How isotropic kernels learn simple invariants

J. Paccolat[1], S. Spigler[1], and M. Wyart[1]

[1]*Institute of Physics École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*
[2]*Email: write everybody emails*

Abstract

We investigate how the training curve of isotropic kernel methods depends on the symmetry of the task to be learned, in several settings. (i) We consider a regression task, where the target function is a Gaussian random field that depends only on $d_{\parallel}$ variables, fewer than the input dimension d. We compute the expected test error ϵ that follows $\epsilon \sim p^{-\beta}$ where p is the size of the training set. We find that $\beta \sim 1/d$ independently of $d_{\parallel}$, supporting previous findings that the presence of invariants does not resolve the curse of dimensionality for kernel regression. (ii) Next we consider support-vector binary classification and introduce the *stripe model* where the data label depends on a single coordinate $y(\underline{x}) = y(x_1)$, corresponding to parallel decision boundaries separating labels of different signs, and consider that there is no margin at these interfaces. We argue and confirm numerically that for large bandwidth, $\beta = \frac{d-1+\xi}{3d-3+\xi}$, where $\xi \in (0, 2)$ is the exponent characterizing the singularity of the kernel at the origin. It implies the surprising result that isotropic kernels can overcome the curse of dimensionality for such invariants. (iii) We confirm these findings for the *spherical model* for which $y(\underline{x}) = y(|\underline{x}|)$. (iv) In the stripe model, we show that if the data are compressed along their invariants by some factor λ (an operation believed to take place in deep networks), the test error is reduced by a factor $\lambda^{-\frac{2(d-1)}{3d-3+\xi}}$.

1 Introduction and related works

Deep neural networks are successful at a variety of tasks, yet understanding why they work remains a challenge. In particular, we do not know a priori how many data are required to learn a given rule — not even the order of magnitude. Specifically, let us denote by p the number of examples in the training set. After learning, performance is quantified by the test error $\epsilon(p)$. Quite remarkably, empirically one observes that $\epsilon(p)$ is often well fitted by a power-law decay $\epsilon \propto p^{-\beta} + o\left(p^{-\beta}\right)$. The exponent β is found to depend on the task, on the dataset and on the learning algorithm (Hestness *et al.* 2017; Spigler, Geiger, and Wyart 2019). General arguments would suggest that β should be extremely small — and learning thus essentially impossible — when the dimension d of the data is large, which is generally the case in practice (e.g. in images where d is the number of pixels times the number of color channels). For example in a regression task, if the only assumption on the target function is that it is Lipschitz continuous, then the test error cannot be guaranteed to decay faster than with an exponent $\beta = O\left(1/d\right)$ (Luxburg and Bousquet 2004). This *curse of dimensionality* (Bach 2017) stems from the geometrical fact that the distance δ among nearest-neighbor data points decays extremely slowly in large d as $\delta \sim p^{1/d}$, so that any interpolation method is very imprecise. The mere observation that deep learning works in large dimension implies that data are very structured (Mallat 2016). Yet how to describe mathematically this structure and to build a quantitative theory for β remains a challenge. Our present goal is to understand the relationship between β and symmetries in the data in simple models.

Recently there has been a considerable interest in studying the infinite-width limit of neural networks, motivated by the observation that performance generally improves with the number of parameters (Neyshabur *et al.* 2017; Neyshabur *et al.* 2018; Bansal *et al.* 2018; Advani and Saxe 2017; Spigler *et al.* 2019; Geiger *et al.* 2020). That limit depends on how the weights at initialization scale with the width. For a specific choice, similar to the LeCun initialization often used in practice, deep learning becomes equivalent to a kernel method (Jacot, Gabriel, and Hongler 2018), which has been coined *neural tangent kernel*. In kernel methods, the learned function $Z(\underline{x})$ is a linear combination of the functions $K(\underline{x}, \underline{x}_\mu)$ where $\underline{x}_\mu$ are the training data and the kernel $K(\underline{x}, \underline{y})$. Kernels achieve performances somewhat inferior but still

comparable to the more refined deep networks (Bruna and Mallat 2013; Arora *et al.* 2019), and are often used both for regression and classification. In this work we study the learning curves of isotropic kernels for which $K(\underline{x}, \underline{y}) = K\left(\left\|\underline{x} - \underline{y}\right\|\right)$, that include the popular Gaussian and Laplace kernels.

When these kernels are used on the image datasets `MNIST` and `CIFAR-10`, one finds that the learning curves decay with respective exponents $\beta_{\text{MNIST}} \approx 0.4$ and $\beta_{\text{CIFAR-10}} \approx 0.1$ that are much larger than $1/d_{\text{MNIST}} \approx 10^{-3}$ and $1/d_{\text{CIFAR10}} \approx 3 \cdot 10^{-4}$ (Spigler, Geiger, and Wyart 2019). Several aspects of the data could explain together these findings.

(i) In the kernel literature, upper bounds on the test error with β independent of d are obtained assuming that the target function lies in the *reproducing-kernel Hilbert space* of the kernel.[1] However for these kernels this assumption is rather extreme: it supposes that the number of derivatives of the target function that are smooth is proportional to the dimension itself (Maiorov 2006; Bach 2017), see (Spigler, Geiger, and Wyart 2019) for a precise statement for Gaussian random functions.

(ii) The data live on a manifold $\mathfrak{M}$ of lower dimensionality $d_{\mathfrak{M}} < d$. This is indeed the case for `MNIST` where $d_{\mathfrak{M}} \approx 15$ (Costa and Hero 2004; Hein and Audibert 2005; Rozza *et al.* 2012; Facco *et al.* 2017; Goldt *et al.* 2019) and `CIFAR-10` where $d_{\mathfrak{M}} \approx 35$ (Spigler, Geiger, and Wyart 2019). This effect is presumably important, yet by itself it may not be the resolution of the problem, since the exponents β are significantly larger than $1/d_{\mathfrak{M}}$.

(iii) The function to be learned present many invariants. For example in the context of classification, some pixels at the edge of the image may be unrelated to the class label. Likewise, smooth deformations of the image may leave the class unchanged. It has been argued that the presence of these invariants is central to the success of deep learning (Mallat 2016). In that view, neural networks corresponds to a succession of non-linear and linear operations where invariant directions are compressed. It is supported by the observations that kernels designed to perform such compression perform well (Mallat 2016) and that compression can indeed occur at intermediate layers of deep networks (Shwartz-Ziv and Tishby 2017). Yet, relating quantitatively these views to the learning-curve exponent β remains a challenge, even for simple isotropic kernels and simple models of data. In (Bach 2017), it was shown for a specific kernel in the context of regression that the presence of invariants did not improve guaranties for β. It is currently unclear if this results holds more generally to other kernels, beyond worst case analysis, and to classification tasks.

1.1 Our contribution

Our work consists of two parts that can be read independently, studying respectively regression and classification for different models.

The first part is presented in Section 2 and focuses on kernel regression. We consider a target function that varies only along a linear manifold of $d_\parallel$ directions of the input space, and is invariant in $d - d_\parallel$ directions. Without loss of generality, we consider that this dependence is on $\underline{x}_\parallel \equiv (x_1, \dots, x_{d_\parallel})^t$, corresponding to the $d_\parallel$ first components of the data vectors $\underline{x} = (x_1, \dots, x_d)^t$. The target function is a Gaussian random function $Z_T(\underline{x}) \equiv Z_T(\underline{x}_\parallel)$ with covariance determined by an isotropic translation-invariant Teacher kernel $K_T(\underline{x})$. Kernel ridgeless regression is then performed using a distinct Student kernel $K_S(\underline{x})$. Using the methods of (Spigler, Geiger, and Wyart 2019) inspired by earlier works on kriging (Stein 1999), we compute the learning curve by calculating the expectation of the mean-squared test error and then we extract the exponent β. We find and confirm numerically that β is independent on $d_\parallel$ and depends only on d: one cannot escape the curse of dimensionality. This result supports that even in a typical, non-worst case analysis, regression using simple kernels does not benefit from invariance in the data.

In the second part of this work we study kernel classification with support-vector machines, and find very different conclusions. We focus on simple models of data that are arguably necessary first steps to

1. Such a Hilbert space is the set of all functions f with finite K-norm: $\|f\|_K < \infty$, see (Scholkopf and Smola 2001).

build quantitative predictions for β in more realistic settings. In Section 3 we introduce the *stripe model*, in which the class label $y(\underline{x}) = \pm 1$ only varies in one direction as illustrated in Fig. 3. Again without loss of generality we consider $y(\underline{x}) = y(x_1)$. This model corresponds to parallel interfaces separating regions where the label changes sign. We further consider the case without margin, where the data distribution $\rho(\underline{x})$ is non zero at interfaces.

The performance of kernel classification depends on the bandwidth σ of the kernel, that is the scale over which it varies significantly. If σ is much smaller than the distance δ between training points, then the support-vector machine is tantamount to a nearest-neighbor algorithm, which inevitably suffers from the curse of dimensionality with an exponent $\beta = O(1/d)$. However in the limit of large σ, we provide scaling (heuristic) arguments that we systematically confirm numerically, showing that $\beta = \frac{d-1+\xi}{3d-3+\xi}$, where ξ is an exponent characterizing the singularity of the kernel at the origin (e.g. $\xi = 1$ for a Laplace kernel). This exponent stays finite even in large dimension. This surprising result indicates that for large bandwidth σ, isotropic translation-invariant kernels correspond to a good prior for flat interfaces.

In Section 4, we show that these results are not restricted to flat interfaces: the same exponent β is found for the *spherical model* in which $y(\underline{x}) = y(|\underline{x}|)$. More generally, our analysis suggests that this result will break down if the boundary separating labels shows significant variation below a length scale $r_c \sim p^{-1/(d-1)}$. Avoiding the curse of dimensionality thus requires to have an increasingly regular boundary separating labels as d increases.

Finally in Section 5 we come back to the stripe model, and study how compressing the input data along its invariants (namely all the directions different from x_1) by a factor λ improves performance - an effect believed to play a key role in the success of deep learning (Mallat 2016). We argue and confirm empirically that when mild, such a compression leaves the exponent β unchanged but reduces the test error by a factor $\lambda^{-\frac{2(d-1)}{3d-3+\xi}}$.

1.2 Related works

There is a long history of works computing the learning curve exponent β in regression or classification tasks where the true function or label depends on a single direction in input space, starting from the perceptron model (Gardner 1988) and including support vector classification (Dietrich, Opper, and Sompolinsky 1999). More recently random features models have received a lot of attention, and can be analytically resolved in some cases using random matrix or replica theories (Advani and Saxe 2017; Montanari *et al.* 2019; Gerace *et al.* 2020; d'Ascoli *et al.* 2020). Yet these results for classification generally consider linearly separable data [2] and most importantly for both regression and classification tasks apply in the limit $d \to \infty$ and $p \to \infty$ with $\alpha = p/d$ fixed. In (Dietrich, Opper, and Sompolinsky 1999) for a single interface separating labels and kernels similar to ours, the learning curves of support vector classifier was shown to decreases as $\epsilon \sim 1/\alpha$, as also found for the perceptron (Engel and Van den Broeck 2001). Here we consider both linearly and non-linearly separable data, and take the limit of large training set size p at fixed dimension d. It is in our view warranted considering data sets commonly used as benchmarks, such as MNIST or CIFAR for which $d_{\mathfrak{M}} \in [15, 35]$ and $p \approx 6 \cdot 10^4$. The exponent β we find depends on dimension d and does not converge to 1 as $d \to \infty$, indicating that the two limits do not commute.

2 Kernel regression: Teacher-Student framework

We consider kernel ridgeless regression on Gaussian random data that present invariants. Our framework corresponds to a Teacher-Student setting for supervised learning (Saad and Solla 1995; Monasson and Zecchina 1995; Opper and Saad 2001; Engel and Van den Broeck 2001; Aubin *et al.* 2018; Franz, Hwang, and Urbani 2018), where two variants of the same model (here kernels) are used both to generate the data and to learn them. The target function $Z_T(\underline{x})$ is assumed to be a random Gaussian process $\mathcal{N}(0, K_T)$ with zero mean and covariance determined by a strictly positive-definite isotropic

2. See (Dietrich, Opper, and Sompolinsky 1999) for an example of non-linearly separable data lying on a hypercube.

translation-invariant Teacher kernel $K_T(\underline{x}, \underline{x}') = K_T\left(\left|\underline{x} - \underline{x}'\right|\right)$, implying that that $\mathbb{E}_T Z_T(\underline{x}) = 0$ and $\mathbb{E}_T Z_T(\underline{x}) Z_T(\underline{x}') = K_T(\underline{x}, \underline{x}')$, where we denote by $\mathbb{E}_T$ the expectation over the Teacher Gaussian random process. Strictly positive-definiteness is required to generate such a random function. We further assume that the function $Z_T(\underline{x})$ does not depend on all the variables $\underline{x} = (x_1, \dots, x_d)^t$, but only on the first components $\underline{x}_\parallel \equiv (x_1, \dots, x_{d_\parallel})^t$ for some $d_\parallel \leq d$: $Z_T(\underline{x}) = Z_T(\underline{x}_\parallel)$, as sketched in Fig. 1. The Gaussian random process $Z_T(\underline{x})$ is constant along the subspace of $\underline{x}_\perp \equiv (x_{d_\parallel+1}, \dots, x_d)^t$ when it is generated by a Teacher kernel that has the same property, namely $K_T\left(\left|\underline{x} - \underline{x}'\right|\right) = K_T\left(\left|\underline{x}_\parallel - \underline{x}'_\parallel\right|\right)$. Indeed, we have that

$$\mathbb{E}_T \left[Z_T(\underline{x}_\parallel + \underline{x}_\perp) - Z_T(\underline{x}_\parallel) \right]^2 = 2K_T(0) - 2K_T(\underline{x}_\perp) = 0. \tag{1}$$

The (finite) training set is made up by the values of the target function $Z_T(\underline{x}^\mu)$ at p points $\{\underline{x}^\mu\}_{\mu=1}^p$. For our theorem we make the technical assumption that data points lie on a regular square lattice in $\mathbb{R}^d$, that is, they are equispatiated in a hypercube $[-L, L]^d$ of side length $2L$. The lattice spacing δ follows $(2L)^d = p\delta^d$, therefore $\delta \propto p^{-1/d}$. We will confirm numerically below that our results hold more generally, for example if training points are iid and Gaussian distributed.

Kernel (ridgeless) regression is performed with a Student kernel $K_S(\underline{x}, \underline{x}')$, that we also take to be isotropic and translation invariant and that can be different from the Teacher kernel $K_T(\underline{x}, \underline{x}')$. The Student has no prior knowledge of the presence of invariants: its kernel is a function of all the spatial components.

Kernel regression consists in writing the prediction for the function $\hat{Z}_S(\underline{x})$ at a generic point $\underline{x}$ as a linear combination of Student kernel overlaps on the whole training set, namely:

$$\hat{Z}_S(\underline{x}) = \sum_\mu a^\mu K(\underline{x}^\mu, \underline{x}) \equiv \underline{a} \cdot \underline{k}_S(\underline{x}). \tag{2}$$

The vector of coefficients $\underline{a}$ is determined by minimizing the mean-squared loss on the training set:

$$\underline{a} = \arg\min_{\underline{a}} \sum_\mu \left[\hat{Z}_S(\underline{x}^\mu) - Z_T(\underline{x}^\mu) \right]^2. \tag{3}$$

The minimization of such a quadratic loss can be carried out explicitly, and the Student prediction can be written as

$$\hat{Z}_S(\underline{x}) = \underline{k}_S(\underline{x}) \cdot \mathbb{K}_S^{-1} \underline{Z}_T, \tag{4}$$

where the vector $\underline{Z}_T \equiv (Z_T(\underline{x}^\mu))_{\mu=1}^n$ contains all the samples in the training set and $\mathbb{K}_S^{\mu\nu} \equiv K_S(\underline{x}^\mu, \underline{x}^\nu)$ is

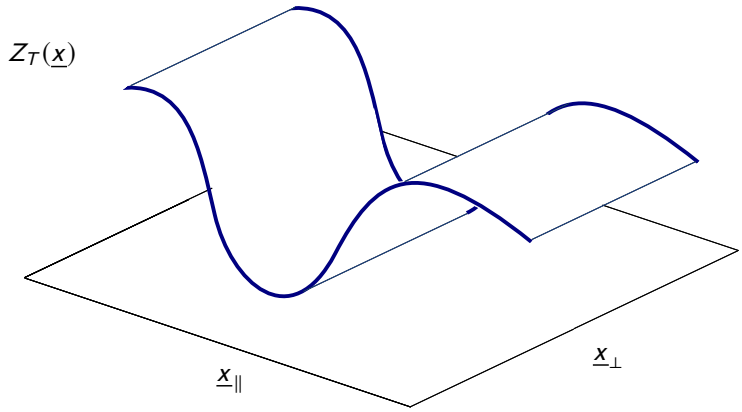

Figure 1. Sketch of a realization of the Gaussian random process $Z_T(\underline{x}) \sim \mathcal{N}(0, K_T)$. The kernel K_T, and consequently the random function Z_T, is constant along the direction $\underline{x}_\perp$ and only depends on $\underline{x}_\parallel$.

the Gram matrix. By definition, the Gram matrix is always invertible for any training set if the kernel K_S is strictly positive definite. The generalization error is then evaluated as the expected mean-squared error on out-of-sample data that were not used for training: numerically, it is estimated by averaging over a test set composed of p_{test} newly-sampled data points, taken uniformly at random in the hypercube $[-L, L]^d$:

$$\epsilon_T = \mathbb{E}_{\underline{x}} \left[\hat{Z}_S(\underline{x}) - Z_T(\underline{x})\right]^2 = \frac{1}{p_{\text{test}}} \sum_{\mu=1}^{p_{\text{test}}} \left[\hat{Z}_S(\underline{x}^\mu) - Z_T(\underline{x}^\mu)\right]^2 . \tag{5}$$

This quantity is a random variable, and we take the expectation also with respect to the Teacher process to define an average test error $\epsilon = \mathbb{E}_T \epsilon_T$ — in the numerical simulations that we discuss later, we simply average over several runs of the Teacher Gaussian process.

We study how the expected test error ϵ decays with the size p of the training set. Asymptotically for large p, this decay follows a power law $\epsilon = O\left(p^{-\beta}\right)$. In (Spigler, Geiger, and Wyart 2019), β was derived in the absence of invariant $d_\parallel = d$, building on results from the kriging literature (Stein 1999). It was found that β depends on three quantities: the dimension d and two exponents $\alpha_T(d)$, $\alpha_S(d)$ related to the two kernels. These exponents describe how the Fourier transform of the kernels decay at large frequencies: $\tilde{K}_T(\underline{w}) = O\left(\|w\|^{-\alpha_T(d)}\right)$, and similarly for the Student K_S. Notice that since the kernels are translation invariant, their Fourier transform is a function of only one frequency vector $\underline{w}$. Moreover, the exponents $\alpha_T(d)$, $\alpha_S(d)$ depend on the dimension of the space where the Fourier transform is computed. Because of the isotropic nature of the kernels, the dependence on the dimension d is fairly simple: we can write the exponents as $\alpha_T(d) = d + \theta_T$ and $\alpha_S(d) = d + \theta_S$, where θ_T and θ_S are quantities that only depend on the kernels taken as functions $K_T(z)$, $K_S(z)$ of the scalar argument $z = \left|\underline{x} - \underline{x}'\right|$. They represent the first non-smooth term in the Taylor expansion of the Kernels near the origin ($z = 0$) and do not depend on the dimension d. A detailed derivation of the relation $\alpha_T(d) = d + \theta_T$ (and analogously for the Student) is carried out in Appendix A.

Our main theorem, formally presented with its proof in Appendix A, is as follows:

Theorem: (informal) *Let ϵ be the average mean-squared error of the regression made with a Student kernel K_S on the data generated by a Teacher kernel K_T, sampled at p points $\{\underline{x}^\mu\}_{\mu=1}^p$ taken on a regular d-dimensional square lattice in $[-L, L]^d$. Assume that the Teacher kernel only varies in a lower dimensional space: $K_T(\underline{x}) = K_T(x_\parallel)$, with $\underline{x}_\parallel = (x_1, \ldots, x_{d_\parallel})^t$ a vector in $d_\parallel \leq d$ dimensions. The student kernel on the contrary varies along all d-dimensional directions in input space. Let the two kernels be singular at the origin with associated exponents θ_T, θ_S as described above. Then as $p \to \infty$, $\epsilon = O\left(p^{-\beta}\right)$ with*

$$\beta = \frac{1}{d} \min(\theta_T, 2d + 2\theta_S). \tag{6}$$

The key result is that β is independent of the dimension $d_\parallel$: indeed we find the same exponent as derived in (Spigler, Geiger, and Wyart 2019) in the absence of invariant. In particular, we recover the curse of dimensionality since $\beta \leq \theta_T / d$.

Numerical Test: We now test the result that kernel regression is blind to the lower-dimensional nature of the task ($d_\parallel < d$). To test robustness with respect to our technical assumption of data points lying on a lattice, we consider instead i.i.d. data sampled uniformly at random on a $d-1$-dimensional sphere of unit radius $\mathbb{S}^{d-1}$ embedded in $\mathbb{R}^d$. The component x^i of each point is generated as a standard Gaussian $\mathcal{N}(0, 1)$ and then the vector $\underline{x}^\mu$ is normalized by dividing it by its norm. For the numerical verification we take the Student to be a Laplace kernel:

$$K_S(z) = \exp\left(-\frac{z}{\sigma}\right), \tag{7}$$

that is characterized by $\theta_S = 1$. As Teacher we use Matérn kernels, which are a family of kernels

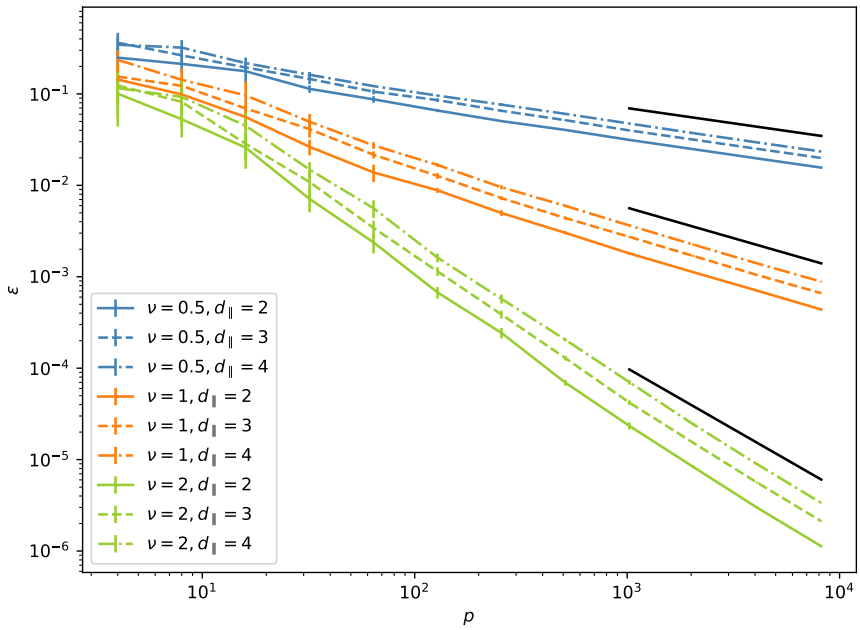

Figure 2. Test error ϵ *vs* the size p of training set for Gaussian data with Matérn covariance regressed using a Laplace kernel. Identical colors correspond to the same parameter ν of the Teacher Matérn kernel but varying dimension $d_{\parallel}$ as indicated in legend. $d_{\parallel}$ has no effect on the exponent β. The solid black lines represent the predicted power law with exponent $\beta = \frac{2}{3}\min(\nu, 4)$.

parametrized by one parameter ν:

$$K_{T,\nu}(z) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{z}{\sigma}\right)^{\nu} \mathcal{K}_{\nu}\left(\sqrt{2\nu}\frac{z}{\sigma}\right), \tag{8}$$

where $\mathcal{K}_{\nu}(z)$ is the modified Bessel function of the second kind with parameter ν, and Γ is the Gamma function. Varying ν one can change the smoothness of the instances of the Gaussian random process, and in particular $\theta_T = 2\nu$. The spatial dimension is $d = 4$ and we vary the amount of invariants in the task by taking $d_{\parallel} = 2, 3, 4$. In order to fix $d_{\parallel}$ we simply use $z = \left|\underline{x}_{\parallel} - \underline{x}'_{\parallel}\right|$ instead of $z = \left|\underline{x} - \underline{x}'\right|$ when computing the Teacher kernel. The scale of the kernel is fixed by the constant σ, that we have taken equal to 4 for both the Teacher and the Student. Notice that in Theorem 2 the value of σ does not play any role since it does not enter in the asymptotic behavior of the test error (at leading order).

In Fig. 2 we show that the numerical simulations match our predictions. Indeed, in this specific case the predicted exponent is

$$\beta = \frac{2}{3}\min(\nu, 4), \tag{9}$$

(the factor 3 comes from the fact that we are generating points on the $d-1$-dimensional sphere embedded in $\mathbb{R}^d$, so even if $d = 4$ the dimension that enters β must be replaced by $d - 1 = 3$, and analogously the second argument of the minimum, for a Laplace Student, is $2(d - 1 + 1) = 8$). Notice that the exponent that characterizes the learning curves is indeed independent of $d_{\parallel}$. Its prefactor may however depend on $d_{\parallel}$ in general.

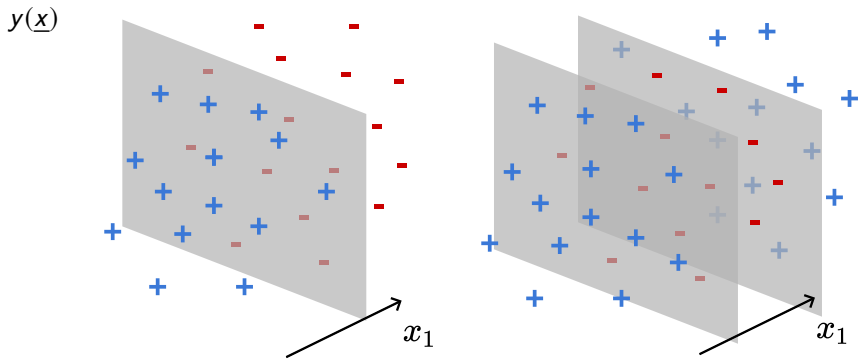

Figure 3. Example of decision boundaries considered in the stripe model, where the label $y(\underline{x})$ of a point $\underline{x}$ depends only on its first component x_1. On the left is the **single-interface setup** where the label function is $y = +1$ on one side of the interface and $y = -1$ on the other. Points labeled in such a way compose a linearly separable dataset. On the right is the **double-interface setup**, where points are labeled $y = -1$ in between the two parallel hyperplanes and $y = +1$ on the outside.

3 Support Vector Classification and stripe model

3.1 The stripe model

We consider a binary classification task where the labels depend only on one direction in the data space, namely with $y(\underline{x}) = y(x_1)$. Layers of $y = +1$ and $y = -1$ regions alternate along the direction x_1, separated by parallel planes. Two examples of this setting are sketched in Fig. 3, corresponding to a single and double interface. The points $\underline{x}$ that constitute the training and test set are iid of distribution $\rho(\underline{x})$. To lighten the notation, we assume that $\rho(\underline{x})$ is uniform on a square box Ω of linear extension γ. Yet we expect our arguments to apply more generally if $\rho(\underline{x})$ is continuous and does not vanish at the location of the interfaces (no margin). To confirm this view we will test and confirm below our predictions when $\rho(\underline{x})$ is Gaussian distributed, with each component $x_i \sim \mathcal{N}(0, \gamma^2)$ with some variance γ^2.

3.2 Definition of margin SVC

In this section we consider margin support-vector classification (margin SVC). This algorithm maximizes the margin between a decision boundary and the points in the training set that are closest to it. The prediction of the label $\hat{y}(\underline{x})$ of a new point $\underline{x}$ is then made according to the sign of the estimated decision function (Scholkopf and Smola 2001):

$$f(\underline{x}) = \sum_{\mu=1}^{p} \alpha^\mu y^\mu K\left(\frac{|\underline{x}^\mu - \underline{x}|}{\sigma}\right) + b, \qquad \hat{y}(\underline{x}) = \operatorname{sign} f(\underline{x}). \tag{10}$$

where the kernel K is conditionally strictly positive definite (Smola, Schölkopf, and Müller 1998) — a condition defined in Appendix C, less stringent than strictly positive definite. In Eq. (10) we write explicitly the kernel bandwidth σ since it will soon play an important role. The formulation of the margin-SVC algorithm presented below is what is referred to as the *dual formulation*, but it can be equivalently recast as an attempt to maximize a (signed) distance between training points and the decision boundary (Scholkopf and Smola 2001). In this dual formulation, the variables α^μ are fixed by maximizing

$$\max_{\underline{\alpha}} \mathcal{L}(\underline{\alpha}), \text{ with } \mathcal{L}(\underline{\alpha}) = \sum_{\mu=1}^{p} \alpha^\mu - \frac{1}{2} \sum_{\mu,\nu=1}^{p} \alpha^\mu \alpha^\nu y^\mu y^\nu K\left(\frac{|\underline{x}^\mu - \underline{x}^\nu|}{\sigma}\right), \tag{11}$$

subject to the constraints

$$\alpha^\mu \geq 0, \tag{12}$$

$$\alpha^\mu > 0 \text{ if and only if } y^\mu f(\underline{x}^\mu) = 1, \tag{13}$$

$$Q \equiv \sum_{\mu=1}^{p} \alpha^\mu y^\mu = 0 \quad \text{(charge conservation)}. \tag{14}$$

The bias b is set to satisfy

$$\min_{1 \leq \mu \leq p} |f(\underline{x}^\mu)| = 1 \quad \text{(canonical condition)}. \tag{15}$$

Eq. (13) states that a dual variable α^μ is strictly positive if and only if its associated vector $\underline{x}^\mu$ lies on the margin, that is $y^\mu f(\underline{x}^\mu) = 1$, otherwise it is zero. Vectors with $\alpha^\mu > 0$ are called *support vectors* (SVs) and are the only ones that enter in the expansion of the decision function Eq. (10).

3.3 Some limiting cases of SVC

Vanishing bandwidth: If the kernel function $K(z)$ decreases exponentially fast with some power of z, then in the limit $\sigma \ll \delta$ where δ is the average nearest-neighbor distance in the training set, the support-vector machine becomes akin to a nearest-neighbor algorithm. A detailed analysis of this regime for the stripe model is presented in Appendix B, here we provide a qualitative argument assuming that the bias b is negligible. If so, as $\sigma \to$ one has for any training point that $f(\underline{x}^\mu) \approx \alpha^\mu y^\mu K(0)$, implying that $\alpha^\mu \neq 0$ to satisfy $|f(\underline{x}^\mu)| \geq 1$: every point is a support vector with identical α^μ. $f(\underline{x})$ at a random test point $\underline{x}$ is dominated by the closest support vector. The classification error is susceptible to the curse of dimensionality for such an algorithm, and one expects generically $\epsilon \sim \delta \sim p^{-1/d}$, as tested numerically in Fig. 12 for the stripe model.

Diverging bandwidth: In this work we focus on the other extreme case where $\sigma \gg \gamma$. In this regime the kernel is always evaluated close to the origin. Assuming that the kernel has a finite derivative in the neighborhood of the origin, we approximate it by its truncated Taylor expansion:

$$K\left(\frac{|\underline{x} - \underline{x}'|}{\sigma}\right) \approx K(0) - \text{const} \cdot \left(\frac{|\underline{x} - \underline{x}'|}{\sigma}\right)^\xi + o(\gamma/\sigma)^\xi. \tag{16}$$

The exponent ξ is related to the exponent θ introduced in Section 2 by $\xi = \min(\theta, 2)$, and varies from kernel to kernel. For instance, we have $\xi = 1$ for Laplace kernels, $\xi = 2$ for Gaussian kernels, $\xi = \tilde{\gamma}$ for $\tilde{\gamma}$-exponential kernels[3] and $\xi = \min(2\nu, 2)$ for Matérn kernels. In Appendix C we show that for $0 < \xi < 2$ the right-hand side is conditionally strictly positive definite (CSPD), which is the necessary condition for the SVC algorithm to converge. In what folows, we consider $0 < \xi < 2$, which excludes the Gaussian case. A proof that in that case the margin-SVC algorithm with the truncated kernel in Eq. (16) leads to the same solution as with the full kernel in the limit $\sigma \gg \gamma$ is presented in Appendix D. Also, due to the charge conservation in Eq. (14), the constant term $K(0)$ in Eq. (16) may safely be ignored.

The decision function Eq. (10) associated to the considered radial power kernel hence becomes

$$f(\underline{x}) = b - \sum_{\mu=1}^{p} \alpha^\mu y^\mu \left(\frac{|\underline{x} - \underline{x}^\mu|}{\sigma}\right)^\xi. \tag{17}$$

where the positive constant in Eq. (16) has been removed by rescaling the bias and the α^μ.

3. We use $\tilde{\gamma}$ to distinguish it from the variance of the data points.

3.4 Single interface

We consider a single interface at location $x_1 = 0$, with negative labels for $x_1 < 0$ and positive ones for $x_1 > 0$. Already in that case, computing analytically the test error remains a challenge, and we resort to a scaling (asymptotic) analysis to compute β. As p increases, support vectors will be present on a narrower and narrower band around the interface. We denote by Δ the characteristic extension of that band. Δ will depend in general on the position $\underline{x}_\perp$ along the interface. Here we will not study this dependence, as we are interested on its asymptotic behavior with p, γ and σ and only track how quantities depend on these variables. From the definition of support vectors we have that the function f varies of order one from one side of the band to the other:

$$f(\underline{x}_\perp + \Delta \underline{e}_1) - f(\underline{x}_\perp - \Delta \underline{e}_1) \sim 1 \tag{18}$$

where $\underline{e}_1$ is the unit vector orthogonal to the interface and $\underline{x}_\perp \in \Omega$.

Another useful quantity is the distance r_c between nearest support vectors. It can be estimated by counting the number of points lying within a cylinder of height Δ (along x_1) and radius r_c centered on a SVs, whose volume follows $\sim \Delta r_c^{d-1}$. Using that the density of data points is $\sim p/\gamma^d$, and imposing that the cylinder contains only one additional SV yields our first scaling relation:

$$\frac{p}{\gamma^d} \cdot \Delta r_c^{d-1} \sim 1 \quad \Longrightarrow \quad \boxed{p \Delta r_c^{d-1} \sim \gamma^d.} \tag{19}$$

Finally, because the distance between one SV and the interface varies between zero and the scale Δ, and because at a SV the value of the function is fixed by $|f(\underline{x}^\mu)| = 1$, fluctuations of f must be of order one inside the band as one moves parallel to the plane on a scale r_c:

$$f(\underline{x}_\perp + r_c \underline{e}_\perp) - f(\underline{x}_\perp) \sim 1. \tag{20}$$

where $\underline{e}_\perp$ is any unit vector parallel to the plane. Due to these fluctuations, test points inside the band have a finite probability to be incorrectly classified, and at fixed d [4] the test error must be proportional to the fraction Δ/γ of points falling in that band:

$$\epsilon \sim \Delta/\gamma.$$

We now show that from these considerations alone β can be computed. Starting from Eq. (17) we estimate the gradient of f along the normal direction $\underline{e}_1$ at any point on the interface:

$$\partial_{x_1} f(\underline{x}_\perp) = \frac{\xi}{\sigma} \sum_{\mu \in \Omega_\Delta} \alpha^\mu y^\mu \left(\frac{\left| \underline{x}_\perp - \underline{x}^\mu \right|}{\sigma} \right)^{\xi - 1} \frac{x_1^\mu}{\left| \underline{x}_\perp - \underline{x}^\mu \right|} \approx \xi \sigma^{-\xi} p \frac{\Delta}{\gamma} \left\langle \alpha^\mu y^\mu x_1^\mu \left| \underline{x}_\perp - \underline{x}^\mu \right|^{\xi - 2} \right\rangle_{\mu \in \Omega_\Delta} \tag{21}$$

where the sum is over all SVs $\underline{x}_\mu$ indicated by the set Ω_Δ. The sum is replaced by its central-limit theorem value valid for large p, and we use that the number of terms in that sum goes as $p\Delta/\gamma$. The average in Eq. (21) scales as $\bar{\alpha} \Delta \gamma^{\xi - 2}$ where $\bar{\alpha}$ is the characteristic value of the dual variables α^μ. Imposing that $\Delta \partial_{x_1} f(\underline{x}_\perp) \sim 1$ as follows from Eq. (18) then leads to our second scaling relation:

$$\boxed{p \, \bar{\alpha} \left(\frac{\Delta}{\gamma} \right)^3 \sim \left(\frac{\sigma}{\gamma} \right)^\xi.} \tag{22}$$

4. The value of $f(\underline{x})$ in the band is governed by the neighboring support vectors, whose characteristic number is independent of p but should grow with d. We believe this effect to be responsible for the non-commutativity of the limits $\lim_{p/d \to \infty} \lim_{d \to \infty}$ and $\lim_{d \to \infty} \lim_{p \to \infty}$. Indeed in the former case, it is found (Dietrich, Opper, and Sompolinsky 1999) that ϵ and the fraction of support vectors Δ/γ scale differently with α, unlike what we argue and confirm numerically in the second limit. We have checked numerically (not shown) that the ratio of these two quantities is indeed decaying with d at fixed p.

Next we compute the consequences of Eq. (20), by recasting it in a more suitable format. We define a smoothed function $\bar{f}(\underline{x}_\perp)$ of $f(\underline{x}_\perp)$ on a scale r_c:

$$\bar{f}(\underline{x}_\perp) = \int \mathrm{d}^{d-1}\underline{x}'_\perp \, f(\underline{x}'_\perp)\, G(\underline{x}_\perp - \underline{x}'_\perp), \tag{23}$$

where the function G is the Fourier transform of $\theta(1/r_c - \left\|\underline{k}_\perp\right\|)$ (which is thus small when $\left\|\underline{x}_\perp - \underline{x}'_\perp\right\| \gg r_c$):

$$G(\underline{x}_\perp) = \int_{\|k_\perp\|<1/r_c} \mathrm{d}^{d-1}\underline{k}_\perp e^{-i\underline{k}_\perp\cdot x_\perp}. \tag{24}$$

Thus $\bar{f}(\underline{x}_\perp)$ is obtained by removing from $f(\underline{x}_\perp)$ the Fourier components $\|k_\perp\| > 1/r_c$. The constraint of Eq. (20) is equivalent to imposing that the fluctuations between f and $\bar{f}$ are of order one. Integrated on space it means that:

$$\gamma^{-d+1} \int \mathrm{d}^{d-1}\underline{x}_\perp \left[f(\underline{x}_\perp) - \bar{f}(\underline{x}_\perp)\right]^2 \sim 1, \tag{25}$$

that can be Fourier-transformed as:

$$\int \mathrm{d}^{d-1}\underline{k}_\perp \left[\tilde{f}(\underline{k}_\perp) - \tilde{f}(\underline{k}_\perp)\,\tilde{G}(\underline{k}_\perp)\right]^2 = \int_{\|\underline{k}_\perp\|>1/r_c} \mathrm{d}^{d-1}\underline{k}_\perp \tilde{f}^2(\underline{k}_\perp) \sim \gamma^{d-1}. \tag{26}$$

The Fourier transform of the decision function along the transverse components can be computed as

$$\tilde{f}\left(\underline{k}_\perp\right) = \int \mathrm{d}^{d-1}\underline{x}_\perp \, e^{-i\underline{k}_\perp\cdot\underline{x}_\perp} f(\underline{x}_\perp) = \sum_{\mu\in\Omega_\Delta} \alpha^\mu y^\mu \int \mathrm{d}^{d-1}\underline{x}_\perp \, e^{-i\underline{k}_\perp\cdot\underline{x}_\perp} K\left(\frac{\left|\underline{x}^\mu - \underline{x}_\perp\right|}{\sigma}\right). \tag{27}$$

Using that $\left|\underline{x}^\mu - \underline{x}_\perp\right| \approx \left|\underline{x}^\mu_\perp - \underline{x}_\perp\right|$ and changing variables one obtains

$$\tilde{f}\left(\underline{k}_\perp\right) \approx \sum_{\mu\in\Omega_\Delta} \alpha^\mu y^\mu e^{-i\underline{k}_\perp\cdot\underline{x}^\mu_\perp} \cdot \int \mathrm{d}^{d-1}\underline{x}_\perp \, e^{-i\underline{k}_\perp\cdot\underline{x}_\perp} K\left(\frac{\left|\underline{x}_\perp\right|}{\sigma}\right) \equiv \tilde{Q}\left(\underline{k}_\perp\right)\cdot\tilde{K}_\perp\left(\underline{k}_\perp\right), \tag{28}$$

where we have defined the kernel (transverse) Fourier transform $\tilde{K}_\perp\left(\underline{k}_\perp\right)$ and the "charge" structure factor $\tilde{Q}\left(\underline{k}_\perp\right)$. The former can be readily computed as shown in Appendix E, and at large frequencies behaves as $\tilde{K}_\perp\left(\underline{k}_\perp\right) \sim \sigma^{-\xi}\left|\underline{k}_\perp\right|^{-(d-1+\xi)}$. Concerning the charge structure factor, for $\left|\underline{k}_\perp\right| \gg 1/r_c$, the phases associated to each term in the sum defining it vary significantly even between neighboring SVs. From a central-limit argument the factor $\tilde{Q}$ then tends to a random variable with 0 mean and variance $\bar{\alpha}^2 p\Delta/\gamma$. It is verified in Appendix F.

We can now estimate the integral in Eq. (26):

$$\int_{\|\underline{k}_\perp\|>1/r_c} \mathrm{d}^{d-1}\underline{k}_\perp \, \tilde{f}^2\left(\underline{k}_\perp\right) \sim \bar{\alpha}^2 p\frac{\Delta}{\gamma}\sigma^{-2\xi}\int_{\|\underline{k}_\perp\|>k_c} \mathrm{d}^{d-1}\underline{k}_\perp \left|\underline{k}_\perp\right|^{-2(d-1+\xi)} \sim \bar{\alpha}^2 p\frac{\Delta}{\gamma}\sigma^{-2\xi} r_c^{d-1+2\xi}. \tag{29}$$

The condition Eq. (26) leads to the last scaling relation:

$$\boxed{\bar{\alpha}^2 p\frac{\Delta}{\gamma}\left(\frac{r_c}{\gamma}\right)^{d-1+2\xi} \sim \left(\frac{\sigma}{\gamma}\right)^{2\xi}.} \tag{30}$$

Putting all the scaling relations together we find:

$$\boxed{\Delta \sim \gamma\, p^{-\frac{d-1+\xi}{3d-3+\xi}}, \qquad \bar{\alpha} \sim \left(\frac{\sigma}{\gamma}\right)^{\xi} p^{\frac{2\xi}{3d-3+\xi}}, \qquad r_c \sim \gamma\, p^{-\frac{2}{3d-3+\xi}}.} \tag{31}$$

And consequently the asymptotic behavior of the test error is given by

$$\boxed{\epsilon \sim \frac{\Delta}{\gamma} \sim p^{-\beta}, \qquad \text{with } \beta = \frac{d-1+\xi}{3d-3+\xi}.} \tag{32}$$

Note 1: The second scaling argument leading to Eq. (30) can be readily obtained by making a "minimal-disturbance hypothesis". Assuming that adding a new training point $\underline{x}^*$ within the domain Ω_Δ will only affect the dual variables of the few closest SVs, the correction of the decision function on the new SV is given by:

$$\sum_{\|\underline{x}^\mu - \underline{x}^*\| \leq r_c} d\alpha^\mu y^\mu \left(\frac{\left|\underline{x}^\mu - \underline{x}^*\right|}{\sigma} \right)^\xi, \tag{33}$$

where $d\alpha^\mu$ is the charge correction. One must have that $\sum_{\|\underline{x}^\mu - \underline{x}^*\| \leq r_c} d\alpha^\mu y^\mu \approx -y^* \alpha^*$ to ensure that SVs further away are not affected by this perturbation. Thus $d\alpha^\mu \sim \alpha^* \sim \bar{\alpha}$, where the last equivalence stems from the fact that the added SV is statistically identical to any other one. Finally, requiring that the new point $\underline{x}^*$ must also be a SV implies that the correction represented by Eq. (33) must be of order one to set $|f(\underline{x}^*)| = 1$. Hence, we obtain the scaling relation (that implies Eq. (30) from Eq. (19) and Eq. (22)):

$$\bar{\alpha} \left(\frac{r_c}{\sigma} \right)^\xi \sim 1. \tag{34}$$

Note 2: The above scaling arguments may also be carried out in the intermediate regime $\delta \ll \sigma < \gamma$. In that case, the kernel Eq. (16) introduces a cutoff to the volume of interaction in the transverse space. In particular, the number of terms in the sum of Eq. (21) now goes as $(\sigma/\gamma)^{d-1} p\Delta/\gamma$ and the average scales as $\bar{\alpha}\Delta\sigma^{\xi-2}$. The discussion on the fluctuations is however unaltered as $r_c \ll \sigma$ by definition. Assembling all the pieces yields the following scaling relations:

$$\Delta \sim \gamma \left(\frac{\sigma}{\gamma} \right)^{-(d-1)\frac{d-3+\xi}{3d-3+\xi}} p^{-\frac{d-1+\xi}{3d-3+\xi}}, \qquad \bar{\alpha} \sim \left(\frac{\sigma}{\gamma} \right)^{\frac{2\xi d}{3d-3+\xi}} p^{\frac{2\xi}{3d-3+\xi}}, \qquad r_c \sim \gamma \left(\frac{\sigma}{\gamma} \right)^{\frac{d-3+\xi}{3d-3+\xi}} p^{-\frac{2}{3d-3+\xi}} \tag{35}$$

and

$$\epsilon \sim \frac{\Delta}{\gamma} \sim \left(\frac{\sigma}{\gamma} \right)^{-(d-1)\frac{d-3+\xi}{3d-3+\xi}} p^{-\frac{d-1+\xi}{3d-3+\xi}}. \tag{36}$$

Note that when this approach breaks down, namely when $\sigma \sim r_c$, the predictions of the vanishing bandwidth are recovered.

3.5 Multiple interfaces

The scaling analysis considered for the single interface can be directly extended to multiple interfaces. Let us denote by w the distance between the two closest planes. In the limit $\Delta << w$, the arguments presented between Eq. (23) and Eq. (30) that rely on local considerations apply identically. Moreover, it is straightforward to see that the gradient computation at one interface Eq. (21) now gets similar contributions from all interfaces in the system, whose asymptotic dependence with relevant variables is identical. Overall, we thus predict that the scaling relations Eqs.(31,32) hold for $\Delta << w$ in that case as well, as we shall confirm empirically below [5].

3.6 Numerical results

In this section, we present the numerical simulations with which we verify the scalings predicted in the two previous sections. Both the single and the double-interface setups have been considered with

5. A qualitative difference with the more symmetric single interface case is that charges or dual variable will in general differ in magnitude on each side of interfaces. Our arguments however do not rely on this difference being zero.

data points sampled from an isotropic Gaussian distribution of variance $\gamma^2 = 1$ along each component. In the single-interface setup the hyperplane is centered at $x_1 = 0$, while in the double-interface setup one hyperplane was located at $x_{\text{min}} = -0.3$ and the other at $x_{\text{max}} \approx 1.18549$.[6] In both setups, the probability of positive and negative labels are equal. The margin-SVC algorithm is run using the class `svm.SVC` from the python library *scikitlearn*, which is a soft margin algorithm. To recover the hard margin algorithm presented in Section 3.2, the regularization parameter C which bounds from above the dual variables (see for example chapter 7 of Smola, Schölkopf, and Müller 1998) is set to $C = 10^{20}$. All results presented in this section have been obtained with the Laplace kernel of bandwidth $\sigma = 100 \gg \gamma$. Further results with the Matérn kernel are displayed in Appendix G.

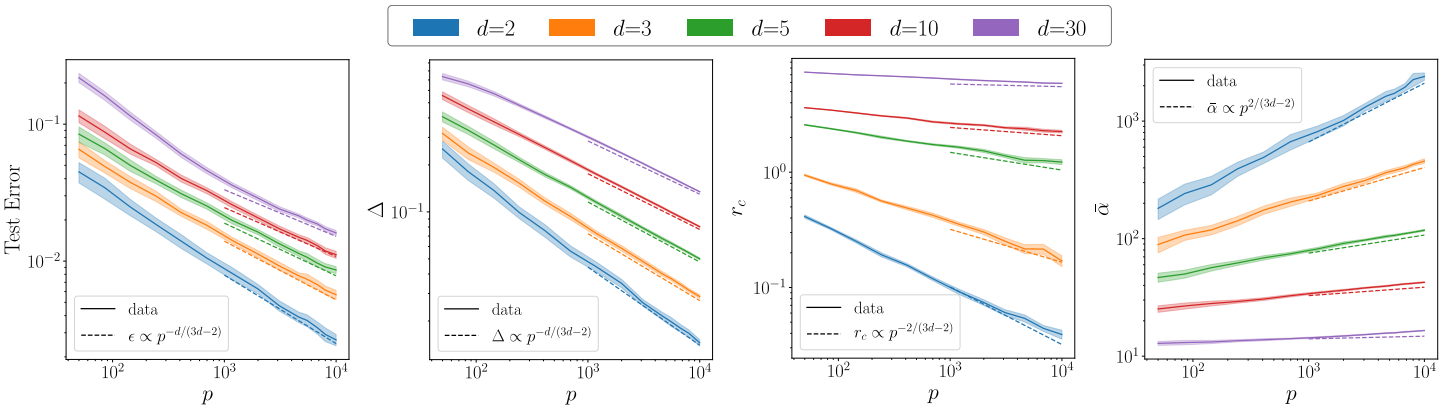

Figure 4. For the **single-interface setup**, we show the dependence on the training-set size p of the test error, the SV band thickness Δ, the scale r_c and the SV mean dual variable $\bar{\alpha}$ (from left to right). The points in the dataset are drawn from the standard normal distribution in dimension d (see the color legend); their labels are defined according to the single-interface setup and learned with the margin-SVC algorithm with the Laplace kernel ($\xi = 1$) of bandwidth $\sigma = 100$. The solid lines correspond to the average over 25 initializations, while the shaded region are the associated standard deviations. The dashed lines illustrate the power law predicted in Eq. (31).

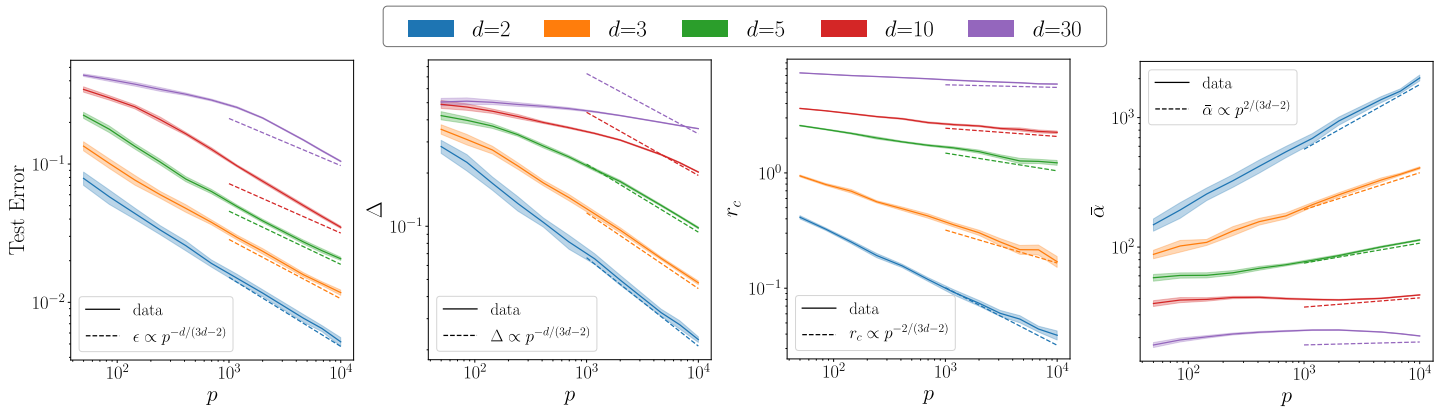

Figure 5. Same plots as in Fig. 4, but for the **double-interface setup**: we show the dependence on the training-set size p of the test error, the SV band thickness Δ, the scale r_c and the SV mean dual variable $\bar{\alpha}$ (from left to right). The points in the dataset are again drawn from the standard normal distribution in dimension d (see the color legend); their labels are defined according to the double-interface setup and learned with the margin-SVC algorithm with the Laplace kernel ($\xi = 1$) of bandwidth $\sigma = 100$. The solid lines correspond to the average over 25 initializations, while the shaded region are the associated standard deviations. The dashed lines illustrate the power law predicted in Eq. (31).

6. This value is $x_{\text{max}} = \sqrt{2}\,\text{erf}^{-1}(1 + \text{erf}(x_{\text{min}})) \approx 1.18549$. This value is chosen in such a way that the expected number of $y = \pm 1$ points is the same.

The power law predictions of Section 3.4 are verified in Fig. 4 (for the single interface) and Fig. 5 (for the double interface). The considered numerical observables are defined as follows: the test error is the fraction of mislabeled points in a test set of size $p_{\text{test}} = 10000$; the typical $\bar{\alpha}$ is the average SV dual variable; the band thickness Δ is the average distance of a SV to the closest interface; the procedure to estimate the SV nearest-neighbor scale r_c is described in Appendix H. The exponents of the power laws are extracted by fitting the numerical curves in the asymptotic regime and compared to the theoretical predictions of Section 3.4 in Fig. 6. Note that in large dimensions, we observe that the system has not yet fully reached the asymptotic regime in the considered range of training-set sizes p.

We also observe that in the double-interface setup, the system only enters the scaling regime when Δ becomes small enough compared to the distance w between the two hyperplanes. The crossover between the regime dominated by the interference between the two bands of SVs and the scaling regime of an independent interface is illustrated in Fig. 7. The test error vs Δ displayed on the left figure for multiple values of w confirms that $\epsilon \sim \Delta$, when $\Delta \ll w$, as expected from the discussion of Section 3.5. We show on the right figure that the transition to the asymptotic regime occurs when $\Delta \sim w$ by rescaling the horizontal axis: $\Delta \to \Delta/w$. Because $\epsilon \sim \Delta$ in the asymptotic regime, it is necessary to also rescale the vertical axis for the curves to collapse, namely $\epsilon \to \epsilon/w$.

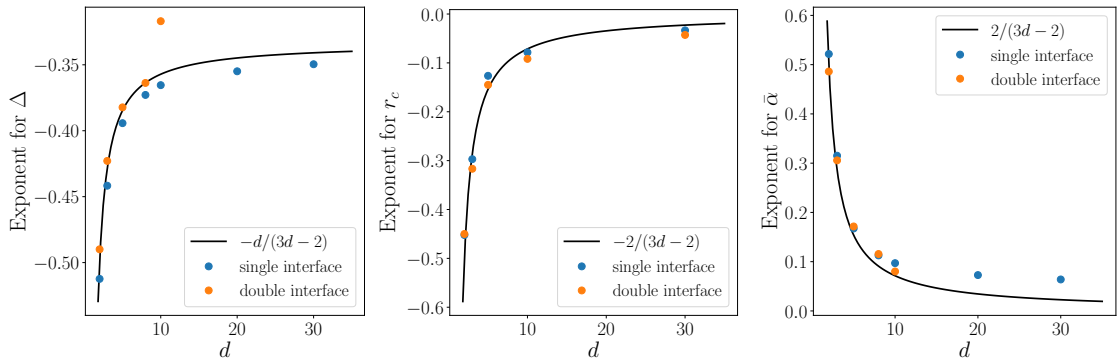

Figure 6. We extract the exponents by fitting the curves in Fig. 4 (for the **single-interface setup**) and in Fig. 5 (for the **double-interface setup**). We then plot the exponents for the SV band thickness Δ (left), the SV nearest-neighbor scale r_c (middle) and the SV mean dual variable $\bar{\alpha}$ (right) against the dimension d of the data. The black solid line is the prediction of Section 3.4, while the dots correspond to the numerical data (blue points for the single-interface setup and orange points for the double-interface setup).

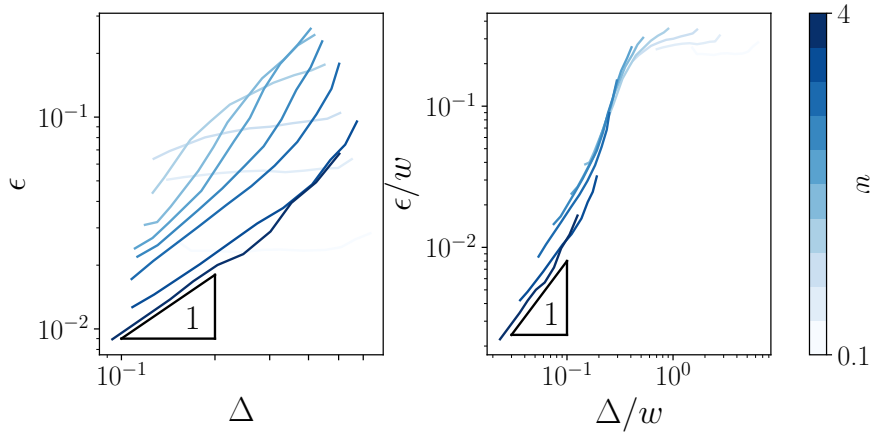

Figure 7. Left: Test error ϵ vs the SV band thickness Δ for multiple values of the distance between the two hyperplanes w for the **double-interface setup** in $d = 5$. The left interface is located at $x_{\text{min}} = -1$ and the right interface at $x_{\text{max}} = x_{\text{min}} + w$. Right: The left plot is rescaled by w^{-1} both horizontally and vertically. The inset triangles indicate a slope of one in log-log scale.

4 Spherical model

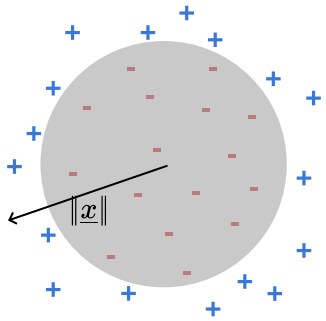

Figure 8. Nonlinear decision boundary for the spherical setup. The label function is $y = +1$ inside the hypersphere and $y = -1$ outside. Note that the label only depends on the norm of the data, $\|\underline{x}\|$.

We consider a spherical interface separating $y = +1$ points outside a sphere of radius R from $y = -1$ points inside. The relevant direction is therefore $x_{\parallel} = \|\underline{x}\|$, and the label is given by $y(\underline{x}) = \text{sign}(\|\underline{x}\| - R)$. We still assume that the SV are distributed along the interface, thus forming a shell of radius R and thickness Δ. Once again, previous arguments presented between Eq. (23) and Eq. (30) that rely on local considerations apply identically. Furthermore, we compute in Appendix I the gradient $\partial f / \partial x_{\parallel}$ and find again the same asymptotic result as for planar interface specified in Eq.21. Thus our predictions for the spherical model are identical to the stripe one.

We test these results numerically for a sphere of radius $R = \sqrt{d}$[7] with a Laplace kernel of variance $\sigma = 100$. The results displayed on Fig. 9 and Fig. 10 confirm our analysis.

7. It guarantees that the fraction of positive and negative labels remain finite. In particular, in the limit $d \to \infty$, this fraction goes to 1/2.

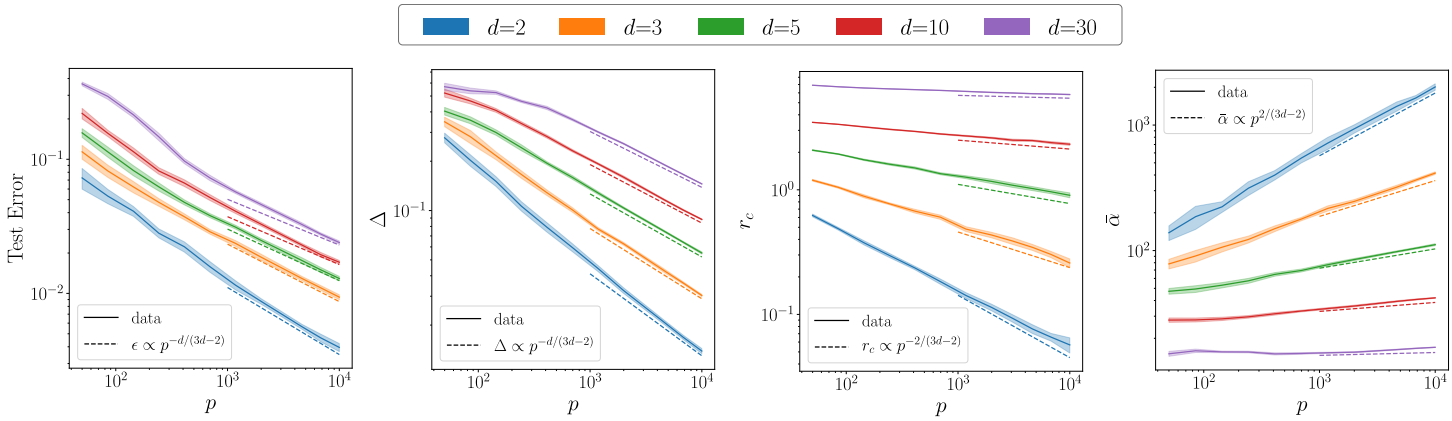

Figure 9. For the **spherical setup**, we show the dependence on the training-set size p of the test error, the SV band thickness Δ, the scale r_c and the SV mean dual variable $\bar{\alpha}$ (from left to right). The points in the dataset are drawn from the standard normal distribution in dimension d (see the color legend); their labels are defined according to the spherical setup of radius $R = \sqrt{d}$ and learned with the margin-SVC algorithm with the Laplace kernel ($\xi = 1$) of bandwidth $\sigma = 100$. The solid lines correspond to the average over 25 initializations, while the shaded region are the associated standard deviations. The dashed lines illustrate the power law predicted in Eq. (31).

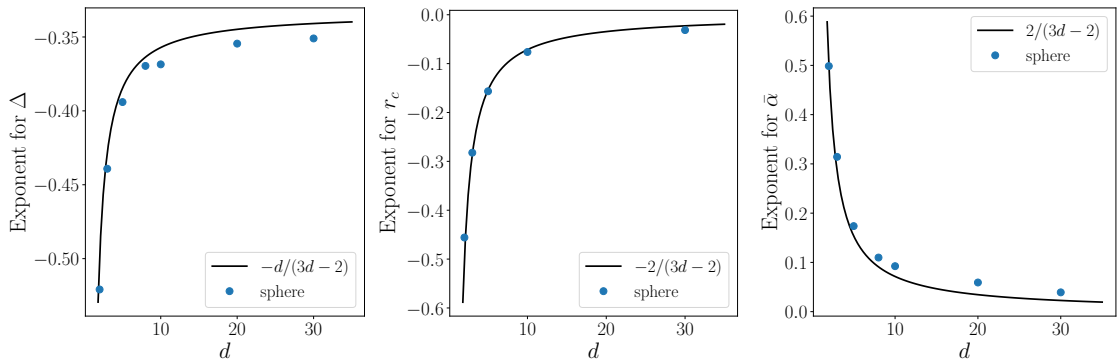

Figure 10. We extract the exponents by fitting the curves in Fig. 9 for the **spherical setup**. We then plot the exponents for the SV band thickness Δ (left), the SV nearest-neighbor scale r_c (middle) and the SV mean dual variable $\bar{\alpha}$ (right) against the dimension d of the data. The black solid line is the prediction of section Section 3.4, while the dots correspond to the numerical data (blue points for the single-interface setup and orange points for the double-interface setup).

5 Improving kernel performance by compressing invariants

In this section, we investigate how compressing the data along the irrelevant directions $\underline{x}_\perp$ affects the performance of kernel classification. This analysis is of particular interest for neural networks, where it is now argued (see for instance Mallat 2016) that a progressive capability to compress invariants in the data is built up moving through the layers of deep networks.

We consider the stripe model of Section 3 with one additional parameter: the amplification factor λ. If the original distribution was characterized by the scales $\gamma_1, \dots, \gamma_d$ along each space direction, we now apply a contraction in the transverse space: $\gamma_i \to \gamma_i/\lambda$ for $i = 2, \dots, d$. Following the same reasoning as in Section 3.4, we can track the effect of the additional amplification parameter. It is not sufficient to merely rescale γ, since the compression is not isotropic. Nevertheless, it is easy to see that the first scaling becomes

$$\lambda^{d-1}\, r_c^{d-1}\, \Delta\, p \sim \gamma^d, \tag{37}$$

since the density of points inside the SV band is now $\sim p\lambda^{d-1}/\gamma$. Then, for the second scaling relation, we

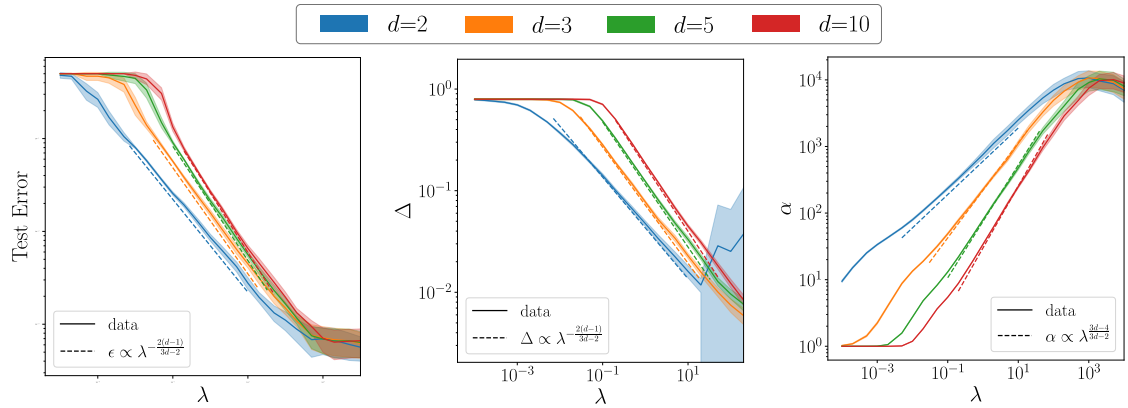

Figure 11. Dependence on the amplification factor λ of the test error (left), the SV band thickness Δ (middle) and the SV mean dual variable $\bar{\alpha}$ (right) for the **single-interface setup** with $p = 1000$ in different dimensions (see the color legend). The SVC algorithm is run with the Laplace kernel ($\xi = 1$) of bandwidth $\sigma = 100 \gg \delta$. The solid lines correspond to the average over 20 initializations and the shaded region are the associated standard deviations. The dashed lines illustrate the power law predictions of Eq. (40).

need to rescale the gradient $\partial_{x_1} f$ defined in Eq. (21). The amplification factors only alters the transverse space: when approximating the average by an integral, the boundaries are rescaled to γ/λ in each transverse direction. The second scaling is thus

$$\lambda^{2-\xi}\, p\, \bar{\alpha} \left(\frac{\Delta}{\gamma}\right)^3 \left(\frac{\gamma}{\sigma}\right)^{\xi} \sim 1. \tag{38}$$

Finally, when imposing that the fluctuations between f and its smoothed version $\tilde{f}$ are of order one, one only needs to update the volume of the transverse space in Eq. (25): $\gamma^{d-1} \to (\gamma/\lambda)^{d-1}$, which leads to the last scaling,

$$\lambda^{d-1}\, \bar{\alpha}^2\, p\, \frac{\Delta}{\gamma} \left(\frac{r_c}{\gamma}\right)^{d-1+2\xi} \sim \left(\frac{\sigma}{\gamma}\right)^{2\xi}. \tag{39}$$

Assembling all the scaling relations yields:

$$\epsilon \sim \Delta \sim \gamma\, \lambda^{-\frac{2(d-1)}{3d-3+\xi}}\, p^{-\frac{d-1+\xi}{3d-3+\xi}}, \qquad \bar{\alpha} \sim \left(\frac{\sigma}{\gamma}\right)^{\xi} \lambda^{\xi\frac{3d-5+\xi}{3d-3+\xi}}\, p^{\frac{2\xi}{3d-3+\xi}}, \qquad r_c \sim \gamma\, \lambda^{-\frac{3d-5+\xi}{3d-3+\xi}}\, p^{-\frac{2}{3d-3+\xi}}. \tag{40}$$

These power laws are assessed numerically for the Laplace kernel ($\xi = 1$) of variance $\sigma = 100$ and a training set of size $p = 1000$ generated from the Gaussian distribution of variance $\gamma^2 = 1$. Varying the amplification factor over eight orders of magnitude (see Fig. 11), our predictions hold in a broad range of λ but break down at large and small values, as we now explain.

In the limit $\lambda \to 0$, the relevant direction x_1 is negligibly small compared to the other directions, information is thus suppressed and points are classified at random: the test error goes to $1/2$. Furthermore, all training points must be SVs, and indeed $\Delta \to \langle |x| \rangle_{x\sim\mathcal{N}(0,1)} = \sqrt{2/\pi}$ (which is the average distance from any point in the dataset to the interface) and $\bar{\alpha} \to 1$.

In the opposite limit $\lambda \to \infty$ the setup lives in dimension one (seeing only x_1) and all curves converge independently of the space dimension d. These relations allow us to identify a critical scale λ_c at which the multidimensional system reduces effectively to a one dimensional system. It occurs when the test error of the compressed multidimensional kernel is equal to the test error of the kernel that only sees the component x_1. Using our scalings, we find

$$\lambda^{-\frac{2(d-1)}{3d-3+\xi}}\, p^{-\frac{d-1+\xi}{3d-3+\xi}} \sim p^{-1} \qquad \Longrightarrow \qquad \lambda_c \sim p. \tag{41}$$

6 Conclusion

We have studied the learning curve exponent β of isotropic kernel in the presence of invariants. For regression on gaussian fields, we find that invariants do not increase β that behaves as $\sim d^{-1}$ in large dimension: methods based on isotropic kernels suffer from the curse of dimensionality, as already argued in (Bach 2017). Surprisingly, for a binary classification and simple models of invariants we find the opposite result. For a planar interface separating labels, $\beta \geq 1/3$ for all dimensions. This result holds when several interfaces are present, or in the spherical case where the interface continuously bends. Thus, isostropic kernels can beat the curse of dimensionality even for non-planar boundaries between labels. For which class of boundaries is this result true? The geometry of the spatial distribution of support vectors suggests an answer. The curse of dimensionality is beaten because a very narrow (i.e. rapidly decaying with p) layer of width Δ is sufficient to fit all data, despite the fact that the distance between support vectors r_c is much larger (and indeed subjected to the curse of dimensionaliy). Thus if the boundary displays significant variations below the scale r_c, it presumably cannot be detected by isotropic kernels. In that view, beating the curse of dimensionality is only possible if the boundary is more and more regular as the dimension increases. Yet, sufficient regularity may be achieved in practical settings at least along some invariants, such as completely uninformative pixels near the boundary of images. Under which conditions other invariants, e.g. related to translation, can be exploited by isotropic kernels remains to be understood.

Acknowledgments

We acknowledge L. Chizat, M. Geiger, L. Petrini for the discussions on the topics treated in this article. This work was partially supported by the grant from the Simons Foundation (#454953 Matthieu Wyart). M.W. thanks the Swiss National Science Foundation for support under Grant No. 200021-165509.

Supplementary Material

The code used to generate the data with the margin-SVC algorithm is available online at https://gitlab.com/jonas.paccolat/svc-for-simple-invariants.

References

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

Arora, S., S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang. 2019. "On Exact Computation with an Infinitely Wide Neural Net." *arXiv preprint arXiv:1904.11955.*

Aubin, B., A. Maillard, F. Krzakala, N. Macris, L. Zdeborová, *et al.* 2018. "The committee machine: Computational to statistical gaps in learning a two-layers neural network." In *Advances in Neural Information Processing Systems,* 3223–3234.

Bach, F. 2017. "Breaking the curse of dimensionality with convex neural networks." *The Journal of Machine Learning Research* 18 (1): 629–681.

Bansal, Y., M. Advani, D. D. Cox, and A. M. Saxe. 2018. "Minnorm training: an algorithm for training over-parameterized deep neural networks." *CoRR.*

Bruna, J., and S. Mallat. 2013. "Invariant scattering convolution networks." *IEEE transactions on pattern analysis and machine intelligence* 35 (8): 1872–1886.

Costa, J. A., and A. O. Hero. 2004. "Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets." In *2004 12th European Signal Processing Conference,* 369–372. IEEE.

d'Ascoli, S., M. Refinetti, G. Biroli, and F. Krzakala. 2020. "Double Trouble in Double Descent: Bias and Variance (s) in the Lazy Regime." *arXiv preprint arXiv:2003.01054.*

Dietrich, R., M. Opper, and H. Sompolinsky. 1999. "Statistical mechanics of support vector networks." *Physical review letters* 82 (14): 2975.

Engel, A., and C. Van den Broeck. 2001. *Statistical mechanics of learning.* Cambridge University Press.

Facco, E., M. d'Errico, A. Rodriguez, and A. Laio. 2017. "Estimating the intrinsic dimension of datasets by a minimal neighborhood information." *Scientific reports* 7 (1): 12140.

Franz, S., S. Hwang, and P. Urbani. 2018. "Jamming in multilayer supervised learning models." *arXiv preprint arXiv:1809.09945.*

Gardner, E. 1988. "The space of interactions in neural network models." *Journal of physics A: Mathematical and general* 21 (1): 257.

Geiger, M., A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d'Ascoli, G. Biroli, C. Hongler, and M. Wyart. 2020. "Scaling description of generalization with number of parameters in deep learning." *Journal of Statistical Mechanics: Theory and Experiment* 2020 (2): 023401.

Gerace, F., B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová. 2020. "Generalisation error in learning with random features and the hidden manifold model." *arXiv preprint arXiv:2002.09339.*

Goldt, S., M. Mézard, F. Krzakala, and L. Zdeborová. 2019. "Modelling the influence of data structure on learning in neural networks." *arXiv preprint arXiv:1909.11500.*

Hein, M., and J.-Y. Audibert. 2005. "Intrinsic dimensionality estimation of submanifolds in R d." In *Proceedings of the 22nd international conference on Machine learning,* 289–296. ACM.

Hestness, J., S. Narang, N. Ardalani, G. F. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou. 2017. "Deep Learning Scaling is Predictable, Empirically." *CoRR* abs/1712.00409.

Jacot, A., F. Gabriel, and C. Hongler. 2018. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks." In *Advances in Neural Information Processing Systems 31,* 8580–8589.

Luxburg, U. v., and O. Bousquet. 2004. “Distance-based classification with Lipschitz functions.” *Journal of Machine Learning Research* 5 (Jun): 669–695.

Maiorov, V. 2006. “Approximation by neural networks and learning theory.” *Journal of Complexity* 22 (1): 102–117.

Mallat, S. 2016. “Understanding deep convolutional networks.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, no. 2065 (April): 20150203. ISSN: 1471-2962. doi:10.1098/rsta.2015.0203. http://dx.doi.org/10.1098/rsta.2015.0203.

Monasson, R., and R. Zecchina. 1995. “Weight space structure and internal representations: a direct approach to learning and generalization in multilayer neural networks.” *Physical review letters* 75 (12): 2432.

Montanari, A., F. Ruan, Y. Sohn, and J. Yan. 2019. “The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime.” *arXiv preprint arXiv:1911.01544.*

Neyshabur, B., Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro. 2018. “Towards Understanding the Role of Over-Parametrization in Generalization of Neural Networks.” *arXiv preprint arXiv:1805.12076.*

Neyshabur, B., R. Tomioka, R. Salakhutdinov, and N. Srebro. 2017. “Geometry of optimization and implicit regularization in deep learning.” *arXiv preprint arXiv:1705.03071.*

Opper, M., and D. Saad. 2001. *Advanced mean field methods: Theory and practice.* MIT press.

Rozza, A., G. Lombardi, C. Ceruti, E. Casiraghi, and P. Campadelli. 2012. “Novel high intrinsic dimensionality estimators.” *Machine learning* 89 (1-2): 37–65.

Saad, D., and S. A. Solla. 1995. “On-line learning in soft committee machines.” *Physical Review E* 52 (4): 4225.

Scholkopf, B., and A. J. Smola. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press.

Shwartz-Ziv, R., and N. Tishby. 2017. “Opening the Black Box of Deep Neural Networks via Information.” *arXiv preprint arXiv:1703.00810.*

Smola, A. J., B. Schölkopf, and K.-R. Müller. 1998. “The connection between regularization operators and support vector kernels.” *Neural networks* 11 (4): 637–649.

Spigler, S., M. Geiger, S. d’Ascoli, L. Sagun, G. Biroli, and M. Wyart. 2019. “A jamming transition from under-to over-parametrization affects generalization in deep learning.” *Journal of Physics A: Mathematical and Theoretical* 52 (47): 474001.

Spigler, S., M. Geiger, and M. Wyart. 2019. “Asymptotic learning curves of kernel methods: empirical data vs Teacher-Student paradigm.” *arXiv preprint arXiv:1905.10843.*

Stein, M. L. 1999. “Predicting random fields with increasing dense observations.” *The Annals of Applied Probability* 9 (1): 242–273.

Wendland, H. 2004. *Scattered data approximation.* Vol. 17. Cambridge university press.

A Kernel regression with invariant dimensions

Lemma *Let $K(\underline{x}) = K(|\underline{x}|)$ be an isotropic function on $\mathbb{R}^d$, continuous everywhere and differentiable everywhere but at the origin. Assume that its Taylor expansion in $\underline{x} = 0$ is*

$$K(\underline{x}) = K(0) - C\,|\underline{x}|^{\theta} + o\left(|\underline{x}|^{\theta}\right), \tag{42}$$

with an exponent $0 < \theta < 2$. Since the function K is isotropic we can define its Fourier transform in any dimension, as:

$$\mathcal{F}_d\left[K(\underline{x})\right](\underline{w}) = L^{-d/2}\int_{\mathcal{V}_d} \mathrm{d}^d\underline{x}\, e^{-i\underline{w}\cdot\underline{x}} K\left(|\underline{x}|\right). \tag{43}$$

At large frequencies these Fourier transforms decay algebraically with a dimension-dependent exponent:

$$\mathcal{F}_d\left[K(\underline{x})\right](\underline{w}) = O\left(|\underline{w}|^{-\alpha(d)}\right). \tag{44}$$

The exponent is then

$$\alpha(d) = d + \theta. \tag{45}$$

Remarks.

(i) The exponent characterizing the decay of the Fourier transform at large frequencies is strictly related to the local behavior of the kernel function at the origin: the same exponent is found if such a function is defined in a compact volume $\mathcal{V}_d$ containing the origin; this is easily verified numerically and is arguably provable, however the proof would be slightly more involved because when using the spherical coordinates one would need to keep track of the shape of the volume $\mathcal{V}_d$.

(ii) What happens when the cusp exponent is not $\theta < 2$? This is the case, for instance, of Matérn kernels $K_\nu(\underline{x})$ with $\nu \geq 1$, that include the Gaussian case as the $\nu \to \infty$ limit. When $\nu < 1$, the first terms in the Taylor expansion around the origin are $K_\nu(\underline{x}) \sim 1 - C\,|\underline{x}|^{2\nu}$, but when $\nu \geq 1$, the truncated Taylor expansion becomes $K_\nu(\underline{x}) \sim 1 - C\,|\underline{x}|^2$. The exponent $\alpha(d)$ is however always $d + 2\nu$. As we argue in the proof, this is due to the fact that the exponent θ that enters in $\alpha(d)$ is indeed related to the first non-smooth term in the Taylor expansion, but the cusp does not necessarily appear in the lowest order terms. Therefore, the exponent ξ that we introduce in Eq. (16) is defined as $\xi = \min(\theta, 2)$.

Proof.

We start by writing the Fourier transform in spherical coordinates (we consider only $d \geq 2$, but a similar calculation holds true for $d = 1$):

$$\begin{aligned}
\int \mathrm{d}^d\underline{x}\, e^{-i\underline{w}\cdot\underline{x}} K(|\underline{x}|) &\propto \int_0^\infty \mathrm{d}x\, x^{d-1} K(x) \int_0^\pi \mathrm{d}\phi_1\, e^{-i\|\underline{w}\| x\cos\phi_1} \sin^{d-2}\phi_1 = \\
&= |\underline{w}|^{-d} \int_0^\infty \mathrm{d}t\, t^{d-1} K\left(\frac{t}{|\underline{w}|}\right) I_{d-2}(t) = \\
&= |\underline{w}|^{-d} \int_0^\infty \mathrm{d}t\, t^{d-1} I_{d-2}(t) \left[K(0) - C\left(\frac{t}{|\underline{w}|}\right)^{\theta} + o\left(\left(\frac{t}{|\underline{w}|}\right)^{\theta}\right)\right] = \\
&= \text{const}\cdot\delta(\underline{w}) - \text{const}'\cdot|\underline{w}|^{-d-\theta}\int_0^\infty \mathrm{d}t\, t^{d+\theta-1} I_{d-2}(t) + o\left(|\underline{w}|^{-d-\theta}\right).
\end{aligned} \tag{46}$$

The first passage follows from the rotational invariance of $K(|\underline{x}|)$, that implies that its Fourier transform only depends on $|\underline{w}|$ and therefore $\underline{w}$ can be assumed to be be parallel to the first component x_1. The function $I_d(t)$ is defined as

$$I_{d-2}(t) = \int_0^\pi \mathrm{d}\phi_1\, e^{-it\cos\phi_1} \sin^{d-2}\phi_1. \tag{47}$$

To show that at large $|\underline{w}|$ the Fourier transform decays as $|\underline{w}|^{-d-\theta}$ we have to show that the integral

multiplying this term is finite. This follows easily considering that $I_{d-2}(0) = \int_0^\pi \mathrm{d}\phi\ \sin^{d-2}\phi < \infty$ and that the behavior at large t is exponentially bounded:

$$|I_{d-2}(t)| \le \frac{2}{t}\int_0^t \mathrm{d}u\,\left(1-\frac{u^2}{t^2}\right)^{\frac{d-3}{2}} \le \frac{1}{t}\int_{-\infty}^{\infty}\mathrm{d}u\,e^{-\frac{d-3}{2}\frac{u^2}{t^2}} \sim e^{-\mathrm{const}\cdot t}. \tag{48}$$

Thus we have the link between the cusp exponent θ and the Fourier exponent α when $\theta < 2$. When $\theta \ge 2$ others scenarios are possible, as one can see with Matérn kernels. A Matérn kernel with parameter ν close to $x = 0$ behaves as $K_\nu(x) \sim 1 - Cx^{2\nu}$ if $\nu < 1$ and as $K_\nu(x) \sim 1 - Cx^2$ if $\nu \ge 1$ (there is no cusp). Nonetheless, the Fourier transform decays in both cases with an exponent $\alpha(d) = d + 2\nu$. The reason is that one can Taylor expand the kernel around $x = 0$ as

$$K_\nu(x) = \sum_{n\ge 0} a_{2n}x^{2n} + x^{2\nu}\sum_{n\ge 0} b_{2n}x^{2n} = k^\infty(x) + \mathrm{const}\cdot x^{2\nu} + o\left(x^{2\nu}\right), \tag{49}$$

where $k^\infty(x) \in C^\infty$ is a smooth function. Smooth functions have a Fourier transform that at large frequency decays faster than any power law, and therefore we recover the relation $\alpha(d) = d + 2\nu$.

Theorem *Let $K_T(\underline{x})$ and $K_S(\underline{x})$ be two translation-invariant kernels (called the* Teacher *and* Student *respectively) defined on $\mathcal{V}_d \equiv [-L/2, L/2]^d$, and let $\tilde{K}_T(\underline{w})$ and $\tilde{K}_S(\underline{w})$ be their Fourier transforms in $\mathcal{V}_d$. Assume that*

- *$K_T(\underline{x})$, $K_S(\underline{x})$ are continuous everywhere and differentiable everywhere except at the origin $\underline{x} = 0$;*
- *$K_T(\underline{x})$ and $K_S(\underline{x})$ are positive definite and isotropic, that is, they only depend on $|\underline{x}|$;*
- *$K_T(\underline{x})$ and $K_S(\underline{x})$ have a cusp at the origin and its d-dimensional Fourier transform decays at high frequencies as in the previous Lemma, with dimensional-dependent exponents $\alpha_T(d_\parallel)$ and $\alpha_S(d)$, respectively (we will evaluate them at $d_\parallel$ for the Teacher and at d for the Student);*
- *$\lim_{\underline{x}\to 0} K_T(0) < \infty$ and $\lim_{\underline{x}\to 0} K_S(0) < \infty$;*
- *$\lim_{\underline{w}\to 0} \tilde{K}_T(\underline{w}) < \infty$ and $\lim_{\underline{w}\to 0} \tilde{K}_S(\underline{w}) < \infty$.*

Assume furthermore that the Teacher kernel lives in a reduced space of dimension $d_\parallel \le d$, in the sense that

- *$K_T(\underline{x}) \equiv K_T(x_1,\dots,x_d) = K_T(x_1,\dots,x_{d_\parallel}) \equiv K_T\left(\left|\underline{x}_\parallel\right|\right)$ (where we have defined $\underline{x}_\parallel \equiv (x_1,\dots,x_{d_\parallel})^t$).*

We use the Teacher kernel to sample a Gaussian random field $Z_T(\underline{x}) \sim \mathcal{N}(0, K_T)$ at p points that lie on a d-dimensional regular lattice in $\mathcal{V}_d$, with spacing $\delta = Lp^{-1/d}$, and we use the Student kernel to infer $\hat{Z}_S(\underline{x})$ at a new point $\underline{x} \in \mathcal{V}_d$ via regression, that is, by minimizing the mean-squared error on the training dataset

$$\sum_{\mu=1}^{p}\left[Z_T(\underline{x}_\mu) - \hat{Z}_S(\underline{x}_\mu)\right]^2. \tag{50}$$

The performance is evaluated by computing the expected mean-squared error on points independent from those used for training:

$$\mathbb{E}\,\mathrm{MSE} = L^{-d}\mathbb{E}\int_{\mathcal{V}_d}\mathrm{d}^d\underline{x}\,\left[Z_T(\underline{x}) - \hat{Z}_S(\underline{x})\right]^2. \tag{51}$$

(The expectation value is taken with respect to the Teacher random process). Then, as $p \to \infty$,

$$\mathbb{E}\,\mathrm{MSE} = p^{-\beta} + o\left(p^{-\beta}\right) \quad \text{with} \quad \beta = \frac{1}{d}\min(\alpha_T(d_\parallel) - d_\parallel, 2\alpha_S). \tag{52}$$

Moreover, because of the previous Lemma the quantity $\alpha_T(d_\parallel) - d_\parallel = \theta_T$ is independent of both d and $d_\parallel$: the exponent β is therefore independent of the intrinsic dimension $d_\parallel$ and depends only on d.

Proof.

(i) Set-up.

Given a function $F(\underline{x})$ on the d-dimensional box $\mathcal{V}_d = [-L/2, L/2]^d$, we denote its Fourier transform (series) and antitransform by

$$\tilde{F}(\underline{w}) \equiv \mathcal{F}_d\left[F(\underline{x})\right](\underline{w}) = L^{-d/2}\int_{\mathcal{V}_d} \mathrm{d}^d\underline{x}\, e^{-i\underline{w}\cdot\underline{x}}F(\underline{x}), \quad \text{where } \underline{w} \in \mathbb{L}_d \equiv \frac{2\pi}{L}\mathbb{Z}^d, \tag{53}$$

$$F(\underline{x}) \equiv \mathcal{F}_d^{-1}\left[\tilde{F}(\underline{w})\right](\underline{x}) = L^{-d/2}\sum_{\underline{w}\in\mathbb{L}} e^{i\underline{w}\cdot\underline{x}}\tilde{F}(\underline{w}). \tag{54}$$

Given the structure of the Teacher kernel we can write

$$\tilde{K}_T(\underline{w}) = L^{-d_\parallel/2}\int_{[-L/2,L/2]^{d_\parallel}} \mathrm{d}^{d_\parallel}\underline{x}_\parallel e^{-i\underline{w}_\parallel\cdot\underline{x}_\parallel} K_T(\left|\underline{x}_\parallel\right|)\cdot L^{-d_\perp/2}\int_{[-L/2,L/2]^{d_\perp}} \mathrm{d}^{d_\perp}\underline{x}_\perp e^{-i\underline{w}_\perp\cdot\underline{x}_\perp} = \\ = \mathcal{F}_{d_\parallel}\left[K_T(\left|\underline{x}_\parallel\right|)\right](\underline{w}_\parallel)\cdot L^{d_\perp/2}\delta_{\underline{w}_\perp}. \tag{55}$$

This formula says that the Fourier transform of the Teacher kernel has frequencies that also live in the corresponding $d_\parallel$-dimensional subspace in the frequency domain. The term $\delta_{\underline{w}_\perp}$ is a discrete delta (not a Dirac delta): this will be important later because it implies that it is scale invariant: $\delta_{a\underline{w}_\perp} = \delta_{\underline{w}_\perp}$. The first term, that is the Fourier transform of the Teacher kernel restricted to the $d_\parallel$-dimensional space, decays at large frequencies with an exponent $\alpha_T(d_\parallel)$ that depends on the intrinsic dimension $d_\parallel$:

$$\tilde{K}_T^\parallel(\underline{w}_\parallel) \equiv \mathcal{F}_{d_\parallel}\left[K_T\left(\left|\underline{x}_\parallel\right|\right)\right](\underline{w}_\parallel) = c_T(d_\parallel)\left|\underline{w}_\parallel\right|^{-\alpha_T(d_\parallel)} + o\left(\left|\underline{w}_\parallel\right|^{-\alpha_T(d_\parallel)}\right). \tag{56}$$

(ii) Regression.

The solution to the regression problem can be computed in closed form:

$$\hat{Z}_S(\underline{x}) = \underline{k}_S(\underline{x})\cdot\mathbb{K}_S^{-1}\underline{Z}_T, \tag{57}$$

where where $\underline{Z}_T = \left(Z_T(\underline{x}_\mu)\right)_{\mu=1}^P$ are the training data (the points $\underline{x}_\mu$ lie on the regular lattice), $\underline{k}_S(\underline{x}) = \left(K_S(\underline{x}_\mu, \underline{x})\right)_{\mu=1}^P$ and $\mathbb{K}_S = \left(K_S(\underline{x}_\mu, \underline{x}_\nu)\right)_{\mu,\nu=1}^P$ is the Gram matrix, that is invertible since the kernel K_S is assumed to be positive definite. This formula can be written in Fourier space as

$$\tilde{Z}_S(\underline{w}) = \tilde{Z}^\star(\underline{w})\frac{\tilde{K}_S(\underline{w})}{\tilde{K_S}^\star(\underline{w})}, \tag{58}$$

where we have defined $F^\star(\underline{w}) \equiv \sum_{\underline{n}\in\mathbb{Z}^d} F\left(\underline{w} + \frac{2\pi\underline{n}}{\delta}\right)$ for a generic function F.

The mean-squared error can then be written using the Parseval-Plancherel identity. After some calculations we find:

$$\mathbb{E}\,\mathrm{MSE} = L^{-d}\mathbb{E}\int_{\mathcal{V}_d}\mathrm{d}^d\underline{x}\,[Z_T(\underline{x}) - \hat{Z}_S(\underline{x})]^2 = L^{-d}\mathbb{E}\sum_{\underline{w}\in\mathbb{L}_d}\left|\tilde{Z}_T(\underline{w}) - \tilde{Z}_T^\star(\underline{w})\frac{\tilde{K}_S(\underline{w})}{\tilde{K}_S^\star(\underline{w})}\right|^2 = \\ = L^{-d/2}\sum_{\underline{w}\in\mathbb{L}_d\cap\mathcal{B}_d}\tilde{K}_T^\star(\underline{w}) - 2\frac{[\tilde{K}_T\tilde{K}_S]^\star(\underline{w})}{\tilde{K}_S^\star(\underline{w})} + \frac{\tilde{K}_T^\star(\underline{w})[\tilde{K}_S^2]^\star(\underline{w})}{\tilde{K}_S^\star(\underline{w})^2}, \tag{59}$$

where $\mathbb{L}_d = \frac{2\pi}{L}\mathbb{Z}^d$ and $\mathcal{B}_d = \left[-\frac{\pi}{\delta}, \frac{\pi}{\delta}\right]^d$ is the Brillouin zone.

In order to simplify this expression in the case where $d_\parallel \le d$, let us also introduce

$$F^{\star_\parallel}(\underline{w}_\parallel) \equiv \sum_{\underline{n}_\parallel\in\mathbb{Z}^{d_\parallel}} F\left(\underline{w}_\parallel + \frac{2\pi\underline{n}_\parallel}{\delta}\right). \tag{60}$$

Using Eq. (55) it follows that

$$\tilde{K}_T^\star(\underline{w}) \propto \delta_{\underline{w}_\perp} \tilde{K}_T^{\star\parallel}(\underline{w}_\parallel), \tag{61}$$

$$[\tilde{K}_T \tilde{K}_S]^\star(\underline{w}) \propto \delta_{\underline{w}_\perp} [\tilde{K}_T \tilde{K}_S]^{\star\parallel}(\underline{w}_\parallel). \tag{62}$$

Plugging the last two equations in Eq. (59) we see that, because of the terms $\delta_{\underline{w}_\perp}$, we find

$$\mathbb{E}\,\mathrm{MSE} \propto \sum_{\underline{w}_\parallel \in \mathbb{L}_\parallel \cap \mathcal{B}_\parallel} \tilde{K}_T^{\star\parallel}(\underline{w}_\parallel) \left\{ 1 + \frac{[\tilde{K}_S^2]^\star(\underline{w}_\parallel)}{\tilde{K}_S^\star(\underline{w})^2} \right\} - 2 \frac{[\tilde{K}_T \tilde{K}_S]^{\star\parallel}(\underline{w}_\parallel)}{\tilde{K}_S^\star(\underline{w}_\parallel)}. \tag{63}$$

Notice that $\tilde{K}_S^\star$ and $[\tilde{K}_S^2]^\star$ do not turn into $[\tilde{K}_S]^{\star\parallel}$ and $[\tilde{K}_S^2]^{\star\parallel}$: this is because the Student kernel does not has the same invariants as the Teacher, and it depends on all the components. Here $\mathbb{L}_\parallel = \frac{2\pi}{L}\mathbb{Z}^{d_\parallel}$, $\mathcal{B}_\parallel = \left[-\frac{\pi}{\delta}, \frac{\pi}{\delta}\right]^{d_\parallel}$.

(iii) Expansion.

Using the high-frequency behavior of the Fourier transforms of the two kernels we can write:

$$\tilde{K}_T^{\star\parallel}(\underline{w}_\parallel) = \tilde{K}_T(\underline{w}_\parallel) + \delta^{\alpha_T(d_\parallel)} c_T(d_\parallel)\, \psi^\parallel_{\alpha_T(d_\parallel)}(\underline{w}_\parallel \delta) + o\left(\left\| \underline{w}_\parallel \right\|^{-\alpha_T(d_\parallel)} \right), \tag{64}$$

$$[\tilde{K}_T \tilde{K}_S]^{\star\parallel}(\underline{w}_\parallel) = \tilde{K}_T(\underline{w}_\parallel)\tilde{K}_S(\underline{w}_\parallel) + \delta^{\alpha_T(d_\parallel)+\alpha_S} c_T(d_\parallel) c_S\, \psi^\parallel_{\alpha_T(d_\parallel)+\alpha_S}(\underline{w}_\parallel \delta) + o\left(\left\| \underline{w}_\parallel \right\|^{-\alpha_T(d_\parallel)} \right), \tag{65}$$

$$\tilde{K}_S^\star(\underline{w}_\parallel) = \tilde{K}_S(\underline{w}_\parallel) + \delta^{\alpha_S} c_S\, \psi_{\alpha_S}(\underline{w}_\parallel \delta) + o\left(\left\| \underline{w}_\parallel \right\|^{-\alpha_S} \right). \tag{66}$$

We have introduced the functions

$$\psi_\alpha(\underline{w}_\parallel) = \sum_{\underline{n} \in \mathbb{Z}^d \setminus \{0\}} \left\| \underline{w}_\parallel + 2\pi \underline{n} \right\|^{-\alpha}, \tag{67}$$

$$\psi^\parallel_\alpha(\underline{w}_\parallel) = \sum_{\underline{n}_\parallel \in \mathbb{Z}^{d_\parallel} \setminus \{0\}} \left\| \underline{w}_\parallel + 2\pi \underline{n}_\parallel \right\|^{-\alpha}. \tag{68}$$

The hypothesis $K_T(\underline{0}) \propto \int \mathrm{d}\underline{w}\, \tilde{K}_T(\underline{w}) < \infty$ and $K_S(\underline{0}) < \infty$ imply $\alpha_T(d_\parallel) > d_\parallel$ and therefore $\sum_{\underline{n}_\parallel \in \mathbb{Z}^{d_\parallel}} \left\| \underline{n}_\parallel \right\|^{-\alpha_T(d_\parallel)} < \infty$. We can argue similarly that $\psi^\parallel_{\alpha_T(d_\parallel)}(\underline{0}), \psi^\parallel_{\alpha_T(d_\parallel)+\alpha_S}(0), \psi_{\alpha_S}(0)$ are finite. Furthermore, the $\underline{w}_\parallel$'s in the sums are at most of order $O\left(\delta^{-1}\right)$, therefore the terms $\psi_\alpha(\underline{w}\delta)$ are $O(\delta^0)$ and do not influence how Eq. (59) scales with δ.

Expanding Eq. (59) and keeping only the highest orders we find:

$$\begin{aligned}
&\mathbb{E}\,\mathrm{MSE} \propto \\
&\propto \sum_{\underline{w}_\parallel \in \mathbb{L}_\parallel \cap \mathcal{B}_\parallel} \left[2c_T(d_\parallel) \psi^\parallel_{\alpha_T(d_\parallel)}(\underline{w}_\parallel \delta) \delta^{\alpha_T(d_\parallel)} + c_S^2 \psi_{2\alpha_S}(\underline{w}_\parallel \delta) \frac{\tilde{K}_T^\parallel(\underline{w}_\parallel)}{\tilde{K}_S^2(\underline{w}_\parallel)} \delta^{2\alpha_S} + o\left(\left\| \underline{w} \right\|^{-\alpha_T(d_\parallel)} \right) + o\left(\left\| \underline{w} \right\|^{-2\alpha_S} \right) \right] = \\
&= \sum_{\underline{w}_\parallel \in \mathbb{L}_\parallel \cap \mathcal{B}_\parallel} \left[2c_T(d_\parallel) \psi^\parallel_{\alpha_T(d_\parallel)}(\underline{w}_\parallel \delta) \delta^{\alpha_T(d_\parallel)} + c_S^2 \psi_{2\alpha_S}(\underline{w}_\parallel \delta) \frac{\tilde{K}_T^\parallel(\underline{w}_\parallel)}{\tilde{K}_S^2(\underline{w}_\parallel)} \delta^{2\alpha_S} \right] + o\left(\left\| \underline{w} \right\|^{-\alpha_T(d_\parallel)-d_\parallel} \right) + o\left(\left\| \underline{w} \right\|^{-2\alpha_S-d_\parallel} \right).
\end{aligned} \tag{69}$$

We have neglected terms proportional to, for instance, $\delta^{\alpha_T(d_\parallel)+\alpha_S}$, since they are subleading with respect to $\delta^{\alpha_T(d_\parallel)}$, but we must keep both $\delta^{\alpha_T(d_\parallel)}$ and δ^{α_S} since we do not know a priori which one is dominant. The additional term δ^{-d} in the subleading terms comes from the fact that $|\mathbb{L} \cap \mathcal{B}| = O\left(\delta^{-d}\right)$.

The first term in Eq. (69) is the simplest to deal with: since $\left\| \underline{w}_\parallel \delta \right\|$ is smaller than some constant for all

$\underline{w}_\parallel \in \mathbb{L}_\parallel \cap \mathcal{B}_\parallel$ and the function $\psi^\parallel_{\alpha_T(d_\parallel)}(\underline{w}_\parallel \delta)$ has a finite limit, we have

$$\delta^{\alpha_T(d_\parallel)} \sum_{\underline{w}_\parallel \in \mathbb{L}_\parallel \cap \mathcal{B}_\parallel} 2c_T(d_\parallel)\psi^\parallel_{\alpha_T(d_\parallel)}(\underline{w}_\parallel \delta) = O\left(\delta^{\alpha_T(d_\parallel)} |\mathbb{L}_\parallel \cap \mathcal{B}_\parallel|\right) = O\left(\delta^{\alpha_T(d_\parallel)-d_\parallel}\right). \tag{70}$$

We then split the second term in Eq. (69) in two contributions:

Small $\left|\underline{w}_\parallel\right|$ We consider "small" all the terms $\underline{w}_\parallel \in \mathbb{L}_\parallel \cap \mathcal{B}_\parallel$ such that $\left|\underline{w}_\parallel\right| < \Gamma$, where $\Gamma \gg 1$ is $O(\delta^0)$ but large. As $\delta \to 0$, $\psi_{2\alpha_S}(\underline{w}_\parallel \delta) \to \psi_{2\alpha_S}(0)$ which is finite because $K_S(0) < \infty$. Therefore

$$\delta^{2\alpha_S} \sum_{\substack{\underline{w}_\parallel \in \mathbb{L}_\parallel \cap \mathcal{B}_\parallel \\ \left|\underline{w}_\parallel\right| < \Gamma}} c_S^2 \psi_{2\alpha_S}(\underline{w}_\parallel \delta) \frac{\tilde{K}_T^\parallel(\underline{w}_\parallel)}{\tilde{K}_S^2(\underline{w}_\parallel)} \to \delta^{2\alpha_S} c_S^2 \psi_{2\alpha_S}(0) \sum_{\substack{\underline{w}_\parallel \in \mathbb{L}_\parallel \cap \mathcal{B}_\parallel \\ \left|\underline{w}_\parallel\right| < \Gamma}} \frac{\tilde{K}_T^\parallel(\underline{w}_\parallel)}{\tilde{K}_S^2(\underline{w}_\parallel)}. \tag{71}$$

The summand is real and strictly positive because the positive definiteness of the kernels implies that their Fourier transforms are strictly positive. Moreover, as $\delta \to 0$, $\mathbb{L}_\parallel \cap \mathcal{B}_\parallel \cap \{\left|\underline{w}_\parallel\right| < \Gamma\} \to \mathbb{L}_\parallel \cap \{\left|\underline{w}_\parallel\right| < \Gamma\}$, which contains a finite number of elements, independent of δ. Therefore

$$\delta^{2\alpha_S} \sum_{\substack{\underline{w}_\parallel \in \mathbb{L}_\parallel \cap \mathcal{B}_\parallel \\ \left|\underline{w}_\parallel\right| < \Gamma}} c_S^2 \psi_{2\alpha_S}(\underline{w}_\parallel \delta) \frac{\tilde{K}_T^\parallel(\underline{w}_\parallel)}{\tilde{K}_S^2(\underline{w}_\parallel)} = O\left(\delta^{2\alpha_S}\right). \tag{72}$$

Large $\left|\underline{w}\right|$ "Large" $\underline{w}$ are those with $\left|\underline{w}\right| > \Gamma$: we recall that $\Gamma \gg 1$ is $O(\delta^0)$ but large. This allows us to approximate $\tilde{K}_T^\parallel, \tilde{K}_S$ in the sum with their asymptotic behavior:

$$\begin{aligned}\delta^{2\alpha_S} \sum_{\substack{\underline{w}_\parallel \in \mathbb{L}_\parallel \cap \mathcal{B}_\parallel \\ \left|\underline{w}_\parallel\right| > \Gamma}} c_S^2 \psi_{2\alpha_S}(\underline{w}_\parallel \delta) \frac{\tilde{K}_T^\parallel(\underline{w}_\parallel)}{\tilde{K}_S^2(\underline{w}_\parallel)} &\propto \delta^{2\alpha_S} \sum_{\substack{\underline{w}_\parallel \in \mathbb{L}_\parallel \cap \mathcal{B}_\parallel \\ \left|\underline{w}_\parallel\right| > \Gamma}} \left|\underline{w}_\parallel\right|^{-\alpha_T(d_\parallel)+2\alpha_S} \approx \\ &\approx \delta^{2\alpha_S} \int_\Gamma^{1/\delta} \mathrm{d}w_\parallel \, w_\parallel^{d_\parallel - 1 - \alpha_T(d_\parallel)+2\alpha_S} = O\left(\delta^{\min(\alpha_T(d_\parallel)-d_\parallel, 2\alpha_S)}\right). \end{aligned} \tag{73}$$

Therefore in the end

$$\mathbb{E}\,\mathrm{MSE} = O\left(\delta^{\min(\alpha_T(d_\parallel)-d_\parallel, 2\alpha_S)}\right) = O\left(p^{-\beta}\right). \tag{74}$$

In order to compute the exponent β we recall that $p \propto \delta^{-1/d}$, where the dimension d of the whole space appears:

$$\beta(d_\parallel) = \frac{1}{d} \min(\alpha_T(d_\parallel) - d_\parallel, 2\alpha_S). \tag{75}$$

Our claim that $\beta(d_\parallel) = \beta(d)$ regardless of the value of $d_\parallel$ follows from the fact that $\alpha(d) - d$ does not depend on the spatial dimension, as proved in the previous Lemma.

B Regime $\sigma \ll \delta$: curse of dimensionality

We consider here the case where the kernel bandwidth σ is much smaller than the nearest-neighbor distance δ. In this limit the contributions in the expansion of the decision boundary in Eq. (10) are significantly suppressed because the kernel is supposed to decay when its argument is large, and the decision boundary is dominated by the charge of training pattern $\underline{x}^\mu$ that is closest to $\underline{x}$. The sign of the decision function is thus fixed by the sign of the nearest neighbor's charge and the accuracy is driven by the nearest neighbor distance, which is susceptible to the curse of dimensionality.

We can see this more precisely if we approximate the kernel interaction between two points $\underline{x}$ and $\underline{x}'$ as

$$K\left(\frac{\left|\underline{x}-\underline{x}'\right|}{\sigma}\right) \approx \begin{cases} a_0 = K(0) \text{ if } \underline{x} = \underline{x}', \\ a_1 = K\left(\frac{\delta}{\sigma}\right) \ll a_0 \text{ if } \underline{x}' \text{ is one of the nearest neighbors of } \underline{x}^{8}, \\ 0 \text{ otherwise.} \end{cases} \tag{76}$$

Hence, the decision function at a point $\underline{x}^\mu$ reads

$$f(\underline{x}^\mu) \approx a_0 \alpha^\mu y^\mu + a_1 \sum_{\nu \in \partial \underline{x}^\mu} \alpha^\nu y^\nu + b \approx (a_0 + a_1')(\alpha_0 + y^\mu \Delta\alpha) y^\mu + b, \tag{77}$$

where the sum runs over the nearest neighbors of $\underline{x}^\mu$. We use that all points are SV, which results from the hierarchy $a_1 \ll a_0$. Indeed, the interaction term alone is never sufficient for $\left|f(\underline{x}^\mu)\right|$ to exceed one. The second equality is justified by the following reasoning. First, in the limit $\delta \to 0$, the nearest neighbors typically share the same sign, so that all the y^ν's in the sum can be replaced by y^μ. a_1' is thus a_1 times the number of terms in the sum. Then, because the distribution is assumed smooth and the kernel is blind to the data structure coming from distant patterns, the SV charge may only depend on its label: $\alpha^\mu = \alpha_0 + y^\mu \Delta\alpha$. $\Delta\alpha$ is taken independent of the associated label y^μ, as we assume the labels to be balanced. The charge conservation Eq. (14) implies immediately that $\Delta\alpha = -\alpha_0 \langle y \rangle$, where $\langle y \rangle = \frac{1}{p}\sum_\mu y^\mu \sim p^{-1/2}$ and imposing the condition $y^\mu f(\underline{x}^\mu) = 1$ on each points $\underline{x}^\mu$ yields $\alpha_0 = 1/(a_0 + a_1')$ and $b = \langle y \rangle$.

We can now compute the test error of the SVC in the limit $\sigma \ll \delta$. The prediction on a test point $\underline{x}$ is

$$\hat{y}(\underline{x}) \approx \text{sign}\left(a_1 \sum_{\nu \in \partial \underline{x}} \alpha^\nu y^\nu + b\right) \approx \text{sign}\left[\frac{a_1'}{a_0} y_{\text{NN}} + b\right], \tag{78}$$

where with a slight abuse of notation we take the sum over the points $\underline{x}^\nu$ in the training set that are nearest neighbors of the test point $\underline{x}$, and y_{NN} is their label (as before, assumed to be constant among nearest neighbors). We observe two distinct behaviors according to the ratio between the bias $b = \langle y \rangle$ and the nearest-neighbor contribution a_1'. If $\langle y \rangle \sim p^{-1/2}$ is much larger than a_1', the above prediction yields $\hat{y}(\underline{x}) = \text{sign}\ \langle y \rangle$ (for any $\underline{x}$): this estimator cannot beat a 50% accuracy. On the contrary, if $\langle y \rangle$ is much smaller than a_1', the prediction yields $\hat{y}(\underline{x}) = \text{sign}(y_{\text{NN}})$: the classifier acts as a nearest-neighbor algorithm, and consequently its test error scales as the nearest-neighbor distance, $\epsilon \sim \delta \sim p^{-1/d}$ — namely, it is susceptible to the curse of dimensionality — as we show in figure Fig. 12.

C Proof that power kernels are CSPD

The margin-SVC algorithm presented on Section 3.4 relies on the assumption that the Gram matrix is conditionally strictly positive definite (CSPD). In this appendix, we prove that the power kernel $K(\underline{x}, \underline{x}') = -\left(\frac{\|\underline{x}-\underline{x}'\|}{\sigma}\right)^\xi$ indeed belongs to the CSPD class for $0 < \xi < 2$ and for any space dimension, by introducing the following definitions and theorems:

Definition: *A real function k is called conditionally strictly positive definite (CSPD) in $\mathbb{R}^d$, if*

$$\sum_{\mu=1}^{p}\sum_{\nu=1}^{p} c_\mu c_\nu k\left(\left|\underline{x}_\mu - \underline{x}_\nu\right|\right) > 0, \tag{79}$$

8. For the derivation of the following scalings the notion of "nearest neighbors" could be relaxed to include points that lie in a thin shell. In any case we assume that the number of nearest neighbors of a given point if finite.

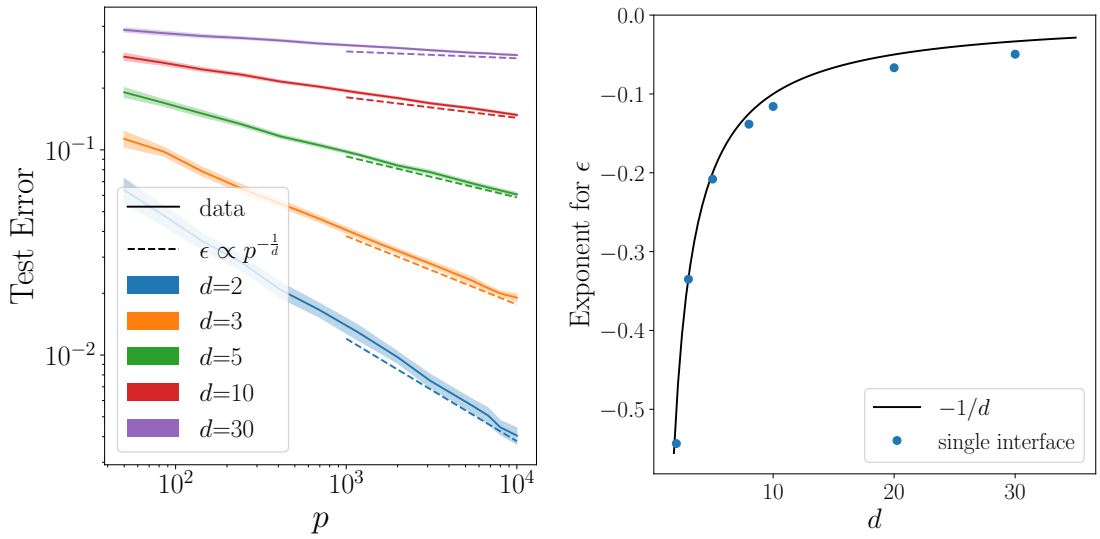

Figure 12. Left: Test error vs the size of the training-set size p for the **single-interface setup** in the vanishing bandwidth regime. The points in the dataset are drawn from the standard normal distribution in dimension d (see the color legend) and learned with the margin-SVC algorithm with the Laplace kernel ($\xi = 1$) of bandwidth $\sigma = 10^{-2}$. The solid lines correspond to the average over 50 initializations, while the shaded regions are the associated standard deviations. The dashed lines illustrate the power law $\epsilon \sim p^{-1/d}$. The bias of the SVC decision function has been removed by hand to avoid that the test error remains stuck at 50% as discussed at the end of Appendix B. Right: The power law exponents are extracted by fitting the curves on the left plot and compared to the nearest neighbor prediction.

for any set of p distinct points $\underline{x}_1, \ldots, \underline{x}_p \in \mathbb{R}^d$ and any choice of p variables $c_1, \ldots, c_p$, satisfying

$$\sum_{\mu=1}^{p} c^\mu = 0. \tag{80}$$

Definition: *A function ϕ is said completely monotone in $(0, \infty)$ if is satisfies $\phi \in C^\infty(0, \infty)$ and $(-1)^n \partial^{(n)} \phi(r) \geq 0$, for all $n \in \mathbb{N}_0$ and all $r > 0$.*

Theorem: *Let $\phi \in C[0, \infty) \cap C^\infty(0, \infty)$. The function $k(\bullet) = \phi(\|\bullet\|^2)$ is CSPD in $\mathbb{R}^d$ for all d, if and only if its negative derivative $-\phi'$ is completely monotone on $(0, \infty)$ and ϕ is not a polynomial of degree at most one. A proof can be found in chapter 8 of (Wendland 2004).*

The introductory statement arises naturally when considering the univariate function $\phi(r) = -r^{\xi/2}$ defined on $\mathbb{R}_+$. Following the theorem and the definitions, one easily show that the function $-\phi'(r) = \frac{\xi}{2} r^{\xi/2-1}$ is completely monotone on $(0, \infty)$ for $0 \leq \xi \leq 2$. The condition that ϕ be not a polynomial of degree at most one excludes further the cases $\xi = 0$ and $\xi = 2$, which proves that the function $k(r) = -r^\xi$ is CSPD for $0 < \xi < 2$. Note that a radial kernel is defined as the multivariate function $K(\underline{x}, \underline{x}') = k\left(\left\|\underline{x} - \underline{x}'\right\|\right)$, and that if the kernel generator k is CSPD, the kernel K is also called CSPD.

D Large σ convergence of the SVC algorithm

In section 3.4, it is loosely argued that in the limit of large σ one could replace the actual kernel $K(r/\sigma)$ by its truncated Taylor expansion $\hat{K}(r/\sigma)$. Here, we prove that in the limit $\sigma \to \infty$, the SVC solution with the truncated kernel converges to the actual SVC solution: $\{\hat{\alpha}^\mu\} \xrightarrow{\sigma \to \infty} \{\alpha^\mu\}$.

We assume that the kernel K can be written as:

$$K\left(\frac{r}{\sigma}\right) = \hat{K}\left(\frac{r}{\sigma}\right) + O\left(\sigma^{-\xi}\right) \text{ , with } \hat{K}\left(\frac{r}{\sigma}\right) = c_0 + c_1\left(\frac{r}{\sigma}\right)^{\xi}$$

For a given classification problem $\{(\underline{x}^\mu, y^\mu)\}$, the SVC algorithm converges to a set of dual variables $\{\alpha^\mu\}$, respectively $\{\hat{\alpha}^\mu\}$ provided that the associated kernel is conditionally strictly positive definite (CSPD). $\hat{K}$ is proved to be CSPD in appendix Appendix C if $c_1 < 0$ and $0 < \xi < 2$, while K is assumed to be CSPD from the start. This condition guarantees that the Lagrangian in Eq. (11) defines a strictly convex problem. Rescaling the dual variables $\alpha^\mu \to \alpha^\mu/\sigma^\xi$ yields the following rescaled Lagrangians:

$$\hat{\mathcal{L}}(\alpha) = \sum_{\mu=1}^{p} \alpha^\mu - \frac{c_1}{2}\sum_{\mu,\nu=1}^{p} \alpha^\mu \alpha^\nu y^\mu y^\nu \left|\underline{x}^\mu - \underline{x}^\nu\right|^\xi \text{ and } \mathcal{L}(\alpha) = \hat{\mathcal{L}}(\alpha) + \epsilon(\sigma), \tag{81}$$

The rescaled solution $\{\hat{\alpha}^\mu\}$ of the maximizing problem with the Lagrangian $\hat{\mathcal{L}}$ is well defined in the limit $\sigma \to \infty$, hence the strict convexity of both Lagrangian ensures that $\{\hat{\alpha}^\mu\} \to \{\alpha^\mu\}$, when the perturbation $\epsilon(\sigma)$ vanishes.

E The kernel Fourier transform

The (transverse) Fourier transform of the Kernel decays at large frequencies with a power law with an exponent that depends on the exponent ξ in Eq. (16):

$$\begin{aligned}\hat{K}_\perp\left(\underline{k}_\perp\right) &= \int \mathrm{d}^{d-1}\underline{x}_\perp\, e^{-i\underline{k}_\perp\cdot\underline{x}_\perp} K\left(\frac{\left|\underline{x}_\perp\right|}{\sigma}\right) \sim -\int \mathrm{d}^{d-1}\,\underline{x}_\perp e^{-i\underline{k}_\perp\cdot\underline{x}_\perp}\left(\frac{\left|\underline{x}_\perp\right|}{\sigma}\right)^\xi \sim \\ &\sim -\sigma^{-\xi}\int_0^\infty \mathrm{d}r\, r^{d-2+\xi}\int_0^\pi \mathrm{d}\phi\, e^{-i\left\|\underline{k}_\perp\right\| r\cos\phi}\sin^{d-2}\phi \sim -\sigma^{-\xi}\left|\underline{k}_\perp\right|^{-(d-1+\xi)}\cdot\int_0^\infty \mathrm{d}t\, t^{d-2+\xi}\Phi(t).\end{aligned} \tag{82}$$

The function that appears in the integral is

$$\Phi(t) = \int_0^\pi \mathrm{d}\phi\, e^{-it\cos\phi}\sin^{d-2}\phi = \frac{1}{t}\int_{-t}^{t}\mathrm{d}u\, e^{-iu}\left(1-\frac{u^2}{t^2}\right)^{\frac{d-3}{2}}, \tag{83}$$

and it is easy to see that $\Phi(0)$ is finite and that $\Phi(t)$ decays exponentially fast at infinity (see for instance the proof of the Lemma in Appendix A). Therefore the integral in $\hat{K}\left(\underline{k}_\perp\right)$ is just a finite constant.

F The charge structure factor

The charge structure factor $\tilde{Q}$ introduced in Eq. (28) is a good measure of the fluctuations in the system and, in particular, of the cutoff occurring at the scale r_c. It is argued in Section 3.4 that $\tilde{Q}^2(\underline{k}_\perp) \sim \bar{\alpha}^2 p\Delta/\gamma$ at large frequencies, namely $\left|\underline{k}_\perp\right| > r_c^{-1}$. This scaling is verified numerically in Fig. 13.

The data are obtained as follows: for each $\left|\underline{k}_\perp\right|$, a set of $N = 2000$ random wave vectors are generated on the interface; the associated factor is computed by summing over the SV of the considered setup and then averaged. The fluctuations observed at large $\left|\underline{k}_\perp\right|$ decrease when N increases. The insets illustrate the expected asymptotic behavior $\tilde{Q}^2_\infty \approx \bar{\alpha}^2 p\Delta/\gamma$, while the vertical dotted lines correspond to the typical nearest-neighbor distance r_c.

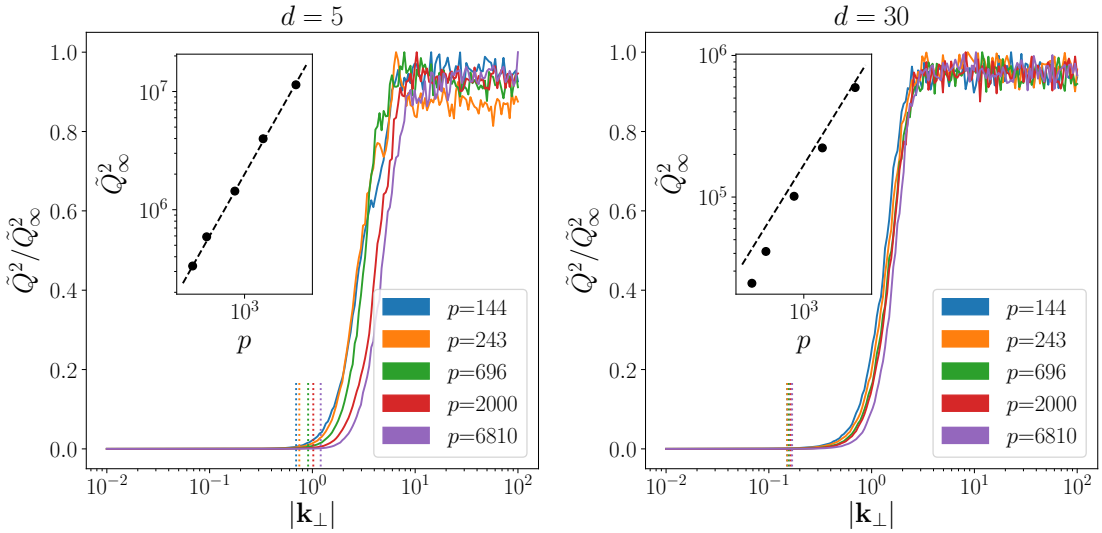

Figure 13. Charge structure factor as a function of the (transverse) wave vector amplitude $|\underline{k}_\perp|$, for different training set sizes p and dimensions $d = 5, 30$. We plot the square $\tilde{Q}^2(\underline{k}_\perp)$ averaged over $N = 2000$ samples, normalized by the expected high-frequency variance $\tilde{Q}^2_\infty = \bar{\alpha}^2 p\Delta/\gamma$.

G Matérn kernel

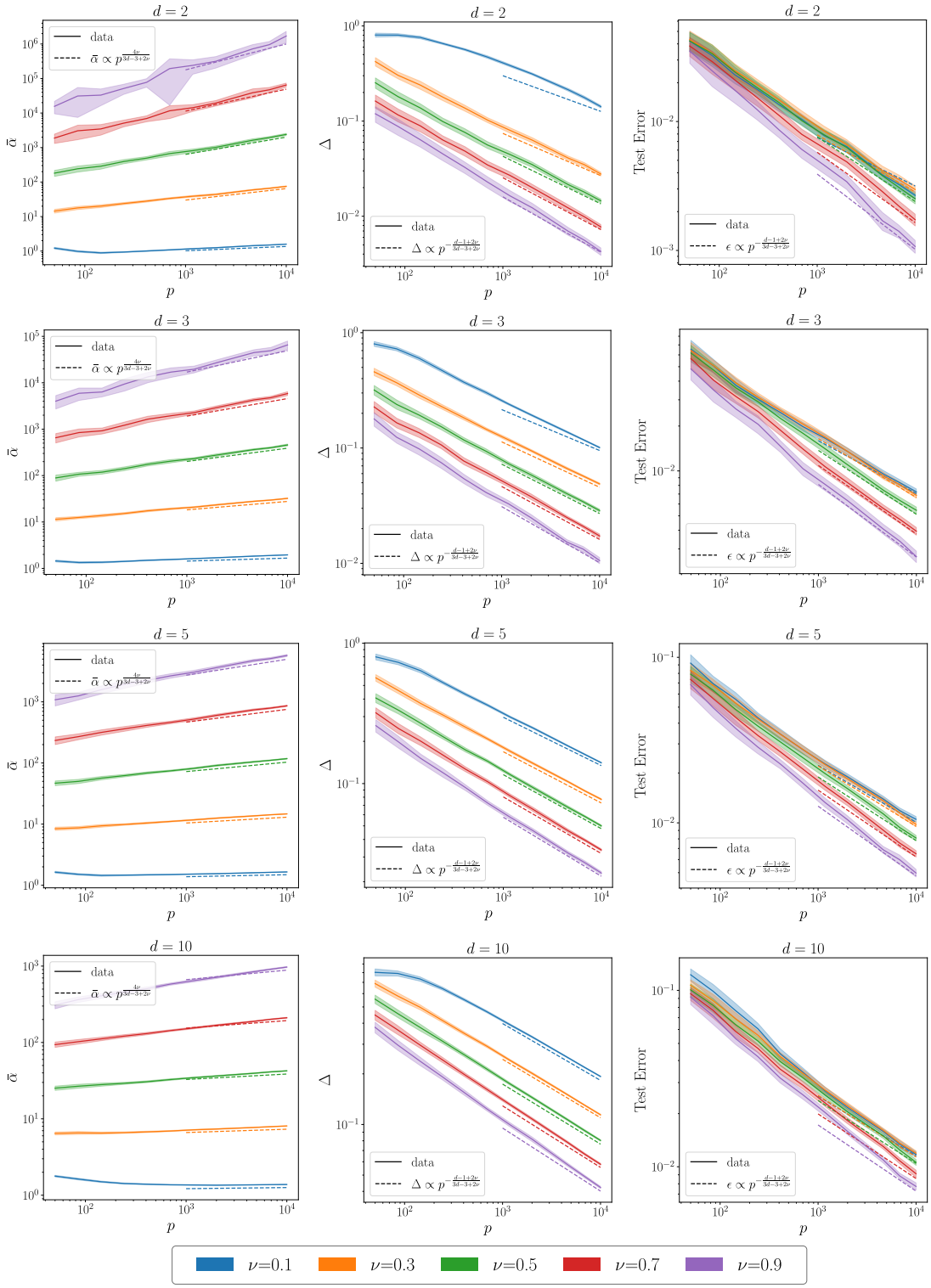

Figure 14. Dependence on the training set size p of the SV mean dual variable $\bar{\alpha}$ (left), the SV band thickness Δ (middle) and the test error (right) for the **single-interface setup** in dimensions $d = 2, 3, 5, 10$. The SVC algorithm is run with the Matérn kernel with bandwidth $\sigma = 100 \gg \delta$ and parameter $\nu = 0.1, 0.3, 0.5, 0.7, 0.9$ for which the kernel is conditionally strictly positive definite. The solid lines are averaged over 50 initializations and the shaded regions represent the standard deviation. Dashed lines illustrate the power-law predictions of Eq. (31).

H Numerical definition of the scale r_c

In Section 3.4, the scale r_c is defined geometrically as the distance between nearest support vectors. The numerical definition of r_c is different as it aims at confirming the "minimal disturbance hypothesis" presented in the note at the end of Section 3.4. From this point of view, the scale r_c is also the scale behind which the charge of two SVs are not correlated. To test this idea, the solution of the margin-SVC problem is computed once for a benchmark training set and a second time for the same training set with one additional point close enough to the interface to be a SV. We then calculate the cumulative distribution of the charge variations $d\alpha^\mu = \|\alpha^\mu - \alpha'^\mu\|$ as function of their distance to the additional point r^μ. The resulting distribution is displayed on Fig. 15 for multiple realizations of the interface setup with $d = 5$ and $p = 6810$. The scale r_c is then defined as the distance for which the cumulative distribution reaches a given value $C < 1$. The particular choice of C leads to the same power law behavior.

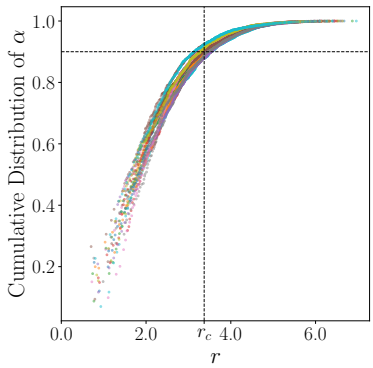

Figure 15. Example of the cumulative distribution of the amplitude of the dual variable variation as a function of the distance r to the additional point (see the text above). Each color corresponds to a different realization of the interface setup with $d = 5$ and $p = 6810$. The vertical dashed line stands for the scale r_c averaged over all realizations obtained with $C = 0.9$ (horizontal dashed line).

I Scaling arguments for the spherical setup

In this appendix, we sketch how the scaling relations in Section 3.4 may be derived for the spherical interface setup discussed in Section 4, where the label only depends on the norm of the vector: $y(\underline{x}) = \text{sign}(|\underline{x}| - R)$, with R the radius of the sphere. In the same line as for the linear interface, it is assumed that all SVs lie within a shell of thickness $\Delta \ll 1$ around the interface. The decision function on the vector $\underline{x}$,

$$f(\underline{x}) = b - \sum_{\mu=1}^{p} \alpha^\mu y^\mu \left(\frac{|\underline{x} - \underline{x}^\mu|}{\sigma} \right)^\xi , \tag{84}$$

is better apprehended in a Cartesian frame such that $\underline{x} = (x_1 = |\underline{x}|, \underline{0})$, which requires to rotate all SVs: $\underline{x}^\mu \longrightarrow x'^\mu = \mathcal{R}\underline{x}^\mu$. In the large p limit, the charge conservation, $Q = \sum_{\mu=1}^{p} \alpha^\mu y^\mu = 0$, reads

$$0 = \int \mathrm{d}^d \underline{x} \rho(\underline{x}) \alpha(\underline{x}) y(\underline{x}) = S_{d-1} \int_{-\Delta}^{\Delta} \mathrm{d}u (R+u)^{d-1} \rho(R+u) \alpha(R+u) y(R+u). \tag{85}$$

Spherical coordinates are used in the second equality: the angular variables trivially integrate to the unit $(d-1)$-sphere surface, S_{d-1}, and the variable $u = r - R$ is used instead of the radius $r = |\underline{x}|$. For simplicity, we assume that the population distribution is radial: $\rho(\underline{x}) = \rho(r)$. Were it not the case, the angular integral would merely yield a different finite factor.

As for the linear interface, the first scaling relation stems from the condition $\Delta \cdot \partial_{x_\parallel} f(\underline{x}^\star) \sim 1$. According to the change of frame introduced above, the relevant direction correspond to the first coordinate, namely $\underline{x}_\parallel = \underline{x}_1$. The gradient expression (21) can thus be expressed as an integral in spherical coordinate with

the north pole $\underline{x}\star = (R, \underline{0})$:

$$\partial_{x_\parallel} f(\underline{x}^\star) = \xi \sigma^{-\xi} p S_{d-2} \int_{-\Delta}^{\Delta} \mathrm{d}u (R+u)^{d-1} \int_0^\pi \mathrm{d}\phi \sin^{d-2}\phi \rho(R+u)\alpha(R+u)y(R+u)I(u,\phi), \tag{86}$$

where the vector of integration norm is $r = R + u$ and its angle with respect to the north pole is ϕ. All other angles simply integrate to the $(d-2)$-sphere surface, S_{d-2}, since they don't contribute to the integrand

$$I(u,\phi) = (x_1 - x_1^\star)\left|\underline{x} - \underline{x}^\star\right|^{\xi-2} = a_0(\phi) + a_1(\phi)u + O(u^2), \tag{87}$$

with

$$a_0(\phi) = \frac{1}{2R}\left[2R^2(1-\cos\phi)\right]^{\xi/2} \quad \text{and} \quad a_1(\phi) = \left[1 - \frac{\xi}{2}(1-\cos\phi)\right]\left[2R^2(1-\cos\phi)\right]^{\xi/2-1}. \tag{88}$$

The leading order contribution a_0 vanishes because of the charge conservation (Eq. (85)), so that the gradient reads

$$\partial_{x_1} f(\underline{x}^\star) \sim p \int_{-\Delta}^{\Delta} \mathrm{d}u (R+u)^{d-1} \rho(R+u)\alpha(R+u)y(R+u)u \int_0^\pi \mathrm{d}\phi \sin^{d-2}\phi\, a_1(\phi) \sim p\Delta^2\bar{\alpha} \tag{89}$$

and the second scaling relation $p\bar{\alpha}\Delta^3 \sim 1$ is identical as for the stripe model. Since the other relations are obtained from local arguments, they are independent on the global shape of the classification task. The scaling laws for the spherical model are thus also given by Eq. (31) and Eq. (32).