). The exponent β depends on the data distribution, the task, and the choice of kernel.

Recent theoretical efforts have characterized the test error in the noiseless setting considered here. In (Spigler et al., 2020), f^* was assumed to be Gaussian and the training set was assumed to be on a lattice. In (Bordelon et al., 2020; Canatar et al., 2021; 2020; Loureiro et al., 2021; Cui et al., 2021), the replica method (Mezard et al., 1987) was used, assuming that the predictor f_P is self-averaging (i.e. concentrates) and using a Gaussian assumption: a tuple of kernel eigenfunctions $(\phi_1, ..., \phi_P)$, once evaluated on P training points, behaves as a Gaussian vector. Random matrix theory was used in (Jacot et al., 2020) with the same

[1]Department of Physics, EPFL, Lausanne, Switzerland.. Correspondence to: Umberto M. Tomasini <umberto.tomasini@epfl.ch>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Gaussian assumption (with results not guaranteed to hold in the ridge-less case), or in (Mei et al., 2021) with a 'spectral gap' assumption. None of these assumptions should hold in practical applications [1]. It is thus important to understand the universality of these results, and when they break down.

Spectral bias: These predictions for ε_t rely on the exact eigendecomposition of the kernel:

$$\int p(y)K(y,x)\phi_\rho(y)dy = \lambda_\rho \phi_\rho(x), \tag{4}$$

with $\{\phi_\rho\}$ the normalised eigenfunctions and $\{\lambda_\rho\}$ the eigenvalues in decreasing order. In particular, the true function f^* can be written as:

$$f^*(x) = \sum_{\rho=1}^{\infty} c_\rho \phi_\rho(x), \tag{5}$$

The key result is that KRR learns faster the eigenmodes corresponding to the P largest eigenvalues, and makes an error on the following ones. Specifically, in the noiseless case with no ridge ($\lambda = 0$) and assuming $c_\rho^2 \sim \rho^{-a}$ and $\lambda_\rho \sim \rho^{-b}$ with $2b > (a-1)$, the prediction of the typical test error ε_B in (Spigler et al., 2020; Bordelon et al., 2020) yields:

$$\varepsilon_B \sim \sum_{\rho=P}^{\infty} c_\rho^2 \sim P^{-a+1}, \tag{6}$$

These predictions are validated on the binary classification (corresponding to $f^*(x) = \pm 1$) of image data sets (Bordelon et al., 2020; Spigler et al., 2020; Jacot et al., 2020). Why it is so is not well understood, since real data do not follow the assumptions made, whose universality class is not characterized. To understand the limit of validity of these theories, we seek to test them in simple models. It requires diagonalizing the kernel and having full control over the test error. Unfortunately, explicit diagonalisations of kernels is difficult, except if the data distribution $P(x)$ is uniform on the sphere (Bordelon et al., 2020) or on the torus (Gretton, 2019). For non-uniform data, the only settings that the authors are aware of are (i) a Gaussian kernel with a Gaussian data distribution (Gretton, 2019) and (ii) the work of (Basri et al., 2020) where $p(x)$ is piece-wise uniform.

1.1. Related works

The decay rate of the generalization error in KRR is a thoroughly studied topic in the statistical learning literature. Classical studies consider the noisy optimally regularized case (Caponnetto & De Vito, 2005; 2007; Steinwart et al., 2009; Fischer & Steinwart, 2020) under the assumptions of power-law behaviour for the kernel spectrum and the eigenexpansion coefficients of the target function, called capacity and source conditions. These asymptotic results under general source and capacity conditions have been obtained also using the replica method in teacher-student models (Loureiro et al., 2021) and rigorously for ridge regression (Wu & Xu, 2020; Richards et al., 2021). A unifying picture for the crossover between noise and regularization has been studied in (Cui et al., 2021). Using the equivalence between KRR and Gaussian process regression (Kanagawa et al., 2018), power-law rates of the generalization error have been rigorously established in (Jin et al., 2021). Other works, instead, use cross-validation estimators and random matrix theory to study the generalization error in KRR (Jacot et al., 2020) and high dimensional linear regression (Xu et al., 2021; Patil et al., 2021). A recent line of works has considered the generalization error in the noiseless case with constant ridge regularization (Bordelon et al., 2020; Spigler et al., 2020; Jacot et al., 2020; Jun et al., 2019), which corresponds to our setting.

1.2. This Paper

We consider data $x \in \mathbb{R}^d$ where the first component x_1 is distributed as $p(x_1) \sim |x_1|^\chi$ when $x_1 \to 0$ for $\chi \geq 0$. We use the Laplacian kernel $K(x,y) = K(|x-y|) = \exp(-||x-y||_2/\sigma)$, where $||.||_2$ is the L_2 norm and $\sigma > 0$ defines the width of the kernel, and consider functions $f^*(x) = f^*(x_1)$ that depend only on the first component x_1 and can be singular or not at $x_1 = 0$. We first study the one-dimensional case where we are able to rigorously prove results by eigendecomposition of the kernel. We then extend these results to generic dimension d by scaling arguments which are validated by numerical simulations.

- In Section 3, for $d = 1$, we compute the scaling of the generalization error with respect to the number of training points P for vanishing ridge.
- In the same section, inspired by (Basri et al., 2020) we derive an exact differential equation for the eigenfunctions of the kernel K, which holds for a general data distribution $p(x)$. This equation is related to the Schrodinger equation in quantum mechanics. We solve it using methods developed in that field, to obtain the asymptotic behavior of the kernel eigenfunctions and the eigenvalues.
- In Section 4, in the one-dimensional case, we find that there exists a cross-over ridge $\lambda^*_{1,\chi}(P) \sim P^{-\frac{1}{1+\chi}}$ such that for $\lambda \gg \lambda^*_{1,\chi}(P)$ spectral bias (6) holds, but not for $\lambda \ll \lambda^*_{1,\chi}(P)$ where the exponent of the training curve is different. We repeat the same analysis for the test error prediction provided by (Jacot et al., 2020) in Appendix B, and we observe the same crossover. We show that when $\lambda \ll \lambda^*_{1,\chi}(P)$, the predictor is not self-

[1] The spectral gap assumption used in (Mei et al., 2021) may hold for Gaussian data in high dimension, but breaks down for real data which are highly anisotropic, see e.g. (Spigler et al., 2020).

averaging: its relative variance does not vanish even for very large P. Our result on $\lambda^*_{1,\chi}$ has an interesting connection with the rigorous results of (Jin et al., 2021), since it saturates the lower bound on the ridge required by their prediction of the KRR test error. Therefore, our result proves that their lower bound on the ridge is tight, since no better bound can be obtained for the general case.

- In Section 5, we generalize these results to any dimension d by scaling arguments that extend the proved results in $d = 1$. One finds a cross-over ridge $\lambda^*_{d,\chi}(P) \sim P^{-\frac{1}{d+\chi}}$ for any d such that the spectral bias (6) does not hold for $\lambda \ll \lambda^*_{d,\chi}(P)$, because the predictor is not self-averaging near the decision boundary. We confirm our results numerically and show that our model captures well the performance of KRR on CIFAR-10.

2. Our models

One dimension. We consider a one-dimensional class of problems, where the data $x \in \mathbb{R}$ are distributed according to the probability distribution:

$$p(x) = \frac{1}{\Gamma\left(\frac{1+\chi}{2}\right)}|x|^\chi e^{-x^2}, \tag{7}$$

where $\chi \geq 0$ and Γ is the Euler gamma function $\Gamma(t) = \int_0^\infty dx\, x^{t-1}e^{-x}$. Our true function $f^*_\xi(x)$ depends on a parameter ξ and it is defined as:

$$f^*_\xi(x) = \text{sign}(x)|x|^{-\xi} \tag{8}$$

We restrict to ξ such that $\xi < \frac{\chi+1}{2}$, to have the L_2 norm with respect to $p(x)$ finite. Note that for $\xi = 0$ the task (8) boils down to a binary classification problem. For $\chi = 0$ the data distribution is uniform, while the case $\chi > 0$ is meant to model the presence of diminished density of data between data of different labels. Such a reduction of density is apparent in low-dimensional representations of real datasets, as for the t-SNE visualization of MNIST in (Van der Maaten & Hinton, 2008).

Generic dimension. We generalize the one-dimensional setting above to a generic dimension d. We consider a cylindrical embedding of the data $x = [x_1, ..., x_{d+1}]$ so that the first coordinate x_1 is distributed according to $p(x_1) \propto x_1^\chi e^{-x_1^2}$, while the other coordinates $x_2, ..., x_{d+1}$ are uniformly randomly distributed on the sphere $\sum_{i=2}^{d+1} x_i^2 = 1$. In this setting, we consider the true function $f^*(x) = \text{sign}(x_1)$, so that the hyper-plane $x_1 = 0$ corresponds to the decision boundary of a binary classification problem.

3. Test Error analysis

We now state our result about the generalisation error in the setting described in Section 2 for $d = 1$.

Theorem 3.1 (Test error)**.** *Consider a training set $\{x_i, f^*(x_i)\}_{i=1...P}$, where the samples x_i are i.i.d. with respect to the PDF (7) and the true function f^* is (8). In the limit of large P, the following asymptotic relation for the test error (3) of KRR with Laplacian kernel with width σ and vanishing ridge $\lambda \to 0^+$ holds:*

$$\varepsilon_t \sim P^{-1+\left(\frac{2\xi}{\chi+1}\right)} \tag{9}$$

The full proof is reported in Appendix C. The intuition behind (9) is the following. If we call $x_A < 0$ and $x_B > 0$ the points of the sampled training set which are closest to $x = 0$, we have that their typical value is the following:

$$\langle |x_A| \rangle \sim \langle |x_B| \rangle \sim P^{-\frac{1}{\chi+1}}. \tag{10}$$

This is given by the fact that $\langle x_B \rangle$ is defined as the extremal point such that in the interval $[0, x_B]$ there is just one sampled point on average:

$$\frac{1}{P} \sim \int_0^{\langle x_B \rangle} dx\, p(x), \tag{11}$$

which yields (10). The same holds for x_A. We then consider the asymptotic limit of $\sigma \to \infty$, where the Laplacian kernel becomes a cone in x. For $\lambda \to 0^+$, the predictor f_P is then given by the following piece-wise linear function for $\xi = 0$:

$$f_P(x) = \begin{cases} \text{sign}(x), & \text{for } x \geq x_B \text{ or } x \leq x_A \\ \dfrac{2x}{x_B - x_A} - \dfrac{x_A + x_B}{x_B - x_A}, & \text{for } x_A < x < x_B \end{cases} \tag{12}$$

A representation of (12) is given by the blue line in Fig. 1. For $\xi > 0$, the predictor for $x_A < x < x_B$ will be as in Eq. (12), and it will approximate f^*_ξ with a piece-wise function otherwise. The leading contribution to the test error in the asymptotic limit of large P is given by the interval $[x_A, x_B]$:

$$\begin{aligned} \varepsilon_t &\sim \int_{x_A}^{x_B} dx\, p(x)(f_P(x) - f^*_\xi(x))^2 \sim \\ &\sim \int_0^{x_B} x^{\chi - 2\xi} dx \sim P^{-1+\left(\frac{2\xi}{\chi+1}\right)}, \end{aligned} \tag{13}$$

in accordance to (9). Considering a generic finite σ, (9) still holds, as we prove and numerically test in Appendix C.

3.1. Eigendecomposition of the kernel

To effectively test the spectral bias prediction for the KRR test error (6) in our context, we need to solve the eigenproblem (4) for the Laplacian kernel with width σ and the probability distribution (7). All the proofs and more detailed statements of what follows are provided in Appendix D, except for Thm. 3.2.

We first show a general result regarding the problem of finding the eigenfunctions ϕ_ρ of the Laplacian kernel using a

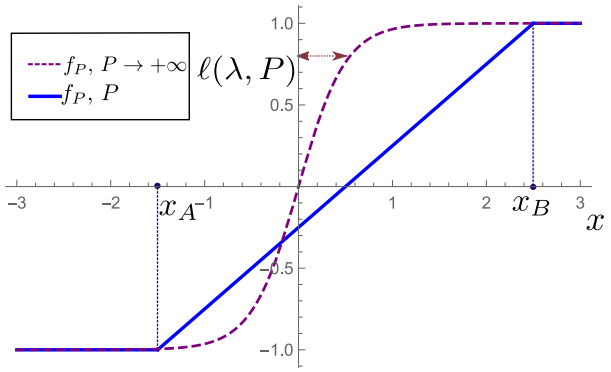

Figure 1. Representations of KRR predictors f_P (2) for $\xi = 0$ and fixed λ/P. The blue line is the predictor f_P for finite P, in the case of the extremal point $x_B \sim P^{-1/(\chi+1)}$ (10) being much larger than the characteristic scale $\ell(\lambda, P) \sim \left(\frac{\lambda\sigma}{P}\right)^{\frac{1}{(2+\chi)}}$ of the predictor (28). In the limit $P \to \infty$, the predictor f_P is represented by the dashed purple line.

generic $p(x)$, which is recast in solving a differential equation. We will then use this result in the particular context of (7).

Theorem 3.2. *Let K be the Laplacian kernel with width σ. Consider a one-dimensional input space $x \in \mathbb{R}$. Then the eigenfunctions ϕ_ρ of the kernel, defined in (4), solve the following differential equation for $\lambda_\rho \neq 0$:*

$$\partial_x^2 \phi_\rho(x) = \left(-2\frac{p(x)}{\lambda_\rho \sigma} + \frac{1}{\sigma^2}\right)\phi_\rho(x). \tag{14}$$

Proof. Let's rewrite the eigendecomposition relation as follows, writing explicitly the kernel K:

$$\int_{-\infty}^{x} p(y)\phi_\rho(y) e^{-\frac{(x-y)}{\sigma}} dy + \\ + \int_{x}^{\infty} p(y)\phi_\rho(y) e^{-\frac{(y-x)}{\sigma}} dy = \lambda_\rho \phi_\rho(x). \tag{15}$$

We derive two times the relation (15) with respect to x , following an idea similar to (Basri et al., 2020), getting:

$$-\frac{2}{\sigma}p(x)\phi_\rho(x) + \frac{1}{\sigma^2}\left(\int_{-\infty}^{x} p(y)\phi_\rho(y) e^{-\frac{(x-y)}{\sigma}} dy + \right. \\ \left. + \int_{x}^{\infty} p(y)\phi_\rho(y) e^{-\frac{(y-x)}{\sigma}} dy\right) = \lambda_\rho \partial_x^2 \phi_\rho(x). \tag{16}$$

Substituting (15) into (16) and dividing by $\lambda_\rho \neq 0$, we get (93). □

The functional operator entering (14) is symmetric with respect to x because $p(x) = p(-x)$. Thus there is always an eigenbasis for the space of solutions for which the ϕ_ρ are either even or odd functions in x. From the definition of the ϕ_ρ in (4) and from (14) in the limit of $|x| \to \infty$, we get the boundary condition $\phi_\rho(x) \to 0$ for $|x| \to \infty$.

We asymptotically solve the equation (14) for ϕ_ρ, in the limit of small λ_ρ. We use the so-called Wentzel–Kramers–Brillouin (WKB) method (A.K.Ghatak, 1991), designed to solve the following differential equation:

$$\partial_x^2 \psi(x) + \Gamma^2(x)\psi(x) = 0. \tag{17}$$

This equation is encountered for example in quantum mechanics: the Schroedinger equation has the same form of (17), with ψ being the wave function of a particle and $\Gamma^2(x) = \frac{2m}{\hbar}[E - V(x)]$, with m the mass particle, $\hbar$ the rescaled Planck's constant, E the total energy and $V(x)$ the potential energy function of the system. In the KRR case of (14), ψ is the eigenfunction ϕ_ρ and $\Gamma^2(x)$ is related to the PDF $p(x)$.

The WKB solution of (17) is obtained as follows. It is crucial to identify a small parameter $\lambda_0 \ll 1$ and a function $\tilde{\Gamma}$ finite in the limit $\lambda_0 \to 0^+$ such that we can rewrite Γ^2:

$$\Gamma^2(x) = \frac{1}{\lambda_0}\tilde{\Gamma}^2(x). \tag{18}$$

The limit of $\lambda_0 \to 0^+$ is equivalent to consider the function $\Gamma^2(x)$ as slowly changing in x. The role of λ_0 is played in quantum mechanics by $\hbar$ and in the KRR setting of (14) by the eigenvalue λ_ρ. One then seeks a solution of (17) of the form:

$$\psi(x) = e^{\frac{iS(x)}{\lambda_0}}, \quad S(x) = S_0(x) + \lambda_0 S_1(x) + \lambda_0^2 S_2(x) + ... \tag{19}$$

where the function $S(x)$ is expanded in series of λ_0. If we substitute the solution (19) into (17), we can get expressions for each function $S_i(x)$ for i arbitrarily large. At the first order in λ_0, we get the following solution:

$$\psi_1(x) = \frac{C}{\left(\tilde{\Gamma}^2(x)\right)^{1/4}} \exp\left(\frac{i}{\sqrt{\lambda_0}}\int^x dy \sqrt{\tilde{\Gamma}^2(y)}\right), \tag{20}$$

which is essentially an oscillatory or exponential function multiplied by an amplitude dependent on x. The contributions to $S(x)$ from $S_2(x)$ onwards are negligible with respect to the others provided that:

$$\left|\frac{1}{2\Gamma}\partial_x^2\Gamma - \frac{3}{4\Gamma^2}(\partial_x\Gamma)^2\right| \ll \Gamma^2(x), \tag{21}$$

which holds in the case of (14) except in the proximity of the two points x_1 and x_2 where $\Gamma^2(x) = 0$.

In the Appendix D, in the Lemmas D.2 and D.3, we derive at leading order in λ_ρ the full form of the eigenfunctionsϕ_ρ for all $x \in \mathbb{R}$. Close to the points x_1 and x_2, we linearize the

function $\Gamma^2(x)$ to solve analytically the differential equation (17) using the Airy functions (Florentin et al., 1966). Then we patch together the solution around the points x_1 and x_2 and the WKB solution using the Modified Airy Functions (MAF) (A.K.Ghatak, 1991).

Once we solve the differential equation (14) and we get the eigenfunctions ϕ_ρ at the leading order in λ_ρ, we can compute the coefficients c_ρ by projecting the true function (8) on the eigenfunctions. In particular, we are interested in the coefficients c_ρ at the leading order in λ_ρ.

Proposition 3.3. *(Coefficients) Let K be the Laplacian kernel with width σ. Let $p(x)$ be (7) and the true function f^* (8). Consider a small eigenvalue $\lambda_\rho \ll 1$. Let ϕ_ρ be the solution of (14). We impose that $\phi_\rho(x) \to 0$ for $|x| \to \infty$. Then the following holds for the coefficient $|c_\rho|$ defined in (94), in the limit $\lambda_\rho \ll 1$:*

$$\begin{aligned} |c_\rho| &\sim \lambda_\rho^{\frac{\frac{3}{4}\chi+1-\xi}{\chi+2}} && \text{if } \phi_\rho \text{ is odd} \\ |c_\rho| &= 0 && \text{if } \phi_\rho \text{ is even.} \end{aligned} \tag{22}$$

To get the scaling of the coefficients c_ρ with respect to the eigenvalue rank ρ, we need to compute the eigenvalues λ_ρ at the leading order in ρ.

We first find a close formula satisfied by the eigenvalues λ_ρ, requiring that they are such that the eigenfunctions ϕ_ρ respect the boundary condition $|\phi_\rho(x)| \to 0$ for $|x| \to \infty$. In other words, we find the eigenvalues λ_ρ such that any not-decaying exponential contribution in the WKB solution (20) is identically zero for large x.

In particular, for odd ϕ_ρ and $\chi > 0$ we find the following self-consistent relation satisfied by $\lambda_\rho \ll 1$, or equivalently by $\rho \gg 1$:

$$\lambda_\rho = \left(\frac{\int_{x_1}^{x_2} dx \sqrt{2\frac{p(x)}{\sigma} - \frac{\lambda_\rho}{\sigma^2}}}{\arctan(-\gamma_1^{-1}) + \frac{\rho-1}{2}\pi} \right)^2 + o(\rho^{-2}) \tag{23}$$

where $\gamma_1 = \mathrm{Ai}(\mu)/\mathrm{Bi}(\mu)$, with Ai and Bi the Airy function of the first and second kind (Florentin et al., 1966) and $\mu = \left(\frac{\chi(\lambda_\rho \Gamma[\frac{1+\chi}{2}])^{\frac{2}{\chi}}}{2^{\frac{2}{\chi}} \sigma^{2(1+\chi)}} \right)^{1/3}$. Similar relations hold for even ϕ_ρ and $\chi = 0$, as presented in Appendix D. The self-consistent relation (23) yields the following asymptotic scaling for the eigenvalues λ_ρ for large ρ:

$$\lambda_\rho \sim \rho^{-2} \tag{24}$$

Now that we have the scaling of the eigenvalues λ_ρ, we can get the scaling of the coefficients c_ρ with respect to their ranks ρ.

Theorem 3.4. *Let K be the Laplacian kernel with width σ. Let $p(x)$ be (7) and the true function f^* (8). As a consequence of (22) and (24), the following aymptotic relation holds for large ρ such that ϕ_ρ is odd in x, for any $\chi \geq 0$:*

$$c_\rho^2 \sim \rho^{-\frac{3\chi+4-4\xi}{\chi+2}}. \tag{25}$$

Using (25), we are finally able to get the prediction of the test error in the ridgless limit $\lambda \to 0^+$ via the spectral bias theory (26). This entails summing the coefficients squared c_ρ^2 from the $P-$th one onwards:

$$\varepsilon_B \sim \sum_{\rho=P}^{\infty} c_\rho^2 \sim P^{-1-\left(\frac{\chi-4\xi}{\chi+2}\right)}. \tag{26}$$

Comparing with Thm. 3.1, we thus conclude that the spectral bias prediction (26) is incorrect.

4. Role of ridge λ

The replica method (Bordelon et al., 2020; Canatar et al., 2021) assumes that the predictor is a self-averaging quantity. Approaches based on random matrix theory only apply under the same condition, which can be guaranteed only for a finite ridge (Jacot et al., 2020) (under the Gaussian assumption). In our model in the ridge-less case, the test error is explicitly a function of two data points x_A and x_B, and thus cannot be self-averaging. Thus we expect that these methods will work only when the ridge increases past some characteristic value $\lambda^*_{1,\chi}(P)$ to make the test error self-averaging, or equivalently if the training set is larger than some characteristic value $P^*(\lambda)$.

To estimate $P^*(\lambda)$, our strategy is to compute the KRR predictor f_P in the limit of $P \to \infty$ and $\frac{\lambda}{P}$ finite. This solution will apply for $P \gg P^*(\lambda)$. In the other limit $P \ll P^*(\lambda)$, the KRR predictor must be similar to the case $\lambda = 0$ studied above, for which it is piece-wise linear.

Proposition 4.1. *Let K be the Laplacian kernel with width σ. The KRR predictor f_P with kernel K, in the limit of $P \to \infty$ and $\frac{\lambda}{P}$ finite, satisfies the following differential equation:*

$$\sigma^2 \partial_x^2 f_P(x) = \left(\frac{\sigma}{\lambda/P} p(x) + 1 \right) f_P(x) - \frac{\sigma}{\lambda/P} p(x) f^*(x). \tag{27}$$

The equation (27) is obtained by noticing that, for the Laplace kernel in one dimension, the kernel norm $||f_P||_K^2$ corresponds to $||f_P||_K^2 = \frac{1}{\sigma}(\int dt f_P(t)^2 + \sigma^2 \int dt f_P'(t)^2)$. Therefore, minimizing the training loss (1) by taking the functional derivative with respect to f_P yields the linear differential equation (27) for $f_P(x)$ (proof in Appendix E.1).

Considering the $p(x)$ introduced in Section 2, the relation (27) yields the following characteristic scale for the function

f_P:

$$\ell(\lambda, P) \sim \left(\frac{\lambda\sigma}{P}\right)^{\frac{1}{(2+\chi)}} . \tag{28}$$

This scale is obtained by noticing that the homogeneous equation of Eq. (27) has the same form as the Schroedinger equation (17). Therefore, the WKB expansion for small λ/P can be used as discussed in Section 3.1, yielding Eq. (28) for the characteristic scale ℓ at small x. The proof is reported in Appendix E.1.

The function f_P is sketched in Fig.1 for fixed $\frac{\lambda}{P}$ and P finite, and compared with the KRR predictor in the limit $P \to \infty$. For large P, the latter limit must be a good approximation of the KRR predictor (27). However, this approximation will break down when P is small: in that case, the first data point x_B will be much larger than $\ell(\lambda, P)$, and the solution will be approximately piece-wise linear, as in Fig.1.

The cross-over between the two regimes must occur when $x_B \sim \ell(\lambda, P)$, leading to a characteristic ridge:

$$\lambda^*_{1,\chi} \sim P^{-\frac{1}{1+\chi}} . \tag{29}$$

We remark that $\lambda^*_{1,\chi}$ saturates the lower bound on the ridge required by (Jin et al., 2021) for the validity of their rigorous analysis of the generalization error of KRR. [2]

Numerical test: To confirm that $\lambda^*_{1,\chi}$ marks the point where

[2] In Theorem 12 in (Jin et al., 2021) it is shown that if the eigenvalues λ_ρ and the coefficients c_ρ scale as $\lambda_\rho \sim \rho^{-\alpha}$ and $c_\rho \sim \rho^{-\beta}$, and assuming that the sup of the infinite norm of the eigenfunctions is bounded as $||\phi_\rho||_\infty \leq \rho^\tau$ (their Assumption 6), then for ridges $\lambda \sim P^t$ such that $\left(1 - \frac{\alpha}{1+2\tau} < t < 1\right)$ the test error of KRR scales as $\varepsilon_t \sim P^{\frac{(1-2\beta)(1-t)}{\alpha}}$. Notice that they use a rescaled λ with respect to ours: in their Remark 13 their KRR predictor (2) has λP instead of λ. We recasted their results in our convention for the ridge, as above. In our case we have $\alpha = 2$

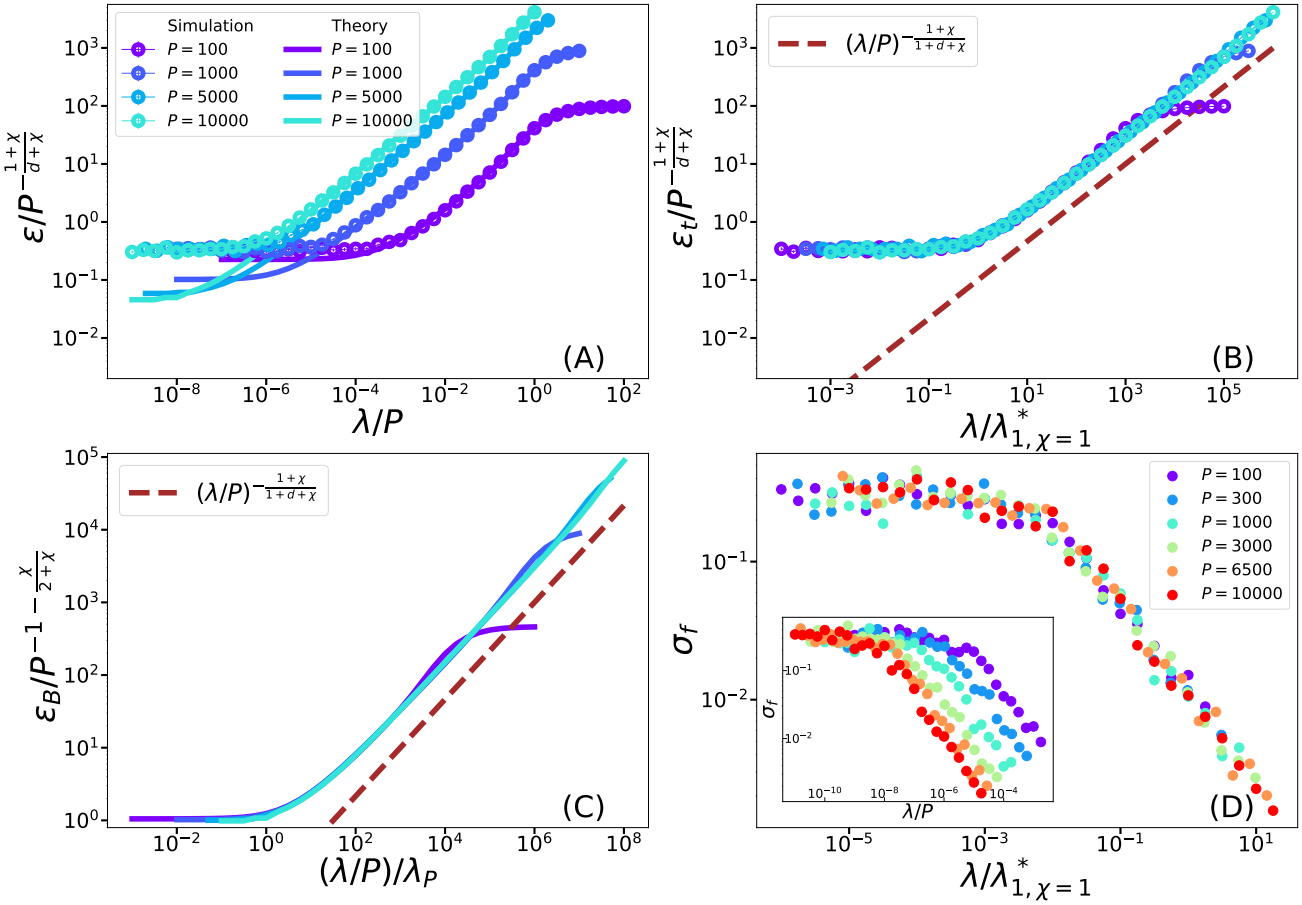

Figure 2. $d = 1$, $\chi = 1$ and $\xi = 0$. (A) Open symbols: empirical test error ε_t (averaging over 200 realisations) rescaled by its ridgeless prediction (9). The error bars of the average test error are within the symbols. Full lines: replica prediction ε_B for fixed training set size P and varying ridge λ/P. (B): the ridge has been rescaled by $\lambda^*_{1,\chi}$, defined in (29). Brown line: asymptotic behavior of ε_B with λ as predicted from Eq. (26). (C) ε_B, rescaled by its ridgeless prediction (26), for fixed P and varying rescaled ridge $(\lambda/P)/\lambda_P$, where λ_P is defined in the main text. (D): Inset: Variance of the predictor σ_f as defined in (30) (averaged over 50 realisations) as a function of the rescaled ridge λ/P. Main plot: after rescaling the x-axis by $\lambda/\lambda^*_{1,\chi}$, the curves collapse as predicted.

replica theory ε_B breaks down, we compare it with the empirical test error ε_t numerically obtained for $\chi = 1$ and $\xi = 0$ in Fig. 2 (A). For small λ/P, the prediction ε_B and the numerical results reach a different plateau, while for large λ/P they coincide. Hence there is a crossover in λ, for fixed P, between values of λ where the prediction ε_B works and where it does not. After rescaling λ by $P^{-\frac{1}{1+\chi}}$, the empirical curves ε_t for different P collapse as shown in Fig. 2 (B). It is true in particular for the location where ε_t starts flattening and departs from ε_B, confirming that replica theory breaks down for $\lambda \ll \lambda^*_{1,\chi}$. In Fig. 2 (C), we confirm that in replica theory ε_B reaches a plateau when $\lambda/P \ll \lambda_P$. In fact, in Appendix A, we show that ε_B has small relative changes when the rescaled ridge λ/P goes from zero to λ_P, where λ_P is the rank P eigenvalue of the kernel.

Finally, we confirm that replica theory breaks down when the predictor is not self-averaging near the decision boundary. To do so, we consider the variance of the predictors f_P obtained from different training sets.

We define σ_f as:

$$\sigma_f = \frac{1}{N_P}\sum_{i=1}^{N_P}[f_{P,1}(x_i) - f_{P,2}(x_i)]^2 \tag{30}$$

from (24), $\beta = \frac{\frac{3}{2}\chi+2-2\xi}{\chi+2}$ from (25) and $\tau = \frac{\chi/2}{\chi+2}$ as it can be extrapolated from the form of the eigenvectors ϕ_ρ in Lemma D.2 in App D. From these coefficients, we get that the minimal exponent t of the ridge $\lambda \sim P^t$ such that the analysis in (Jin et al., 2021) is valid is $-1/(\chi+1)$, which is exactly our crossover ridge $\lambda^*_{1,\chi}$ (29). This implies that the lower bound they get for the ridge is tight.

where $f_{P,1}$ and $f_{P,2}$ are two different predictors obtained by two different training sets of same size P and $x_{\{i=1,\dots,N_P\}}$ are the test points where the signs of the two predictors $\text{sign}(f_{P,1}(x_i))$ and $\text{sign}(f_{P,2}(x_i))$ are different. In the inset of Fig. 2 (D), σ_f *v.s.* λ/P is shown: for small ridges, the variance of the predictors does not decrease for increasing P, and the predictor is not self-averaging. We observe in the main plot that the curves collapse if λ is rescaled by $\lambda^*_{1,\chi}$ as predicted in Eq. (29).

5. Higher dimension setting and real data

We generalize the previous results to higher dimension d using scaling (non-rigorous) arguments, that make stringent predictions that we test numerically.

Ridgless case: The typical distance r_{min} between training points at the decision boundary can be estimated as the size of the ball in which in average one data point lies. It leads to:

$$r_{min} \sim P^{-\frac{1}{d+\chi}} \tag{31}$$

In the absence of ridge, $f_P(x)$ will display fluctuations of order one for $|x_1| \sim r_{min}$. Thus the test error must be of order of the probability for a test point to fall within a distance r_{min} from the interface:

$$\varepsilon_t \sim r_{min}^{1+\chi} \sim P^{-\frac{1+\chi}{d+\chi}} \tag{32}$$

Finite ridge: In the limit λ/P fixed and large P, the predictor will vary near the decision boundary on some length scale $\ell(\lambda, P)$.

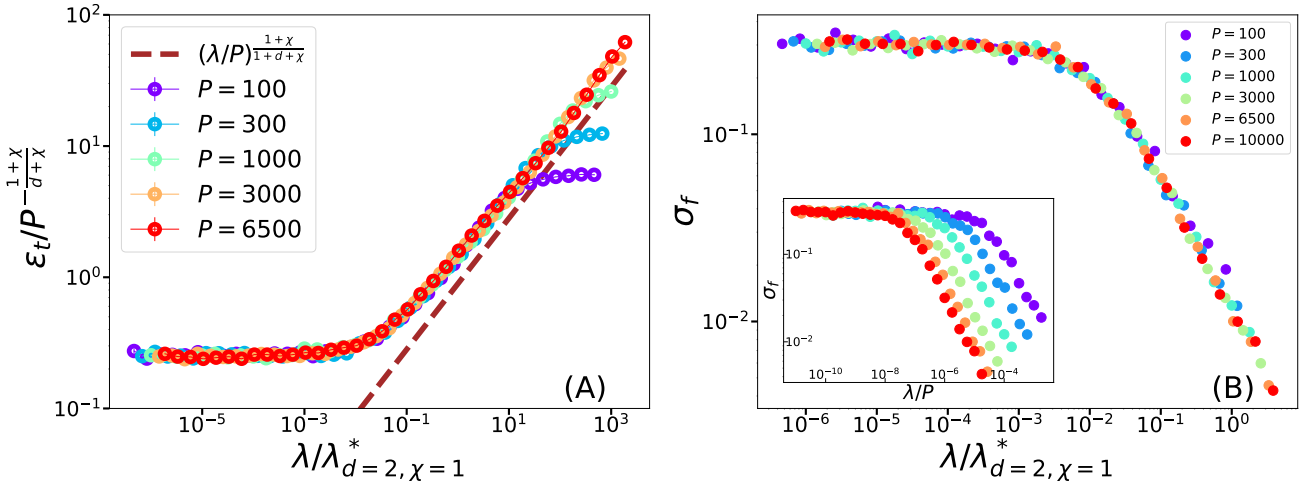

Figure 3. $d = 2$ and $\chi = 1$. (A): Empirical test error ε_t rescaled by its ridgeless prediction (9) for fixed training set size P and varying rescaled ridge $\lambda/\lambda^*_{d,\chi}$, with $\lambda^*_{d,\chi}$ defined in Eq. (36). The error bars of the average test error are within the symbols. Brown line: predicted scaling of ε_B with respect to λ, as follows from Eq. (34). (B) Inset: σ_f, defined in (30) for fixed P and varying λ/P. At small ridge, σ_f does not decrease with P. Main plot: σ_f collapses as a function of $\lambda/\lambda^*_{d,\chi}$ as predicted.

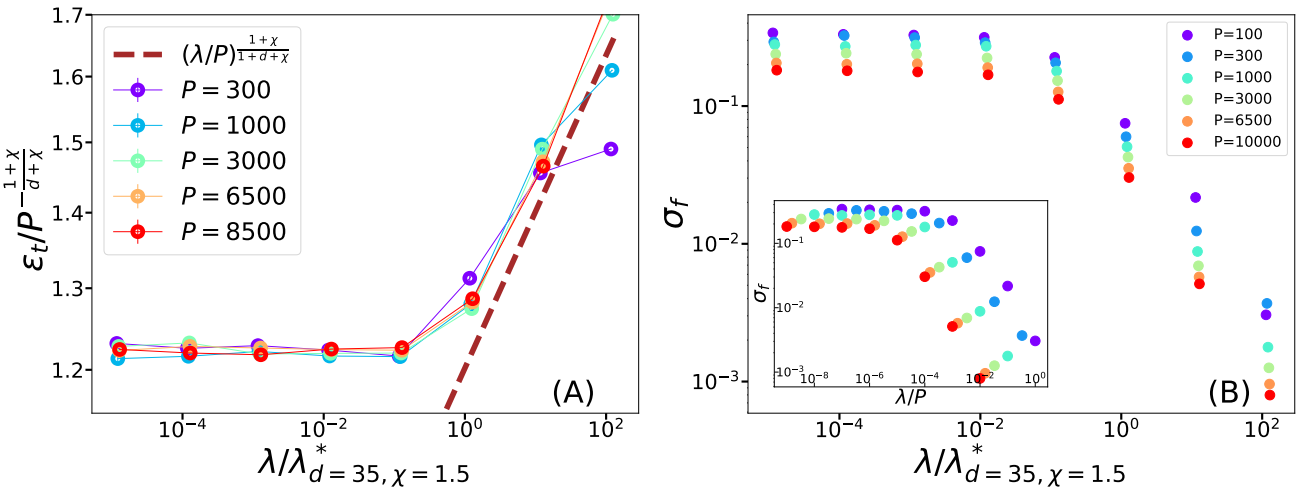

Figure 4. Binary CIFAR10. (A): Empirical test error ε_t *v.s.* ridge. Each quantity is rescaled by our predictions (32) and (36) for $d = 35$ and $\chi = 1.5$. The error bars of the average test error are within the symbols. Brown line: predicted scaling of ε_B with respect to λ, as follows from Eq. (34). (B) Inset: variance of the predictor σ_f *v.s.* re-scaled ridge λ/P. Main plot: After rescaling the ridge by $\lambda^*_{d=35,\chi=1.5}$, curves nearly collapse.

In Appendix E.2 we argue that:

$$\ell(\lambda, P) \sim \left(\frac{\lambda}{P}\right)^{\frac{1}{1+d+\chi}} \tag{33}$$

The test error predicted by the replica method then follows [3]:

$$\varepsilon_B \sim \ell(\lambda, P)^{1+\chi} \sim \left(\frac{\lambda}{P}\right)^{\frac{1+\chi}{1+d+\chi}} \tag{34}$$

In Appendix A, we show that the replica solution has only mild relative changes when the rescaled ridge λ/P goes from zero to λ_P, where λ_P is the rank P eigenvalue of the covariant operator. For a Laplace kernel, $\lambda_P \sim P^{-1-\frac{1}{d}}$ [4]. Substituting λ/P by λ_P in Eq. (34), we obtain the spectral bias prediction:

$$\varepsilon_B \sim \lambda_P^{\frac{1+\chi}{1+d+\chi}} \sim P^{-(1+\frac{1}{d})\frac{1+\chi}{1+d+\chi}} \tag{35}$$

Comparing (32) and (35), we obtain the following key results: (i) for $\chi = 0$, the spectral bias predicts the correct asymptotic training curve exponent. (ii) For $\chi > 0$, the spectral bias predicts a wrong exponent. However, the prediction is correct in the limit $d \to \infty$, and is already excellent at intermediary dimensions (say $d = 10$). (iii) The replica prediction breaks down when $\ell(\lambda, P) \sim r_{min}$, which implies a cross-over ridge:

$$\lambda^*_{d,\chi} \sim P^{-\frac{1}{d+\chi}} \tag{36}$$

[3]The same prediction for the test error is obtained in Theorem 12 in (Jin et al., 2021)

[4]Using the Fourier variable q, we have in that case $\lambda_P \sim q_{max}^{-1-d}$ and $q_{max} \sim P^{1/d}$, see e.g. (Spigler et al., 2020).

Numerical tests: We consider the case $d = 2$ and $\chi = 1$. Fig. 3 (A) shows the test error *v.s.* the ridge, both rescaled by our predictions Eqs. (36), (32). The collapse is excellent, supporting the validity of both predictions. The prediction of Eq. (35) is also indicated, and still shows an excellent agreement with observation. Fig. 3 (B) reveals that once again, the cross-over ridge $\lambda^*_{d,\chi}$ where the replica method breaks down corresponds to a predictor f_P that does not self-average near the decision boundary.

Real data: In Fig. 4, we show the same quantities for the binary CIFAR-10 dataset (the 10 classes are grouped in two). The behavior of the test error as a function of the ridge is well-fitted by our model of decision boundaries, taking $d = 35$ (the intrinsic dimension of CIFAR (Spigler et al., 2020)) and $\chi = 1.5$ as shown in Fig. 4 (A). Remarkably, as shown in the inset of Fig. 4 (B), we also find that there exists a ridge-less regime where relative fluctuations of the predictor near decision boundaries remain large ($\sigma_f > 0.1$) for all P, from a regime where these fluctuations decay rapidly with increasing P. The curves σ_f for different P all collapse when the ridge is rescaled by $\lambda^*_{d,\chi}$ as predicted. Note that in the ridge-less regime, we observe a very slight decay of σ_f (twofold) as P increases 100 folds, which signals that the geometry of decision boundaries is presumably more complex than in our model (which assumes, for example, that its properties are invariant when moving along them). A similar behaviour is shown for the binary MNIST dataset in Appendix G.

6. Conclusion

We have shown that recent results based on replica or random matrix theory (Bordelon et al., 2020; Jacot et al., 2020; Loureiro et al., 2021) can give excellent results even if data lie in low-dimension if the ridge is large enough. However, together with other approaches (Spigler et al., 2020) in the ridge-less case they lead to a spectral bias prediction. We showed that the latter does not apply for regression of a classification task if the density of data between classes vanishes, except for $d \to \infty$. Ultimately, these methods fail because the predictor is not self-averaging near the decision boundaries. Quantitatively, however, predictions are already accurate for moderate dimensions.

Finally, it is interesting to note that a vanishing density of data points implies a significant departure from the Gaussian assumption used in these approaches. Following (20), in $d = 1$ the eigenfunctions $\phi_\rho(x)$ are oscillating functions with envelope $\sim |x|^{-\chi/4}$ for small x. Thus, the probability distribution $P(\phi_\rho(x) = \phi)$ behaves as a power law $\sim \phi^{-5-\frac{4}{\chi}}$. Moreover, the eigenfunctions are not independent for different ρ: they all have large values for small x, since their envelope is $|x|^{-\chi/4}$ for any ρ.

Our results open the question of what happens when a classification loss (such the hinge loss) is applied to our learning problem. It is known that the test performances of regression and classification can be very different and we leave this analysis for future work.

References

A.K.Ghatak. *Modified Airy Function and WKB Solutions to the Wave Equation*. 1991.

Basri, R., Galun, M., Geifman, A., Jacobs, D., Kasten, Y., and Kritchman, S. Frequency bias in neural networks for input of non-uniform density. volume PartF168147-1, 2020.

Bordelon, B., Canatar, A., and Pehlevan, C. Spectrum dependent learning curves in kernel regression andwide neural networks. volume PartF168147-2, 2020.

Canatar, A., Bordelon, B., and Pehlevan, C. Statistical mechanics of generalization in kernel regression. *arXiv:2006.13198*, 2020.

Canatar, A., Bordelon, B., and Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12, 2021. ISSN 20411723. doi: 10.1038/s41467-021-23103-1.

Caponnetto, A. and De Vito, E. Fast rates for regularized least-squares algorithm. Technical report, Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, 2005.

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

Cui, H., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *ArXiv, 2105.15004*, 2021.

Fischer, S. and Steinwart, I. Sobolev norm learning rates for regularized least-squares algorithms. *J. Mach. Learn. Res.*, 21:205–1, 2020.

Florentin, J. J., Abramowitz, M., and Stegun, I. A. Handbook of mathematical functions. *The American Mathematical Monthly*, 73, 1966. ISSN 00029890. doi: 10.2307/2314682.

Gretton. Introduction to rkhs, and some simple kernel algorithms. *link*, 2019.

Hestness, J., Narang, S., Ardalani, N., Diamos, G. F., Jun, H., Kianinejad, H., Patwary, M. M., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *CoRR 1712.00409*, 2017.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. volume 2018-December, 2018.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Kernel alignment risk estimator: Risk prediction from training data. volume 2020-December, 2020.

Jin, H., Banerjee, P. K., and Montúfar, G. Learning curves for gaussian process regression with power-law priors and targets. *arXiv preprint arXiv:2110.12231*, 2021.

Jun, K.-S., Cutkosky, A., and Orabona, F. Kernel truncated randomized ridge regression: Optimal rates and low noise acceleration. *Advances in neural information processing systems*, 32, 2019.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mézard, M., and Zdeborová, L. Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model. *ArXiv, 2102.08127*, 2021.

Mei, S., Misiakiewicz, T., and Montanari, A. Generalization error of random features and kernel methentration. *arXiv preprint arXiv:2101.10588*, 2021.

Mezard, M., Parisi, G., and Virasoro, M. A. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Olver, S. *Numerical Approximation of Highly Oscillatory Integrals*. 2008.

Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 3178–3186. PMLR, 2021.

Richards, D., Mourtada, J., and Rosasco, L. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, pp. 3889–3897. PMLR, 2021.

Smola, A. and Scholkopf, B. *Learning with Kernels*. 2002.

Spigler, S., Geiger, M., and Wyart, M. Asymptotic learning curves of kernel methods: Empirical data versus teacher-student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020, 2020. ISSN 17425468. doi: 10.1088/1742-5468/abc61d.

Steinwart, I., Hush, D. R., Scovel, C., et al. Optimal rates for regularized least squares regression. In *COLT*, pp. 79–93, 2009.

Teschl, G. Ordinary differential equations and dynamical systems. *Lecture Notes from http://www. mat. univie. ac. at/gerald*, 2004.

Thomas-Agnan, C. Computing a family of reproducing kernels for statistical applications. *Numerical Algorithms*, 13, 1996. ISSN 10171398. doi: 10.1007/BF02143124.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wu, D. and Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Xu, J., Maleki, A., Rad, K. R., and Hsu, D. Consistent risk estimation in moderately high-dimensional linear regression. *IEEE Transactions on Information Theory*, 67(9):5997–6030, 2021.

A. Statistical mechanics of generalisation: spectral bias

In (Bordelon et al., 2020) a general formula for the test error (3) has been derived , which requires the exact eigendecomposition of the kernel K (4). To obtain a prediction ε_B for the generalization error (3), the authors make two assumptions. First, they assume the test error ε_t to be a self-averaging quantity with respect to the sampling of the training set. Second, they assume the probability distribution for the values of the eigenfunctions ϕ_ρ over the training to be Gaussian. Given these assumptions, they derive via the replica method the following prediction for the test error:

$$\varepsilon_B = \sum_{\rho=1}^{\infty} \frac{c_\rho^2}{\lambda_\rho^2} \left(\frac{1}{\lambda_\rho} + \frac{P}{\lambda + t(P)} \right)^{-2} \left(1 - \frac{P\gamma(P)}{(\lambda + t(P))^2} \right)^{-1}, \tag{37}$$

where λ is the ridge and:

$$t(P) = \sum_{\rho} \left(\frac{1}{\lambda_\rho} + \frac{P}{\lambda + t(P)} \right)^{-1}, \qquad \gamma(P) = \sum_{\rho} \left(\frac{1}{\lambda_\rho} + \frac{P}{\lambda + t(P)} \right)^{-2} \tag{38}$$

It is important to notice that this prediction in the ridge-less case $\lambda = 0$ is equivalent, for the scaling at large P, to choosing a ridge λ/P that is of the same order of magnitude of the smallest eigenvalue λ_P of the Gram matrix. To see this from Eq. (37) it is sufficient to show that $t(P)/P \sim \lambda_P$ when $\lambda = 0$. In this case, calling $\tilde{t}(P) = t(P)/P$, we can rewrite the definition of $t(P)$ in Eq. (38) as

$$P = \sum_{\rho} \frac{1}{1 + \frac{\tilde{t}(P)}{\lambda_\rho}} \tag{39}$$

The sum in the right hand side of Eq. (39) takes contributions of $O(1)$ for $\lambda_\rho \gg \tilde{t}(P)$ and contributions of $O(\lambda_\rho/\tilde{t}(P))$ for $\lambda_\rho \ll \tilde{t}(P)$. Since λ_ρ decreases with ρ, this suggests that, for large P, $\tilde{t}(P)$ should be of the same order of λ_P to have a sum of order P at the right hand side of Eq. (39). To see it more explicitly, we can consider an eigenvalue spectrum decaying as $\lambda_\rho \sim \rho^{-a}$ and we can approximate the sum in Eq. (39) with an integral:

$$P = \sum_{\rho} \frac{1}{1 + \frac{\tilde{t}(P)}{\lambda_\rho}} \sim \int_0^{\infty} d\rho \frac{1}{1 + \tilde{t}(P)\rho^a} \propto \tilde{t}(P)^{-\frac{1}{a}} \tag{40}$$

which gives $\tilde{t}(P) \sim P^{-a}$, that is $\tilde{t}(P) \sim \lambda_P$.

We have seen that the spectral bias prediction (26) does not work for vanishing λ. We may wonder about what happens for larger ridges. We compare the empirical test error ε_t obtained from the experiments and the full prediction provided by (37) for a large range of ridges λ. To compute the prediction (37) we need:

(i) The exact eigenvalues λ_ρ, found via the self-consistent numerical scheme (23). We computed them for ranks ρ up to $5.1 \cdot 10^4$. Since the scheme is valid for small λ_ρ, we replaced the first 10^3 eigenvalues with the ones obtained diagonalising a large Gram matrix.

(ii) The exact coefficients c_ρ, found projecting the solution ϕ_ρ of the differential equation (14) onto the true function f^* (8). We found them exactly for ranks $\rho < 10^4$, then for ranks between 10^4 and $5.1 \cdot 10^4$ we extrapolated the value of c_ρ^2 doing a linear fit of c_ρ^2 with respect to ρ for the first 10^4 rank.

Once we have these ingredients, we can compute the prediction ε_B provided by the full formula in (37) for different training set sizes P, and compare it with the the empirical test error with respect to the ridge λ, as we do for $\xi = 0$ iFig. 2 (A) in the main text for $\chi = 1$ and in Fig. 5 (A) for $\chi = 0$. We can notice that:

(1) The prediction (37) for the scaling of ε_t for fixed P works for large ridge λ and it breaks down lowering it. In section 4 we argue that the crossover happens at $\lambda^*_{1,\chi} \sim P^{-\frac{1}{1+\chi}}$, as shown in Fig. 2 (B) in the main text for $\chi = 1$ and in Fig. 5 (B) for $\chi = 0$.

(2) The scaling of the prediction ε_B with respect to P, given by (26), captures the behaviour of the numerical results of ε_B for small λ.

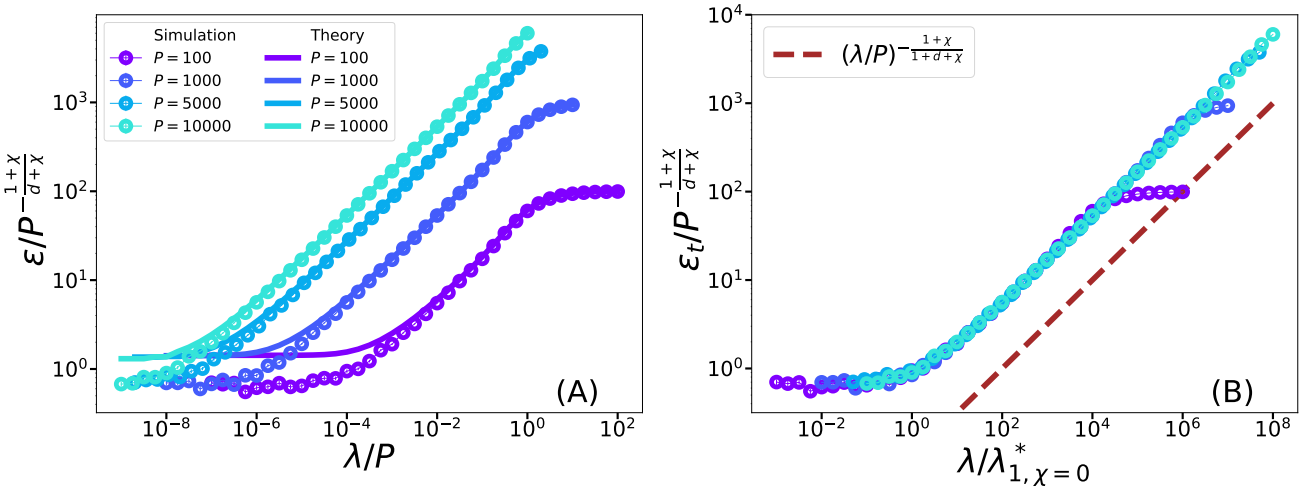

Figure 5. (A) Open symbols: empirical test error ε_t (averaging over 200 realisations) rescaled by its ridgeless prediction (9). The error bars of the average test error are within the symbols. Full lines: replica prediction ε_B for fixed training set size P and varying ridge λ/P. (B): the ridge has been rescaled by $\lambda^*_{1,\chi}$, defined in (29). Brown line: asymptotic behavior of ε_B with λ as predicted from Eq. (26).

B. Kernel Alignment Risk Estimator

In this section we look at the results shown in (Jacot et al., 2020). In their work, the authors assume that, as far as one is interested in just the first two moments of the predictor f_P (2), for any tuple of functions $(f_1, ..., f_P)$ the vector of observations of these functions $(f_1(x_1), ..., f_P(x_P))$ over P points $\{x_i\}_{i=1,...,P}$ is a Gaussian vector. This Gaussianity Assumption includes also the eigenfunctions ϕ_ρ. As a consequence, it is possible to use rigorous Random Matrix Theory techniques for Gaussian matrices to obtain an estimate, called Kernel Alignment Risk Estimator (KARE), of the test error (3) which depends just on the training data:

$$\varepsilon_K \sim \left\langle \frac{\frac{1}{P}(\vec{y})^T (K+\lambda \mathbb{1})^{-2} \vec{y}}{\left(\frac{1}{P}\mathrm{Tr}\left[(K+\lambda\mathbb{1})^{-1}\right]\right)^2} \right\rangle, \tag{41}$$

where the index K stands for "KARE", the average is over different sampled sets, $\vec{y}$ is the vector of the labels in the training set and K is the Gram matrix related to the P samples $\{x_i\}$. The relation (41) has a different prefactor in front of the Gram matrix with respect to the formula in (Jacot et al., 2020), which is due to our different definition of the training loss (1). To obtain the relation (41) they rely on some concentration results, whose fluctuations are controlled for values of the ridge $\lambda \to 0^+$ and training set size $P \to \infty$ such that $1/(\lambda\sqrt{P}) \to 0^+$.

We then test the prediction ε_K, comparing it with the empirical test error ε_t with respect to the ridge λ for fixed training set size P in Fig. 6 (A) for $\chi = 0$ and in Fig. 7 (A) for $\chi = 0$. Both ε_K and ε_t are obtained averaging over 200 sampling realisations. We can see that the KARE prediction works for large λ, then it breaks down for small ridges, for fixed P. In section 4 we argue that the crossover between the ridges where the KARE prediction works and where it does not is at $\lambda^*_{1,\chi} \sim P^{-\frac{1}{1+\chi}}$, as shown in Fig. 6 (B) and in Fig. 7 (B).

C. No ridge test error Proofs and Numerics

C.1. Proofs

Theorem C.1 (Test error)**.** *Consider a training set $\{x_i, f^*(x_i)\}_{i=1...P}$, where the samples x_i are i.i.d. with respect to the PDF (7) and the true function f^*_ξ is (8). In the limit of $P \to \infty$, the following asymptotic relation for the test error (3) of KRR with Laplacian kernel $K(|x-y|) = \exp(-||x-y||_2/\sigma)$ and ridge $\lambda \to 0^+$ holds:*

$$\varepsilon_t \sim P^{-1+\frac{2\xi}{\chi+1}} \tag{42}$$

Proof. The sketch of the proof is the following. We first find the form of the KRR predictor f_P between any couple of sampled points $\{x_i, x_{i+1}\}$ which are neighbours. Then we estimate the amount of test error in the interval $[x_i, x_{i+1}]$, in the

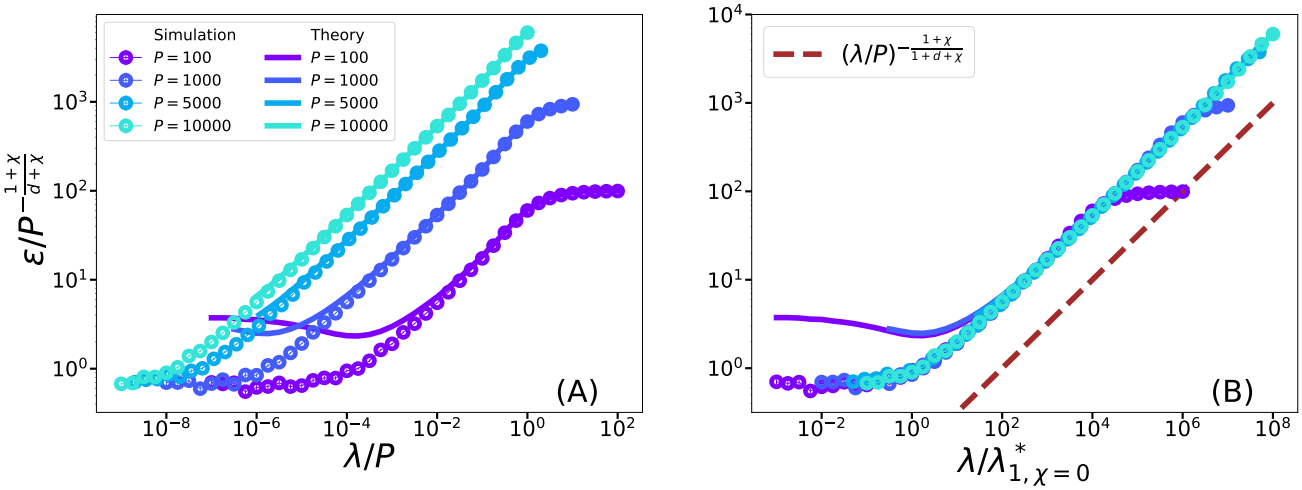

Figure 6. $d = 1$, $\xi = 0$, $\chi = 0$. (A) Open symbols: empirical test error ε_t (averaging over 200 realisations) rescaled by its ridgeless prediction (9). The error bars of the average test error are within the symbols. Full lines: prediction ε_K by (Jacot et al., 2020) for fixed training set size P and varying ridge λ/P. (B): the ridge has been rescaled by $\lambda^*_{1,\chi}$, defined in (29). Brown line: asymptotic behavior of ε_t with λ as predicted by replica prediction in Eq. (26).

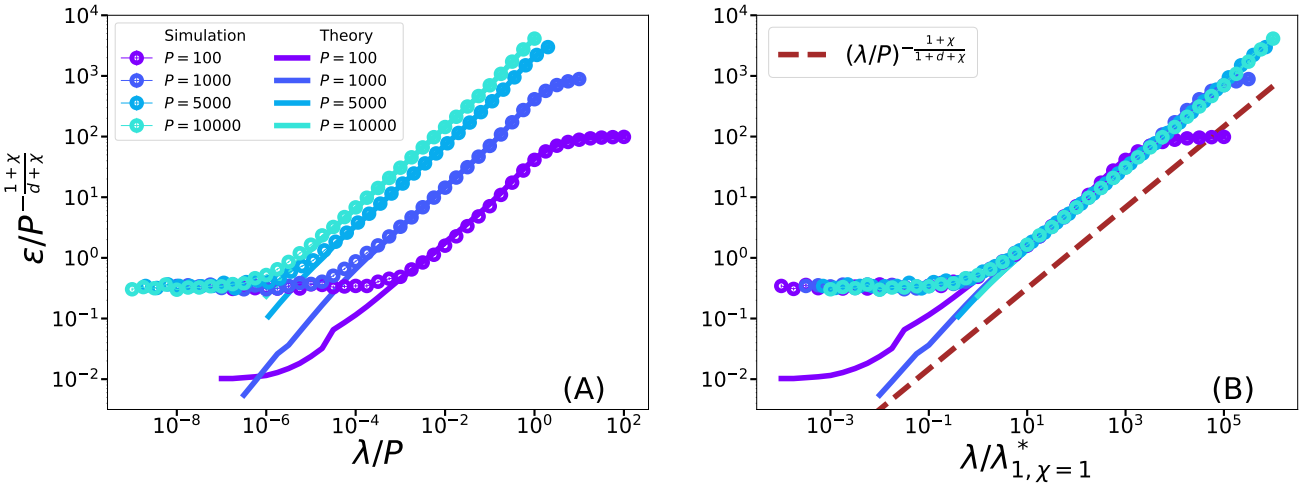

Figure 7. $d = 1$, $\xi = 0$, $\chi = 1$. (A) Open symbols: empirical test error ε_t (averaging over 200 realisations) rescaled by its ridgeless prediction (9). The error bars of the average test error are within the symbols. Full lines: prediction ε_K by (Jacot et al., 2020) for fixed training set size P and varying ridge λ/P. (B): the ridge has been rescaled by $\lambda^*_{1,\chi}$, defined in (29). Brown line: asymptotic behavior of ε_t with λ as predicted by replica prediction in Eq. (26).

limit of large P. Lastly, we get the test error as the sum of all the contributions in such intervals.

Without loss of generality, we say that the sampled points are such that $x_1 < x_2 < ... < x_M < 0$ and $0 < x_{M+1} < ... < x_P$. Note that, in the asymptotic limit of large P, we expect:

$$M \sim P/2 \sim P, \tag{43}$$

due to the symmetry of the PDF (7). It is now relevant to look at the scaling of the typical value of x_M and x_P with P, since it will be important to get the scaling of ε_t.

Lemma C.2. *Let's consider the sampled point x_M and x_{M+1}, which are the closest to $x = 0$. In the asymptotic limit of large P, the following scaling holds for their averages over the sampling:*

$$\langle |x_M| \rangle \sim \langle |x_{M+1}| \rangle \sim P^{-\frac{1}{\chi+1}}. \tag{44}$$

Let's consider the sampled points x_1 and x_P, which are the most far from $x = 0$. In the asymptotic limit of large P, it holds for their averages:

$$\langle |x_1| \rangle \sim \langle |x_P| \rangle \sim \sqrt{\log P} \tag{45}$$

Proof. Let's start from x_{M+1}. We have that its typical value $\langle x_{M+1} \rangle$ is such that:

$$\frac{1}{P} \sim \int_0^{\langle x_{M+1} \rangle} dx\, p(x) \sim \int_0^{\langle x_{M+1} \rangle} dx\, x^\chi e^{-x^2}, \tag{46}$$

where $p(x)$ is given by (7). The relation (46) can be interpreted as if we expect to sample on average one point out of P in the interval $[0, \langle x_{M+1} \rangle]$. For large P, we have $\langle x_{M+1} \rangle \ll 1$, hence we can write:

$$\frac{1}{P} \sim \int_0^{\langle x_{M+1} \rangle} dx\, x^\chi \propto \langle x_{M+1} \rangle^{\chi+1}, \tag{47}$$

which gives $\langle x_{M+1} \rangle = P^{-1/(\chi+1)}$, as in (44). The same logic can be applied to x_M.

Now we consider x_P. Its typical value $\langle x_P \rangle$ will be such that:

$$\frac{1}{P} \sim \int_{\langle x_P \rangle}^{\infty} p(x)\, dx. \tag{48}$$

Since we expect $\langle x_P \rangle$ to increase with P, we can rewrite the quantity $\int_{\langle x_P \rangle}^{\infty} p(x)\, dx$ as:

$$\int_{\langle x_P \rangle}^{\infty} p(x)\, dx \sim \int_{\langle x_P \rangle}^{\infty} e^{-x^2}\, dx, \tag{49}$$

which is the erfc function evaluated in $\langle x_P \rangle$. We can now make use of the following asymptotic expansion of the erfc function for large x:

$$\text{erfc}(x) = \frac{e^{-x^2}}{x\sqrt{\pi}} \left[1 + \sum_{n=1}^{\infty} (-1)^n \frac{(2n)!}{n!(2x)^{2n}} \right], \tag{50}$$

which at the leading order for large x gives $\text{erfc}(x) \sim e^{-x^2}/x$. We can then rewrite (48) as

$$\frac{1}{P} \sim \frac{e^{-\langle x_P \rangle^2}}{\langle x_P \rangle}. \tag{51}$$

Looking at the leading behaviour of $\langle x_P \rangle$ with respect to P, we find:

$$\langle x_P \rangle \sim \sqrt{\log P}, \tag{52}$$

which is (45). The same procedure can be applied to x_1. □

Now we consider the form of the predictor f_P.

Lemma C.3. *Given a couple of neighboring points $\{x_i, x_{i+1}\}$, we define $\Delta x_i = x_{i+1} - x_i$. The form of the predictor $f_P(x)$ (2) in the interval $x \in [x_i, x_i + \Delta x_i]$ depends on x_i:*

- *For $x_i > 0$:*

$$f_P(x) = |x_i|^{-\xi} \left[A(x_i) e^{\frac{(x-x_i)}{\sigma}} + B(x_i) e^{-\frac{(x-x_i)}{\sigma}} \right], \tag{53}$$

where:

$$A(x_i) = \left[\frac{\left(1 - e^{-\frac{\Delta x_i}{\sigma}}\right)}{\left(1 - e^{\frac{\Delta x_i}{\sigma}}\right)} \frac{\left((1 + \frac{\Delta x_i}{x_i})^{-\xi} - e^{\frac{\Delta x_i}{\sigma}}\right)}{2 \sinh\left(\frac{\Delta x_i}{\sigma}\right)} + \frac{1 - (1 + \frac{\Delta x_i}{x_i})^{-\xi}}{\left(1 - e^{\frac{\Delta x_i}{\sigma}}\right)} \right] \tag{54}$$

and:

$$B(x_i) = -\frac{\left((1 + \frac{\Delta x_i}{x_i})^{-\xi} - e^{\frac{\Delta x_i}{\sigma}}\right)}{2 \sinh\left(\frac{\Delta x_i}{\sigma}\right)} \tag{55}$$

- *For $x_i < 0$ and $x_i \neq x_M$, the predictor $f_P(x)$ has the form (53), with the functions $A(x_i)$ and $B(x_i)$ defined as in (54) and (55) with changed sign.*
- *For $x_i = x_M$, the predictor $f_P(x)$ has the same form of (53) but with the functions $A(x_i)$ and $B(x_i)$ defined as:*

$$A(x_i) = \left[\frac{\left(1 - e^{-\frac{\Delta x_i}{\sigma}}\right)}{\left(1 - e^{\frac{\Delta x_i}{\sigma}}\right)} \frac{\left((1 + \frac{\Delta x_i}{x_i})^{-\xi} + e^{\frac{\Delta x_i}{\sigma}}\right)}{2 \sinh\left(\frac{\Delta x_i}{\sigma}\right)} + \frac{1 + (1 + \frac{\Delta x_i}{x_i})^{-\xi}}{\left(1 - e^{\frac{\Delta x_i}{\sigma}}\right)} \right] \tag{56}$$

and:

$$B(x_i) = -\frac{\left((1 + \frac{\Delta x_i}{x_i})^{-\xi} + e^{\frac{\Delta x_i}{\sigma}}\right)}{2 \sinh\left(\frac{\Delta x_i}{\sigma}\right)} \tag{57}$$

Proof. The general form of the KRR predictor is:

$$f_P(x) = \sum_{i=1}^{P} \alpha_i K(x, x_i) = \sum_{i=1}^{P} \alpha_i e^{-||x - x_i||_2/\sigma}, \tag{58}$$

with the coefficients α_i fixed by minimizing the training loss (1). Since we are in the ridgeless limit $\lambda \to 0^+$, the minimisation problem boils down to having the predictor f_P fit the training set $\{x_i\}_{i=1\dots P}$.

Let's consider $x \in]x_i, x_i + \Delta x_i[$. The predictor f_P can be then rewritten as follows:

$$f_P(x) = \sum_{j=1}^{i} \alpha_j e^{-\frac{(x - x_j)}{\sigma}} + \sum_{j=i+1}^{P} \alpha_j e^{\frac{-(x_j - x)}{\sigma}}. \tag{59}$$

If we derive (59) two times with respect to x, we find the following differential equation satisfied by f_P:

$$\begin{aligned} f_P''(x) &= \frac{1}{\sigma^2} \left(\sum_{j=1}^{i} \alpha_j e^{-\frac{(x - x_j)}{\sigma}} + \sum_{j=i+1}^{P} \alpha_j e^{\frac{-(x_j - x)}{\sigma}} \right) \\ &= \frac{1}{\sigma^2} f_P(x) \end{aligned} \tag{60}$$

The solution of (60) is given by the sum of two exponential functions, with coefficients A_i and B_i:

$$f_P(x) = A_i e^{\frac{x}{\sigma}} + B_i e^{-\frac{x}{\sigma}}. \tag{61}$$

The coefficients A_i and B_i are fixed by requesting that the predictor f_P perfectly fits the true function $f^*_\xi(x)$ (8) on the training set $\{x_i\}$. This requirement amounts to imposing the following boundary conditions:

$$f_P(x_i) = f^*_\xi(x_i), \qquad f_P(x_i + \Delta x_i) = f^*_\xi(x_i + \Delta x_i). \tag{62}$$

Imposing these boundary conditions, the previously stated relations are found. □

We now look at the amount of test error (3) done by the predictor given by (53) in a generic interval $[x_i, x_i + \Delta x_i]$.

Lemma C.4. *We define the amount of test error in the interval $[x_i, x_i + \Delta x_i]$, for $i < P$, as follows:*

$$\varepsilon_{x_i} = \int_{x_i}^{x_i+\Delta x_i} dx p(x) |f_P(x) - f^*(x)|^2. \tag{63}$$

where $p(x)$ is given by (7). In the asymptotic limit of small Δx_i (which is equivalent to the asymptotic limit of large P), we have for $\xi > 0$ that:

$$\varepsilon_{x_i} \sim p(x_i) \frac{(\Delta x_i)^3}{x_i^2} |x_i|^{-2\xi}, \tag{64}$$

while for $\xi = 0$:

- *For $x_i \neq x_M$:*

$$\varepsilon_{x_i} \sim p(x_i) \frac{(\Delta x_i)^5}{\sigma^4} |x_i|^{-2\xi}. \tag{65}$$

- *For $x_i = x_M$:*

$$\varepsilon_{x_M} \sim |x_M|^{\chi+1-\xi}. \tag{66}$$

Proof. Let's first consider the $\xi > 0$ case. The contribution ε_{x_i} (63) for $x_i > 0$ can be written as follows, using (53):

$$\begin{aligned} \varepsilon_{x_i} &= \int_{x_i}^{x_i+\Delta x_i} dx\, p(x) |f_P(x) - f^*_\xi(x)|^2 \\ &\sim \int_{x_i}^{x_i+\Delta x_i} dx\, p(x)\, |x_i|^{-2\xi} \left| \left[\frac{\left((1 + \frac{\Delta x_i}{x_i})^{-\xi} - 1 \right)}{\frac{\Delta x_i}{\sigma}} \right] \sinh\left(\frac{x - x_i}{\sigma} \right) + \right. \\ &\qquad \left. + \cosh\left(\frac{x - x_i}{\sigma} \right) - \left| \frac{x}{x_i} \right|^{-\xi} \right|^2, \end{aligned} \tag{67}$$

where the second equation has been obtained expanding the function (53) for small Δx_i with respect to σ. We now change variable $y = x - x_i$, obtaining:

$$\varepsilon_{x_i} \sim \int_0^{\Delta x_i} dy\, p(y + x_i)\, |x_i|^{-2\xi} \left| \frac{y}{\sigma} \left[\frac{\left((1 + \frac{\Delta x_i}{x_i})^{-\xi} - 1 \right)}{\frac{\Delta x_i}{\sigma}} \right] + 1 + \frac{y^2}{\sigma^2} - \left| \frac{y + x_i}{x_i} \right|^{-\xi} \right|^2 \tag{68}$$

where we have expanded the $\sinh$ and the $\cosh$ for small y with respect to σ. Now we have two cases:

- One case where the increment Δx_i is small with respect to x_i. This happens for a number of order P of sampled points in the training set. In this case we can expand $p(y + x_i)$, $(1 + \frac{\Delta x_i}{x_i})^{-\xi}$ and $(1 + \frac{y}{x_i})^{-xi}$ in y or Δx_i with respect to x_i in (68), obtaining at the leading order in Δx_i:

$$\begin{aligned} \varepsilon_{x_i} &\sim \int_0^{\Delta x_i} dy\, p(x_i)\, |x_i|^{-2\xi} \left(\xi \frac{y}{x_i} + \frac{y^2}{\sigma^2} \right)^2 \\ &\sim p(x_i) \frac{(\Delta x_i)^3}{x_i^2} |x_i|^{-2\xi} \end{aligned} \tag{69}$$

  which is the relation (64).

- Another case where the increment Δx_i is of the same order in P with respect to x_i. This happens just for a few points around $x = 0$, and the number of these points is of order $P^0 = \mathcal{O}(1)$. Since these points x_i are close to 0, we have:

$$\begin{aligned} p(x_i + y) &\sim |x_i + y|^\chi = x_i^\chi \left|1 + \frac{y}{x_i}\right|^\chi \\ &\sim p(x_i) \left|1 + \frac{y}{x_i}\right|^\chi \end{aligned} \tag{70}$$

  Plugging (70) in (68) and noticing that the quantity $\left((1 + \frac{\Delta x_i}{x_i})^{-\xi} - 1\right)$ is a constant in P, we rewrite (68) at the leading order in $\Delta x_i \sim x_i$:

$$\varepsilon_{x_i} \sim \int_0^{\Delta x_i} dy\, p(x_i)\, |x_i|^{-2\xi} \left|1 + \frac{y}{\Delta x_i} - (1 + \frac{y}{x_i})^{-\xi}\right|^2, \tag{71}$$

  which yields:

$$\varepsilon_{x_i} \sim p(x_i)\, |x_i|^{-2\xi} \Delta x_i, \tag{72}$$

  which is consistent with the wanted relation (64), since $\Delta x_i \sim x_i$.

For $\xi > 0$ and $x_i < 0$ we can repeat the same procedure, getting relation (64) for the contribution ε_{x_i}.

For $\xi = 0$ and $x_i \neq x_M$ we can repeat the procedure done above for $\xi > 0$ up to (68). Then, we notice that imposing $\xi = 0$ in (68), we get:

$$\varepsilon_{x_i} \sim \int_0^{\Delta x_i} dy\, p(y + x_i) \frac{y^4}{\sigma^4}. \tag{73}$$

We can then repeat the study of the two different cases done for $\xi > 0$, depending on whether the increment Δx_i is of the same order or not with respect to x_i. We then get (65).

For $\xi \geq 0$ and $x_i = x_M$ the contribution ε_{x_M} (63) can be rewritten using the form (56) and (57) of the predictor and expanding for small $\Delta x_M / \sigma$:

$$\begin{aligned} \varepsilon_{x_M} \sim \int_{x_M}^{x_M + \Delta x_M} dx\, p(x)\, |x_M|^{-2\xi} \Bigg| \left[\frac{\left((1 + \frac{\Delta x_i}{x_i})^{-\xi} + 1\right)}{\frac{\Delta x_i}{\sigma}}\right] \sinh\left(\frac{x - x_i}{\sigma}\right) + \\ - \cosh\left(\frac{x - x_i}{\sigma}\right) - f_\xi^*(x)\Bigg|^2. \end{aligned} \tag{74}$$

We can split the integral in (74) in two parts, one from x_M to 0 and another one from 0 to x_{M+1}. We analyze the first part, the second part can be analysed similarly. Since $x_M \sim P^{-\frac{1}{\chi+1}} \ll 1$, we can write (74) in the following form, with the change of variable $y = x - x_M$:

$$\varepsilon_{x_M} \sim \int_0^{-x_M} dy\, (y + x_M)^\chi \left|2\frac{y}{\Delta x_i} + \left|\frac{y + x_M}{x_M}\right|^{-\xi} - 1\right|^2, \tag{75}$$

which yields at the leading order in P:

$$\varepsilon_{x_M} \sim |x_M|^{\chi - 2\xi} \frac{|x_M|^3}{(\Delta x_M)^2}, \tag{76}$$

which is consistent with both (64) and (66). □

Now that we have the contributions (63) to the test error in a given interval $[x_i, x_i + \Delta x_i]$, we want to sum over them to get the behaviour of the full test error with respect to P. Before getting to that, we prove an intermediate result about the average spacing between two neighbouring points x_i and $x_i + \Delta x_i$:

Lemma C.5. *Given a couple of neighbouring points x_i and $x_i + \Delta x_i$, for $i < P$ and $i \neq M$, the average distance between them in the asymptotic limit of large P is given by:*

$$\langle \Delta x_i \rangle \sim \frac{1}{P p(x_i)}, \tag{77}$$

where the average is over the sampling of the training set.

Proof. On average, we expect that between $\langle x_i \rangle$ and $\langle x_i + \Delta x_i \rangle$ there is one sampled point out of P:

$$\frac{1}{P} \sim \int_{\langle x_i \rangle}^{\langle x_i + \Delta x_i \rangle} p(x)\, dx. \tag{78}$$

Since we are considering large P, we have:

$$\frac{1}{P} \sim p(x_i) \langle \Delta x_i \rangle, \tag{79}$$

which gives (77). □

Lemma C.6. *In the asymptotic limit of large P, the test error (3) can be rewritten as follows:*

$$\varepsilon_t = \sum_{i=1}^{P-1} \varepsilon_{x_i} + \int_{x_P}^{\infty} p(x) |f_P(x) - f_\xi^*(x)|^2\, dx + \int_{-\infty}^{x_1} p(x) |f_P(x) - f_\xi^*(x)|^2\, dx, \tag{80}$$

where ε_{x_i} is defined in (63). Then, the following holds:

$$\varepsilon_t \sim P^{-\left(\frac{\chi+1-2\xi}{\chi+1}\right)}. \tag{81}$$

Proof. The relation (80) is immediate from the the definition of the test error (3) and of the contributions (63).

Consider the dependence on P of the second term in (80). It is always possible to bound from above $|f_P(x) - f_\xi^*|^2$ with a positive constant C. Then:

$$\int_{x_P}^{\infty} p(x) |f_P(x) - f_\xi^*(x)|^2\, dx \leq C \int_{x_P}^{\infty} p(x) dx. \tag{82}$$

The right hand side of this relation is exactly (up to the constant C) the definition (48) of the typical value of x_P. As a consequence, we have that the contribution of the second term to the test error is of order smaller or equal to P^{-1}. The same applies for the third term in (80):

$$\int_{x_P}^{\infty} p(x) |f_P(x) - f_\xi^*(x)|^2\, dx + \int_{-\infty}^{x_1} p(x) |f_P(x) - f_\xi^*(x)|^2\, dx \leq \frac{C_1}{P}, \tag{83}$$

where $C_1 > 0$.

Now we consider the contributions corresponding to the first term in (80). We start from $\xi = 0$. The scaling with respect to P of the contributions ε_{x_i} can be of four different kinds.

- We have a number of order $\mathcal{O}(P)$ of contributions ε_{x_i} such that $p(x_i)$ does not scale with P. These contributions correspond to points x_i sampled in the bulk of the distribution (7). Combining (65) and (77) we have then that the contribution of these terms to the full test error in (80) is:

$$P \cdot \frac{1}{P^5 \sigma^4} \sim \frac{1}{P^4 \sigma^4}, \tag{84}$$

  where the first factor P stands for the number of contributions we are considering.

- We have a number of order $\mathcal{O}(1)$ of contributions ε_{x_i} related to points x_i sampled with (7) close to $x = 0$, and they are different from x_M. We expect that their typical value scales like $\langle x_i \rangle \sim P^{-\frac{1}{\chi+1}}$, as x_M in (44). As a consequence, we have that for these points:

$$p(x_i) \sim |x_i|^\chi \sim P^{-\frac{\chi}{\chi+1}}. \tag{85}$$

Combining (65), (77) and (85) we obtain that the contribution of these terms to the full test error in (80) is:

$$\frac{1}{P^5\sigma^4 p^4(x_i)} \sim \frac{1}{P^5\sigma^4 P^{-\frac{4\chi}{\chi+1}}} \sim \frac{1}{\sigma^4 P^{\left(1+\frac{4}{\chi+1}\right)}}. \tag{86}$$

- There is a number of order $\mathcal{O}(1)$ of contributions ε_{x_i} related to points x_i sampled with (7) in the tail of the Gaussian. Their typical value will scale with P as $\sqrt{\log P}$, as x_P in (45). Then, disregarding logarithmic factors in P:

$$p(x_i) \sim e^{-x_i^2} \sim \frac{1}{P}. \tag{87}$$

  Combining (65), (77) and (87) we have then that the contribution of these terms to the full test error in (80) is, disregarding logarithmic factors in P:

$$\frac{1}{P^5\sigma^4 p^4(x_i)} \sim \frac{1}{P^5\sigma^4 P^{-4}} \sim \frac{1}{\sigma^4 P}. \tag{88}$$

- Now we look at the contribution ε_{x_M}. Combining (66) and (44) we have:

$$\varepsilon_{x_M} \sim \frac{1}{P} \tag{89}$$

Summing over the contributions (83), (84), (86), (88) and (89), we obtain that the leading contribution in P to the test error is given by the relation (81).

Now we look at the $\xi > 0$ case. In this case we have three different types of contributions ε_{x_i}.

- For the number $\mathcal{O}(P)$ of contributions ε_{x_i} where $p(x_i)$ and x_i do not scale with P, we obtain combining (64) and (77) the following contribution to ε_t:

$$P \cdot \frac{1}{P^3} \sim \frac{1}{P^2}. \tag{90}$$

- For the number $\mathcal{O}(1)$ of contributions ε_{x_i} where $\langle x_i \rangle \sim P^{-\frac{1}{\chi+1}}$, we have that the contribution to the test error is given by:

$$\frac{1}{P^3}\frac{1}{x_i^{2\chi+2+2\xi}} \sim P^{-\frac{\chi+1-2\xi}{\chi+1}}. \tag{91}$$

- For the number $\mathcal{O}(1)$ of contributions ε_{x_i} where $\langle x_i \rangle \sim \sqrt{\log P}$, the contribution to ε_t is (disregarding logarithmic factors in P):

$$\frac{p(x_i)}{P^3 p^3(x_1)} \sim \frac{1}{P}, \tag{92}$$

  using (87).

Combining the contributions (90), (91) and (92) we get that the leading contribution in P to the test error ε_t is given by the relation (81).

□

□

C.2. Numerics

Considering $\frac{\chi+1}{2} > \xi > 0$ and $\chi = 2$ and $\chi = 4$, the prediction (9) still holds, as shown in Fig. 8, realised for $\sigma = 100$ and $\lambda = 10^{-12}$. e use as a ridge $\lambda = 10^{-12}$ and not exactly 0 to avoid numerical instabilities due to the inversion of the matrix in (2). We choose $\lambda = 10^{-12}$ because it is smaller than the values of the eigenvalues of the Gram matrices used in (2), but it is large enough to avoid the instabilities. In Appendix F there are further details for the sampling scheme used for the training and test sets.

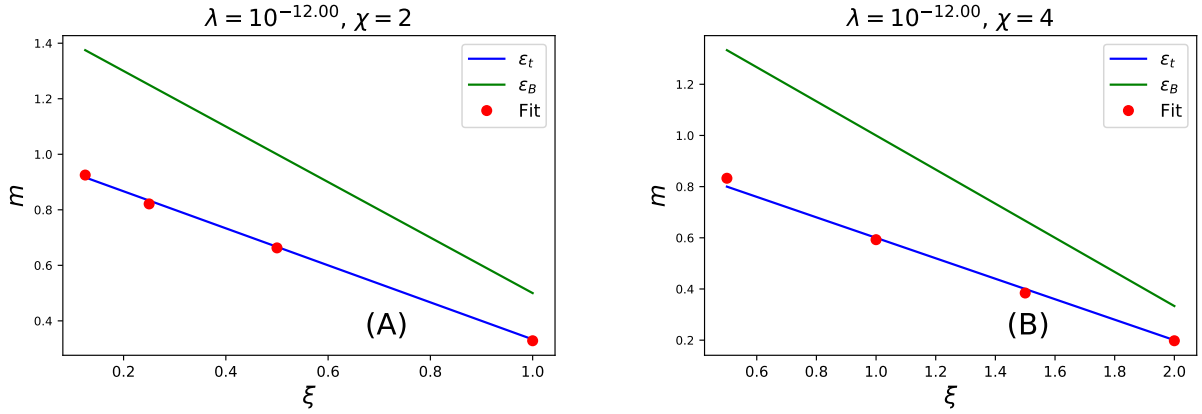

Figure 8. $d = 1$, ridge $\lambda = 10^{-12}$, $\chi = 2$ (A) and $\chi = 4$ (B). Exponent m of the relation $\varepsilon \sim P^{-m}$, where ε is the test error and P the training set size, with respect to ξ of the true function (8). Red points: exponent from a fit of the simulations, for $P \in [10, 10^4]$, averaged over 20 realizations. Blue crosses: exponent of ε_t from the prediction (9). Green crosses: exponent of ε_B from the spectral bias prediction (26).

D. Eigendecomposition Proofs and Numerics

D.1. Proofs

We remind, as presented in Th. 3.2 in the main text, that the eigenfunctions ϕ_ρ of the Laplace kernel satisfy the following differential equation:

$$\phi_\rho''(x) = \left(-2\frac{p(x)}{\lambda_\rho \sigma} + \frac{1}{\sigma^2}\right)\phi_\rho(x). \tag{93}$$

We proceed to solve the differential equation (14) to get an explicit form of the ϕ_ρ. Then we compute the coefficients c_ρ obtained decomposing the true function f^* onto the eigenfunctions:

$$c_\rho = \int dx\, p(x) f_\xi^*(x)\phi_\rho(x) \tag{94}$$

Lastly, we obtain a numerical scheme to get small eigenvalues λ_ρ.

Proposition D.1. *(Coefficients) Let K be the Laplacian kernel with width $\sigma > 0$: $K(x, y) = K(|x - y|) = \exp(-||x - y||_2/\sigma)$. Let $p(x)$ be (7). Consider a small eigenvalue $\lambda_\rho \ll 1$. Let ϕ_ρ be the solution of (93). We impose that $\phi_\rho(x) \to 0$ for $|x| \to \infty$. Then the following holds for the coefficient $|c_\rho|$ defined in (94), in the limit $\lambda_\rho \ll 1$:*

$$\begin{aligned} |c_\rho| &\sim \lambda_\rho^{\frac{\frac{3}{4}\chi+1-2\xi}{\chi+2}} && \text{if } \phi_\rho \text{ is odd} \\ |c_\rho| &= 0 && \text{if } \phi_\rho \text{ is even.} \end{aligned} \tag{95}$$

Proof. Let's consider the differential equation (93), satisfied by the eigenfunctions ϕ_ρ. We will solve that differential equation for small λ_ρ, then we will compute the integral which defines c_ρ (94) at the leading order in λ_ρ.

Lemma D.2 (Eigenfunctions for $\chi > 0$)**.** *Let ϕ_ρ be a solution of (93) for a given λ_ρ. For $\chi > 0$ and $x > 0$, let x_1 and x_2 be the roots of the function:*

$$\Gamma^2(x) = \frac{2}{\lambda_\rho \sigma} p(x) - \frac{1}{\sigma^2}, \tag{96}$$

where $p(x)$ is given by (7). Let Ai and Bi be the Airy functions of first and second kind (Florentin et al., 1966). We impose that $\phi_\rho(x) \to 0$ for $|x| \to \infty$. If ϕ_ρ is odd in x, then some positive coefficients α, β, ζ, δ_1, δ_2 exist, independent on λ_ρ,

such that ϕ_ρ is approximated at the leading order in λ_ρ by the following definition by parts, for $x > 0$:

$$\phi_\rho(x) \simeq \begin{cases} \phi_\rho^{(I)}(x), & \text{for } x \in [0, \beta\lambda_\rho^{\frac{1}{\chi}}] \\ \phi_\rho^{(II)}(x), & \text{for } x \in [\beta\lambda_\rho^{\frac{1}{\chi}}, x_2 - \frac{\delta_1}{\sqrt{-\log\lambda_\rho}}] \\ \phi_\rho^{(IV)}(x), & \text{for } x \in [x_2 - \frac{\delta_1}{\sqrt{-\log\lambda_\rho}}, x_2 + \frac{\delta_2}{\sqrt{-\log\lambda_\rho}}] \\ \phi_\rho^{(V)}(x), & \text{for } x \in [x_2 + \frac{\delta_2}{\sqrt{-\log\lambda_\rho}}, \infty] \end{cases} \tag{97}$$

with:

$$\phi_\rho^{(I)}(x) = \frac{\alpha}{\lambda_\rho^{1/12}} \left(Ai(\mu - \nu x) - \gamma_1 Bi(\mu - \nu x)\right) \tag{98}$$

$$\phi_\rho^{(II)}(x) \simeq \frac{\alpha}{\lambda_\rho^{1/4}} \left(Ai(\xi(x)) - \gamma_1 Bi(\xi(x))\right) \frac{|\xi(x)|^{1/4}}{|\Gamma^2(x)|^{1/4}} \tag{99}$$

$$\phi_\rho^{(IV)}(x) = W_1 Ai\left[\left(\frac{2x_2}{\sigma^2}\right)^{1/3}(x - x_2)\right] + W_2 Ai\left[\left(\frac{2x_2}{\sigma^2}\right)^{1/3}(x - x_2)\right] \tag{100}$$

$$\phi_\rho^{(V)}(x) \sim \frac{\alpha(\sin\theta - \gamma_1\cos\theta)}{2\sqrt{\pi}(-p(x) + \lambda_\rho)^{1/4}} \exp\left(-\int_{x_2}^{x} \sqrt{-\Gamma^2(z)}dz\right) \tag{101}$$

where we have introduced the notation $\mu = \left(\frac{\chi(\lambda_\rho\Gamma[\frac{1+\chi}{2}])^{\frac{2}{\chi}}}{2^{\frac{2}{\chi}}\sigma^{2(1+\chi)}}\right)^{1/3}$, $\nu = \left(\frac{2^{1-\frac{1}{\chi}}\chi}{\sigma^{2-\frac{1}{\chi}}\lambda_\rho^{\frac{1}{\chi}}}\right)^{1/3}$, $\gamma_1 = Ai(\mu)/Bi(\mu)$, $\xi(x) = -\left[\frac{3}{2}\int_{x_1}^{x}\sqrt{\Gamma^2(z)}dz\right]^{2/3}$, $\Gamma^2(x)$ defined in (96), $\theta = \int_{x_1}^{x_2}\sqrt{\Gamma^2(z)}dz + \frac{\pi}{4}$. The coefficients W_1 and W_2 are found matching the solution parts (99), (100) and (101). They are such that $W_1 \sim W_2 \sim \lambda_\rho^{-1/4}$.
This definition by parts is to be interpreted as made of two matching parts $\phi_\rho^{(I)}$ and $\phi_\rho^{(IV)}$ around the roots x_1 and x_2 of the function $\Gamma^2(x)$, a bulk part $\phi_\rho^{(II)}$ for $x_1 < x < x_2$, and a tail part $\phi_\rho^{(V)}$ for $x \gg x_2$. In particular, the bulk part $\phi_\rho^{(II)}$ (99) can be simplified for x far from the roots x_1 and x_2. More precisely, in the interval $[\zeta\lambda_\rho^{\frac{1}{(2+\chi)}}, x_2 - \frac{\delta_1}{\sqrt{-\log\lambda_\rho}}]$, it holds $\phi_\rho^{(II)}(x) \simeq \phi_\rho^{(III)}(x)$ with:

$$\phi_\rho^{(III)}(x) \sim \frac{\alpha}{(p(x) - \lambda_\rho)^{1/4}} \left(\sin\left(\int_{x_1}^{x}\sqrt{\Gamma^2(z)}dz + \frac{\pi}{4}\right) - \ + \gamma_1\cos\left(\int_{x_1}^{x}\sqrt{\Gamma^2(z)}dz + \frac{\pi}{4}\right)\right). \tag{102}$$

If ϕ_ρ is instead even in x, and such that $\phi_\rho^{'}(0) = 0$, then it has the same form as the equations (98)-(101) for $x > 0$, but the coefficient γ_1 is defined as $Ai'(\mu)/Bi'(\mu)$.

Proof. In this proof we will consider for simplicity the following form of the differential equation (93):

$$\phi_\rho^{''}(x) + \tilde{\Gamma}^2(x)\phi_\rho(x) = 0, \tag{103}$$

with $\tilde{\Gamma}^2$ defined as follows:

$$\tilde{\Gamma}^2(x) = \frac{1}{\lambda_\rho}|x|^\chi e^{-x^2} - 1, \tag{104}$$

where we get rid of the numerical coefficients 2, σ and $\Gamma[\frac{1+\chi}{2}]$ present in (93) and (96). Then in the final results (98)-(101) we put back these coefficients.

We will start with the odd eigenfunctions ϕ_ρ. The first thing we remark is that the function $\tilde{\Gamma}^2$ has two roots for $\chi > 0$:

$$x_1 \sim (\lambda_\rho)^{1/\chi}, \qquad x_2 \sim \sqrt{-\log \lambda_\rho}. \tag{105}$$

To obtain the first piece of the eigenfunction (98), we expand the function $\tilde{\Gamma}^2(x)$ in x around x_1:

$$\begin{aligned}\tilde{\Gamma}^2(x) \sim & e^{-\lambda_\rho^{2/\chi}} \left(\chi - 2\lambda_\rho^{2/\chi}\right) \lambda_\rho^{-\frac{1}{\chi}} \left(x - \lambda_\rho^{1/\chi}\right) + \\ & + \frac{1}{2} e^{-\lambda_\rho^{2/\chi}} \left(-2(2\chi+1)\lambda_\rho^{2/\chi} + 4\lambda_\rho^{4/\chi} + (\chi-1)\chi\right) \lambda_\rho^{-\frac{2}{\chi}} \left(x - \lambda_\rho^{1/\chi}\right)^2\end{aligned} \tag{106}$$

We want to truncate this expansion at first order in $\left(x - \lambda_\rho^{1/\chi}\right)$. The solution of (103) with the truncated expansion will be equal to the full solution form 0 up to a certain x^*. We find x^* as the point such that the second order of the expansion (106) is of the same order of the first order, at the leading order in λ_ρ. For $\chi \neq 1$, the comparison of the two orders in (106) yields:

$$\lambda_\rho^{-1/\chi}(x - \lambda_\rho^{1/\chi}) \sim \lambda_\rho^{-2/\chi}(x - \lambda_\rho^{1/\chi})^2, \tag{107}$$

from which we have:

$$x^* \sim \lambda_\rho^{1/\chi} \tag{108}$$

For $\chi = 1$ the relation (108) holds again, since x^* depends continuously on χ. In other words, it exists a constant β, independent on λ_ρ, such that $x^* = \beta\lambda_\rho^{1/\chi} > x_1$.

For $x \in [0, x^*]$, the equation (103) becomes, at the leading order in λ_ρ:

$$\phi_\rho''(x) + \lambda_\rho^{-\frac{1}{\chi}} \left(x - \lambda_\rho^{1/\chi}\right) \phi_\rho(x) = 0, \tag{109}$$

whose solution $\phi_\rho^{(I)}$, with boundary condition $\phi_\rho(0) = 0$, is given by (98). The factor $\lambda_\rho^{-1/12}$ is due to the normalisation to 1 of the full ϕ_ρ, as we will see later in the proof.

The solution of (103), for x distant from the roots x_1 and x_2 of $\tilde{\Gamma}^2$, can be found by means of the technique of the Modified Airy Function (MAF) (A.K.Ghatak, 1991). The solution of (103) is given by:

$$\phi_\rho^{(II)}(x) \sim \left(Q_1 \mathrm{Ai}(\xi(x)) + Q_2 \mathrm{Bi}(\xi(x))\right) \frac{|\xi(x)|^{1/4}}{|\Gamma^2(x)|^{1/4}}, \tag{110}$$

where Q_1 and Q_2 are constants to be determined and $\xi(x) = -\left[\frac{3}{2}\int_{x_1}^{x} \sqrt{\tilde{\Gamma}^2(z)}dz\right]^{2/3}$. The solution (110) holds up for the points x such that the following inequality is valid:

$$\left|(\xi')^{-3}\left(\frac{3(\xi'')^2}{4\xi'} - \frac{1}{2}\xi'''\right)\right| \ll |\xi|. \tag{111}$$

The relation (111) holds for $x \in [x^*, x_2 - \frac{\delta_1}{\sqrt{\log \lambda_\rho}}]$, since in that interval we have, at the leading order in λ_ρ:

$$\xi(x) \sim \xi'(x) \sim \xi''(x) \sim \xi'''(x) \sim \lambda_\rho^{-1/3}. \tag{112}$$

The constants Q_1 and Q_2 can be found matching the solution (110) with (98), taking care of the fact that:

$$\frac{\xi^{1/4}(x)}{\left(\tilde{\Gamma}^2(x)\right)^{1/4}} \xrightarrow[x \to (x^*)^+]{} \lambda_\rho^{1/6}, \tag{113}$$

then obtaining:

$$\phi_\rho^{(II)}(x) \sim \frac{\alpha}{\lambda_\rho^{1/4}} \left(\mathrm{Ai}(\xi(x)) - \gamma_1 \mathrm{Bi}(\xi(x))\right) \frac{|\xi(x)|^{1/4}}{|\Gamma^2(x)|^{1/4}}, \tag{114}$$

The relation (114) can be approximated in the following way. The Airy function Ai has the following asymptotic approximation for large negative y (A.K.Ghatak, 1991):

$$\begin{aligned}\mathrm{Ai}(y) \xrightarrow[y\to-\infty]{} \frac{1}{\sqrt{\pi} y^{1/4}} \Bigg[&\sin\left(y^{\frac{3}{2}} + \frac{\pi}{4}\right) \sum_{k=0}^{\infty} (-1)^k c_{2k} (y^{3/2})^{-2k} + \\ &- \cos\left(y^{\frac{3}{2}} + \frac{\pi}{4}\right) \sum_{k=0}^{\infty} (-1)^k c_{2k+1} (y^{3/2})^{-2k-1} \Bigg],\end{aligned} \tag{115}$$

where:

$$c_0 = 1, \quad c_k = \frac{\Gamma(3k + \frac{1}{2})}{54^k k! \Gamma(k + \frac{1}{2})}, \; k \geq 1. \tag{116}$$

As a consequence, for large y:

$$\mathrm{Ai}(y) \sim \frac{1}{\sqrt{\pi} y^{1/4}} \sin\left(y^{\frac{3}{2}} + \frac{\pi}{4}\right). \tag{117}$$

Similarly, it holds for Bi:

$$\mathrm{Bi}(y) \sim \frac{1}{\sqrt{\pi} y^{1/4}} \cos\left(y^{\frac{3}{2}} + \frac{\pi}{4}\right). \tag{118}$$

Using these asymptotic relations with $y = \xi(x)$, we find that when they are valid the solution $\phi_\rho^{(II)}$ (114) can be approximated by $\phi_\rho^{(III)}$ (102). We remark that the solution (102) can be found in the literature under the name of WKB approximation (A.K.Ghatak, 1991) and it is of interest in the field of quantum mechanics.

We want to define better the interval where we can use the approximation (102) of ϕ_ρ. The upper bound of that integral will be given by the limit of validity of (114), hence $\frac{\delta_1}{\sqrt{-\log \lambda_\rho}}$. The lower bound of that interval will be given by the $x = \hat{x}$ such that the zero and first order in the expansion (115) are of the same order, hence when:

$$1 \sim \frac{1}{(\xi(\hat{x}))^{3/2}} \tag{119}$$

For small x and small λ_ρ, using (105), we have the following asymptotic relation:

$$\int_{x_1}^{\hat{x}} \left(\frac{x^\chi}{\lambda_\rho} e^{-x^2} - 1\right)^{1/2} \sim \frac{\hat{x}^{\frac{\chi}{2}+1}}{\sqrt{\lambda_\rho}}. \tag{120}$$

Combining (119) and (120), we get:

$$\hat{x} \sim \lambda_\rho^{\frac{1}{2+\chi}}, \tag{121}$$

hence it exists a $\zeta > 0$, independent on λ_ρ, such that $\hat{x} = \zeta \lambda_\rho^{\frac{1}{2+\chi}}$.

Since the MAF solution (99) is not valid around x_2, to get the form of ϕ_ρ in that region we linearize $\tilde{\Gamma}^2(x)$ around x_2, and then we solve exactly the differential equation in the region where this approximation is valid. Since we are looking at large x (since $x_2 \sim \sqrt{-\log \lambda_\rho}$), we can approximate $x^\chi e^{-x^2}$ in (104) with e^{-x^2}. The expansion of $\tilde{\Gamma}^2(x)$ around x_2 up to second order then yields:

$$\tilde{\Gamma}^2(x) \sim -2\sqrt{-\log(\lambda_\rho)} \left(x - \sqrt{-\log(\lambda_\rho)}\right) + (-2\log(\lambda_\rho) - 1) \left(x - \sqrt{-\log(\lambda_\rho)}\right)^2. \tag{122}$$

The truncation at the first order of (122) is valid in a region around x_2 such that the at the boundaries of that region the second order in (122) is of the same order in λ_ρ of the first order. This happens for x such that:

$$|x - \sqrt{-\log \lambda_\rho}| \sim \frac{1}{\sqrt{-\log \lambda_\rho}} \tag{123}$$

Consequently, there exist coefficients δ_1 and δ_2 independent on λ_ρ such that the differential equation (103) with $\tilde{\Gamma}^2(x)$ approximated up to the first order in (122) has a solution (100):

$$\phi_\rho^{(IV)}(x) = W_1 \text{Ai}\left[\left(\frac{2x_2}{\sigma^2}\right)^{1/3}(x+x_2)\right] + W_2 \text{Ai}\left[\left(\frac{2x_2}{\sigma^2}\right)^{1/3}(x+x_2)\right], \quad \text{for } x \in [x_2 - \frac{\delta_1}{\sqrt{-\log\lambda_\rho}}, x_2 + \frac{\delta_2}{\sqrt{-\log\lambda_\rho}}] \tag{124}$$

The coefficients W_1 and W_2 are found matching (124) with the solution parts of ϕ_ρ before and after its interval of validity. Since $\phi_\rho^{(III)}(x)$ in $x = x_2 - \frac{\delta_1}{\sqrt{-\log\lambda}}$ has amplitude $\sim e^{\frac{1}{4}x_2^2} \sim \lambda_\rho^{-1/4}$, then $W_1 \sim W_2 \sim \lambda_\rho^{-1/4}$.

To find the solution for $x \geq \left(x_2 + \frac{\delta_2}{\sqrt{-\log\lambda_\rho}}\right)$, we can make use of some formulae in the literature, called "connection formulae" (A.K.Ghatak, 1991), which map the WKB solution for $x < x_2$ (at left of the arrow) into the one for $x > x_2$ (at right):

$$\begin{aligned}
&\frac{2}{(\tilde{\Gamma}^2(x))^{1/4}}\sin\left[\int_x^{x_2} dx(\tilde{\Gamma}^2(x))^{1/2} + \frac{\pi}{4}\right] \to \frac{2}{(-\tilde{\Gamma}^2(x))^{1/4}}\exp\left[-\int_{x_2}^{x} dx(-\tilde{\Gamma}^2(x))^{1/2}\right] \\
&\frac{1}{(\tilde{\Gamma}^2(x))^{1/4}}\cos\left[\int_x^{x_2} dx(\tilde{\Gamma}^2(x))^{1/2} + \frac{\pi}{4}\right] \to \frac{2}{(-\tilde{\Gamma}^2(x))^{1/4}}\exp\left[\int_{x_2}^{x} dx(-\tilde{\Gamma}^2(x))^{1/2}\right]
\end{aligned} \tag{125}$$

The idea behind the proof of these formulae is to take the exact solution of the equation (103) in the interval close to x_2, and then expand the Airy functions at left and at right of x_2. Using the relations (125), the solution for $x \geq \left(x_2 + \frac{\delta_2}{\sqrt{-\log\lambda_\rho}}\right)$ is:

$$\begin{aligned}
\phi_\rho^{(V)}(x) \sim &\frac{\alpha(\sin\theta - \gamma_1\cos\theta)}{2\sqrt{\pi}(-x^\chi e^{-x^2} + \lambda_\rho)^{1/4}}\exp\left(-\int_{x_2}^{x}\sqrt{-\tilde{\Gamma}^2(z)}dz\right) + \\
&+ \frac{\alpha(\gamma_1\sin\theta + \cos\theta)}{\sqrt{\pi}(-x^\chi e^{-x^2} + \lambda_\rho)^{1/4}}\exp\left(\int_{x_2}^{x}\sqrt{-\tilde{\Gamma}^2(z)}dz\right),
\end{aligned} \tag{126}$$

where $\theta = \int_{x_1}^{x_2}\sqrt{\Gamma^2(z)}dz + \frac{\pi}{4}$. Now we impose that $\phi_\rho(x) \to 0$ for $x \to \infty$, in order not to have exponentially divergent terms which would make the norm of ϕ_ρ infinite. This request is equivalent to imposing the condition:

$$\gamma_1\sin\theta + \cos\theta = 0, \tag{127}$$

which we will see in a different Proposition that it fixes the eigenvalues λ_ρ. Imposing this boundary condition, we find $\phi_\rho^{(V)}$ as in (101).

What it is left is the proof of the factor $\lambda_\rho^{-1/12}$ present in (98), which then fixes the dependence on λ_ρ of the normalisation coefficients of the relations (99)-(101). We show now that the factor is such that the eigenfunction ϕ_ρ is normalised to 1. More specifically, we show that the norm of ϕ_ρ does not depend on λ_ρ at the leading order in λ_ρ, and then the constant α in (98)-(101), independently on λ_ρ, fixes the norm of ϕ_ρ to 1.

The norm $||\phi_\rho^2||_p$ of ϕ_ρ is defined as follows:

$$||\phi_\rho||_p = \int_{-\infty}^{\infty} dx\, p(x)\phi_\rho^2(x). \tag{128}$$

We restrict to $x \geq 0$ since the integrand is even in x and we divide this norm in five pieces, analysing them one by one, disregarding the numerical factors and looking just at the behaviour with respect to λ_ρ for clarity purposes.

- For $x \in [0, \beta\lambda_\rho^{\frac{1}{\chi}}]$ we use (98). The contribution to the norm (128) is the following :

$$\int_0^{\beta\lambda_\rho^{\frac{1}{\chi}}} dx\, x^\chi e^{-x^2}\left(\phi_\rho^{(I)}(x)\right)^2. \tag{129}$$

Squaring (98) we get four terms. We analyze one of them and the same logic can be applied to the other three terms, since at the leading order in λ_ρ the factor γ_1 is of order $\mathcal{O}(1)$.

$$\frac{1}{\lambda_\rho^{1/6}}\int_0^{\beta\lambda_\rho^{\frac{1}{\chi}}} dx\, x^\chi e^{-x^2}\mathrm{Ai}^2(\lambda_\rho^{\frac{2}{3\chi}} - x\lambda_\rho^{-\frac{1}{3\chi}}) \tag{130}$$

We do the change of variable $y = x\lambda_\rho^{-\frac{1}{3\chi}}$:

$$\lambda_\rho^{\frac{1}{6}+\frac{1}{3\chi}}\int_0^{\beta\lambda_\rho^{\frac{2}{3\chi}}} dy y^\chi e^{-y^2\lambda_\rho^{\frac{2}{3\chi}}}\mathrm{Ai}^2(\lambda_\rho^{\frac{2}{3\chi}} - y). \tag{131}$$

Since we are interested at the leading order of (131) in λ_ρ:

$$\lambda_\rho^{\frac{1}{6}+\frac{1}{3\chi}}\int_0^{\beta\lambda_\rho^{\frac{2}{3\chi}}} dy y^\chi e^{-y^2\lambda_\rho^{\frac{2}{3\chi}}}\mathrm{Ai}^2(\lambda_\rho^{\frac{2}{3\chi}} - y) \sim \lambda_\rho^{\frac{1}{6}+\frac{1}{3\chi}}\int_0^{\beta\lambda_\rho^{\frac{2}{3\chi}}} dy y^\chi \mathrm{Ai}^2(-x) \tag{132}$$

Since the function Ai is continuous in the integration interval, we can bound (129) from above with a constant $F > 0$. Then we have that:

$$\lambda_\rho^{\frac{1}{6}+\frac{1}{3\chi}}\int_0^{\beta\lambda_\rho^{\frac{2}{3\chi}}} dy y^\chi \mathrm{Ai}^2(-x) \leq F\lambda_\rho^{\frac{5}{6}+\frac{1}{\chi}}. \tag{133}$$

Hence:

$$\int_0^{\beta\lambda_\rho^{\frac{1}{\chi}}} dx\, x^\chi e^{-x^2}\left(\phi_\rho^{(I)}(x)\right)^2 \leq F_1\lambda_\rho^{\frac{5}{6}+\frac{1}{\chi}}, \tag{134}$$

with $F_1 > 0$.

- For $x \in [\beta\lambda_\rho^{\frac{1}{\chi}}, \zeta\lambda_\rho^{\frac{1}{2+\chi}}]$ we use (99). As for $\phi^{(I)}$, we study one of the four terms we get doing the square of $\phi^{(II)}$:

$$\begin{aligned}&\frac{1}{\lambda_\rho^{1/2}}\int_{\beta\lambda_\rho^{\frac{1}{\chi}}}^{\zeta\lambda_\rho^{\frac{1}{2+\chi}}} dx\, x^\chi e^{-x^2}\mathrm{Ai}^2(\xi(x))\frac{|\xi(x)|^{1/2}}{|\Gamma^2(x)|^{1/2}} \sim \\ &\frac{1}{\lambda_\rho^{1/6}}\int_{\beta\lambda_\rho^{\frac{1}{\chi}}}^{\zeta\lambda_\rho^{\frac{1}{2+\chi}}} dx\, x^\chi \frac{\left(\int_{x_1}^x (z^\chi - \lambda_\rho)^{\frac{1}{2}}dz\right)^{1/3}}{(x^\chi - \lambda_\rho)^{1/2}}\mathrm{Ai}^2\left(\frac{1}{\lambda_\rho^{1/3}}\left(\int_{x_1}^x (z^\chi - \lambda_\rho)^{\frac{1}{2}}dz\right)^{2/3}\right)\end{aligned} \tag{135}$$

At the leading order in λ_ρ we can do the following approximation, recalling (105):

$$\int_{x_1}^x (z^\chi - \lambda_\rho)^{\frac{1}{2}}dz \sim x^{\frac{\chi}{2}+1}, \tag{136}$$

which we insert in (135). Then we perform in (135) the substitution $y = x - \beta\lambda_\rho^{\frac{1}{\chi}}$, getting at the leading order in λ_ρ:

$$\frac{1}{\lambda_\rho^{1/6}}\int_0^{\zeta\lambda_\rho^{\frac{1}{2+\chi}}} dy\, y^{\chi+\frac{1}{3}(\frac{\chi}{2}+1)-\frac{\chi}{2}}\mathrm{Ai}^2\left(\frac{1}{\lambda_\rho^{1/3}}y^{\frac{2}{3}(\frac{\chi}{2}+1)}\right). \tag{137}$$

We then do the substitution $w = \frac{y^{\frac{2}{3}(\frac{\chi}{2}+1)}}{\lambda_\rho^{1/3}}$ in (137), getting:

$$\lambda_\rho^{\frac{1}{2}}\int_0^1 dw\, w^{\frac{\chi}{2+\chi}}\mathrm{Ai}^2(w) \sim \lambda_\rho^{\frac{1}{2}}, \tag{138}$$

hence obtaining:

$$\int_{\beta\lambda_\rho^{\frac{1}{\chi}}}^{\zeta\lambda_\rho^{\frac{1}{2+\chi}}} dx\, x^\chi e^{-x^2}\left(\phi_\rho^{(II)}(x)\right)^2 \sim \lambda_\rho^{\frac{1}{2}}. \tag{139}$$

- For $x \in [\zeta \lambda_\rho^{\frac{1}{2+\chi}}, x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}]$ we use (102). Squaring $\phi_\rho^{(III)}$ we get four terms. We focus on one of them, and the logic we will use can be applied also to the other three terms. We consider then:

$$\int_{\zeta \lambda_\rho^{\frac{1}{2+\chi}}}^{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}} dx\, x^\chi e^{-x^2} \frac{1}{(x^\chi e^{-x^2} - \lambda_\rho)^{1/2}} \sin^2\left(\int_{x_1}^{x} \sqrt{\Gamma^2(z)} dz + \frac{\pi}{4}\right). \tag{140}$$

We can replace the $\sin^2(...)$ with $\frac{1}{2}(1 - \cos(2...))$, where $(...)$ is the argument of the $\sin^2$ in (140). We look at the first term which comes from this substitution, at the leading order in λ_ρ:

$$\frac{1}{2} \int_{\zeta \lambda_\rho^{\frac{1}{2+\chi}}}^{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}} dx\, x^\chi e^{-x^2} \frac{1}{(x^\chi e^{-x^2} - \lambda_\rho)^{1/2}} \sim \int_0^\infty x^{\chi/2} e^{-\frac{1}{2}x^2} \sim \mathcal{O}(1). \tag{141}$$

For the second term:

$$\begin{aligned} &\frac{1}{2} \int_{\zeta \lambda_\rho^{\frac{1}{2+\chi}}}^{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}} dx\, x^\chi e^{-x^2} \frac{1}{(x^\chi e^{-x^2} - \lambda_\rho)^{1/2}} \sin\left(\int_{x_1}^{x} \sqrt{\Gamma^2(z)} dz + \frac{\pi}{4}\right) \\ &\qquad \leq \frac{1}{2} \int_{\zeta \lambda_\rho^{\frac{1}{\chi}}}^{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}} dx\, x^\chi e^{-x^2} \frac{1}{(x^\chi e^{-x^2} - \lambda_\rho)^{1/2}} = \mathcal{O}(1). \end{aligned} \tag{142}$$

Putting together (141) and (142), and repeating the logic for the other terms coming from squaring $\phi_\rho^{(III)}$, we have:

$$\int_{\zeta \lambda_\rho^{\frac{1}{2+\chi}}}^{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}} dx\, x^\chi e^{-x^2} \left(\phi_\rho^{(III)}(x)\right)^2 = \mathcal{O}(1) \tag{143}$$

- For $x \in [x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}, x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}]$ we use (100). As above, we consider just one of the four terms we get doing the square of $\phi_\rho^{(IV)}$, and the same procedure can be applied to the other three:

$$W_1^2 \int_{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}}^{x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}} dx\, x^\chi e^{-x^2} \mathrm{Ai}^2\left[x_2^{1/3}(x - x_2)\right]. \tag{144}$$

Since $x_2 \sim \sqrt{-\log \lambda_\rho} \gg 1$, we can do the following approximation, where we also made the substituion $y = x - x_2$ and used the fact that $W_1 \sim \lambda_\rho^{-1/4}$:

$$\begin{aligned} &\frac{1}{\lambda_\rho^{1/2}} \int_{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}}^{x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}} dx\, e^{-x^2} \mathrm{Ai}^2\left[x_2^{1/3}(x - x_2)\right] \sim \\ &\qquad \frac{1}{\lambda_\rho^{1/2}} \int_{-\frac{\delta_1}{\sqrt{-\log \lambda_\rho}}}^{\frac{\delta_2}{\sqrt{-\log \lambda_\rho}}} dy\, e^{-(y+x_2)^2} \mathrm{Ai}^2\left[x_2^{1/3} y\right]. \end{aligned} \tag{145}$$

Noticing that $y \ll x_2$ and doing the substitution $w = x_2^{1/3} y$ we get:

$$\lambda_\rho^{1/2} \int_{-(\sqrt{-\log \lambda_\rho})^{2/3}}^{(\sqrt{-\log \lambda_\rho})^{2/3}} dw\, w^2 \mathrm{Ai}^2[w] \sim \lambda_\rho^{1/2}. \tag{146}$$

Consequently:

$$\int_{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}}^{x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}} dx\, x^\chi e^{-x^2} \left(\phi_\rho^{(IV)}(x)\right)^2 \sim \lambda_\rho^{1/2}. \tag{147}$$

- For $x \geq x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}$ we use (101), getting the following:

$$\int_{x_2+\frac{\delta_2}{\sqrt{-\log\lambda_\rho}}}^{\infty} dx\, x^\chi e^{-x^2} \left(\phi_\rho^{(V)}(x)\right)^2 \leq \int_{x_2+\frac{\delta_2}{\sqrt{-\log\lambda_\rho}}}^{\infty} dx\, x^{\frac{\chi}{2}} e^{-\frac{1}{2}x^2} \sim \int_{x_2}^{\infty} dx\, e^{-\frac{1}{2}x^2}. \tag{148}$$

  Then we use the expansion of the erfc function for large x (50), recalling that $x_2 \sim \sqrt{-\log\lambda_\rho}$ (105), getting:

$$\int_{x_2+\frac{\delta_2}{\sqrt{-\log\lambda_\rho}}}^{\infty} dx\, x^\chi e^{-x^2} \left(\phi_\rho^{(V)}(x)\right)^2 \leq H\lambda_\rho^{1/2}, \tag{149}$$

  for $H > 0$.

Combining the contributions (134), (139), (143), (147) and (149) to the norm (128), we get that the norm of ϕ_ρ is of order $\mathcal{O}(1)$ independently on ρ, as wanted.

For the even eigenfunctions $\phi_\rho(x) = \phi_\rho(-x)$ everything in the proof above applies equally, expect for the definition of γ_1 in (98), which becomes:

$$\gamma_2 = \mathrm{Ai}'(\mu)/\mathrm{Bi}'(\mu), \tag{150}$$

with $\mu = \left(\frac{\chi(\lambda_\rho \Gamma[\frac{1+\chi}{2}])^{\frac{2}{\chi}}}{2^{\frac{2}{\chi}}\sigma^{2(1+\chi)}}\right)^{1/3}$. This change in definition is due to the new the boundary condition we impose on ϕ_ρ, which is no more $\phi_\rho(0) = 0$ but $\phi'_\rho(0) = 0$. □

Lemma D.3 (Eigenfunctions for $\chi = 0$)**.** *The eigenfunctions ϕ_ρ have the same form as the relations (99)-(101), but with the following replacement:*

$$x_1 \to 0, \quad \beta \to 0, \quad \mu = 0 \tag{151}$$

Proof. In the case $\chi = 0$ the function $\Gamma(\tilde{x})$ has just one root x_2. Then there is not the solution $\phi_\rho^{(I)}$ and the part of the solution $\phi_\rho^{(II)}$ applies also up to $x = 0$, which gives $\beta = 0$. As a consequence, the quantity γ_1 and γ_2 are defined replacing the argument μ of the Airy functions and their derivatives with 0. Lastly, the integral in $\xi(x)$ in (99) is defined starting from $x = 0$ and not x_1. □

Now that we have an approximated form for the eigenfunctions ϕ_ρ, we can proceed to prove (95). We start considering $\chi > 0$ and odd eigenfunctions $\phi_\rho(x) = -\phi_\rho(-x)$. We can restrict the analysis of the integral (94) which defines c_ρ to $x \geq 0$, since the integrand is an even function, given by $p(x)f^*(x)\phi_\rho(x)$, where p is given by (7) and f^* by (8).

The logic to compute c_ρ is the same as the one used to compute the leading order of the norm $||\phi_\rho||_p$ in (128): we split the integral (94) in five pieces for $x > 0$ and we compute their value at the leading order in λ_ρ. Into each piece of the integral we will use the relative approximation for ϕ_ρ found in relations (98)-(101). We do the computations disregarding numerical factors and looking just at the main behaviour in λ_ρ.

- For $x \in [0, \beta\lambda_\rho^{\frac{1}{\chi}}]$ we use (98). The contribution to the coefficient (94) is the following :

$$\int_0^{\beta\lambda_\rho^{\frac{1}{\chi}}} dx\, x^{\chi-\xi} e^{-x^2} \phi_\rho^{(I)}(x). \tag{152}$$

  We analyze the scaling in ρ of just one of the two terms we get inserting (98) into (152). The same logic can be applied to the other term, since the factor γ_1 is of order $\mathcal{O}(1)$ in λ_ρ.

$$\frac{1}{\lambda_\rho^{1/12}} \int_0^{\beta\lambda_\rho^{\frac{1}{\chi}}} dx\, x^{\chi-\xi} e^{-x^2} \mathrm{Ai}(\lambda_\rho^{\frac{2}{3\chi}} - x\lambda_\rho^{-\frac{1}{3\chi}}) \tag{153}$$

  We do the change of variable $y = x\lambda_\rho^{-\frac{1}{3\chi}}$ and we look at the leading order in λ_ρ:

$$\lambda_\rho^{\frac{1}{4}+\frac{1}{3\chi}-\frac{\xi}{3\chi}} \int_0^{\beta\lambda_\rho^{\frac{2}{3\chi}}} dy y^\chi \mathrm{Ai}(-x) \tag{154}$$

Since the function Ai is continuous in the integration interval, we can bound (129) from above with a constant $K > 0$. Then we have that:

$$\lambda_\rho^{\frac{1}{4}+\frac{1}{3\chi}-\frac{\xi}{3\chi}} \int_0^{\beta\lambda_\rho^{\frac{2}{3\chi}}} dy y^\chi \mathrm{Ai}^2(-x) \leq K\lambda_\rho^{\frac{11}{12}+\frac{1}{\chi}-\frac{\xi}{3\chi}}. \tag{155}$$

Hence:

$$\int_0^{\beta\lambda_\rho^{\frac{1}{\chi}}} dx\, x^\chi e^{-x^2} \phi_\rho^{(I)}(x) \leq K_1 \lambda_\rho^{\frac{11}{12}+\frac{1}{\chi}-\frac{\xi}{3\chi}}, \tag{156}$$

with $K_1 > 0$. The exponent in the right hand side of (156) is positive, since we have imposed in (8) the condition $\xi < \frac{\chi+1}{2}$.

- For $x \in [\beta\lambda_\rho^{\frac{1}{\chi}}, \zeta\lambda_\rho^{\frac{1}{2+\chi}}]$ we use (99). As for $\phi^{(I)}$, we study one of the two terms we get using (99) into (94):

$$\begin{aligned}
&\frac{1}{\lambda_\rho^{1/4}} \int_{\beta\lambda_\rho^{\frac{1}{\chi}}}^{\zeta\lambda_\rho^{\frac{1}{2+\chi}}} dx\, x^{\chi-\xi} e^{-x^2} \mathrm{Ai}(\xi(x)) \frac{|\xi(x)|^{1/4}}{|\Gamma^2(x)|^{1/4}} \sim \\
&\frac{1}{\lambda_\rho^{1/12}} \int_{\beta\lambda_\rho^{\frac{1}{\chi}}}^{\zeta\lambda_\rho^{\frac{1}{2+\chi}}} dx\, x^{\chi-\xi} \frac{\left(\int_{x_1}^x (z^\chi - \lambda_\rho)^{\frac{1}{2}} dz\right)^{1/6}}{(x^\chi - \lambda_\rho)^{1/4}} \mathrm{Ai}\left(\frac{1}{\lambda_\rho^{1/3}} \left(\int_{x_1}^x (z^\chi - \lambda_\rho)^{\frac{1}{2}} dz\right)^{2/3}\right)
\end{aligned} \tag{157}$$

We plug (136) in (157) and we substitute $y = x - \beta\lambda_\rho^{\frac{1}{\chi}}$, getting at the leading order in λ_ρ:

$$\frac{1}{\lambda_\rho^{1/12}} \int_0^{\zeta\lambda_\rho^{\frac{1}{2+\chi}}} dy\, y^{\chi-\xi+\frac{1}{6}(\frac{\chi}{2}+1)-\frac{\chi}{4}} \mathrm{Ai}\left(\frac{1}{\lambda_\rho^{1/3}} y^{\frac{2}{3}(\frac{\chi}{2}+1)}\right). \tag{158}$$

Then we substitute $w = \frac{y^{\frac{2}{3}(\frac{\chi}{2}+1)}}{\lambda_\rho^{1/3}}$ in (158), obtaining:

$$\lambda_\rho^{\frac{\frac{3}{4}\chi+1-\xi}{\chi+2}} \int_0^1 dw\, w^{\frac{3}{2}\frac{\chi+1}{2+\chi}} \mathrm{Ai}(w) \sim \lambda_\rho^{\frac{\frac{3}{4}\chi+1-\xi}{\chi+2}}, \tag{159}$$

hence obtaining:

$$\int_{\beta\lambda_\rho^{\frac{1}{\chi}}}^{\zeta\lambda_\rho^{\frac{1}{2+\chi}}} dx\, x^\chi e^{-x^2} \phi_\rho^{(II)}(x) \sim \lambda_\rho^{\frac{\frac{3}{4}\chi+1-\xi}{\chi+2}}. \tag{160}$$

The exponent in the right hand side of (160) is positive, since we have imposed in (8) the condition $\xi < \frac{\chi+1}{2}$.

- For $x \in [\zeta\lambda_\rho^{\frac{1}{2+\chi}}, x_2 - \frac{\delta_1}{\sqrt{-\log\lambda_\rho}}]$ we use (102). Plugging $\phi_\rho^{(III)}$ in (94) we get two terms. As before, we focus on just one of them:

$$\int_{\zeta\lambda_\rho^{\frac{1}{2+\chi}}}^{x_2-\frac{\delta_1}{\sqrt{-\log\lambda_\rho}}} dx\, x^{\chi-\xi} e^{-x^2} \frac{1}{(x^\chi e^{-x^2} - \lambda_\rho)^{1/4}} \sin\left(\int_{x_1}^x \sqrt{\Gamma^2(z)} dz + \frac{\pi}{4}\right), \tag{161}$$

which becomes, at the leading order in λ_ρ:

$$\int_{\zeta\lambda_\rho^{\frac{1}{2+\chi}}}^{x_2-\frac{\delta_1}{\sqrt{-\log\lambda_\rho}}} dx\, x^{\frac{3}{4}\chi-\xi} e^{-\frac{3}{4}x^2} \sin\left(\frac{1}{\lambda_\rho^{1/2}} \int_0^x z^{\frac{\chi}{2}} e^{-\frac{1}{2}z^2} dz\right). \tag{162}$$

We now make use of the following result (Olver, 2008) for oscillating integrals. Given an integral of the following kind:

$$I[f] = \int_a^b f(x) e^{i\omega g(x)} dx, \tag{163}$$

with f and g sufficently differentiable functions. If $g'(x) \neq 0$ for $x \in [a,b]$, then the following expansion holds:

$$I[f] \sim \sum_{k=1}^{\infty} \frac{1}{(-i\omega)^k} \left[\sigma_k(b) e^{i\omega g(b)} - \sigma_k(a) e^{i\omega g(a)}\right], \tag{164}$$

where $\sigma_1 = \frac{f}{g'}$ and $\sigma_{k+1} = \frac{\sigma_k'}{g'}$ for $k \geq 1$. The relation (164) can be proved integrating by parts.

In our case, we have:

$$\omega = \frac{1}{\lambda_\rho^{1/2}}, \quad f(x) = x^{\frac{3}{4}\chi - \xi} e^{-\frac{3}{4}x^2}, \quad g(x) = \int_0^x z^{\frac{\chi}{2}} e^{-\frac{1}{2}z^2} dz, \tag{165}$$

and:

$$a \sim \lambda_\rho^{\frac{1}{2+\chi}}, \qquad b \sim \sqrt{-\log \lambda_\rho}. \tag{166}$$

Since $g'(x) = x^{\frac{1}{2}\chi} e^{-\frac{1}{2}x^2}$ is never 0 in the interval given by a and b, we can apply (164). Since we are interested in the limit of small λ_ρ (hence of highly oscillating integrals), we can stop at the first order in the expansion (164), getting:

$$\int_{\zeta \lambda_\rho^{\frac{1}{2+\chi}}}^{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}} dx\, x^{\frac{3}{4}\chi - \xi} e^{-\frac{3}{4}x^2} \sin\left(\frac{1}{\lambda_\rho^{1/2}} \int_0^x z^{\frac{\chi}{2}} e^{-\frac{1}{2}z^2} dz\right) \sim \lambda_\rho^{\frac{\frac{3}{4}\chi + 1 - \xi}{\chi+2}} + \lambda_\rho^{3/4}, \tag{167}$$

where the first term comes from $x = a$, and it dominates the second term, which comes from $x = b$. Consequently, we have that:

$$\int_{\zeta \lambda_\rho^{\frac{1}{2+\chi}}}^{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}} dx\, x^{\chi - \xi} e^{-x^2} \left(\phi_\rho^{(III)}(x)\right)^2 \sim \lambda_\rho^{\frac{\frac{3}{4}\chi + 1 - \xi}{\chi+2}} \tag{168}$$

- For $x \in [x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}, x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}]$ we use (100). As above, we consider just one of the two terms we get plugging $\phi_\rho^{(IV)}$ into (94), and the same procedure can be applied to the other one:

$$W_1 \int_{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}}^{x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}} dx\, x^{\chi - \xi} e^{-x^2} \mathrm{Ai}\left[x_2^{1/3}(x - x_2)\right]. \tag{169}$$

Exploiting the fact that $x_2 \sim \sqrt{-\log \lambda} \gg 1$, we do the following approximation, in addition to substituting $y = x - x_2$ and using the fact that $W_1 \sim \lambda_\rho^{-1/4}$:

$$\begin{aligned} \frac{1}{\lambda_\rho^{1/4}} \int_{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}}^{x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}} dx\, e^{-x^2} \mathrm{Ai}\left[x_2^{1/3}(x - x_2)\right] \sim \\ \frac{1}{\lambda_\rho^{1/4}} \int_{-\frac{\delta_1}{\sqrt{-\log \lambda_\rho}}}^{\frac{\delta_2}{\sqrt{-\log \lambda_\rho}}} dy\, e^{-(y + x_2)^2} \mathrm{Ai}\left[x_2^{1/3} y\right]. \end{aligned} \tag{170}$$

Noticing that $y \ll x_2$ and doing the substitution $w = x_2^{1/3} y$ we get:

$$\lambda_\rho^{3/4} \int_{-(\sqrt{-\log \lambda_\rho})^{2/3}}^{(\sqrt{-\log \lambda_\rho})^{2/3}} dw\, w^2 \mathrm{Ai}\,[w] \sim \lambda_\rho^{3/4}. \tag{171}$$

Hence:

$$\int_{x_2 - \frac{\delta_1}{\sqrt{-\log \lambda_\rho}}}^{x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}} dx\, x^{\chi - \xi} e^{-x^2} \phi_\rho^{(IV)}(x) \sim \lambda_\rho^{3/4}. \tag{172}$$

- In the last interval $x \geq x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}$ we plug (101) into (94), obtaining:

$$\int_{x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}}^{\infty} dx\, x^{\chi - \xi} e^{-x^2} \phi_\rho^{(V)}(x) \leq \int_{x_2 + \frac{\delta_2}{\sqrt{-\log \lambda_\rho}}}^{\infty} dx\, x^{\frac{3\chi}{4} - \xi} e^{-\frac{3}{4}x^2} \sim \int_{x_2}^{\infty} dx\, e^{-\frac{3}{4}x^2}. \tag{173}$$

Then we use the expansion of the erfc function for large x (50) and, using (105), we get for a constant $H_1 > 0$:

$$\int_{x_2+\frac{\delta_2}{\sqrt{-\log\lambda_\rho}}}^{\infty} dx\, x^{\chi-\xi} e^{-x^2} \phi_\rho^{(V)}(x) \leq H_1 \lambda_\rho^{3/4}. \tag{174}$$

As it is implied by Lemma D.3, the relations (160), (168), (172) and (174) hold also for $\chi = 0$, since $\phi_\rho^{(II)}$ is valid up to $x = 0$. Combining the contributions (156), (160), (168), (172) and (174) to the integral (94) defining c_ρ, we get the following asymptotic relation for odd eigenfunctions ϕ_ρ:

$$|c_\rho| \sim \lambda_\rho^{\frac{\frac{3}{4}\chi+1-\xi}{\chi+2}}. \tag{175}$$

For even eigenfunctions ϕ_ρ, we have that the coefficient c_ρ is 0, since it is an integral over all $\mathbb{R}$ of the odd function $p(x)f^*(x)\phi_\rho(x)$, with p given by (7) and f^* given by (8). We then get (95).

□

In the following proposition we find a numerical scheme to get small eigenvalues λ_ρ.

Proposition D.4. *(Eigenvalues) Let x_1 and x_2 be the solutions for $x > 0$ of the equation $\frac{2p(x)}{\sigma\lambda_\rho} - \frac{1}{\sigma^2} = 0$ for $p(x)$ given by (7). Let ϕ_ρ be the eigenfunction solution of (93). We impose $\phi_\rho \to 0$ for $|x| \to \infty$. If ϕ_ρ is an odd function in x, then the eigenvalue λ_ρ satisfies the following self-consistent equation for $\chi > 0$:*

$$\lambda_\rho = \left(\frac{\int_{x_1}^{x_2} dx \sqrt{2\frac{p(x)}{\sigma} - \frac{\lambda_\rho}{\sigma^2}}}{\arctan(-\gamma_1^{-1}) + n\pi} \right)^2, \quad \rho = (2n+1) \tag{176}$$

where $n \geq 0$ is an integer, $\gamma_1 = Ai(\mu)/Bi(\mu)$, with Ai and Bi the Airy function of the first and second kind (Florentin et al., 1966) and $\mu = \left(\frac{\chi(\lambda_\rho \Gamma[\frac{1+\chi}{2}])^{\frac{2}{\chi}}}{2^{\frac{2}{\chi}}\sigma^{2(1+\chi)}} \right)^{1/3}$. If ϕ_ρ is an even function in x such that $\phi_\rho'(0) = 0$, then the eigenvalue λ_ρ satisfies the following for $\chi > 0$:

$$\lambda_\rho = \left(\frac{\int_{x_1}^{x_2} dx \sqrt{2\frac{p(x)}{\sigma} - \frac{\lambda_\rho}{\sigma^2}}}{\arctan(-\gamma_2^{-1}) + n\pi} \right)^2, \quad \rho = (2n+2) \tag{177}$$

where $\gamma_2 = Ai'(\mu)/Bi'(\mu)$ and $n \geq 0$ integer.

If $\chi = 0$, the eigenvalues λ_ρ satisfy the same equations (176) and (177) with the following replacements:

$$x_1 \to 0, \quad \mu \to 0. \tag{178}$$

Proof. We start from considering $\chi > 0$ and odd eigenfunctions $\phi_\rho(x)$. In the proof of Lemma D.2, we find that imposing the boundary condition $\phi_\rho(x) \to 0$ for $|x| \to \infty$, the following condition (127) holds:

$$\gamma_1 \sin\theta + \cos\theta = 0, \tag{179}$$

where $\gamma_1 = \text{Ai}(\mu)/\text{Bi}(\mu)$, $\mu = \left(\frac{\chi(\lambda_\rho \Gamma[\frac{1+\chi}{2}])^{\frac{2}{\chi}}}{2^{\frac{2}{\chi}}\sigma^{2(1+\chi)}} \right)^{1/3}$ and:

$$\theta = \theta(\lambda_\rho) = \int_{x_1}^{x_2} \sqrt{2\frac{p(x)}{\lambda_\rho\sigma} - \frac{1}{\sigma^2}} dx + \frac{\pi}{4}. \tag{180}$$

The relation (179) translates into a condition for the eigenvalues λ_ρ. The condition (179) can be solved by the following values of θ:

$$\theta(\lambda_{\rho_1}) = -\arctan(-\gamma_1^{-1}) + n_1\pi, \tag{181}$$

with $n_1 \geq 0$ and $\rho_1 \geq 1$ integers. A similar relation can be found for even eigenfunctions:

$$\theta(\lambda_{\rho_2}) = -\arctan(-\gamma_2^{-1}) + n_2\pi, \tag{182}$$

where $\gamma_2 = \text{Ai}'(\mu)/\text{Bi}'(\mu)$, with $n_2 \geq 0$ and and $\rho_2 \geq 1$ integers. We want now to find a relation between the integers ρ_1 and n_1 and between ρ_2 and n_2. We make the choice that, given a value of $n_1 = n_2 = n$, the eigenvalue of the odd eigenfunction has rank $\rho_1 = 2n + 1$ and the eigenvalue of the even eigenfunction has rank $\rho_2 = 2n + 2$. In that way, the integer n is the index of an eigenvalue doublet. Moreover, the eigenvalues of the odd eigenfunctions have odd rank, while the even ones have even rank.

After this numbering choice, we develop the relation (181) plugging (180) into it, getting then (176). The same can be done to obtain (177).

Thanks to Lemma (D.3), the same logic can be applied to the eigenfunctions for $\chi = 0$, just making the following replacements:

$$x_1 \to 0, \mu \to 0. \tag{183}$$

□

Corollary D.5. *Let λ_ρ satisfy either the relation (176) or (177), with γ defined accordingly to χ as above. Then for large ρ, the following asymptotic relation holds:*

$$\lambda_\rho \sim \rho^{-2} \tag{184}$$

Proof. Since we are looking at the limit of large ranks ρ (small eigenvalues λ_ρ), we can just look at the relation (176) and the same logic can be applied to (177). We rewrite (176) expliciting the relation between ρ and n:

$$\lambda_\rho = \left(\frac{\int_{x_1}^{x_2} dx \sqrt{2\frac{p(x)}{\sigma} - \frac{\lambda_\rho}{\sigma^2}}}{\arctan(-\gamma_1^{-1}) + (\rho - 1)\frac{\pi}{2}} \right)^2. \tag{185}$$

Using (105) we have that, at the leading order in small λ_ρ, the numerator in (185) is given by:

$$\left(\int_{x_1}^{x_2} dx \sqrt{2\frac{p(x)}{\sigma} - \frac{\lambda_\rho}{\sigma^2}} \right)^2 = \mathcal{O}(1), \tag{186}$$

while the denominator, for large ρ:

$$\left(\arctan(-\gamma_1^{-1}) + (\rho - 1)\frac{\pi}{2} \right)^2 \sim \rho^2, \tag{187}$$

then giving the asymptotic relation (184). □

D.2. Numerics

The comparison between the eigenvalues obtained with the self-consistent numerical scheme (176) and (177) and the eigenvalues obtained diagonalising a large Gram matrix shows a good agreement. This is shown in Fig. 9, realised in log-log scale and for $\sigma = 100$ and for $\chi = 0$ and $\chi = 1$. For very large ρ, the eigenvalues of the Gram matrix decay abruptly because of finite-size effects. The scaling (24) captures the asymptotic behaviour of the eigenvalues λ_ρ, as we can notice in Fig. 9.

We can compute exactly the coefficients c_ρ^2 projecting the true function f^* onto the eigenfunctions ϕ_ρ, obtained solving numerically the differential equation (93), using as eigenvalues λ_ρ the ones obtained from the numerical scheme (23) for ranks $\rho \geq 10^3$. For $\rho \leq 10^3$ we use the eigenvalues obtained diagonalising a large Gram matrix $31k \times 31k$. To solve the differential equation (93), we use the method NDSolve in Mathematica. Once we compute them, we can compare their scaling with respect to ρ with the one predicted in (25) and the spectral bias prediction:

$$c_\rho^2 \sim \rho^{-\frac{2\chi+2-\xi}{\chi+1}}, \tag{188}$$

obtained combining the test error scaling (9) and the spectral bias formula (6). The comparison of the simulations shows a better agreement with the prediction (25) than the spectral bias prediction, as shown in Fig. 10, for $\xi = 0$ and $\chi = 0$ and 2. This suggests the *non* applicability of the theory presented in (Bordelon et al., 2020) in our setting, in the ridgeless case.

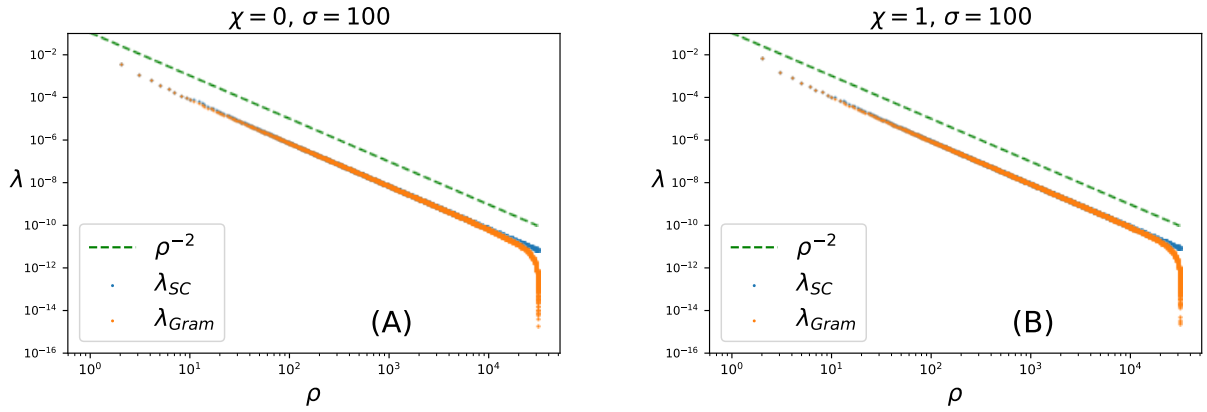

Figure 9. $d = 1$, $\sigma = 100$. Comparison of the eigenvalues λ_ρ obtained via the self-consistent numerical scheme (176) and (177), with label λ_{SC} (blue points), with the eigenvalues obtained diagonalizing a large Gram Matrix $31k \times 31k$, with label λ_{Gram} (orange points), for (A) $\chi = 0$ and (B) $\chi = 1$. The dashed green line ρ^{-2} indicates the predicted scaling $\lambda_\rho \sim \rho^{-2}$ in (24).

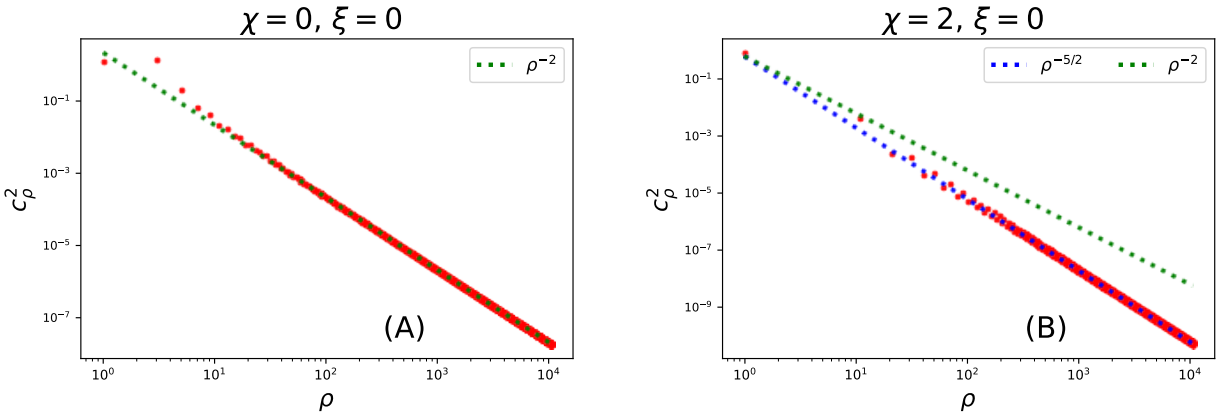

Figure 10. $d = 1$, $\sigma = 100$, $\xi = 0$. Coefficients c_ρ^2 obtained projecting the true function f^* onto the normalized eigenfunction Ψ_ρ solution of the PDE (93). The coefficients are plotted with respect to the rank ρ for (A) $\chi = 0$ and (B) $\chi = 2$. The green dashed line is the spectral bias prediction (188) and the blue dashed line is the theoretical prediction (25). For (A) $\chi = 0$ the two predictions coincide, while for (B) $\chi = 2$ they are different.

E. Proofs for finite ridge λ

E.1. Case $d = 1$

Proposition E.1. *Let K be the Laplacian kernel with width $\sigma > 0$: $K(x, y) = K(|x - y|) = \exp(-||x - y||_2/\sigma)$. Let $f^*(x)$ be the true function and $p(x)$ the data distribution. Then the KRR predictor f_P found via the minimisation problem (1) with ridge λ, in the limit of $P \to \infty$ and $\frac{\lambda}{P}$ finite, is given by the following differential equation:*

$$\sigma^2 \partial_x^2 f_P(x) = \left(\frac{\sigma}{\lambda/P} p(x) + 1 \right) f_P(x) - \frac{\sigma}{\lambda/P} p(x) f^*(x). \tag{189}$$

Proof. We start looking at a way to express the kernel norm $||.||_K$. Then, we find the KRR predictor f_P taking the functional derivative of the minimisation problem (1) and imposing it to be zero.

Let's now look at the kernel norm of the Laplace kernel $K(|x - y|) = \exp(-||x - y||_2/\sigma)$. For a trial function $u(x)$ in the RKHS of the kernel K, the kernel norm is given by:

$$||u||_K^2 = \int dx \int dy\, u(x) K^{-1}(x - y) u(y), \tag{190}$$

where $K^{-1}(x - y)$ is the inverse kernel which satisfies $\int dy K^{-1}(x - y) K(y - z) = \delta(x - z)$. In (Thomas-Agnan, 1996) it is proven that the reproducing kernel $K_0(x, y)$ of the Sobolev space $S_{1,1}$ is given by $K_0(x, y) = e^{-|x-y|}$. This means that the kernel norm $||.||_{K_0}$ is given by the norm of $S_{1,1}$:

$$||u||_{K_0}^2 = ||u||_{S_{1,1}}^2 = \int dt\, u^2(t) + \int dt\, (u'(t))^2, \tag{191}$$

for any function $u(x)$ in $S_{1,1}$. Following the proof of (191) in (Thomas-Agnan, 1996), it is possible to prove that the kernel norm $||.||_K$ with K given by the Laplace kernel $K(|x - y|) = \exp(-||x - y||_2/\sigma)$ is very similar to (191):

$$||u||_K^2 = \frac{1}{\sigma} \left(\int dt\, u^2(t) + \sigma^2 \int dt\, (u'(t))^2 \right). \tag{192}$$

For $P \to \infty$ and λ/P fixed, we can restate the functional (1) which we want to minimise in KRR as follows:

$$\frac{\lambda/P}{\sigma} \left(\int dt\, u^2(t) + \sigma^2 \int dt\, (u'(t))^2 \right) + \int dx\, p(x) \left(f^*(x) - u \right)^2, \tag{193}$$

for a trial function $u(x)$ in the RKHS of the kernel K. If we now take the functional derivative of (193) with respect to u and we put it equal to zero, we get the following differential equation for the KRR predictor f_P:

$$\sigma^2 f_P{}''(x) = \left(\frac{\sigma}{\lambda/P} p(x) + 1 \right) f_P(x) - \frac{\sigma}{\lambda/P} p(x) f^*(x). \tag{194}$$

□

We want now to get the characteristic scale in x of the predictor $f_P(x)$ with respect to λ/P. Since we are considering a $p(x)$ that is even in x and an $f^*(x)$ that is odd in x, the predictor $f_P(x)$ obtained from Eq. (189) will be an odd function of x, therefore $f_P(0) = 0$. We consider the characteristic scale ℓ of $f_P(x)$ for small x and vanishing λ/P as the scale over which the predictor grows from $f_P(0) = 0$ to $f_P(\ell) \sim f^*(\ell)$.

Lemma E.2. *Let's consider the KRR predictor f_P which solves the differential equation (189). Its characteristic scale $\ell(\lambda, P)$, for $x \ll 1$ and $\lambda/P \to 0$, is given by:*

$$\ell(\lambda, P) \sim \left(\frac{\lambda \sigma}{P} \right)^{\frac{1}{(2+\chi)}}. \tag{195}$$

Proof. We notice that the differential equation (189) solved by f_P is an inhomogeneous version of the following homogeneous equation:

$$\sigma^2 u"(x) = \left(\frac{\sigma}{\lambda/P} p(x) + 1\right) u(x), \tag{196}$$

solved by a generic function $u(x)$. Using the variation of parameters method (Teschl, 2004), the general solution f_P of the inhomogeneous equation (189) is given by a linear combination in two independent solutions $u_1(x)$ and $u_2(x)$ of the homogeneous (196):

$$f_P(x) = A(x)u_1(x) + B(x)u_2(x), \tag{197}$$

where the functions $A(x)$ and $B(x)$ satisfy the following relation:

$$A'(x)u_1(x) + B'(x)u_2(x) = 0. \tag{198}$$

Imposing that f_P, in the form (197), solves (196), the following expressions for $A(x)$ and $B(x)$ are obtained:

$$\begin{aligned} A(x) &= \frac{\sigma}{\lambda/P}\int_0^x dy\, \frac{1}{W(y)} u_2(y)p(y)f^*(y) + a \\ B(x) &= -\frac{\sigma}{\lambda/P}\int_0^x dy\, \frac{1}{W(y)} u_1(y)\frac{\sigma}{\lambda/P}p(y)f^*(y) + b, \end{aligned} \tag{199}$$

where a and b are integration constants and W is the Wronskian of u_1 and u_2:

$$W(y) = u_1(y)u_2'(y) - u_1'(y)u_2(y), \tag{200}$$

which is different from 0 since u_1 and u_2 are independent. We now obtain an expansion in λ/P for the solutions u_1 and u_2 . The homogeneous equation (196) belongs to the type of second-order equations that can be solved by the WKB method (A.K.Ghatak, 1991) in the limit of small λ/P. It is a method of multi-scale analysis and the idea behind it is described in Section 3. The generic WKB solution which is proposed to solve (196) has the form:

$$u(x) = e^{\pm i\frac{S(x)}{\lambda/P}}, \quad S(x) = S_0(x) + (\lambda/P)S_1(x) + o\left(\frac{\lambda}{P}\right), \tag{201}$$

where $S(x)$ has been expanded in powers of the small parameter λ/P. Looking at the order 0 in λ/P, we get the following two independent solutions:

$$u_{1,2}(x) = Ce^{\pm\sqrt{\frac{\sigma}{\lambda/P}}\int_0^x d\eta\sqrt{p(\eta)+\frac{\lambda/P}{\sigma}}} + O(\lambda/P), \tag{202}$$

where C is a constant. We stop at the order 0 in λ/P in the expansion (201) since we are interested only in the characteristic scale of $u(x)$ with respect to x, which can be extracted by the exponential in (202). Indeed, considering higher orders in λ/P, we would get a polynomial factor multiplying the exponential in (202), as shown in (20) in the main text. Plugging (202) into the expression of $A(x)$ in (199) we get:

$$A(x) = \sqrt{\frac{\sigma}{\lambda/P}}\int_0^x dy\, u_2(y)\sqrt{p(y)}f^*(y) + a. \tag{203}$$

We now extract the characteristic scale of the first term $A(x)u_1(x)$ in the relation defining f_P in (197). The same analysis will apply for the second term in that relation. The exponent in (202), for small λ/P and small x, is given by $\sqrt{\frac{\sigma}{\lambda/P}}\int_0^x dy\, y^{\chi/2} \propto \sqrt{\frac{\sigma}{\lambda/P}}x^{1+\chi/2}$, yielding the following scale for the functions $u_1(x)$ and $u_2(x)$ at small x:

$$\ell(\lambda, P) \sim \left(\frac{\lambda\sigma}{P}\right)^{\frac{1}{(2+\chi)}}. \tag{204}$$

The characteristic scale of $A(x)$ in (203) is given by the scale of u_2, which is again that of (204). Indeed, for small x and y, the factor $\sqrt{p(y)}f^*(y)$ in (203) is just a polynomial factor $y^{\frac{\chi}{2}-\xi}$, which does not affect the fact that the main scale of $A(x)$ is the one given by u_2. As a consequence, the characteristic scale of the predictor f_P is given by (204). □

We remark that it exists a second, and quicker, way to obtain the characteristic scale ℓ with respect to x of f_P. We notice that the left hand side of Eq. (189) scales as:

$$\sigma^2 f_P''(x) \sim \sigma^2 \frac{f^*(\ell)}{\ell^2} \tag{205}$$

and the right hand side of (189) scales as:

$$\frac{\sigma}{\lambda/P} p(x) \left(f_P(x) - f^*(x)\right) \sim \frac{\sigma}{\lambda/P} p(\ell) f^*(\ell) \sim \frac{\sigma}{\lambda/P} \ell^\chi f^*(\ell) \tag{206}$$

Comparing the two sides we obtain again the characteristic scale (204).

E.2. Case $d > 1$

For a given kernel $K(x-y)$, we consider the predictor $f_P(x)$ that minimizes the functional $\lambda/P||f_P||_K + \int dx\, p(x) \left(f^*(x) - f_P(x)\right)^2$. The predictor $f_P(x)$ can be written as $f_P(x) = \int d^d\eta \frac{p(\eta) f^*(\eta)}{\lambda/P} G(x,\eta)$, where the Green function $G(x,\eta) = G_\eta(x)$ satisfies the equation $\int d^d y K^{-1}(x-y) G_\eta(y) = \frac{p(x)}{\lambda/P} G_\eta(x) + \delta(x-\eta)$, where $\int d^d y K^{-1}(x-y) K(y-z) = \delta(x-z)$. Taking the Fourier transform $\mathcal{F}[...]$ we get

$$\mathcal{F}[K](q)^{-1} \mathcal{F}[G_\eta](q) = \frac{1}{\lambda/P} \mathcal{F}[p\, G_\eta](q) + e^{-iq\eta} \tag{207}$$

where q is the Fourier frequency. We now estimate each of this term, in the limit of small x, large q and vanishing λ/P. Since the term $e^{-iq\eta}$ is such that $|e^{-iq\eta}| = 1$, we drop the dependence of G_η from η.

For small $x \ll 1$, the transform on the right hand side of Eq. (207) becomes $\mathcal{F}[p(x)\ G(x)](q) \sim \mathcal{F}[x^\chi\ G(x)](q) \sim \partial_q^\chi \mathcal{F}[G](q)$. Assuming a power-law behavior (to be confirmed self-consistently) of this quantity for large q, we have $\partial_q^\chi \mathcal{F}[G](q) \sim q^{-\chi} \mathcal{F}[G](q)$.

Furthermore, for a Laplace kernel we have $\mathcal{F}[K](q)^{-1} \sim q^{1+d}$.

Comparing these two terms, we obtain two regimes:

$$\mathcal{F}[G](q) \sim q^{-1-d} \quad \text{for} \quad q \gg q_c \tag{208}$$

$$\mathcal{F}[G](q) \sim \frac{\lambda}{P} q^\chi \quad \text{for} \quad q \ll q_c \tag{209}$$

$$\text{with} \quad q_C \sim \left(\frac{\lambda}{P}\right)^{-\frac{1}{1+d+\chi}} \tag{210}$$

Thus in magnitude, $\mathcal{F}[G](q)$ is maximum for $q \sim q_c$. It implies that in real space, $G(x)$ is characterized by a length scale:

$$\ell(\lambda, P) \sim 1/q_c \sim \left(\frac{\lambda}{P}\right)^{\frac{1}{1+d+\chi}} \tag{211}$$

F. Sampling scheme details

In this appendix we give further details about our sampling scheme for the training and test sets used for KRR simulations.

F.1. One dimension

Training set We sample P points from the PDF (7), in the interval $x \in [-x_{\text{max}}, x_{\text{max}}]$. We do it using the rejection sampling algorithm. We choose $x_{\text{max}} = 3$.

Test set Given λ and P, we find the characteristic length of the predictor f_P, computed from the training set, as follows. We compute the derivative of f_P on a fine grid over $x \in [0, x_{\text{max}}]$. We take as estimate of the characteristic length of f_P the point $\tilde{x}$ such that $f_P'(\tilde{x}) = \frac{1}{10} f_P'(0)$. Then, we divide the interval $x \in [0, x_{\text{max}}]$ in m bins $[x_j, x_{j+1}]$ of width given by $\tilde{x}$, with $j \in 0, ..., m-1$. Then we compute the contribution of the test error ε_t as an integral over a grid made by $10^5 \cdot e^{-j} + q$ points, where q is given by $\max[\frac{1}{m} 10^5, 2000]$. The number q states the minimum number of points per bin. If $\frac{1}{m} 10^5 > 2000$ then we sample at least 10^5 points (in addition to $10^5 \cdot e^{-j}$), otherwise we sample at maximum 2000 points per bin. We sum the contributions over j to get the full ε_t.

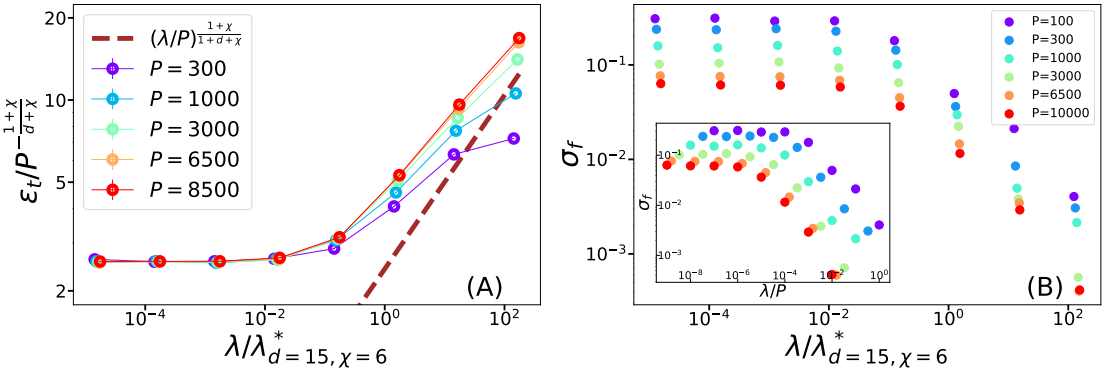

Figure 11. Binary MNIST. (A): Empirical test error ε_t *v.s.* ridge. Each quantity is rescaled by our predictions (32) and (36) for $d = 15$ and $\chi = 6$. The error bars of the average test error are within the symbols. Brown line: asymptotic behavior of ε_B with λ as predicted from Eq. (26). (B) Inset: variance of the predictor σ_f *v.s.* re-scaled ridge λ/P. Main plot: After rescaling the ridge by $\lambda^*_{d=15,\chi=6}$, curves nearly collapse.

F.2. $d > 1$ case

In the case of generic dimension d described in Section 2, the sampling along the informative direction x_1 is the same as in the one-dimensional case above. For the other d coordinates, we first sample from a $d-$dimensional standard Gaussian distribution. Secondly, we normalize these coordinates by their $d-$dimensional L_2 norm, to collocate the points on a cylindrical surface.

G. Additional Figures

In Fig. 11 we repeat the analysis done for CIFAR10 in the main text in Fig. 4 for a binary version of the dataset MNIST.