

---

# Parameter-free Statistically Consistent Interpolation: Dimension-independent Convergence Rates for Hilbert kernel regression

---

**Partha P. Mitra**

\*  
Cold Spring Harbor Laboratory  
Cold Spring Harbor, NY 11724, USA  
mitra@cshl.edu

**Clément Sire**

Laboratoire de Physique Théorique  
CNRS & Université de Toulouse – Paul Sabatier  
31062 Toulouse, France  
clement.sire@univ-tlse3.fr

## Abstract

Previously, statistical textbook wisdom has held that interpolation of noisy training data will lead to poor generalization. However, recent work has shown that this is not true and that good generalization can be obtained with function fits that interpolate training data. This could explain why overparameterized deep nets with zero or small training error do not necessarily overfit and could generalize well. Data interpolation schemes have been exhibited that are provably Bayes optimal in the large sample limit and achieve the theoretical lower bounds for excess risk (Statistically Consistent Interpolation) in any dimension. These interpolation schemes are non-parametric Nadaraya-Watson style estimators with singular kernels, which exhibit statistical consistency in any data dimension for large sample sizes. The recently proposed weighted interpolating nearest neighbors scheme (wiNN) is in this class, as is the previously studied Hilbert kernel interpolation scheme. In the Hilbert scheme, the regression function estimator for a set of labelled data pairs,  $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ ,  $i = 0, \dots, n$ , has the form  $\hat{f}(x) = \sum_i y_i w_i(x)$ , where  $w_i(x) = \|x - x_i\|^{-d} / \sum_j \|x - x_j\|^{-d}$ . This interpolating function estimator is unique in being entirely free of parameters and does not require bandwidth selection. While statistical consistency was previously proven for this scheme, the precise convergence rates for the finite sample risk were not established. Here, we carry out a comprehensive study of the asymptotic finite sample behavior of the Hilbert kernel regression scheme and prove a number of relevant theorems. We prove under broad conditions that the excess risk of the Hilbert regression estimator is asymptotically equivalent pointwise to  $\sigma^2(x) / \ln(n)$  where  $\sigma^2(x)$  is the noise variance. We also show that the excess risk of the plugin classifier is upper bounded by  $2|f(x) - 1/2|^{1-\alpha} (1 + \varepsilon)^\alpha \sigma^\alpha(x) (\ln(n))^{-\frac{\alpha}{2}}$ , for any  $0 < \alpha < 1$ , where  $f$  is the regression function  $x \mapsto \mathbb{E}[y|x]$ . Our proofs proceed by deriving asymptotic equivalents of the moments of the weight functions  $w_i(x)$  for large  $n$ , for instance for  $\beta > 1$ ,  $\mathbb{E}[w_i^\beta(x)] \sim_{n \rightarrow \infty} ((\beta - 1)n \ln(n))^{-1}$ . We further derive an asymptotic equivalent for the Lagrange function and explicitly exhibit the nontrivial extrapolation properties of this estimator. Notably, the convergence rates are independent of data dimension and the excess risk is dominated by the noise variance. The bias term, for which we also give precise asymptotic estimates, is always subleading when the density of data at the considered point is strictly positive. If this local density is zero, we show that the bias term does not vanish in the limit of a large data set and we compute its limit explicitly. Finally, we present heuristic arguments

---

\*Center for Computational Brain Research, IIT Madras, Chennai, India

for a universal  $w^{-2}$  power-law behavior of the probability density of the weights in the large  $n$  limit.

## 1 Introduction

Data interpolation and statistical regression of noisy data are both classical subjects but their domain of application have been disjoint until recently. Scattered data interpolation techniques [1] are generally used for clean data. On the other hand, when supervised learning or statistical regression techniques are applied to noisy data, in general smoothing or regularization methods are applied to prevent training data interpolation, as the latter is believed to lead to poor generalization [2]. However, accumulating empirical evidence from overparameterized deep networks has shown that data interpolation (equivalently, zero error on the training set) does not automatically imply poor generalization [3, 4]. This has in turn given rise to a rapidly growing body of theoretical work to understand how and why noisy data interpolation can still lead to good generalization [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15].

A key observations in this regard is the phenomenon of Statistically Consistent Interpolation [16], i.e., regression function estimation that interpolates training data but also generalizes as well as possible by achieving the Bayes limit for expected generalization error (risk) when the sample size becomes large. This hints at a rich set of theoretical questions at the interface between the disciplines of scattered data interpolation and supervised learning, that have only begun to be addressed. In particular, there has been comparatively little study of the generalization error or risk of interpolating learners. Computation of generalization error bounds in machine learning often relies on the capacity of the class of fitting functions [17], however such model complexity based bounds are not tight enough to be useful for interpolating learners [4]. For nonparametric interpolation approaches such as that considered here, it is also not clear what model complexity means. Thus, there is a need for other approaches to understanding the generalization behavior of nonparametric interpolating learners, including more direct treatments of the generalization error for specific interpolation schemes so as to gain better theoretical understanding. The current paper addresses this need.

We present a detailed analysis of the finite-sample risk of an interpolating learner with intriguing theoretical properties, the Hilbert kernel estimator (Devroye *et. al.* [18]). A unique property of this Nadaraya-Watson (NW) style estimator [19, 20] is that it is fully parameter-free and does not have any bandwidth or scale parameter. It is global and uses all data points for each estimate: the associated kernel is a power law and thus scale-free. Although statistical consistency of this estimator was proven [18] when it was proposed, there has been no systematic analysis of the associated convergence rates and asymptotic finite sample behavior. We provide this analysis in the present study.

**Related work** The only other interpolation scheme we are aware of, that is proven to be statistically consistent in arbitrary dimensions under general conditions, is the recently proposed weighted interpolating nearest neighbors method (wiNN) [7], which is also a NW estimator utilizing a singular power law kernel of a very similar form but with two important differences: a finite number of neighbors  $k$  is utilized (rather than all data points), and the power law exponent  $\delta$  of the NW kernel satisfies  $0 < \delta < d/2$  rather than  $\delta = d$ . To achieve consistency  $k$  has to scale appropriately with sample size. Despite the superficial resemblance, the wiNN and Hilbert Kernel estimators have quite different convergence rates, as we will see from the results of this paper. Also worth mentioning is the Shepard interpolation scheme [21] originally proposed for interpolation of 2D geospatial data sets, also a NW style interpolating estimator, though used in the context of scattered data interpolation. In scattered data interpolation [1], the focus is generally on the approximation error (corresponding to the “bias” term in our analysis below). The approximation error of the Shepard scheme has been analyzed [22] but as we will see below the risk for Hilbert kernel interpolation is dominated by the noise or “variance” term. In contrast with wiNN or Hilbert kernel interpolation, other interpolating learning methods such as simplex interpolation [7] or ridgeless kernel regression [11] are generally not statistically consistent in fixed finite dimension [8].

**Summary of results of this paper** Notation and assumptions pertaining to this summary are defined in the problem setup section below. We prove under broad conditions that the excess risk of the Hilbert regression estimator is asymptotically equivalent pointwise to  $\sigma^2(x)/\ln(n)$  where  $\sigma^2(x)$  is the noise variance. We also show that the excess risk of the plugin classifier is upper bounded by

$2|f(x) - 1/2|^{1-\alpha} (1 + \varepsilon)^\alpha \sigma^\alpha(x) (\ln(n))^{-\frac{\alpha}{2}}$ , for any  $0 < \alpha < 1$ , where  $f$  is the regression function  $x \mapsto \mathbb{E}[y|x]$ . Our proofs proceed by deriving asymptotic equivalents of the moments of the weight functions  $w_i(x)$  for large  $n$ , for instance for  $\beta > 1$ ,  $\mathbb{E}[w_i^\beta(x)] \sim_{n \rightarrow \infty} ((\beta - 1)n \ln(n))^{-1}$ . We further derive an asymptotic equivalent for the Lagrange function and explicitly exhibit the nontrivial extrapolation properties of this estimator. Notably, the convergence rates are independent of data dimension and the excess risk is dominated by the noise variance. The bias term, for which we also give precise asymptotic estimates, is always subleading when the density of data at the considered point is strictly positive. If this local density is zero, we show that the bias term does not vanish in the limit of a large data set and we compute its limit explicitly. Finally, we present heuristic arguments for a universal  $w^{-2}$  power-law behavior of the probability density of the weights in the large  $n$  limit.

## 2 Problem setup

**Notation, Definitions, Statistical Model** We model the labelled training data set  $(x_0, y_0), \dots, (x_n, y_n)$  as  $n + 1$  *i.i.d.* observations of a random vector  $(X, Y)$  with values in  $\mathbb{R}^d \times \mathbb{R}$  for regression, and with values in  $\mathbb{R}^d \times \{0, 1\}$  for binary classification. Due to the independence property, the collection  $X_0, \dots, X_n$  has the product density  $\prod_{i=0}^n \rho(x_i)$ . We will denote by  $\mathbb{E}$  an expectation over the collection of  $n + 1$  random vectors and by  $\mathbb{E}_X$  the expectation over the collection  $X_0, \dots, X_n$ . An expectation over the same collection while holding  $X_i = x_i$  will be denoted  $\mathbb{E}_{X|x_i}$ . The regression function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  is defined as the conditional mean of  $Y$  given  $X = x$ ,  $f(x) := \mathbb{E}[Y | X = x]$  and the conditional variance function is  $\sigma^2(x) := \mathbb{E}[|Y - f(X)|^2 | X = x]$ .  $f$  minimizes the expected value of the mean squared prediction error (risk under squared loss),  $f = \arg \min \mathcal{R}_{\text{sq}}(h)$  where  $\mathcal{R}_{\text{sq}}(h) := \mathbb{E}[(h(X) - Y)^2]$ . Given any regression estimator  $\hat{f}(x)$  the corresponding risk can be decomposed as  $\mathbb{E}[\mathcal{R}_{\text{sq}}(\hat{f}(X))] = \mathcal{R}_{\text{sq}}(f) + \mathbb{E}[(\hat{f}(X) - f(X))^2]$ . The excess risk is given by  $\mathcal{R}_{\text{sq}}(\hat{f}) - \mathcal{R}_{\text{sq}}(f) = \mathbb{E}[(\hat{f}(X) - f(X))^2]$ . For a consistent estimator this excess risk goes to zero as  $n \rightarrow \infty$  and we are interested in characterizing the *rate* at which it goes to zero with increasing  $n$  (note our sample size is  $n + 1$  for notational simplicity but for large  $n$  this does not change the rate).

In the case of binary classification,  $Y \in \{0, 1\}$  and  $f(x) = \mathbb{P}[Y = 1 | X = x]$ . Let  $F: \mathbb{R}^d \rightarrow \{0, 1\}$  denote the Bayes optimal classifier, defined by  $F(x) := \theta(f(x) - 1/2)$  where  $\theta(\cdot)$  is the Heaviside theta function. This classifier minimizes the risk  $\mathcal{R}_{0/1}(h) := \mathbb{E}[\mathbb{1}_{\{h(X) \neq Y\}}] = \mathbb{P}(h(X) \neq Y)$  under zero-one loss. Given the regression estimator  $\hat{f}$ , we consider the plugin classifier  $\hat{F}(x) = \theta(\hat{f}(x) - 1/2)$ . The classification risk for the plugin classifier  $\hat{F}$  is bounded as  $\mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] - \mathcal{R}_{0/1}(F(x)) \leq 2\mathbb{E}[|\hat{f}(x) - f(x)|] \leq 2\sqrt{\mathbb{E}[(\hat{f}(x) - f(x))^2]}$ .

Finally, we define two sequences  $a_n, b_n > 0$ ,  $n \in \mathbb{N}$ , to be asymptotically equivalent for  $n \rightarrow +\infty$ , denoted  $a_n \sim_{n \rightarrow +\infty} b_n$ , if the limit of their ratio exists and  $\lim_{n \rightarrow \infty} a_n/b_n = 1$ .

In summary, our work will focus on the estimation of asymptotic equivalents for  $\mathbb{E}[(\hat{f}(x) - f(x))^2]$  and other relevant quantities as this determines the rate at which the excess risk goes to zero for regression, and bounds the rate at which the excess risk goes to zero for classification.

**Assumptions.** We define the support  $\Omega$  of the density  $\rho$  as  $\Omega = \{x \in \mathbb{R}^d / \rho(x) > 0\}$ , the closed support  $\bar{\Omega}$  as the closure of  $\Omega$ , and  $\Omega^\circ$  as the interior of  $\Omega$ . Our results will not assume any compactness condition on  $\Omega$  or  $\bar{\Omega}$ . The boundary of  $\Omega$  is then defined as  $\partial\Omega = \bar{\Omega} \setminus \Omega^\circ$ . We assume that  $\rho$  has a finite variance  $\sigma_\rho^2$ . In addition, we will most of the time assume that the density  $\rho$  is continuous at the considered point  $x \in \Omega^\circ$ , and in some cases,  $x \in \partial\Omega \cap \Omega$ .

For the regression function  $f$ , we will obtain results assuming either of the following conditions

- $C_{\text{Cont}}^f$ :  $f$  is continuous at the considered  $x$ ,
- $C_{\text{Holder}}^f$ : for all  $x \in \Omega^\circ$ , there exist  $\alpha_x > 0$ ,  $K_x > 0$ , and  $\delta_x > 0$ , such that  $x' \in \Omega$  and  $\|x - x'\| \leq \delta_x \implies |f(x) - f(x')| \leq K_x \|x - x'\|^{\alpha_x}$  (local Hölder smoothness condition),

where condition  $C_{\text{Holder}}^f$  is obviously stronger than  $C_{\text{Cont}}^f$ . In addition, we will always assume a growth condition for the regression function  $f$ :

- $C_{\text{Growth}}^f: \int \rho(y) \frac{f^2(y)}{1+\|y\|^{2\alpha}} d^d y < \infty$ .

As for the variance function  $\sigma$ , we will obtain results assuming either that  $\sigma$  is bounded or satisfies a growth condition similar to the one above

- $C_{\text{Bound}}^\sigma$ : there exists  $\sigma_0^2 \geq 0$ , such that, for all  $x \in \Omega$ , we have  $\sigma^2(x) \leq \sigma_0^2$ ,
- $C_{\text{Growth}}^\sigma: \int \rho(y) \frac{\sigma^2(y)}{1+\|y\|^{2\alpha}} d^d y < \infty$ .

When we will assume condition  $C_{\text{Growth}}^\sigma$  (obviously satisfied when  $\sigma^2$  is bounded), we will also assume a continuity condition  $C_{\text{Cont}}^\sigma$  for  $\sigma$  at the considered  $x$ .

Note that all our results can be readily extended in the case where  $x \in \partial\Omega = \bar{\Omega} \setminus \Omega^\circ$  but keeping the condition  $\rho(x) > 0$  (i.e.,  $x \in \partial\Omega \cap \Omega$ ), and assuming the continuity at  $x$  of  $\rho$  as seen as a function restricted to  $\Omega$ , i.e.,  $\lim_{y \in \Omega \rightarrow x} \rho(y) = \rho(x)$ . Useful examples are when the support  $\Omega$  of  $\rho$  is a  $d$ -dimensional sphere or hypercube and  $x$  is on the surface of  $\Omega$  (but still with  $\rho(x) > 0$ ). To guarantee these results for  $x \in \partial\Omega \cap \Omega$ , we need also to assume the continuity at  $x$  of  $f$ , and assume that  $\Omega$  is smooth enough near  $x$ , so that there exists a strictly positive local solid angle  $\omega_x$  defined by

$$\omega_x = \lim_{r \rightarrow 0} \frac{1}{V_d \rho(x) r^d} \int_{\|x-y\| \leq r} \rho(y) d^d y = \lim_{r \rightarrow 0} \frac{1}{V_d r^d} \int_{y \in \Omega / \|x-y\| \leq r} d^d y, \quad (1)$$

where  $V_d = S_d/d = \pi^{d/2}/\Gamma(d/2 + 1)$  is the volume of the unit ball in  $d$  dimensions, and the second inequality results from the continuity of  $\rho$  at  $x$ . If  $x \in \Omega^\circ$ , we have  $\omega_x = 1$ , while for  $x \in \partial\Omega$ , we have  $0 \leq \omega_x \leq 1$ . For instance, if  $x$  is on the surface of a sphere or on the interior of a face of a hypercube (and in general, when the boundary near  $x$  is locally an hyperplane), we have  $\omega_x = \frac{1}{2}$ . If  $x$  is a corner of a hypercube, we have  $\omega_x = \frac{1}{2^d}$ . From our methods of proof presented in the appendix, it should be clear that all our results for  $x \in \Omega^\circ$  perfectly generalize to any  $x \in \partial\Omega \cap \Omega$  for which  $\omega_x > 0$ , by simply replacing  $V_d$  whenever it appears in our different results by  $\omega_x V_d$ .

**Hilbert kernel interpolating estimator and Bias-Variance decomposition.** The Hilbert kernel regression estimator  $\hat{f}(x)$  is a Nadaraya-Watson style estimator employing a singular kernel:

$$w_i(x) = \frac{\|x - x_i\|^{-d}}{\sum_{j=0}^n \|x - x_j\|^{-d}}, \quad (2)$$

$$\hat{f}(x) = \sum_{i=0}^n w_i(x) y_i. \quad (3)$$

The weights  $w_i(x)$  are also called Lagrange functions in the interpolation literature and satisfy the interpolation property  $w_i(x_j) = \delta_{ij}$ , where  $\delta_{ij} = 1$ , if  $i = j$ , and 0 otherwise. At any given point  $x$ , they provide a partition of unity so that  $\sum_{i=0}^n w_i(x) = 1$ . The mean squared error between the Hilbert estimator and the true regression function has a bias-variance decomposition (using the *i.i.d* condition and the earlier definitions)

$$\hat{f}(x) - f(x) = \sum_{i=0}^n w_i(x) [f(x_i) - f(x)] + \sum_{i=0}^n w_i(x) [y_i - f(x_i)], \quad (4)$$

$$\mathbb{E}[(\hat{f}(x) - f(x))^2] = \mathcal{B}(x) + \mathcal{V}(x), \quad (5)$$

$$(\text{Bias}) \mathcal{B}(x) = \mathbb{E}_X \left[ \left( \sum_{i=0}^n w_i(x) [f(x_i) - f(x)] \right)^2 \right], \quad (6)$$

$$(\text{Variance}) \mathcal{V}(x) = \mathbb{E} \left[ \sum_{i=0}^n w_i^2(x) [y_i - f(x_i)]^2 \right] = \mathbb{E}_X \left[ \sum_{i=0}^n w_i^2(x) \sigma^2(x_i) \right]. \quad (7)$$

The present work derives asymptotic behaviors and bounds for the regression and classification risk of the Hilbert estimator for large sample size  $n$ . These results are derived by analyzing the large  $n$  behaviors of the bias and variance terms, which in turn depend on the behavior of the moments of the weights or the Lagrange functions  $w_i(x)$ . For all these quantities, asymptotically equivalent forms are derived. The proofs exploit a simple integral form of the weight function and details are provided in the appendix, while the body of the paper provides the results and associated discussions.

### 3 Results

#### 3.1 The weights, variance and bias terms

##### 3.1.1 Moments of the weights: large $n$ behavior

In this section, we consider the moments and the distribution of the weights  $w_i(x)$  at a given point  $x$ . The first moment is simple to compute. Since the weights sum to 1 and  $X_i$  are *i.i.d.*, it follows that  $\mathbb{E}_{X_i|x_i}[w_i(x)]$  are all equal and thus  $\mathbb{E}_{X_i|x_i}[w_i(x)] = (n+1)^{-1}$ . The other moments are much less trivial to compute and we prove the following theorem in the appendix A.2:

**Theorem 3.1.** *For  $x \in \Omega^\circ$  (so that  $\rho(x) > 0$ ), we assume  $\rho$  continuous at  $x$ . Then, the moments of the weight  $w_0(x)$  satisfy the following properties:*

- For  $\beta > 1$ :

$$\mathbb{E} \left[ w_0^\beta(x) \right] \underset{n \rightarrow +\infty}{\sim} \frac{1}{(\beta-1)n \ln(n)}. \quad (8)$$

- For  $0 < \beta < 1$ : defining  $\kappa_\beta(x) := \int \frac{\rho(x+y)}{\|y\|^{|\beta|d}} d^d y < \infty$ , we have

$$\mathbb{E} \left[ w_0^\beta(x) \right] \underset{n \rightarrow +\infty}{\sim} \frac{\kappa_\beta(x)}{(V_d \rho(x) n \ln(n))^\beta}. \quad (9)$$

- For  $\beta < 0$ : all moments for  $\beta \leq -1$  are infinite, and the moments of order  $-1 < \beta < 0$  satisfy

$$\mathbb{E} \left[ w_0^\beta(x) \right] \leq 1 + n \kappa_{|\beta|}(x) \kappa_\beta(x), \quad (10)$$

so that a sufficient condition for its existence is  $\kappa_\beta(x) = \int \rho(x+y) \|y\|^{|\beta|d} d^d y < \infty$ .

Heuristically, the behavior of these moments are consistent with the random variable  $W = w_0(x)$  having a probability distribution satisfying a scaling relation  $P(W) = \frac{1}{W_n} p\left(\frac{W}{W_n}\right)$ , with the scaling function  $p$  having the universal tail (i.e., independent of  $x$  and  $\rho$ ),  $p(w) \underset{w \rightarrow +\infty}{\sim} w^{-2}$ , and a scale  $W_n$  expected to vanish with  $n$ , when  $n \rightarrow +\infty$ . With this assumption, we can determine the scale  $W_n$  by imposing the exact condition  $\mathbb{E}[W] = 1/(n+1) \sim 1/n$ :

$$\mathbb{E}[W] = \frac{1}{W_n} \int_0^1 p\left(\frac{W}{W_n}\right) W dW = W_n \int_0^{\frac{1}{W_n}} p(w) w dw \quad (11)$$

$$\sim W_n \int_1^{\frac{1}{W_n}} \frac{dw}{w} \sim -W_n \ln(W_n) \sim \frac{1}{n}, \quad (12)$$

leading to  $W_n \sim \frac{1}{n \ln(n)}$ . Then, the moment of order  $\beta > 1$  is given by

$$\mathbb{E}[W^\beta] = \frac{1}{W_n} \int_0^1 p\left(\frac{W}{W_n}\right) W^\beta dW \sim W_n \int_0^1 W^{\beta-2} dW \underset{n \rightarrow +\infty}{\sim} \frac{1}{(\beta-1)n \ln(n)}, \quad (13)$$

which indeed coincides with the first result of Theorem 3.1. Our heuristic argument also suggests that in the case  $0 < \beta < 1$ , we have

$$\mathbb{E}[W] = \frac{1}{W_n} \int_0^1 p\left(\frac{W}{W_n}\right) W^\beta dW \underset{n \rightarrow +\infty}{\sim} \frac{\int_0^{+\infty} p(w) w^\beta dw}{(n \ln(n))^\beta}, \quad (14)$$

where the last integral converges since  $p(w) \underset{w \rightarrow +\infty}{\sim} w^{-2}$  and  $\beta < 1$ . This result is perfectly consistent with Eq. (9) in Theorem 3.1, and suggests that  $\int_0^{+\infty} p(w) w^\beta dw = \frac{\kappa_\beta(x)}{(V_d \rho(x))^\beta}$ . Interestingly, for  $0 < \beta < 1$ , and contrary to the case  $\beta > 1$ , we find that the large  $n$  equivalent of the moment is not universal and depends explicitly on  $x$  and the density  $\rho$ . As for moments of order  $-1 < \beta < 0$ , we conjecture that they are still given by Eq. (9) (and equivalently, by Eq. (14)) provided they exist, and that the sufficient condition for their existence  $\kappa_\beta(x) < \infty$  is hence also necessary, since  $\kappa_\beta(x)$  also appears in Eq. (9). The fact that moments for  $\beta \leq -1$  do not exist strongly suggests that  $p(0) > 0$ .

In fact, Eq. (14)) also suggests that all moments for  $-1 < \beta < 0$  exist if and only if  $0 < p(0) < \infty$ . In the Fig. 2 of the appendix, we present numerical simulations confirming our scaling ansatz, the fact that  $p(w) \underset{w \rightarrow +\infty}{\sim} w^{-2}$ , and the quantitative prediction for  $W_n$ .

It is shown in Devroye *et al.* [18] that the Hilbert kernel regression estimate does not converge almost surely (*a.s.*) by giving a specific example. Insight can be gained into this lack of almost sure convergence by considering the weight function  $w_0(x)$ , for a sequence of independent training sample sets of increasing size  $n + 1$ . Let the corresponding sequence of weights be denoted as  $\omega_n \in [0, 1]$ . From Theorem 3.1, it is clear that  $\omega_n$  converges to zero in probability, since the following Chebyshev bound holds (analogous to the bound on the regression risk):

$$\mathbb{P}(\omega_n > \varepsilon) \leq \frac{1 + \delta}{\varepsilon^2 n \ln(n)}, \quad (15)$$

for arbitrary  $\varepsilon > 0$  and  $\delta > 0$ , and for  $n$  larger than some constant  $N_{x,\delta}$ . Alternatively, one can exploit the fact that  $\mathbb{E}[\omega_n] = \frac{1}{n+1}$ , leading to  $\mathbb{P}(\omega_n > \varepsilon) \leq \frac{1}{\varepsilon n}$ , which is less stringent than Eq. (15) as far as the  $n$ -dependence is concerned, but is more stringent for the  $\varepsilon$ -dependence of the bounds.

Let us show heuristically that  $\omega_n$  does not converge *a.s.* to zero. Consider the infinite sequence of events  $\mathcal{E}_n \equiv \{\omega_n > \varepsilon\}$ ,  $n \in \mathbb{N}$ , and the corresponding infinite sum  $\sum_n \mathbb{P}(\mathcal{E}_n) = \sum_n \mathbb{P}(\omega_n > \varepsilon)$ . Exploiting our previous heuristic argument for the scaling form of the distribution of weights, we obtain

$$\mathbb{P}(\omega_n > \varepsilon) = \int_{\varepsilon}^1 \frac{1}{W_n} p\left(\frac{W}{W_n}\right) dW \sim \int_{\varepsilon n \ln n}^{n \ln n} \frac{dw}{w^2} \sim \frac{1 - \varepsilon}{\varepsilon n \ln(n)}. \quad (16)$$

Since  $\sum_{n=2}^N \frac{1}{n \ln(n)} \sim \ln(\ln(N))$  is a divergent series, a Borel-Cantelli argument suggests that an infinite number of the events  $\mathcal{E}_n$  (i.e.,  $\omega_n > \varepsilon$ ) must occur, which implies that  $\omega_n$  does not converge *a.s.* to 0. Note that the weights are equal to 1 at the data points due to the interpolation condition, so that large weights occasionally occur, causing the lack of *a.s.* convergence.

### 3.1.2 Lagrange function: scaling limit

The expected value of the Lagrange functions  $w_i(x)$  have a simple form in the large  $n$  limit. Due to the *i.i.d.* condition the indices  $i$  are exchangeable and we set  $i = 0$  for the computation of the expected Lagrange function  $L_0(x) = \mathbb{E}_{X|x_0}[w_0(x)]$ . Thus, one of the sample points (denoted  $x_0$ ) is held fixed and the other ones are averaged over in computing the expected Lagrange function. For  $x_0 \neq x$  kept fixed, we have  $\lim_{n \rightarrow \infty} L_0(x) = 0$ . However, we show in the appendix A.3 that  $L_0(x)$  takes a very simple form when taking a specific scaling limit:

**Theorem 3.2.** *For  $x \in \Omega^\circ$ , we assume  $\rho$  continuous at  $x$ . Then, in the limit (denoted by  $\lim_Z$ ),  $n \rightarrow +\infty$ ,  $\|x - x_0\|^{-d} \rightarrow +\infty$  (i.e.,  $x_0 \rightarrow x$ ), and such that  $z_x(n, x_0) = V_d \rho(x) \|x - x_0\|^d n \log(n) \rightarrow Z$ , the Lagrange function  $L_0(x) = \mathbb{E}_{X|x_0}[w_0(x)]$  converges to a proper limit,*

$$\lim_Z L_0(x) = \frac{1}{1 + Z}. \quad (17)$$

The proof of this theorem shows that the relative error between  $L_0(x)$  and  $\frac{1}{1+Z}$  for finite but large  $n$  and large  $\|x - x_0\|^{-d}$ , such that  $z_x(n, x_0)$  remains close to  $Z$ , is  $O(1/\ln(n))$ .

Exploiting Theorem 3.2, we can use a simple heuristic argument to estimate the tail of the distribution of the random variable  $W = w_0(x)$ . Indeed, approximating  $L_0(x)$  for finite but large  $n$  by its asymptotic form  $\frac{1}{1+z_x(n, x_0)}$ , with  $z_x(n, x_0) = V_d \rho(x) n \log(n) \|x - x_0\|^d$ , we obtain

$$\int_W^1 P(W') dW' \sim \int \rho(x_0) \theta\left(\frac{1}{1 + V_d \rho(x) n \log(n) \|x - x_0\|^d} - W\right) d^d x_0, \quad (18)$$

$$\sim V_d \rho(x) \int_0^{+\infty} \theta\left(\frac{1}{1 + V_d \rho(x) n \log(n) u} - W\right) du, \quad (19)$$

$$\sim \frac{1}{n \ln(n) W} \implies P(W) \sim \frac{1}{n \ln(n) W^2}, \quad (20)$$

where  $\theta(\cdot)$  is the Heaviside function. This heuristic result is again perfectly consistent with our guess of the previous section that  $P(W) = \frac{1}{W_n} p\left(\frac{W}{W_n}\right)$ , with the scaling function  $p$  having the universal

tail,  $p(w) \underset{w \rightarrow +\infty}{\sim} w^{-2}$ , and a scale  $W_n \sim \frac{1}{n \ln(n)}$ . Indeed, in this case and in the limit  $n \rightarrow +\infty$ , we obtain that  $P(W) \sim \frac{1}{W_n} \left( \frac{W_n}{W} \right)^2 \sim \frac{W_n}{W^2} \sim \frac{1}{n \ln(n) W^2}$ , which is identical to the result of Eq. (20).

### 3.1.3 The variance term

A simple application of the result of Theorem 3.1 for  $\beta = 2$  (see appendix A.4) allows us to bound the variance term  $\mathcal{V}(x) = \mathbb{E} \left[ \sum_{i=0}^n w_i^2(x) [y_i - f(x_i)]^2 \right]$  for a bounded variance function  $\sigma^2$ :

**Theorem 3.3.** *For  $x \in \Omega^\circ$ ,  $\rho$  continuous at  $x$ ,  $\sigma^2 \leq \sigma_0^2$ , and for any  $\varepsilon > 0$ , there exists a constant  $N_{x,\varepsilon}$  such that for  $n \geq N_{x,\varepsilon}$ , we have*

$$\mathcal{V}(x) \leq (1 + \varepsilon) \frac{\sigma_0^2}{\ln(n)}. \quad (21)$$

Relaxing the boundedness condition for  $\sigma$ , but assuming the continuity of  $\sigma^2$  at  $x$  along with a growth condition, allows us to obtain a precise asymptotic equivalent of  $\mathcal{V}(x)$ , when  $n \rightarrow +\infty$ :

**Theorem 3.4.** *For  $x \in \Omega^\circ$ ,  $\sigma(x) > 0$ ,  $\rho \sigma^2$  continuous at  $x$ , and assuming the condition  $C_{\text{Growth}}^\sigma$ , i.e.,  $\int \rho(y) \frac{\sigma^2(y)}{1 + \|y\|^{2d}} d^d y < \infty$ , we have*

$$\mathcal{V}(x) \underset{n \rightarrow +\infty}{\sim} \frac{\sigma^2(x)}{\ln(n)}. \quad (22)$$

Note that if the mean variance  $\int \rho(y) \sigma^2(y) d^d y < \infty$ , which is in particular the case when  $\sigma^2$  is bounded over  $\Omega$ , then the condition  $C_{\text{Growth}}^\sigma$  is in fact automatically satisfied.

### 3.1.4 The bias term

In appendix A.5, we prove the following three theorems for the bias term.

**Theorem 3.5.** *For  $x \in \Omega^\circ$  (so that  $\rho(x) > 0$ ), we assume that  $\rho$  is continuous at  $x$ , and the conditions*

- $C_{\text{Growth}}^f$ :  $\int \rho(y) \frac{f^2(y)}{1 + \|y\|^{2d}} d^d y < \infty$ ,
- $C_{\text{Holder}}^f$ : *there exist  $\alpha_x > 0$ ,  $K_x > 0$ , and  $\delta_x > 0$ , such that  $x' \in \Omega$  and  $\|x - x'\| \leq \delta_x \implies |f(x) - f(x')| \leq K_x \|x - x'\|^{\alpha_x}$  (local Hölder condition for  $f$ ).*

Moreover, we define  $\kappa(x) = \int \rho(x+y) \frac{f(x+y) - f(x)}{\|y\|^d} d^d y$ , where we have  $|\kappa(x)| < \infty$ .

Then, for  $\kappa(x) \neq 0$ , the bias term  $\mathcal{B}(x) = \mathbb{E}_X \left[ \left( \sum_{i=0}^n w_i(x) [f(x_i) - f(x)] \right)^2 \right]$  satisfies

$$\mathcal{B}(x) \underset{n \rightarrow +\infty}{\sim} \left( \mathbb{E} [\hat{f}(x)] - f(x) \right)^2, \quad \text{with} \quad \mathbb{E} [\hat{f}(x)] - f(x) \underset{n \rightarrow +\infty}{\sim} \frac{\kappa(x)}{V_d \rho(x) \ln(n)}. \quad (23)$$

In the non generic case  $\kappa(x) = 0$ , we have the weaker result

$$\mathcal{B}(x) = \begin{cases} O \left( n^{-\frac{2\alpha_x}{d}} (\ln(n))^{-1 - \frac{2\alpha_x}{d}} \right), & \text{for } d > 2\alpha_x \\ O \left( n^{-1} (\ln(n))^{-1} \right), & \text{for } d = 2\alpha_x \\ O \left( n^{-1} (\ln(n))^{-2} \right), & \text{for } d < 2\alpha_x \end{cases} \quad (24)$$

Note that  $\kappa(x) = 0$  is non generic but can still happen, even if  $f$  is not constant. For instance, if  $\Omega$  is a sphere centered at  $x$  or  $\Omega = \mathbb{R}^d$ , if  $\rho(x+y) = \hat{\rho}(\|y\|)$  is isotropic around  $x$ , and if  $f_x : y \mapsto f(x+y)$  is an odd function of  $y$ , then we indeed have  $\kappa(x) = 0$  at this symmetric point  $x$ .

Interestingly, for  $\kappa(x) \neq 0$ , Eq. (23) shows that the bias  $\mathcal{B}(x)$  is asymptotically dominated by the square of  $\mathbb{E}[\hat{f}(x)] - f(x)$ , showing that the fluctuations of  $\mathbb{E}[\hat{f}(x)] - \sum_{i=0}^n w_i(x)f(x_i)$  are negligible compared to  $\mathbb{E}[\hat{f}(x)] - f(x)$ , in the limit  $n \rightarrow +\infty$  and for  $\kappa(x) \neq 0$ .

One can relax the local Hölder condition, but at the price of a weaker estimate for  $\mathcal{B}(x)$  which will however be enough to obtain strong results for the regression and classification risks (see below):

**Theorem 3.6.** *For  $x \in \Omega^\circ$ , we assume  $\rho$  and  $f$  continuous at  $x$ , and the growth condition  $C_{\text{Growth}}^f$ :  $\int \rho(y) \frac{f^2(y)}{1+\|y\|^{2d}} d^d y < \infty$ . Then, the bias term satisfies*

$$\mathcal{B}(x) = o\left(\frac{1}{\ln(n)}\right), \quad (25)$$

or equivalently, for any  $\varepsilon > 0$ , there exists  $N_{x,\varepsilon}$ , such that for  $n \geq N_{x,\varepsilon}$

$$\mathcal{B}(x) \leq \frac{\varepsilon}{\ln(n)}. \quad (26)$$

Let us now consider a point  $x \in \partial\Omega$  for which we have  $\rho(x) = 0$  (note that  $x \in \partial\Omega$  does not necessarily imply  $\rho(x) = 0$ ). In appendix A.5, we show the following theorem for the expectation value of the estimator  $\hat{f}(x)$  in the limit  $n \rightarrow +\infty$ :

**Theorem 3.7.** *For  $x \in \partial\Omega$  such that  $\rho(x) = 0$ , we assume that  $f$  and  $\rho$  satisfy the conditions*

- $C_{\text{Growth}}^f$ :  $\int \rho(y) \frac{|f(y)|}{1+\|y\|^d} d^d y < \infty$ ,
- $C_{\text{Holder}}^\rho$ : *there exist  $\alpha_x > 0$ ,  $K_x > 0$ , and  $\delta_x > 0$ , such that  $x' \in \Omega$  and  $\|x - x'\| \leq \delta_x \implies |\rho(x')| \leq K_x \|x - x'\|^{\alpha_x}$  (local Hölder condition for  $\rho$ ).*

Moreover, we define  $\kappa(x) = \int \rho(x+y) \frac{f(x+y)-f(x)}{\|y\|^d} d^d y$  ( $|\kappa(x)| < \infty$  under condition  $C_{\text{Growth}}^f$ ), and  $\lambda(x) = \int \frac{\rho(x+y)}{\|y\|^d} d^d y$  ( $0 < \lambda(x) < \infty$  under condition  $C_{\text{Holder}}^\sigma$ ). Then,

$$\lim_{n \rightarrow +\infty} \mathbb{E}[\hat{f}(x)] - f(x) = \frac{\kappa(x)}{\lambda(x)}. \quad (27)$$

Hence, in the generic case  $\kappa(x) \neq 0$  (see Theorem 3.5 and the discussion below it) and under condition  $C_{\text{Holder}}^\rho$ , we find that the bias does not vanish when  $\rho(x) = 0$ , and that the estimator  $\hat{f}(x)$  does not converge to  $f(x)$ . When  $\rho(x) = 0$ , the scarcity of data near the point  $x$  indeed prevents the estimator to converge to the actual value of  $f(x)$ . In appendix A.5, we show an example of a density  $\rho$  continuous at  $x$  and such that  $\rho(x) = 0$ , but not satisfying the condition  $C_{\text{Holder}}^\rho$ , and for which  $\lim_{n \rightarrow +\infty} \mathbb{E}[\hat{f}(x)] = f(x)$ , even if  $\kappa(x) \neq 0$ .

### 3.2 Asymptotic equivalent for the regression risk

In appendix A.6, we prove the following theorem establishing the asymptotic rate at which the excess risk goes to zero with large sample size  $n$  for Hilbert kernel regression, under mild conditions that do not require  $f$  or  $\sigma$  to be bounded, but only to satisfy some growth conditions:

**Theorem 3.8.** *For  $x \in \Omega^\circ$ , we assume  $\sigma(x) > 0$ ,  $\rho$ ,  $\sigma$ , and  $f$  continuous at  $x$ , and the growth conditions  $C_{\text{Growth}}^\sigma$ :  $\int \rho(y) \frac{\sigma^2(y)}{1+\|y\|^{2d}} d^d y < \infty$  and  $C_{\text{Growth}}^f$ :  $\int \rho(y) \frac{f^2(y)}{1+\|y\|^{2d}} d^d y < \infty$ .*

Then the following statements are true:

- *The excess regression risk at the point  $x$  satisfies*

$$\mathbb{E}[(\hat{f}(x) - f(x))^2] \underset{n \rightarrow +\infty}{\sim} \frac{\sigma^2(x)}{\ln(n)}. \quad (28)$$



- *The Hilbert kernel estimate converges pointwise to the regression function in probability. More specifically, for any  $\delta > 0$ , there exists a constant  $N_{x,\delta}$ , such that for any  $\varepsilon > 0$ , we have the following Chebyshev bound, valid for  $n \geq N_{x,\delta}$*

$$\mathbb{P}[|\hat{f}(x) - f(x)| \geq \varepsilon] \leq \frac{1 + \delta}{\varepsilon^2} \frac{\sigma^2(x)}{\ln(n)}. \quad (29)$$

This theorem is a consequence of the corresponding asymptotically equivalent forms of the variance and bias terms presented above. Note that as long as  $\rho(x) > 0$ , the variance term dominates over the bias term and the regression risk has the same form as the variance term.

### 3.3 Rates for the plugin classifier

In appendix A.7, we prove the following theorem establishing the asymptotic rate at which the classification risk goes to zero with large sample size  $n$  for Hilbert kernel regression:

**Theorem 3.9.** *For  $x \in \Omega^\circ$ , we assume  $\sigma(x) > 0$ ,  $\rho$ ,  $\sigma$ , and  $f$  continuous at  $x$ . Then, the classification risk  $\mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] - \mathcal{R}_{0/1}(F(x))$  vanishes for  $n \rightarrow +\infty$ .*

*More precisely, for any  $\varepsilon > 0$ , there exists  $N_{x,\varepsilon}$ , such that for any  $n \geq N_{x,\varepsilon}$ ,*

$$0 \leq \mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] - \mathcal{R}_{0/1}(F(x)) \leq 2(1 + \varepsilon) \frac{\sigma(x)}{\sqrt{\ln(n)}}, \quad (30)$$

*In addition, for any  $0 < \alpha < 1$ , the general inequality*

$$\mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] - \mathcal{R}_{0/1}(F(x)) \leq 2|f(x) - 1/2|^{1-\alpha} \mathbb{E}[|\hat{f}(x) - f(x)|^2]^{\frac{\alpha}{2}}, \quad (31)$$

*holds unconditionally and, for  $n \geq N_{x,\varepsilon}$ , leads to*

$$0 \leq \mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] - \mathcal{R}_{0/1}(F(x)) \leq 2|f(x) - 1/2|^{1-\alpha} (1 + \varepsilon)^\alpha \frac{\sigma^\alpha(x)}{(\ln(n))^{\frac{\alpha}{2}}}. \quad (32)$$

For  $0 < \alpha < 1$ , Eq. (32) is weaker than Eq. (30) in terms of its dependence on  $n$ , but explicitly shows that the classification risk vanishes for  $f(x) = 1/2$ . This theorem does not require any growth condition for  $f$  or  $\sigma$ , since both functions takes values in  $[0, 1]$  in the classification context.

### 3.4 Extrapolation behavior outside the support of $\rho$

We now take the point  $x$  outside the closed support  $\bar{\Omega}$  of the distribution  $\rho$  (which excludes the case  $\Omega = \mathbb{R}^d$ ). We are interested in the behavior of  $\mathbb{E}[\hat{f}(x)]$  as  $n \rightarrow +\infty$ . In appendix A.8 we prove:

**Theorem 3.10.** *For  $x \notin \bar{\Omega}$ , we assume the growth condition  $\int \rho(y) \frac{|f(y)|}{1+\|y\|^d} d^d y < \infty$ . Then,*

$$\hat{f}_\infty(x) := \lim_{n \rightarrow +\infty} \mathbb{E}[\hat{f}(x)] = \frac{\int \rho(y) f(y) \|x - y\|^{-d} d^d y}{\int \rho(y) \|x - y\|^{-d} d^d y}, \quad (33)$$

*and  $\hat{f}_\infty$  is continuous at all  $x \notin \bar{\Omega}$ .*

*In addition, if  $\int \rho(y) |f(y)| d^d y < \infty$ , and defining  $d(x, \Omega) > 0$  as the distance between  $x$  and  $\Omega$ , we have*

$$\lim_{d(x,\Omega) \rightarrow +\infty} \hat{f}_\infty(x) = \int \rho(y) f(y) d^d y. \quad (34)$$

*Finally, we consider  $x_0 \in \partial\Omega$  such that  $\rho(x_0) > 0$  (i.e.,  $x_0 \in \partial\Omega \cap \Omega$ ), and assume that  $f$  and  $\rho$  seen as functions restricted to  $\Omega$  are continuous at  $x_0$ , i.e.  $\lim_{y \in \Omega \rightarrow x_0} \rho(y) = \rho(x_0)$  and  $\lim_{y \in \Omega \rightarrow x_0} f(y) = f(x_0)$ . We also assume that the local solid angle  $\omega_0 = \lim_{r \rightarrow 0} \frac{1}{V_d \rho(x_0) r^d} \int_{\|x_0 - y\| \leq r} \rho(y) d^d y$  exists and satisfies  $\omega_0 > 0$ . Then,*

$$\lim_{x \notin \bar{\Omega} \rightarrow x_0} \hat{f}_\infty(x) = f(x_0). \quad (35)$$

Eq. (34) shows that far away from  $\Omega$  (which is possible to realize, for instance, when  $\Omega$  is bounded),  $\hat{f}_\infty(x)$  goes smoothly to the  $\rho$ -mean of  $f$ . Moreover, Eq. (35) establishes a continuity property for the extrapolation  $\hat{f}_\infty$  at  $x_0 \in \partial\Omega \cap \Omega$  under the stated conditions (remember that for  $x \in \Omega^\circ$ , we have  $\lim_{n \rightarrow +\infty} \mathbb{E}[\hat{f}(x)] = f(x)$ ; see Theorem 3.5, and in particular Eq. (23)).

## Acknowledgements

PPM gratefully acknowledges the support of the Crick-Clay Professorship (CSHL) and the H N Mahabala Chair (IITM).

## A Proofs

### A.1 Preliminaries

In the following,  $x \in \Omega^\circ$  so that  $\rho(x) > 0$ , and we will assume for simplicity that the distribution  $\rho$  is continuous at  $x$ .

For the proof of our results, we will often exploit the following integral relation, valid for  $\beta > 0$ ,

$$\frac{1}{\Gamma(\beta)} \int_0^{+\infty} t^{\beta-1} e^{-tz} dt = z^{-\beta}. \quad (36)$$

In addition, we define

$$\psi(x, t) := \int \rho(x+y) e^{-\frac{t}{\|y\|^d}} d^d y, \quad (37)$$

which will play a central role. We note that  $\psi(x, 0) = 1$ , and that  $t \mapsto \psi(x, t)$  is a continuous and strictly decreasing function of  $t$ . It is even infinitely differentiable at any  $t > 0$ , but not necessarily at  $t = 0$ . In fact, for a fixed  $x$ , controlling the behavior of  $1 - \psi(x, t)$  when  $t \rightarrow 0$  will be essential to obtain our results.

We show in Fig. 1 an example of the Hilbert kernel regression estimator in one dimension. Both the bias and the variance of the estimator can be visually seen, as well as the extrapolation behavior outside the data domain. Note that in higher dimensions, the sharp peaks would have rounded tops.

### A.2 Moments of the weights: large $n$ behavior

In this section, we provide a complete proof of Theorem 3.1. Several other theorems will use the very same method of proof and some basic steps will not be repeated in their proof.

Using Eq. (36) for  $\beta > 0$ , we can express powers of the weight function as

$$w_0^\beta(x) = \frac{1}{\|x - x_0\|^{\beta d}} \frac{1}{\Gamma(\beta)} \int_0^{+\infty} t^{\beta-1} e^{-t\|x-x_0\|^{-d-t} \sum_{i=1}^n \|x-x_i\|^{-d}} dt. \quad (38)$$

By taking the expected value over the  $n+1$  independent random variables  $X_i$ , we obtain

$$\mathbb{E} [w_0^\beta(x)] = \frac{1}{\Gamma(\beta)} \int_0^{+\infty} t^{\beta-1} \psi^n(x, t) \phi_\beta(x, t) dt, \quad (39)$$

with

$$\phi_\beta(x, t) := \int \rho(x+y) \frac{e^{-\frac{t}{\|y\|^d}}}{\|y\|^{\beta d}} d^d y, \quad (40)$$

which is also a strictly decreasing function of  $t$ , continuous at any  $t > 0$  (in fact, infinitely differentiable for  $t > 0$ ).

Note that the exchange of the integral over  $t$  and over  $\vec{x} = (x_0, x_1, \dots, x_n)$  used to obtain Eq. (39) is justified by the Fubini theorem, by first noting that the function  $\vec{x} \mapsto w_0^\beta(x) \prod_{i=0}^n \rho(x_i)$  is in  $L^1(\mathbb{R}^d)$ , since  $0 \leq w_0^\beta(x) \leq 1$ , and since  $\rho$  is obviously in  $L^1(\mathbb{R}^d)$ . Moreover, the function  $t \mapsto t^{\beta-1} \psi^n(x, t) \phi_\beta(x, t) > 0$  is also in  $L^1(\mathbb{R})$ . Indeed, we will show below that it decays fast enough when  $t \rightarrow +\infty$  (see Eqs. (42-50)), ensuring the convergence of its integral at  $+\infty$ , and that it is bounded (and continuous) near  $t = 0$  (see Eqs. (63-68)), ensuring that this function is integrable at  $t = 0$ .

For  $\beta = 1$ ,  $\phi_1 = -\partial_t \psi$ , and we obtain  $\mathbb{E} [w_0(x)] = \frac{1}{n+1}$ , as expected. In the following, we first focus on the case  $\beta > 1$ , before addressing the cases  $0 < \beta < 1$  and  $\beta < 0$  at the very end of this section.

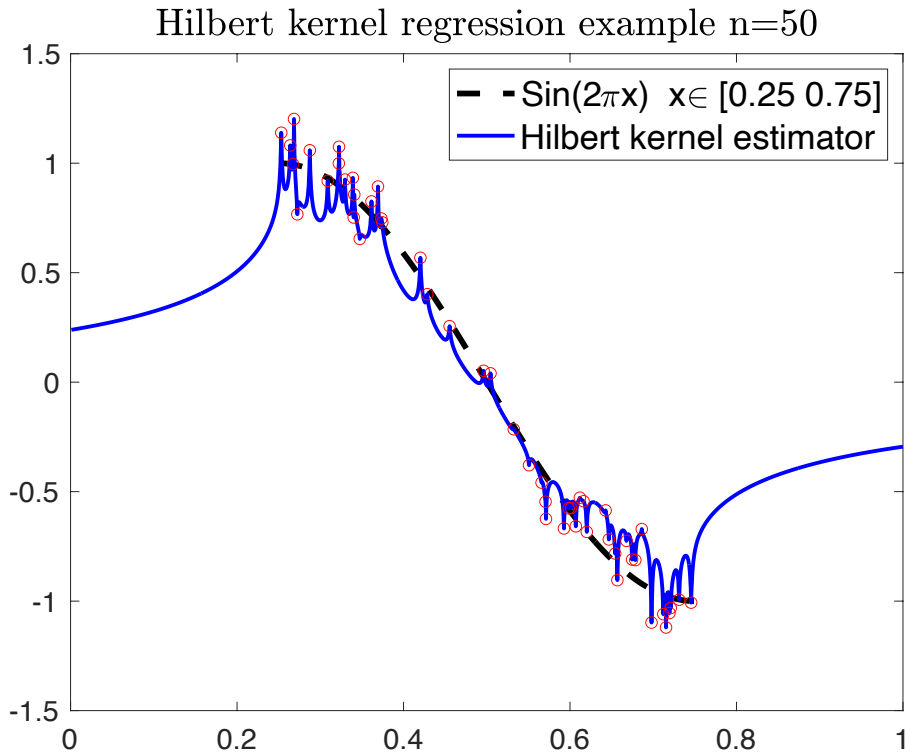


Figure 1: An example is shown of the Hilbert kernel regression estimator in one dimension, both within and outside the input data domain. A total of 50 samples  $x_i$  were chosen uniformly distributed in the interval  $[0.25 \ 0.75]$  and  $y_i = \sin(2\pi x_i) + n_i$  with the noise  $n_i$  chosen *i.i.d.* Gaussian distributed  $\sim N(0, 0.1)$ . The sample points are circled, and the function  $\sin(2\pi x)$  is shown with a dashed line within the data domain. The solid line is the Hilbert kernel regression estimator. Note the interpolation behavior within the data domain and the extrapolation behavior outside the data domain.

We now introduce  $t_1$  and  $t_2$  (to be further constrained later) such that  $0 < t_1 < t_2$ . We then express the integral of Eq. (39) as the sum of corresponding integrals  $I_1 + I_{12} + I_2$ .  $I_1$  is the integral between 0 and  $t_1$ ,  $I_{12}$  the integral between  $t_1$  and  $t_2$ , and  $I_2$  the integral between  $t_2$  and  $+\infty$ . Thus, we have

$$I_1 \leq \mathbb{E} \left[ w_0^\beta(x) \right] \leq I_1 + I_{12} + I_2, \quad (41)$$

provided these integral exists, which we will show below, by providing upper bounds for  $I_2$  and  $I_{12}$ , and tight lower and upper bound for the leading term  $I_1$ .

*Bound for  $I_2$*

For any  $R \geq 1$ , we can write the integral defining  $\psi(x, t)$

$$\psi(x, t) = \int_{\|y\| \leq R} + \int_{\|y\| \geq R} \quad (42)$$

$$\leq e^{-\frac{t}{R^d}} + \int_{\|y\| \geq R} \rho(x+y) \frac{\|y\|^2}{R^2} d^d y, \quad (43)$$

$$\leq e^{-\frac{t}{R^d}} + \frac{C_x}{R^2}, \quad (44)$$

with  $C_x = \sigma_\rho^2 + \|x - \mu_\rho\|^2$  depending on the mean  $\mu_\rho$  and variance  $\sigma_\rho^2$  of the distribution  $\rho$ . Similarly, for  $\phi_\beta(x, t)$ , we obtain the bound

$$\phi_\beta(x, t) \leq \frac{1}{R^{\beta d}} e^{-\frac{t}{R^d}} + \frac{C_x}{R^{2+\beta d}}, \quad (45)$$

valid for  $t \geq \max(1, \beta)$  and  $R \leq r_t$ , where  $r_t = (t/\beta)^{1/d} \geq 1$  is the location of the maximum of the function  $r \mapsto \frac{e^{-\frac{t}{r^d}}}{r^{\beta d}}$ .

We now set  $R = t^{\frac{s}{d}}$ , with  $0 < s < 1$ , and take  $T_2' \geq \max(1, \beta, \beta^{1/(1-s)})$  (so that  $1 \leq R \leq r_t$ ) is large enough such that the following conditions are satisfied for  $t \geq t_2 \geq T_2'$ ,

$$e^{-\frac{t}{R^d}} = e^{-t^{1-s}} \leq \frac{C_x}{t^{\frac{2s}{d}}}, \quad (46)$$

$$\frac{1}{R^{\beta d}} e^{-\frac{t}{R^d}} = \frac{1}{t^{\beta s}} e^{-t^{1-s}} \leq \frac{C_x}{t^{\frac{2s}{d} + 2\beta s}}. \quad (47)$$

Hence, for  $t \geq t_2 \geq T_2'$ , we obtain

$$\psi(x, t) \leq \frac{2C_x}{t^{\frac{2s}{d}}}, \quad (48)$$

$$\phi_\beta(x, t) \leq \frac{2C_x}{t^{\frac{2s}{d} + 2\beta s}}. \quad (49)$$

In addition, we also impose  $t_2 \geq T_2'' = (4C_x)^{d/(2s)}$ , so that  $\frac{2C_x}{t^{\frac{2s}{d}}} \leq \frac{1}{2}$ , for any  $t \geq T_2 = \max(T_2', T_2'')$ . Finally, exploiting the resulting bounds for  $\psi(x, t)$  and  $\phi_\beta(x, t)$  for  $s = 1/2$ , we obtain the convergence of  $I_2$  (which, along with the bounds for  $I_1$  and  $I_{12}$  below, justifies our use of Fubini theorem to obtain Eq. (39)) and the exact bound

$$I_2 = \frac{1}{\Gamma(\beta)} \int_{t_2}^{+\infty} t^{\beta-1} \psi^n(x, t) \phi_\beta(x, t) dt \leq \frac{d}{\Gamma(\beta)} \times \frac{1}{2^{n+1}(n+1)}, \quad (50)$$

for any given  $t_2 \geq T_2$ .

*Bound for  $I_{12}$*

Again, exploiting the fact that  $\psi(x, t)$  and  $\phi_\beta(x, t)$  are strictly decreasing functions of  $t$ , we obtain

$$I_{12} \leq \frac{\phi_\beta(x, t_1) t_2^\beta}{\Gamma(\beta)} \times \psi^n(x, t_1), \quad (51)$$

where we note that  $\psi(x, t_1) < 1$ , for any  $t_1 > 0$ .

*Bound for  $I_1$*

We first want to obtain bounds for  $1 - \psi(x, t)$ , where  $0 \leq t \leq t_1$ , with  $t_1 > 0$  to be constrained below. In addition, exploiting the continuity of  $\rho$  at  $x$  and the fact that  $\rho(x) > 0$ , we introduce  $\varepsilon$  satisfying  $0 < \varepsilon < 1/4$ , and define  $\lambda > 0$  small enough so that the ball  $B(x, \delta) \subset \Omega^\circ$ , and  $\|y\| \leq \lambda \implies |\rho(x+y) - \rho(x)| \leq \varepsilon \rho(x)$ . Exploiting this definition, we obtain the following lower

and upper bounds

$$1 - \psi(x, t) \geq (1 - \varepsilon)\rho(x) \int_{\|y\| \leq \lambda} \left(1 - e^{-\frac{t}{\|y\|^d}}\right) d^d y, \quad (52)$$

$$1 - \psi(x, t) \leq (1 + \varepsilon)\rho(x) \int_{\|y\| \leq \lambda} \left(1 - e^{-\frac{t}{\|y\|^d}}\right) d^d y \quad (53)$$

$$+ \int_{\|y\| \geq \lambda} \rho(x + y) \left(1 - e^{-\frac{t}{\lambda^d}}\right) d^d y, \quad (54)$$

$$\leq (1 + \varepsilon)\rho(x) \int_{\|y\| \leq \lambda} \left(1 - e^{-\frac{t}{\|y\|^d}}\right) d^d y + \frac{t}{\lambda^d}. \quad (55)$$

The integral appearing in these bounds can be simplified by using radial coordinates:

$$\int_{\|y\| \leq \lambda} \left(1 - e^{-\frac{t}{\|y\|^d}}\right) d^d y, = S_d \int_0^\lambda \left(1 - e^{-\frac{t}{r^d}}\right) r^{d-1} dr, \quad (56)$$

$$= V_d t \int_{\frac{t}{\lambda^d}}^{+\infty} \frac{1 - e^{-u}}{u^2} du, \quad (57)$$

where  $S_d$  and  $V_d = \frac{S_d}{d}$  are respectively the surface and the volume of the  $d$ -dimensional unit sphere and we have used the change of variable  $u = \frac{t}{r^d}$ .

We note that for  $0 < z \leq 1$ , we have

$$\int_z^{+\infty} \frac{1 - e^{-u}}{u^2} du = -\ln(z) + \int_z^1 \frac{1 - u - e^{-u}}{u^2} du + \int_1^{+\infty} \frac{1 - e^{-u}}{u^2} du. \quad (58)$$

Exploiting this result and now imposing  $t_1 \leq \lambda^d$ , we have, for any  $t \leq t_1$

$$\ln\left(\frac{C_-}{t}\right) \leq \int_{\frac{t}{\lambda^d}}^{+\infty} \frac{1 - e^{-u}}{u^2} du \leq \ln\left(\frac{C_+}{t}\right), \quad (59)$$

$$\ln(C_-) = d \ln(\lambda) + \int_1^{+\infty} \frac{1 - e^{-u}}{u^2} du, \quad (60)$$

$$\ln(C_+) = \ln(C_-) + \int_0^1 \frac{1 - u - e^{-u}}{u^2} du. \quad (61)$$

Combining these bounds with Eq. (52) and Eq. (55), we have shown the existence of two  $x$ -dependent constants  $D_\pm$  such that, for  $0 \leq t \leq t_1 \leq \lambda^d$ , we have

$$(1 - \varepsilon)V_d \rho(x) t \ln\left(\frac{D_-}{t}\right) \leq 1 - \psi(x, t) \leq (1 + \varepsilon)V_d \rho(x) t \ln\left(\frac{D_+}{t}\right). \quad (62)$$

In addition, we will also chose  $t_1 < D_\pm/3$ , such that the two functions  $t \ln\left(\frac{D_\pm}{t}\right)$  are positive and strictly increasing for  $0 \leq t \leq t_1$ .  $t_1$  is also taken small enough such that the two bounds in Eq. (62) are always less than 1/2, for  $0 \leq t \leq t_1$  (both bounds vanish when  $t \rightarrow 0$ ).

We now obtain efficient bounds for  $\phi_\beta(x, t)$ , for  $0 \leq t \leq t_1$ . Proceeding in a similar manner as above, we obtain

$$\phi_\beta(x, t) \geq (1 - \varepsilon)\rho(x) \int_{\|y\| \leq \lambda} \frac{e^{-\frac{t}{\|y\|^d}}}{\|y\|^{\beta d}} d^d y, \quad (63)$$

$$\phi_\beta(x, t) \leq (1 + \varepsilon)\rho(x) \int_{\|y\| \leq \lambda} \frac{e^{-\frac{t}{\|y\|^d}}}{\|y\|^{\beta d}} d^d y + \frac{1}{\lambda^{\beta d}}. \quad (64)$$

Again, the integral appearing in these bounds can be rewritten as

$$\int_{\|y\| \leq \lambda} \frac{e^{-\frac{t}{\|y\|^d}}}{\|y\|^{\beta d}} d^d y = S_d \int_0^\lambda r^{d(1-\beta)-1} e^{-\frac{t}{r^d}} dr. \quad (65)$$

For  $0 < \beta < 1$ , the integral of Eq. (65) is finite for  $t = 0$ , ensuring the existence of  $\phi_\beta(x, 0)$  and the fact that  $t \mapsto t^{\beta-1}\psi(x, t)\phi_\beta(x, t)$  belongs to  $L^1(\mathbb{R})$  (hence, justifying our use of Fubini theorem for  $0 < \beta < 1$ ). For  $\beta > 1$ , we have

$$\int_{\|y\| \leq \lambda} \frac{e^{-\frac{t}{\|y\|^d}}}{\|y\|^{\beta d}} = V_d t^{1-\beta} \int_{\frac{t}{\lambda^d}}^{+\infty} u^{\beta-2} e^{-u} du. \quad (66)$$

$$\sim_{t \rightarrow 0} V_d \Gamma(\beta - 1) t^{1-\beta}. \quad (67)$$

This integral diverges when  $t \rightarrow 0$  and the constant term  $\lambda^{-\beta d}$  in Eq. (64) can be made as small as necessary (by a factor less than  $\varepsilon$ ) compared to this leading integral term, for a small enough  $t_1$ . Similarly, we can choose  $t_1$  small enough so that the integral Eq. (65) is approached by the asymptotic result of Eq. (67) up to a factor  $\varepsilon$ . Thus, we find that for  $0 \leq t \leq t_1$ , one has

$$(1 - 2\varepsilon)V_d \rho(x) \Gamma(\beta - 1) t^{1-\beta} \leq \phi_\beta(x, t) \leq (1 + 3\varepsilon)V_d \rho(x) \Gamma(\beta - 1) t^{1-\beta}. \quad (68)$$

This shows that  $t^{\beta-1}\phi_\beta(x, t)$  has a smooth limit when  $t \rightarrow 0$  so that, combined with the finite upper bound for  $I_2$ ,  $t \mapsto t^{\beta-1}\psi(x, t)\phi_\beta(x, t)$  belongs to  $L^1(\mathbb{R})$ , for  $\beta > 1$ , and hence for all  $\beta > 0$ . Hence, the use of the Fubini theorem to derive Eq. (39) has been justified.

Now combining the bounds for  $\psi(x, t)$  and  $\phi_\beta(x, t)$ , we obtain

$$I_1 \geq (1 - 2\varepsilon) \frac{1}{\beta - 1} V_d \rho(x) \int_0^{t_1} \left( 1 - (1 + \varepsilon)V_d \rho(x) t \ln \left( \frac{D_+}{t} \right) \right)^n dt, \quad (69)$$

$$I_1 \leq (1 + 3\varepsilon) \frac{1}{\beta - 1} V_d \rho(x) \int_0^{t_1} \left( 1 - (1 - \varepsilon)V_d \rho(x) t \ln \left( \frac{D_-}{t} \right) \right)^n dt. \quad (70)$$

*Asymptotic behavior of  $I_1$  and  $\mathbb{E} [w_0^\beta(x)]$*

We will show below that

$$\int_0^{t_1} \left( 1 - E_\pm t \ln \left( \frac{D_\pm}{t} \right) \right)^n dt \underset{n \rightarrow +\infty}{\sim} \frac{1}{E_\pm n \ln(n)}, \quad (71)$$

where  $E_\pm = (1 \mp \varepsilon)V_d \rho(x)$ . For a given  $x$ , and for  $t_1$  and  $t_2$  satisfying the requirements mentioned above, the upper bounds for  $I_{12}$  (see Eq. (51)) and  $I_2$  (see Eq. (50)) appearing in Eq. (41) both decay exponentially with  $n$  and can hence be made arbitrarily small compared to  $I_1$  which decays as  $1/(n \ln(n))$ .

Finally, assuming for now the result of Eq. (71) (to be proven below), we have obtained the exact asymptotic result

$$\mathbb{E} [w_0^\beta(x)] \underset{n \rightarrow +\infty}{\sim} \frac{1}{(\beta - 1)n \ln(n)}. \quad (72)$$

*Proof of Eq. (71)*

We are then left to prove the result of Eq. (71). First, we will use the fact that, for  $0 \leq z \leq z_1 < 1$ , one has

$$e^{-\mu z} \leq 1 - z \leq e^{-z}, \quad (73)$$

where  $\mu = -\ln(1 - z_1)/z_1$ . We can apply this result to the integral of Eq. (71), using  $z_1^\pm = E_\pm t_1 \ln(D_\pm/t_1) > 0$ . Note that  $0 < t_1 < D_\pm/3$  and hence  $z_1^\pm > 0$  can be made as close to 0 as desired, and the corresponding  $\mu_\pm > 1$  can be made as close to 1 as desired. Thus, in order to prove Eq. (71), we need to prove the following equivalent

$$I_n = \int_0^{t_1} e^{-nEt \ln(\frac{D}{t})} dt \underset{n \rightarrow +\infty}{\sim} \frac{1}{En \ln(n)}, \quad (74)$$

for an integral of the form appearing in Eq. (74). Let us mention again that  $t_1$  has been taken small enough, so that the function  $t \mapsto t \ln \left( \frac{D}{t} \right)$  is positive and strictly increasing (with its maximum at  $t_{\max} = D/e < t_1$ ), for  $0 \leq t \leq t_1$ .

We now take  $n$  large enough so that  $\frac{\ln(n)}{n} < t_1$  and  $E \ln(n) > 1$ . One can then write

$$I_n = \frac{1}{n} \int_0^{\ln(n)} e^{-Eu \ln\left(\frac{Dn}{u}\right)} du + \int_{\frac{\ln(n)}{n}}^{t_1} e^{-nEt \ln\left(\frac{D}{t}\right)} dt = J_n + K_n, \quad (75)$$

$$J_n \leq \frac{1}{n} \int_0^{1/E} e^{-Eu \ln(DEn)} du + \frac{1}{n} \int_{1/E}^{\ln(n)} e^{-Eu \ln\left(\frac{Dn}{\ln(n)}\right)} du, \quad (76)$$

$$\leq \frac{1}{E n \ln(DEn)} + \frac{\ln(n)}{D E n^2 \ln\left(\frac{Dn}{\ln(n)}\right)}, \quad (77)$$

$$K_n \leq \int_{\frac{\ln(n)}{n}}^{+\infty} e^{-nEt \ln\left(\frac{D}{t}\right)} dt \leq \frac{1}{E n^{1+E \ln\left(\frac{D}{t_1}\right)} \ln\left(\frac{D}{t_1}\right)}. \quad (78)$$

When  $n \rightarrow +\infty$ , we hence find that the upper bound  $I_n^+$  of  $I_n$  satisfies

$$I_n^+ \underset{n \rightarrow +\infty}{\sim} \frac{1}{E n \ln(DEn)} \underset{n \rightarrow +\infty}{\sim} \frac{1}{E n \ln(n)}. \quad (79)$$

Let us now prove a similar result for a lower bound of  $I_n$  by considering  $n$  large enough so that  $nEt_1 > 1$ , and by introducing  $\delta$  satisfying  $0 \leq \delta < 1/e$ :

$$I_n = \frac{1}{nE} \int_0^{nEt_1} e^{-u \ln(DEn) + u \ln(u)} du, \quad (80)$$

$$\geq \frac{1}{nE} \int_0^\delta e^{-u \ln(DEn) + \delta \ln(\delta)} du, \quad (81)$$

$$\geq \frac{e^{\delta \ln(\delta)}}{nE \ln(DEn)} \left(1 - (DEn)^{-\delta}\right) = I_n^-(\delta). \quad (82)$$

Hence, for any  $0 \leq \delta < 1/e$  which can be made arbitrarily small, and for  $n$  large enough, we find that  $I_n \geq I_n^-(\delta)$ , with

$$I_n^-(\delta) \sim \frac{e^{\delta \ln(\delta)}}{E n \ln(DEn)} \sim \frac{e^{\delta \ln(\delta)}}{E n \ln(n)}. \quad (83)$$

Eq. (83) combined with the corresponding result of Eq. (79) for the upper bound  $I_n^+$  finally proves Eq. (74), and ultimately, Eq. (72) and Theorem 3.1 for the asymptotic behavior of the moment  $\mathbb{E} \left[ w_0^\beta(x) \right]$ , for  $\beta > 1$ .

*Moments of order  $0 < \beta < 1$*

The integral representation Eq. (36) allows us to also explore moments of order  $0 < \beta < 1$ . In that case  $\kappa_\beta(x) = \phi_\beta(x, 0) < \infty$  is finite, with

$$\kappa_\beta(x) = \int \frac{\rho(x+y)}{\|y\|^{\beta d}} d^d y. \quad (84)$$

By retracing the different steps of our proof in the case  $\beta > 1$ , it is straightforward to show that

$$\mathbb{E} \left[ w_0^\beta(x) \right] \underset{n \rightarrow +\infty}{\sim} \frac{\kappa_\beta(x)}{\Gamma(\beta)} \int_0^{t_1} t^{\beta-1} e^{-nV_d \rho(x)t \ln\left(\frac{D \pm}{t}\right)} dt, \quad (85)$$

$$\underset{n \rightarrow +\infty}{\sim} \frac{\kappa_\beta(x)}{(V_d \rho(x) n \ln(n))^\beta}, \quad (86)$$

where the equivalent for the integral can be obtained by exploiting the very same method used in our proof of Eq. (71) above, hence proving the second part of Theorem 3.1.

We observe that contrary to the universal result of Eq. (72) for  $\beta$ , the asymptotic equivalent for the moment of order  $0 < \beta < 1$  is non universal and explicitly depends on  $x$  and the distribution  $\rho$ .

*Moments of order  $\beta < 0$*

Finally, moments of order  $\beta < 0$  are unfortunately inaccessible to our methods relying on the integral relation Eq. (36), which imposes  $\beta > 0$ . We can however obtain a few rigorous results for these moments (see also the heuristic discussion just after Theorem 3.1).

Indeed, for  $\beta = -1$ , we have

$$\frac{1}{w_0(x)} = 1 + \|x - x_0\|^d \sum_{i=1}^n \frac{1}{\|x - x_i\|^d}. \quad (87)$$

But since we have assumed that  $\rho(x) > 0$ ,  $\mathbb{E}[\|x - x_i\|^{-d}] = \int \frac{\rho(x+y)}{\|y\|^d} d^d y$  is infinite and moments of order  $\beta < -1$  are definitely not defined.

As for the moment of order  $-1 < \beta < 0$ , it can be easily bounded,

$$\mathbb{E} \left[ w_0^\beta(x) \right] \leq 1 + n \int \rho(x+y) \|y\|^{|\beta|d} d^d y \int \frac{\rho(x+y)}{\|y\|^{|\beta|d}} d^d y, \quad (88)$$

and a sufficient condition for its existence is  $\kappa_\beta(x) = \int \rho(x+y) \|y\|^{|\beta|d} d^d y < \infty$  (the other integral, equal to  $\kappa_{|\beta|}(x)$ , is always finite for  $|\beta| < 1$ ), which proves the last part of Theorem 3.1.

*Numerical distribution of the weights*

In the main text below Theorem 3.1, we presented an heuristic argument showing that the results of Theorem 3.1 and Theorem 3.2 (for the Lagrange function; that we prove below) were fully consistent with the weight  $W = w_0(x)$  having a long-tailed scaling distribution,

$$P_n(W) = \frac{1}{W_n} p \left( \frac{W}{W_n} \right). \quad (89)$$

The scaling function  $p$  was shown to have a universal tail  $p(w) \sim w^{-2}$  and the scale  $W_n$  was shown to obey the equation  $-W_n \ln(W_n) = n^{-1}$ . To the leading order for large  $n$ , we have  $W_n \sim \frac{1}{n \ln(n)}$ , and we can solve this equation recursively to find the next order approximation,  $W_n \sim \frac{1}{n \ln(n \ln(n))}$ . In Fig. 2, we present numerical simulations for the scaling distribution  $p$  of the variable  $w = W/W_n$ , for  $n = 65536$ , using the estimate  $W_n \approx \frac{1}{n \ln(n \ln(n))}$ . We observe that  $p(w)$  is very well approximated by the function  $\hat{p}(w) = \frac{1}{(1+w)^2}$ , confirming our non rigorous results. The data were generated by drawing random values of  $r_i^d = \|x - x_i\|^d$  using  $(n+1)$  *i.i.d.* random variables  $a_i$  uniformly distributed in  $[0, 1[$ , with the relation  $r_i = [a_i/(1-a_i)]^{1/d}$ , and by computing the resulting weight  $W = r_i^{-d} / \sum_{j=0}^n r_j^{-d}$ . This corresponds to a distribution of  $\|x - x_i\|$  given by  $\rho(x - x_i) = 1/V_d / (1 + \|x - x_i\|^d)^2$ .

### A.3 Lagrange function: scaling limit

In this section, we prove Theorem 3.2 for the scaling limit of the Lagrange function  $L_0(x) = \mathbb{E}_{X|x_0}[w_0(x)]$ . Exploiting again Eq. (36), the expected Lagrange function can be written as

$$L_0(x) = \|x - x_0\|^{-d} \int_0^{+\infty} \psi^n(x, t) e^{-t\|x-x_0\|^{-d}} dt, \quad (90)$$

where  $\psi(x, t)$  is again given by Eq. (37).

For a given  $t_1 > 0$ , and remembering that  $\psi(x, t)$  is a strictly decreasing function of  $t$ , with  $\psi(x, 0) = 1$ , we obtain

$$L_1 \leq L_0(x) \leq L_1 + L_2, \quad (91)$$

with

$$L_1 = \|x - x_0\|^{-d} \int_0^{t_1} \psi^n(x, t) e^{-t\|x-x_0\|^{-d}} dt, \quad (92)$$

$$L_2 = e^{-t_1\|x-x_0\|^{-d}}. \quad (93)$$



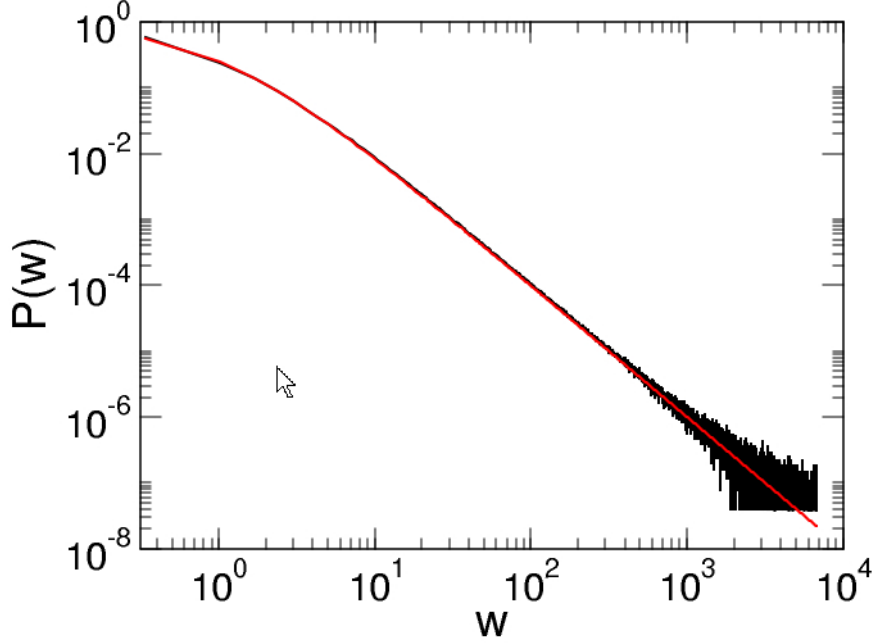


Figure 2: We plot the results of numerical simulations for the distribution  $p$  of the scaling variable  $w = \frac{W}{W_n}$ , with  $W_n \approx \frac{1}{n \ln(n \ln(n))}$ , and for  $n = 65536$  (black line). This is compared to  $\hat{p}(w) = \frac{1}{(1+w)^2}$  (red line), which has the predicted universal tail  $p(w) \sim w^{-2}$  for large  $w$ .

For  $\varepsilon > 0$  and a sufficiently small  $t_1 > 0$  (see section A.2), we can use the bound for  $\psi(x, t)$  obtained in section A.2, to obtain

$$L_1 \geq (1 - 2\varepsilon) \frac{1}{\|x - x_0\|^d} \int_0^{t_1} \left( 1 - (1 + \varepsilon) V_d \rho(x) t \ln \left( \frac{D_+}{t} \right) \right)^n e^{-\frac{t}{\|x - x_0\|^d}} dt, \quad (94)$$

$$L_1 \leq (1 + 3\varepsilon) \frac{1}{\|x - x_0\|^d} \int_0^{t_1} \left( 1 - (1 - \varepsilon) V_d \rho(x) t \ln \left( \frac{D_-}{t} \right) \right)^n e^{-\frac{t}{\|x - x_0\|^d}} dt. \quad (95)$$

Then, proceeding exactly as in section A.2, it is straightforward to show that  $L_1$  can be bounded (up to factors  $1 + O(\varepsilon)$ ) by the two integrals  $L_1^\pm$

$$L_1^\pm = \frac{1}{\|x - x_0\|^d} \int_0^{t_1} e^{-n V_d \rho(x) t \ln \left( \frac{D_\pm}{t} \right) - \frac{t}{\|x - x_0\|^d}} dt. \quad (96)$$

Like in section A.2, we impose  $t_1 < D_\pm/3$ , such that the two functions  $t \ln \left( \frac{D_\pm}{t} \right)$  are positive and strictly increasing for  $0 \leq t \leq t_1$ .

We now introduce the scaling variable  $z(n, x_0) = V_d \rho(x) \|x - x_0\|^d n \log(n)$ , so that

$$L_1^\pm = \frac{1}{\|x - x_0\|^d} \int_0^{t_1} e^{-\frac{t}{\|x - x_0\|^d} \left( 1 + z \frac{\ln(D_\pm/t)}{\ln(n)} \right)} dt = \int_0^{\frac{t_1}{\|x - x_0\|^d}} e^{-u \left( 1 + z \frac{\ln(D_\pm \|x - x_0\|^{-d}/u)}{\ln(n)} \right)} du, \quad (97)$$

where we have used the shorthand notation  $z = z(n, x_0)$ .

For a given real  $Z \geq 0$ , we now want to study the limit of  $L_0(x)$  when  $n \rightarrow \infty$ ,  $\|x - x_0\|^{-d} \rightarrow +\infty$  (i.e.,  $x_0 \rightarrow x$ ), and such that  $z(n, x_0) \rightarrow Z$ , which we will simply denote  $\lim_Z L_0(x)$ . We note that  $\lim_Z L_2 = 0$  (see Eq. (91) and Eq. (93)), so that we are left to show that  $\lim_Z L_1^\pm = \frac{1}{1+Z} = \lim_Z L_0(x)$ , which will prove Theorem 3.2.

Exploiting the fact that  $u \ln(u) > -1/e$ , for  $u > 0$ , we obtain

$$L_1^\pm \geq e^{-\frac{z}{e \ln(n)}} \int_0^{\frac{t_1}{\|x-x_0\|^d}} e^{-u} \left( 1 + z \frac{\ln(D_\pm \|x-x_0\|^{-d})}{\ln(n)} \right) du, \quad (98)$$

$$\geq \frac{1}{1+z} e^{-\frac{z}{e \ln(n)}} \left( 1 - e^{-\frac{t_1}{\|x-x_0\|^d}} \right), \quad (99)$$

which shows that  $L_1^\pm$  is bounded from below by a term for which the  $\lim_Z$  is  $\frac{1}{1+Z}$ .

Anticipating that we will take the  $\lim_Z$  and hence the limit  $x_0 \rightarrow x$ , we can freely assume that  $\|x - x_0\| < 1$  and  $K = \frac{t_1}{\|x-x_0\|^{d/2}} > 1$ , so that we also have  $K < \frac{t_1}{\|x-x_0\|^d}$ . We then obtain

$$L_1^\pm \leq \int_0^K e^{-u} \left( 1 + z \frac{\ln(D_\pm \|x-x_0\|^{-d/u})}{\ln(n)} \right) du + \int_K^{+\infty} e^{-u} du, \quad (100)$$

$$\leq \int_0^1 e^{-u} \left( 1 + z \frac{\ln(D_\pm \|x-x_0\|^{-d})}{\ln(n)} \right) du + \int_1^K e^{-u} \left( 1 + z \frac{\ln(D_\pm \|x-x_0\|^{-d/K})}{\ln(n)} \right) du + e^{-K}, \quad (101)$$

$$\leq \frac{1 - e^{-1-z \frac{\ln(D_\pm \|x-x_0\|^{-d})}{\ln(n)}}}{1 + z \frac{\ln(D_\pm \|x-x_0\|^{-d})}{\ln(n)}} + \frac{e^{-1-z \frac{\ln(D_\pm \|x-x_0\|^{-d/K})}{\ln(n)}}}{1 + z \frac{\ln(D_\pm \|x-x_0\|^{-d/K})}{\ln(n)}} + e^{-K}. \quad (102)$$

For  $Z > 0$ ,  $\lim_Z \frac{\ln(\|x-x_0\|^{-d})}{\ln(n)} = \lim_Z \frac{\ln(\|x-x_0\|^{-d/2})}{\ln(n)} = 1$ , and the  $\lim_Z$  of the upper bound in

Eq. (102) is also  $\frac{1}{1+Z}$ . For  $Z = 0$ , we have  $\lim_Z z \frac{\ln(\|x-x_0\|^{-d})}{\ln(n)} = \lim_Z z \frac{\ln(\|x-x_0\|^{-d/2})}{\ln(n)} = 0$ , so that the  $\lim_Z$  of the upper bound in Eq. (102) is 1. Finally, since  $\lim_Z L_2 = 0$ , we have shown that for any real  $Z \geq 0$ ,  $\lim_Z L_1^\pm = \lim_Z L_0(x) = \frac{1}{1+Z}$ , which proves Theorem 3.2. Note that the two bounds obtained suggest that the relative error between  $L_0(x)$  and  $\frac{1}{1+Z}$  for finite large  $n$  and large  $\|x - x_0\|^{-d}$  with  $z(n, x_0)$  remaining close to  $Z$  is of order  $1/\ln(n)$ , or equivalently, of order  $1/\ln(\|x - x_0\|)$ .

*Numerical simulations for the Lagrange function at finite  $n$*

#### A.4 The variance term

We define the variance term  $\mathcal{V}(x)$  as

$$\mathcal{V}(x) = \mathbb{E} \left[ \sum_{i=0}^n w_i^2(x) [y_i - f(x_i)]^2 \right] = \mathbb{E}_X \left[ \sum_{i=0}^n w_i^2(x) \sigma^2(x_i) \right] = (n+1) \mathbb{E} \left[ w_0^2(x) \sigma^2(x_0) \right]. \quad (103)$$

If we first assume that  $\sigma^2(x)$  is bounded by  $\sigma_0^2$ , we can readily bound  $\mathcal{V}(x)$  using Theorem 3.1 with  $\beta = 2$ :

$$\mathcal{V}(x) \leq (n+1) \sigma_0^2 \mathbb{E} \left[ w_0^2(x) \right]. \quad (104)$$

Hence, for any  $\varepsilon > 0$ , there exists a constant  $N_{x,\varepsilon}$ , such that for  $n \geq N_{x,\varepsilon}$ , we obtain Theorem 3.3

$$\mathcal{V}(x) \leq (1 + \varepsilon) \frac{\sigma_0^2}{\ln(n)}. \quad (105)$$

However, one can obtain an exact asymptotic equivalent for  $\mathcal{V}(x)$  by assuming that  $\sigma^2$  is continuous at  $x$  (with  $\sigma^2(x) > 0$ ), while relaxing the boundedness condition. Indeed, we now assume the growth condition  $C_{\text{Growth}}^\sigma$

$$\int \rho(y) \frac{\sigma^2(y)}{1 + \|y\|^{2d}} d^d y < \infty. \quad (106)$$

Note that this condition can be satisfied even in the case where the mean variance  $\int \rho(y) \sigma^2(y) d^d y$  is infinite.

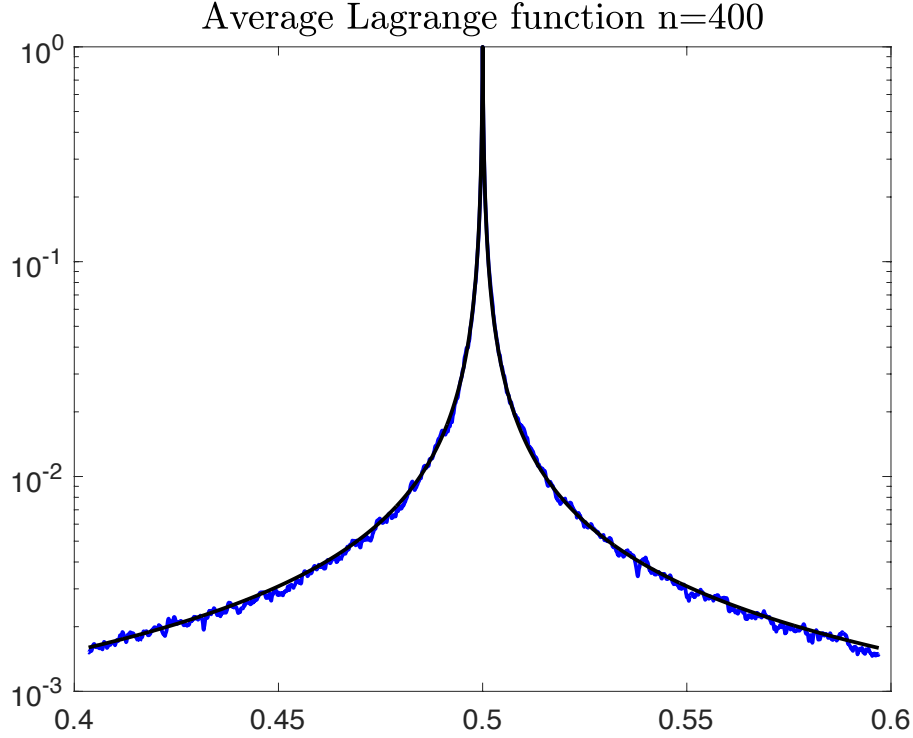


Figure 3: A numerical simulation is shown of the expected value of the Lagrange function of the Hilbert kernel regression estimator in one dimension for a uniform distribution like in Fig. 1. A total of  $n = 400$  samples  $x_i$  were chosen uniformly distributed in the interval  $[0, 1]$  for 100 repeats and the Lagrange function evaluated at  $x_0 = 0.5$  was averaged across these 100 repeats (blue curve). The black curve shows the asymptotic form  $(1 + Z)^{-1}$  with  $Z = 2|x - x_0|/W_n$ . Since  $n = 400$  is not too large, we used the implicit form for the scale  $W_n$  given by  $W_n \ln(1/W_n) = 1/n$  (see main text below Theorem 3.1) leading to  $W_n^{-1} = 3232.39$  (compare with  $400 \ln(400) = 2396.59$ ).

Proceeding along the very same line as the proof of Theorem 3.1 in section A.2, we can write

$$\mathbb{E} \left[ w_0^2(x) \sigma^2(x_0) \right] = \int_0^{+\infty} t \psi^n(x, t) \phi(x, t) dt, \quad (107)$$

with

$$\phi(x, t) := \int \rho(x + y) \sigma^2(x + y) \frac{e^{-\frac{t}{\|y\|^d}}}{\|y\|^{2d}} d^d y, \quad (108)$$

which as a similar form as Eq. (40), with  $\beta = 2$ . The condition of Eq. (106) ensures that the integral defining  $\phi(x, t)$  converges for all  $t > 0$ .

The continuity of  $\sigma^2$  at  $x$  (and hence of  $\rho\sigma^2$ ) and the fact the  $\rho(x)\sigma^2(x) > 0$  implies the existence a small enough  $\lambda > 0$  such that the ball  $B(x, \lambda) \subset \Omega^\circ$  and  $\|y\| \leq \lambda \implies |\rho(x+y)\sigma^2(x+y) - \rho(x)\sigma^2(x)| \leq \varepsilon\rho(x)\sigma^2(x)$ , a property exploited for  $\rho$  in the proof of Theorem 3.1 (see Eq. (52) and the paragraph above it), and which can now be used to efficiently bound  $\phi(x, t)$ . In addition, using the method of proof of Theorem 3.1 (see Eq. (64)) also requires that  $\int_{\|y\| \geq \lambda} \rho(y) \frac{\sigma^2(y)}{\|y\|^{2d}} d^d y < \infty$ , which is ensured by the condition  $C_{\text{Growth}}^\sigma$  of Eq. (106). Apart from these details, one can proceed strictly along the proof and Theorem 3.1, leading to the proof of Theorem 3.4:

$$\mathcal{V}(x) \underset{n \rightarrow +\infty}{\sim} \frac{\sigma^2(x)}{\ln(n)}. \quad (109)$$

Note that if  $\sigma^2(x) = 0$ , one can straightforwardly show that for any  $\varepsilon > 0$ , and for  $n$  large enough, one has

$$\mathcal{V}(x) \leq \frac{\varepsilon}{\ln(n)}, \quad (110)$$

while a more optimal estimate can be easily obtained if one specifies how  $\sigma^2$  vanishes at  $x$ .

## A.5 The bias term

This section aims at proving Theorem 3.5, 3.6, and 3.7.

### Assumptions

We first impose the following growth condition  $C_{\text{Growth}}^f$  for  $f(x) := \mathbb{E}[Y \mid X = x]$ :

$$\int \rho(y) \frac{f^2(y)}{(1 + \|y\|^d)^2} d^d y < \infty, \quad (111)$$

which is obviously satisfied if  $f$  is bounded. Since  $\rho$  is assumed to have a second moment, condition  $C_{\text{Growth}}^f$  is also satisfied for any function satisfying  $|f(x)| \leq A_f \|y\|^{d+1}$  for all  $y$ , such that  $\|y\| \geq R_f$ , for some  $R_f > 0$ . Using the Cauchy-Schwartz inequality, we find that the condition  $C_{\text{Growth}}^f$  also implies that

$$\int \rho(y) \frac{|f(y)|}{1 + \|y\|^d} d^d y < \infty. \quad (112)$$

In addition, for any  $x \in \Omega^\circ$  (so that  $\rho(x) > 0$ ), we assume that there exists a neighborhood of  $x$  such that  $f$  satisfies a local Hölder condition. In other words, there exist  $\delta_x > 0$ ,  $K_x > 0$ , and  $\alpha_x > 0$ , such that the ball  $B(0, \delta_x) \subset \Omega$ , and

$$\|y\| \leq \delta_x \implies |f(x+y) - f(x)| \leq K_x \|y\|^{\alpha_x}, \quad (113)$$

which defines condition  $C_{\text{Holder}}^f$ .

### Definition of the bias term and preparatory results

We define the bias term  $\mathcal{B}(x)$  as

$$\mathcal{B}(x) = \mathbb{E}_X \left[ \left( \sum_{i=0}^n w_i(x) [f(x_i) - f(x)] \right)^2 \right] = (n+1)\mathcal{B}_1(x) + n(n+1)\mathcal{B}_2(x), \quad (114)$$

$$\mathcal{B}_1(x) = \frac{1}{n+1} \mathbb{E}_X \left[ \sum_{i=0}^n w_i^2(x) [f(x_i) - f(x)]^2 \right], \quad (115)$$

$$= \mathbb{E}_X \left[ w_0^2(x) [f(x_0) - f(x)]^2 \right], \quad (116)$$

$$\mathcal{B}_2(x) = \frac{1}{n(n+1)} \mathbb{E}_X \left[ \sum_{0 \leq i < j \leq n} w_i(x) w_j(x) [f(x_i) - f(x)] [f(x_i) - f(x)] \right], \quad (117)$$

$$= \mathbb{E}_X \left[ w_0(x) w_1(x) [f(x_0) - f(x)] [f(x_1) - f(x)] \right]. \quad (118)$$

Exploiting again Eq. (36) for  $\beta = 2$  like we did in section A.2, we obtain

$$\mathcal{B}_1(x) = \int_0^{+\infty} t \psi^n(x, t) \chi_1(x, t) dt, \quad (119)$$

where  $\psi(x, t)$  is again the function defined in Eq. (37), and where

$$\chi_1(x, t) := \int \rho(x+y) e^{-\frac{t}{\|y\|^d}} \frac{(f(x+y) - f(x))^2}{\|y\|^{2d}} d^d y. \quad (120)$$

For any  $t > 0$ , and under condition  $C_{\text{Growth}}^f$ , the integral defining  $\chi_1(x, t)$  is well defined. Moreover,  $\chi_1(x, t)$  is a strictly positive and strictly decreasing function of  $t > 0$ .

Now, defining  $u_i = \|x - x_i\|^{-d}$ ,  $i = 0, \dots, n$  and exploiting again Eq. (36) for  $\beta = 2$ , we can write

$$w_0(x) w_1(x) = u_0 u_1 \int_0^{+\infty} t e^{-(u_0 + u_1)t - (\sum_{i=2}^n u_i)t} dt \quad (121)$$

Now taking the expectation value over the  $n + 1$  independent variables, we obtain

$$\mathcal{B}_2(x) = \int_0^{+\infty} t \psi^{n-1}(x, t) \chi_2^2(x, t) dt, \quad (122)$$

where

$$\chi_2(x, t) := \int \rho(x+y) e^{-\frac{t}{\|y\|^d}} \frac{f(x+y) - f(x)}{\|y\|^d} d^d y. \quad (123)$$

Again, for any  $t > 0$ , and under condition  $C_{\text{Growth}}^f$ , the integral defining  $\chi_2(x, t)$  is well defined. Note that, the integral defining  $\chi_2(x, 0)$  is well behaved at  $y = 0$  under condition  $C_{\text{Holder}}^f$ . Indeed, for  $\|y\| \leq \delta_x$ , we have  $\frac{|f(x+y) - f(x)|}{\|y\|^d} \leq K_x \|y\|^{-d + \alpha_x}$ , which is integrable at  $y = 0$  in dimension  $d$ . Note that, if  $f(x+y) - f(x)$  were only decaying as  $\text{const.}/\ln(\|y\|)$ , then  $|\chi_2(x, t)| \sim \text{const.} \ln(|\ln(t)|) \rightarrow +\infty$ , when  $t \rightarrow 0$ , and  $\chi_2(x, 0)$  would not exist (see the end of this section where we relax the local Hölder condition).

From now, we denote

$$\kappa(x) := \chi_2(x, 0) = \int \rho(x+y) \frac{f(x+y) - f(x)}{\|y\|^d} d^d y. \quad (124)$$

Also note that  $\kappa(x) = 0$  is possible even if  $f$  is not constant. For instance, if  $\Omega$  is a sphere centered at  $x$  or  $\Omega = \mathbb{R}^d$ , if  $\rho(x+y) = \hat{\rho}(\|y\|)$  is isotropic around  $x$  and, if  $f_x : y \mapsto f(x+y)$  is an odd function of  $y$ , then we indeed have  $\kappa(x) = 0$  at the symmetry point  $x$ .

*Upper bound for  $\mathcal{B}_1(x)$*

For  $\varepsilon > 0$ , we define  $\lambda$  like in section A.2 and define  $\eta = \min(\lambda, \delta_x)$ , so that

$$\chi_1(x, t) \leq (1 + \varepsilon) K_x \rho(x) \int_{\|y\| \leq \eta} e^{-\frac{t}{\|y\|^d}} \|y\|^{2(\alpha_x - d)} d^d y + \Lambda_x, \quad (125)$$

$$\Lambda_x = \int_{\|y\| \geq \eta} \rho(x+y) \frac{(f(x+y) - f(x))^2}{\|y\|^{2d}} d^d y, \quad (126)$$

where the constant  $\Lambda_x < \infty$  under condition  $C_{\text{Growth}}^f$ . The integral in Eq. (125), can be written as

$$\int_{\|y\| \leq \eta} e^{-\frac{t}{\|y\|^d}} \|y\|^{2(\alpha_x - d)} d^d y = S_d \int_0^\eta e^{-\frac{t}{r^d}} r^{2\alpha_x - d - 1} dr, \quad (127)$$

$$= V_d t^{\frac{2\alpha_x}{d} - 1} \int_{\frac{t}{\eta^d}}^{+\infty} u^{-\frac{2\alpha_x}{d}} e^{-u} du, \quad (128)$$

Hence, we find that  $\chi_1(x, t)$  is bounded for  $\alpha_x > d/2$ . For  $\alpha_x < d/2$ , and for  $t < t_1$  small enough, there exists a constant  $M(2\alpha_x/d)$  so that  $\chi_1(x, t) \leq M(2\alpha_x/d) t^{\frac{2\alpha_x}{d} - 1}$ . Finally, in the marginal case  $\alpha_x = d/2$  and for  $t < t_1$ , we have  $\chi_1(x, t) \leq M(1) \ln(1/t)$ , for some constant  $M(1)$ .

Now, exploiting again the upper bound of  $\psi(x, t)$  obtained in section A.2 and repeating the steps to bound the integrals involving  $\psi^n(x, t)$ , we find that, for  $\alpha_x \neq d/2$ ,  $\mathcal{B}_1(x)$  is bounded up to a multiplicative constant by

$$\int_0^{t_1} t^{\min(1, \frac{2\alpha_x}{d})} e^{-nV_d\rho(x)t \ln\left(\frac{D-t}{t}\right)} dt \underset{n \rightarrow +\infty}{\sim} M'(2\alpha_x/d) (V_d\rho(x)n \ln(n))^{-\min(2, \frac{2\alpha_x}{d}+1)}, \quad (129)$$

where  $M'(2\alpha_x/d)$  is a constant depending only on  $2\alpha_x/d$ . In the marginal case,  $\alpha_x = d/2$ ,  $\mathcal{B}_1(x)$  is bounded up to a multiplicative constant by  $n^{-2} \ln(n)$ .

In summary, we find that

$$(n+1)\mathcal{B}_1(x) = \begin{cases} O\left(n^{-\frac{2\alpha_x}{d}} (\ln(n))^{-1-\frac{2\alpha_x}{d}}\right), & \text{for } d > 2\alpha_x \\ O\left(n^{-1} (\ln(n))^{-1}\right), & \text{for } d = 2\alpha_x \\ O\left(n^{-1} (\ln(n))^{-2}\right), & \text{for } d < 2\alpha_x \end{cases} \quad (130)$$

*Asymptotic equivalent for  $\mathcal{B}_2(x)$*

Let us first assume that  $\kappa(x) = \chi_2(x, 0) \neq 0$ . Then again, as shown in detail in section A.2, the integral defining  $\mathcal{B}_2(x)$  is dominated by the small  $t$  region, and will be asymptotically equivalent to

$$\mathcal{B}_2(x) = \int_0^{+\infty} t \psi^{n-1}(x, t) \chi_2^2(x, t) dt, \quad (131)$$

$$\underset{n \rightarrow +\infty}{\sim} \kappa^2(x) \int_0^{t_1} t e^{-nV_d\rho(x)t \ln\left(\frac{D-t}{t}\right)} dt, \quad (132)$$

$$\underset{n \rightarrow +\infty}{\sim} \left( \frac{\kappa(x)}{V_d\rho(x)n \ln(n)} \right)^2. \quad (133)$$

On the other hand, if  $\kappa(x) = 0$ , one can bound  $\chi_2(x, t)$  (up to a multiplicative constant) for  $t \leq t_1$  by the integral

$$\int_{\|y\| \leq \eta} \left(1 - e^{-\frac{t}{\|y\|^d}}\right) \|y\|^{\alpha_x-d} d^d y = S_d \int_0^\eta \left(1 - e^{-\frac{t}{r^d}}\right) r^{\alpha_x-d} r^{d-1} dr, \quad (134)$$

$$= V_d t^{\frac{\alpha_x}{d}} \int_{\frac{t}{\eta^d}}^{+\infty} u^{-1-\frac{\alpha_x}{d}} (1 - e^{-u}) du. \quad (135)$$

Hence, for  $\kappa(x) = 0$ , we find that

$$n(n+1)\mathcal{B}_2(x) = O\left(n^{-\frac{2\alpha_x}{d}} (\ln(n))^{-2-\frac{2\alpha_x}{d}}\right). \quad (136)$$

*Asymptotic equivalent for the bias term  $\mathcal{B}(x)$*

In the generic case  $\kappa(x) \neq 0$ , we find that  $(n+1)\mathcal{B}_1(x)$  is always dominated by  $n(n+1)\mathcal{B}_2(x)$ , and we find the following asymptotic equivalent for  $\mathcal{B}(x) = (n+1)\mathcal{B}_1(x) + n(n+1)\mathcal{B}_2(x)$ :

$$\mathcal{B}(x) \underset{n \rightarrow +\infty}{\sim} \left( \frac{\kappa(x)}{V_d\rho(x) \ln(n)} \right)^2. \quad (137)$$

In the non-generic case  $\kappa(x) = 0$ , the bound for  $(n+1)\mathcal{B}_1(x)$  in Eq. (130) is always more stringent than the bound for  $n(n+1)\mathcal{B}_2(x)$  in Eq. (136), leading to

$$\mathcal{B}(x) = \begin{cases} O\left(n^{-\frac{2\alpha_x}{d}} (\ln(n))^{-1-\frac{2\alpha_x}{d}}\right), & \text{for } d > 2\alpha_x \\ O\left(n^{-1} (\ln(n))^{-1}\right), & \text{for } d = 2\alpha_x \\ O\left(n^{-1} (\ln(n))^{-2}\right), & \text{for } d < 2\alpha_x \end{cases}, \quad (138)$$

which prove the statements made in Theorem 3.5.

*Interpretation of the bias term  $\mathcal{B}(x)$  for  $\kappa(x) \neq 0$*

Here, we assume the generic case  $\kappa(x) \neq 0$  and define  $\bar{f}(x) = \mathbb{E} [\hat{f}(x)]$ . We have

$$\Delta(x) := \mathbb{E} \left[ \sum_{i=0}^n w_i(x) (f(x_i) - f(x)) \right] = \bar{f}(x) - f(x), \quad (139)$$

$$\bar{f}(x) = \mathbb{E} \left[ \sum_{i=0}^n w_i(x) f(x_i) \right] = (n+1) \mathbb{E} [w_0(x) f(x_0)]. \quad (140)$$

By using another time Eq. (36), we find that

$$\Delta(x) = (n+1) \int_0^{+\infty} \psi^n(x, t) \chi_2(x, t) dt, \quad (141)$$

$$\underset{n \rightarrow +\infty}{\sim} n \kappa(x) \int_0^{t_1} e^{-nV_d \rho(x)t \ln\left(\frac{D_{\pm}}{t}\right)} dt, \quad (142)$$

$$\underset{n \rightarrow +\infty}{\sim} \frac{\kappa(x)}{V_d \rho(x) \ln(n)}. \quad (143)$$

Comparing this result to the one of Eq. (137), we find that the bias  $\mathcal{B}(x)$  is asymptotically dominated by the square of the difference  $\Delta^2(x)$  between  $\bar{f}(x) = \mathbb{E} [\hat{f}(x)]$  and  $f(x)$ :

$$\mathcal{B}(x) \underset{n \rightarrow +\infty}{\sim} \left( \mathbb{E} [\hat{f}(x)] - f(x) \right)^2, \quad (144)$$

a statement made in Theorem 3.5.

*Relaxing the local Hölder condition*

We now only assume the condition  $C_{\text{Cont.}}^f$  that  $f$  is continuous at  $x$  (but still assuming the growth conditions). We can now define  $\delta_x$  such that the ball  $B(x, \delta) \subset \Omega^\circ$  and  $\|y\| \leq \delta_x \implies |f(x+y) - f(x)| \leq \varepsilon$ . Then, the proof proceeds as above but by replacing  $K_x$  by  $\varepsilon$ ,  $\alpha_x$  by 0, and by updating the bounds for  $\chi_1(x, t)$  (for which this replacement is safe) and  $\chi_2(x, t)$  (for which it is not). We now find that for  $0 < t \leq t_1$ , with  $t_1$  small enough

$$0 \leq \chi_1(x, t) \leq \varepsilon(1 + 2\varepsilon)V_d \rho(x)t^{-1}, \quad (145)$$

$$|\chi_2(x, t)| \leq \varepsilon(1 + 2\varepsilon)V_d \rho(x) \ln\left(\frac{1}{t}\right). \quad (146)$$

As already mentioned below Eq. (123) where we provided an explicit counterexample, we see that relaxing the local Hölder condition does not guarantee anymore that  $\lim_{t \rightarrow 0} |\chi_2(x, 0)| < \infty$ . With these new bounds, and carrying the rest of the calculation as in the previous sections, we ultimately find the following weaker result compared to Eq. (137) and Eq. (138):

$$\mathcal{B}(x) = o\left(\frac{1}{\ln(n)}\right), \quad (147)$$

or equivalently, that for any  $\varepsilon > 0$ , there exists a constant  $N_{x, \varepsilon}$  such that, for  $n \geq N_{x, \varepsilon}$ , we have

$$\mathcal{B}(x) \leq \frac{\varepsilon}{\ln(n)}. \quad (148)$$

*The bias term at a point where  $\rho(x) = 0$*

This section aims at proving Theorem 3.7 expressing the lack of convergence of the estimator  $\hat{f}(x)$  to  $f(x)$ , when  $\rho(x) = 0$ , and under mild conditions. Let us now consider a point  $x \in \partial\Omega$  for which

$\rho(x) = 0$ , let us assume that there exists constants  $\eta_x, \gamma_x > 0$ , and  $G_x > 0$ , such that  $\rho$  satisfies the local Hölder condition at  $x$

$$\|y\| \leq \eta_x \implies \rho(x+y) \leq G_x \|y\|^{\gamma_x}. \quad (149)$$

We will also assume that the growth condition of Eq. (112) is satisfied. With these two conditions,  $\kappa(x)$  defined in Eq. (124) exists. The vanishing of  $\rho$  at  $x$  strongly affects the behavior of  $\psi(x, t)$  in the limit  $t \rightarrow 0$ , which is not singular anymore:

$$1 - \psi(x, t) \underset{t \rightarrow 0}{\sim} t \int \rho(y) \|x - y\|^{-d} d^d y, \quad (150)$$

where the convergence of the integral  $\lambda(x) := \int \rho(y) \|x - y\|^{-d} d^d y$  is ensured by the local Hölder condition of  $\rho$  at  $x$ .

Let us now evaluate  $\bar{f}(x) = \lim_{n \rightarrow +\infty} \mathbb{E}[\hat{f}(x)]$ , the expectation value of the estimator  $\hat{f}(x)$  in the limit  $n \rightarrow +\infty$ , introduced in Eq. (140). First assuming,  $\kappa(x) = \chi_2(x, 0) \neq 0$ , we obtain

$$\bar{f}(x) - f(x) = \lim_{n \rightarrow +\infty} (n+1) \int_0^{+\infty} \psi^n(x, t) \chi_2(x, t) dt, \quad (151)$$

$$= \lim_{n \rightarrow +\infty} n \chi_2(x, 0) \int_0^{t_1} e^{n t \partial_t \psi(x, 0)} dt, \quad (152)$$

$$= \frac{\kappa(x)}{\lambda(x)}, \quad (153)$$

which shows that the bias term does not vanish in the limit  $n \rightarrow +\infty$ . Eq. (153) can be straightforwardly shown to remain valid when  $\kappa(x) = 0$ . Indeed, for any  $\varepsilon > 0$  chosen arbitrarily small, we can choose  $t_1$  small enough such that  $|\chi_2(x, t)| \leq \varepsilon$  for  $0 \leq t \leq t_1$ , which leads to  $|\bar{f}(x) - f(x)| \leq \varepsilon/\lambda(x)$ .

Note that relaxing the local Hölder condition for  $\rho$  at  $x$  and only assuming the continuity of  $f$  at  $x$  and  $\kappa(x) \neq 0$  is not enough to guarantee that  $\bar{f}(x) \neq f(x)$ . For instance, if  $\rho(x+y) \sim_{y \rightarrow 0} \rho_0 / \ln(1/\|y\|)$ , and there exists a local solid angle  $\omega_x > 0$  at  $x$ , one can show that  $1 - \psi(x, t) \sim_{t \rightarrow 0} \omega_x S_d \rho_0 t \ln(\ln(1/t))$ , and the bias would still vanish in the limit  $n \rightarrow +\infty$ , with  $\hat{f}(x) - f(x) \sim_{n \rightarrow +\infty} \kappa(x) / [\omega_x S_d \rho_0 \ln(\ln(n))]$ .

## A.6 Asymptotic equivalent for the regression risk

This sections aim at proving Theorem 3.8. Under conditions  $C_{\text{Growth}}^\sigma$ ,  $C_{\text{Growth}}^f$ , and  $C_{\text{Cont.}}^f$ , the results of Eq. (109) and Eq. (147) show that for  $\rho(x)\sigma^2(x) > 0$  and  $\rho$  and  $\sigma^2$  continuous at  $x$ , the bias term  $\mathcal{B}(x)$  is always dominated by the variance term  $\mathcal{V}(x)$  in the limit  $n \rightarrow +\infty$ . Thus, the excess regression risk satisfies

$$\mathbb{E}[(\hat{f}(x) - f(x))^2] \underset{n \rightarrow +\infty}{\sim} \frac{\sigma^2(x)}{\ln(n)}. \quad (154)$$

As a consequence, the Hilbert kernel estimate converges pointwise to the regression function in probability. Indeed, for  $\delta > 0$ , there exists a constant  $N_{x, \delta}$ , such that

$$\mathbb{E}[(\hat{f}(x) - f(x))^2] \leq (1 + \delta) \frac{\sigma^2(x)}{\ln(n)}, \quad (155)$$

for  $n \geq N_{x, \delta}$ . Moreover, for any  $\varepsilon > 0$ , since  $\mathbb{E}[(\hat{f}(x) - f(x))^2] \geq \varepsilon^2 \mathbb{P}[|\hat{f}(x) - f(x)| \geq \varepsilon]$ , we deduce the following Chebyshev bound, valid for  $n \geq N_{x, \delta}$

$$\mathbb{P}[|\hat{f}(x) - f(x)| \geq \varepsilon] \leq \frac{1 + \delta}{\varepsilon^2} \frac{\sigma^2(x)}{\ln(n)}. \quad (156)$$

## A.7 Rates for the plugin classifier

In the case of binary classification  $Y \in \{0, 1\}$  and  $f(x) = \mathbb{P}[Y = 1 \mid X = x]$ . Let  $F: \mathbb{R}^d \rightarrow \{0, 1\}$  denote the Bayes optimal classifier, defined by  $F(x) := \theta(f(x) - 1/2)$  where  $\theta(\cdot)$  is the Heaviside



theta function. This classifier minimizes the risk  $\mathcal{R}_{0/1}(h) := \mathbb{E}[\mathbb{1}_{\{h(X) \neq Y\}}] = \mathbb{P}[h(X) \neq Y]$  under zero-one loss. Given the regression estimator  $\hat{f}$ , we consider the plugin classifier  $\hat{F}(x) = \theta(\hat{f}(x) - \frac{1}{2})$ , and we will exploit the fact that

$$0 \leq \mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] - \mathcal{R}_{0/1}(F(x)) \leq 2 \mathbb{E}[|\hat{f}(x) - f(x)|] \leq 2\sqrt{\mathbb{E}[(\hat{f}(x) - f(x))^2]} \quad (157)$$

*Proof of Eq. (157)*

For the sake of completeness, let us briefly prove the result of Eq. (157). The rightmost inequality is simply obtained from the Cauchy-Schwartz inequality and we hence focus on proving the first inequality. Obviously, Eq. (157) is satisfied for  $f(x) = 1/2$ , for which  $\mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] = \mathcal{R}_{0/1}(F(x)) = 1/2$ .

If  $f(x) > 1/2$ , we have  $F(x) = 1$ ,  $\mathcal{R}_{0/1}(F(x)) = 1 - f(x)$ , and

$$\mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] = f(x)\mathbb{P}[\hat{f}(x) \leq 1/2] + (1 - f(x))\mathbb{P}[\hat{f}(x) \geq 1/2], \quad (158)$$

$$= \mathcal{R}_{0/1}(F(x)) + (2f(x) - 1)\mathbb{P}[\hat{f}(x) \leq 1/2], \quad (159)$$

which implies  $\mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] \geq \mathcal{R}_{0/1}(F(x))$ . Since  $\mathbb{P}[\hat{f}(x) \leq 1/2] = \mathbb{E}[\theta(1/2 - \hat{f}(x))]$ , and using  $\theta(1/2 - \hat{f}(x)) \leq \frac{|\hat{f}(x) - f(x)|}{f(x) - 1/2}$ , valid for any  $1/2 < f(x) \leq 1$ , we readily obtain Eq. (157).

Similarly, in the case  $f(x) < 1/2$ , we have  $F(x) = 0$ ,  $\mathcal{R}_{0/1}(F(x)) = f(x)$ , and

$$\mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] = \mathcal{R}_{0/1}(F(x)) + (1 - 2f(x))\mathbb{P}[\hat{f}(x) \geq 1/2]. \quad (160)$$

Since  $\mathbb{P}[\hat{f}(x) \geq 1/2] = \mathbb{E}[\theta(\hat{f}(x) - 1/2)]$ , and using  $\theta(\hat{f}(x) - 1/2) \leq \frac{|\hat{f}(x) - f(x)|}{1/2 - f(x)}$ , valid for any  $0 \leq f(x) < 1/2$ , we again obtain Eq. (157) in this case.

In fact, for any  $\alpha > 0$ , the inequalities  $\theta(1/2 - \hat{f}(x)) \leq \left(\frac{|\hat{f}(x) - f(x)|}{f(x) - 1/2}\right)^\alpha$  and  $\theta(\hat{f}(x) - 1/2) \leq \left(\frac{|\hat{f}(x) - f(x)|}{1/2 - f(x)}\right)^\alpha$  hold, respectively for  $f(x) > 1/2$  and  $f(x) < 1/2$ . Combining this remark with the use of the Hölder inequality leads to

$$\mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] - \mathcal{R}_{0/1}(F(x)) \leq 2|f(x) - 1/2|^{1-\alpha} \mathbb{E}\left[|\hat{f}(x) - f(x)|^\alpha\right], \quad (161)$$

$$\leq 2|f(x) - 1/2|^{1-\alpha} \mathbb{E}\left[|\hat{f}(x) - f(x)|^{\frac{\alpha}{\beta}}\right]^\beta, \quad (162)$$

for any  $0 < \beta \leq 1$ . In particular, for  $0 < \alpha < 1$  and  $\beta = \alpha/2$ , we obtain

$$0 \leq \mathbb{E}[\mathcal{R}_{0/1}(\hat{F}(x))] - \mathcal{R}_{0/1}(F(x)) \leq 2|f(x) - 1/2|^{1-\alpha} \mathbb{E}\left[|\hat{f}(x) - f(x)|^2\right]^{\frac{\alpha}{2}}. \quad (163)$$

The interest of this last bound compared to the more classical bound of Eq. (157) is to show explicitly the cancellation of the classification risk as  $f(x) \rightarrow 1/2$ , while still involving the regression risk  $\mathbb{E}\left[|\hat{f}(x) - f(x)|^2\right]$  (to the power  $\alpha/2 < 1/2$ ).

*Bound for the classification risk*

Now exploiting the results of section A.6 for the regression risk, and the two inequalities Eq. (157) and Eq. (163), we readily obtain Theorem 3.9.

## A.8 Extrapolation behavior outside the support of $\rho$

This section aims at proving Theorem 3.10 characterizing the behavior of the regression estimator  $\hat{f}$  outside the closed support  $\Omega$  of  $\rho$  (extrapolation).

*Extrapolation estimator in the limit  $n \rightarrow \infty$*

We first assume the growth condition  $\int \rho(y) \frac{|f(y)|}{1+\|y\|^a} d^d y < \infty$ . For  $x \in \mathbb{R}^d$  (i.e., not necessarily in  $\Omega$ ), we have quite generally

$$\mathbb{E} [\hat{f}(x)] = (n+1) \mathbb{E} [w_0(x)f(x)] = (n+1) \int_0^{+\infty} \psi^n(x, t) \chi(x, t) dt, \quad (164)$$

where  $\psi(x, t)$  is again given by Eq. (37) and

$$\chi(x, t) := \int \rho(x+y) f(x+y) \frac{e^{-\frac{t}{\|y\|^d}}}{\|y\|^d} d^d y, \quad (165)$$

which is finite for any  $t > 0$ , thanks to the above growth condition for  $f$ .

Let us now assume that the point  $x$  is not in the closed support  $\bar{\Omega}$  of the distribution  $\rho$  (which excludes the case  $\Omega = \mathbb{R}^d$ ). Since the integral in Eq. (164) is again dominated by its  $t \rightarrow 0$  behavior, we have to evaluate  $\psi(x, t)$  and  $\chi(x, t)$  in this limit, like in the different proofs above. In fact, when  $x \notin \bar{\Omega}$ , the integral defining  $\psi(x, t)$  and  $\chi(x, t)$  are not singular anymore, and we obtain

$$1 - \psi(x, t) \underset{t \rightarrow 0}{\sim} t \int \rho(y) \|x - y\|^{-d} d^d y, \quad (166)$$

$$\chi(x, 0) = \int \rho(y) f(y) \|x - y\|^{-d} d^d y. \quad (167)$$

Note that  $\psi(x, t)$  has the very same linear behavior as in Eq. (150), when we assumed  $x \in \partial\Omega$  with  $\rho(x) = 0$ , and a local Hölder condition for  $\rho$  at  $x$ .

Finally, by using the same method as in the previous sections to evaluate the integral of Eq. (164) in the limit  $n \rightarrow +\infty$ , we obtain

$$\int_0^{+\infty} \psi^n(x, t) \chi(x, t) dt \underset{n \rightarrow +\infty}{\sim} \chi(x, 0) \int_0^{t_1} e^{n t \partial_t \psi(x, 0)} dt, \quad (168)$$

$$\underset{n \rightarrow +\infty}{\sim} \frac{1}{n} \frac{\chi(x, 0)}{|\partial_t \psi(x, 0)|}, \quad (169)$$

which leads to the first result of Theorem 3.10:

$$\hat{f}_\infty(x) := \lim_{n \rightarrow +\infty} \mathbb{E} [\hat{f}(x)] = \frac{\int \rho(y) f(y) \|x - y\|^{-d} d^d y}{\int \rho(y) \|x - y\|^{-d} d^d y}. \quad (170)$$

Note that since the function  $(x, y) \mapsto \|x - y\|^{-d}$  is continuous at all points  $x \notin \bar{\Omega}$ ,  $y \in \Omega$ , and thanks to the absolute convergence of the integrals defining  $\hat{f}_\infty(x)$ , standard methods show that  $\hat{f}_\infty$  is continuous (in fact, infinitely differentiable) at all  $x \notin \bar{\Omega}$ .

### Extrapolation far from $\Omega$

Let us now investigate the behavior of  $\hat{f}_\infty(x)$  when the distance  $L := d(x, \Omega) = \inf\{\|x - y\|, y \in \Omega\} > 0$  between  $x$  and  $\Omega$  goes to infinity, which can only happen for certain  $\Omega$ , in particular, when  $\Omega$  is bounded. We now assume the stronger condition,  $\langle |f| \rangle := \int \rho(y) |f(y)| d^d y < \infty$ , such that the  $\rho$ -mean of  $f$ ,  $\langle f \rangle := \int \rho(y) f(y) d^d y$ , is finite. We consider a point  $y_0 \in \Omega$ , so that  $\|x - y_0\| \geq L > 0$ , and we will exploit the following inequality, valid for any  $y \in \Omega$  satisfying  $\|y - y_0\| \leq R$ , with  $R > 0$ :

$$0 \leq 1 - \frac{L^d}{\|x - y\|^d} \leq \frac{\|x - y\|^d - L^d}{L^d} \leq \frac{(L + R)^d - L^d}{L^d} \leq e^{\frac{dR}{L}} - 1. \quad (171)$$

Now, for a given  $\varepsilon > 0$ , there exist  $R > 0$  large enough such that  $\int_{\|y - y_0\| \geq R} \rho(y) d^d y \leq \varepsilon/2$  and  $\int_{\|y - y_0\| \geq R} \rho(y) |f(y)| d^d y \leq \varepsilon/2$ . Then, for such a  $R$ , we consider  $L$  large enough such that the above bound satisfies  $e^{\frac{dR}{L}} - 1 \leq \varepsilon \min(1/\langle |f| \rangle, 1)/2$ . We then obtain

$$\left| L^d \int \rho(y) f(y) \|x - y\|^{-d} d^d y - \langle f \rangle \right| \leq \left( e^{\frac{dR}{L}} - 1 \right) \int_{\|y - y_0\| \leq R} \rho(y) |f(y)| d^d y \quad (172)$$

$$+ \int_{\|y - y_0\| \geq R} \rho(y) |f(y)| d^d y, \quad (173)$$

$$\leq \frac{\varepsilon}{2 \langle |f| \rangle} \times \langle |f| \rangle + \frac{\varepsilon}{2} \leq \varepsilon, \quad (174)$$

which shows that under the condition  $\langle |f| \rangle < \infty$ , we have

$$\lim_{d(x, \Omega) \rightarrow +\infty} d^d(x, \Omega) \int \rho(y) f(y) \|x - y\|^{-d} d^d y = \langle f \rangle. \quad (175)$$

Similarly, one can show that

$$\lim_{d(x, \Omega) \rightarrow +\infty} d^d(x, \Omega) \int \rho(y) \|x - y\|^{-d} d^d y = \int \rho(y) d^d y = 1. \quad (176)$$

Finally, we obtain the second result of Theorem 3.10,

$$\lim_{d(x, \Omega) \rightarrow +\infty} \hat{f}_\infty(x) = \langle f \rangle. \quad (177)$$

### Continuity of the extrapolation

We now consider  $x \notin \bar{\Omega}$  and  $y_0 \in \partial\Omega$ , but such that  $\rho(y_0) > 0$  (i.e.,  $y_0 \in \partial\Omega \cap \Omega$ ), and we note  $l := \|x - y_0\| > 0$ . We assume the continuity at  $y_0$  of  $\rho$  and  $f$  as seen as functions restricted to  $\Omega$ , i.e.,  $\lim_{y \in \Omega \rightarrow y_0} \rho(y) = \rho(y_0)$  and  $\lim_{y \in \Omega \rightarrow y_0} f(y) = f(y_0)$ . Hence, for any  $0 < \varepsilon < 1$ , there exists  $\delta > 0$  small enough such that  $y \in \Omega$  and  $\|y - y_0\| \leq \delta \implies |\rho(y_0) - \rho(y)| \leq \varepsilon$  and  $|\rho(y_0)f(y_0) - \rho(y)f(y)| \leq \varepsilon$ . Since we intend to take  $l > 0$  arbitrary small, we can impose  $l < \delta/2$ .

We will also assume that  $\partial\Omega$  is smooth enough near  $y_0$ , such that there exists a strictly positive local solid angle  $\omega_0$  defined by

$$\omega_0 = \lim_{r \rightarrow 0} \frac{1}{V_d \rho(y_0) r^d} \int_{\|y - y_0\| \leq r} \rho(y) d^d y = \lim_{r \rightarrow 0} \frac{1}{V_d r^d} \int_{y \in \Omega / \|y - y_0\| \leq r} d^d y, \quad (178)$$

where the second inequality results from the continuity of  $\rho$  at  $y_0$  and the fact that  $\rho(y_0) > 0$ . If  $y_0 \in \Omega^\circ$ , we have  $\omega_0 = 1$ , while for  $y_0 \in \partial\Omega$ , we have generally  $0 \leq \omega_0 \leq 1$ . Although we will assume  $\omega_0 > 0$  for our proof below, we note that  $\omega_0 = 0$  or  $\omega_0 = 1$  can happen for  $y_0 \in \partial\Omega$ . For instance, we can consider  $\Omega_0, \Omega_1 \subset \mathbb{R}^2$  respectively defined by  $\Omega_0 = \{(x_1, x_2) \in \mathbb{R}^2 / x_1 \geq 0, |x_2| \leq x_1^2\}$  and  $\Omega_1 = \{(x_1, x_2) \in \mathbb{R}^2 / x_1 \leq 0\} \cup \{(x_1, x_2) \in \mathbb{R}^2 / x_1 \geq 0, |x_2| \geq x_1^2\}$ . Then, it is clear that the local solid angle at the origin  $O = (0, 0)$  is respectively  $\omega_0 = 0$  and  $\omega_0 = 1$ . Also note that if  $x$  is on the surface of a sphere or on the interior of a face of a hypercube (and in general, when the boundary near  $x$  is locally an hyperplane; the generic case), we have  $\omega_x = \frac{1}{2}$ . If  $x$  is a corner of the hypercube, we have  $\omega_x = \frac{1}{2^d}$ .

Returning to our proof, and exploiting Eq. (178), we consider  $\delta$  small enough such that for all  $0 \leq r \leq \delta$ , we have

$$\left| \int_{y \in \Omega / \|y - y_0\| \leq r} d^d y - \omega_0 V_d r^d \right| \leq \varepsilon \omega_0 V_d r^d. \quad (179)$$

We can now use these preliminaries to obtain

$$(\rho(y_0)f(y_0) - \varepsilon)J(x) - C \leq \int \rho(y)f(y)\|x - y\|^{-d} d^d y \leq (\rho(y_0)f(y_0) + \varepsilon)J(x) + C, \quad (180)$$

$$(\rho(y_0) - \varepsilon)J(x) - C' \leq \int \rho(y)\|x - y\|^{-d} d^d y \leq (\rho(y_0) + \varepsilon)J(x) + C', \quad (181)$$

with

$$J(x) := \int_{y \in \Omega / \|y - y_0\| \leq \delta} \|x - y\|^{-d} d^d y, \quad (182)$$

$$C = \left(\frac{2}{\delta}\right)^2 \int_{\|y - y_0\| \geq \delta} \rho(y)|f(y)| d^d y, \quad (183)$$

$$C' = \left(\frac{2}{\delta}\right)^2. \quad (184)$$

Let us now show that  $\lim_{l \rightarrow 0} J(x) = +\infty$ . We define  $N := [\delta/l] \geq 2$ , where  $[\cdot]$  is the integer part, and we have  $N \geq 2$ , since we have imposed  $l < \delta/2$ . For  $n \in \mathbb{N} \geq 1$ , we define,

$$I_n := \int_{y \in \Omega / \|y - y_0\| \leq \delta/n} d^d y, \quad (185)$$

and note that we have

$$I_n - I_{n+1} = \int_{\substack{y \in \Omega / \|y - y_0\| \leq \delta/n, \\ \|y - y_0\| \geq \delta/(n+1)}} d^d y, \quad (186)$$

$$\left| I_n - \omega_0 V_d \left( \frac{\delta}{n} \right)^d \right| \leq \varepsilon \omega_0 V_d \left( \frac{\delta}{n} \right)^d. \quad (187)$$

We can then write

$$J(x) \geq \sum_{n=1}^N \frac{1}{\left( l + \frac{\delta}{n} \right)^d} (I_n - I_{n+1}), \quad (188)$$

$$\geq \sum_{n=1}^N \left( \frac{1}{\left( l + \frac{\delta}{n+1} \right)^d} - \frac{1}{\left( l + \frac{\delta}{n} \right)^d} \right) I_{n+1} + \frac{I_1}{(l + \delta)^d} - \frac{I_{N+1}}{\left( l + \frac{\delta}{N+1} \right)^d}. \quad (189)$$

We have

$$\frac{I_1}{(l + \delta)^d} - \frac{I_{N+1}}{\left( l + \frac{\delta}{N+1} \right)^d} \geq \omega_0 V_d \left( (1 - \varepsilon) \frac{1}{\left( 1 + \frac{l}{\delta} \right)^d} - (1 + \varepsilon) \frac{1}{\left( 1 + \frac{(N+1)l}{\delta} \right)^d} \right), \quad (190)$$

$$\geq \omega_0 V_d \left( (1 - \varepsilon) \frac{2^d}{3^d} - (1 + \varepsilon) \right) =: C'', \quad (191)$$

which defines the constant  $C''$ . Now using Eq. (187),  $l < \delta/2$ ,  $N = \lceil \delta/l \rceil$ , and the fact that  $(1 + u)^d - 1 \geq du$ , for any  $u \geq 0$ , we obtain

$$J(x) \geq (1 - \varepsilon) \omega_0 V_d \sum_{n=1}^N \frac{1}{\left( 1 + \frac{(n+1)l}{\delta} \right)^d} \left( \left( \frac{l + \frac{\delta}{n}}{l + \frac{\delta}{n+1}} \right)^d - 1 \right) + C'', \quad (192)$$

$$\geq (1 - \varepsilon) \omega_0 S_d \sum_{n=1}^N \frac{1}{\left( 1 + \frac{(n+1)l}{\delta} \right)^{d+1}} \frac{1}{n} + C'', \quad (193)$$

$$\geq \frac{(1 - \varepsilon) \omega_0 S_d}{\left( 1 + \frac{(N+1)l}{\delta} \right)^{d+1}} \ln(N - 1) + C'', \quad (194)$$

$$\geq (1 - \varepsilon) \omega_0 \left( \frac{2}{5} \right)^{d+1} S_d \ln \left( \frac{\delta}{l} - 2 \right) + C''. \quad (195)$$

We hence have shown that  $\lim_{l \rightarrow 0} J(x) = +\infty$ . Note that we can obtain an upper bound for  $J(x)$  similar to Eq. (193) in a similar way as above, and with a bit more work, it is straightforward to show that we have in fact  $J(x) \sim_{l \rightarrow 0} \omega_0 S_d \ln \left( \frac{\delta}{l} \right)$ , a result that we will not need here.

Now, using Eq. (180) and Eq. (181) and the fact that  $\lim_{l \rightarrow 0} J(x) = +\infty$ , we find that

$$\int \rho(y) f(y) \|x - y\|^{-d} d^d y \underset{l \rightarrow 0}{\sim} \rho(y_0) f(y_0) J(x), \quad (196)$$

$$\int \rho(y) \|x - y\|^{-d} d^d y \underset{l \rightarrow 0}{\sim} \rho(y_0) J(x), \quad (197)$$

for  $f(y_0) \neq 0$  (remember that  $\rho(y_0) > 0$ ), while for  $f(y_0) = 0$ , we obtain  $\int \rho(y) f(y) \|x - y\|^{-d} d^d y = o(J(x))$ . Finally, we have shown that

$$\lim_{x \notin \Omega, x \rightarrow y_0} \hat{f}_\infty(x) = f(y_0), \quad (198)$$

establishing the continuity of the extrapolation and the last part of Theorem 3.10.

## References

- [1] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.
- [2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.
- [3] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.
- [4] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 541–549, 2018.
- [5] Adele Cutler and Guohua Zhao. Pert-perfect random tree ensembles. *Computing Science and Statistics*, 33:490–497, 2001.
- [6] Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, 18(48):1–33, 2017.
- [7] Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *arXiv preprint arXiv:1806.05161*, 2018.
- [8] Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. *arXiv preprint arXiv:1812.11167*, 2018.
- [9] Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.
- [10] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- [11] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.
- [12] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.
- [13] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.
- [14] Mina Karzand and Robert D Nowak. Active learning in the overparameterized and interpolating regime. *arXiv preprint arXiv:1905.12782*, 2019.
- [15] Yue Xing, Qifan Song, and Guang Cheng. Statistical optimality of interpolated nearest neighbor algorithms. *arXiv preprint arXiv:1810.02814*, 2018.
- [16] Partha P. Mitra. Fitting elephants in modern machine learning by statistically consistent interpolation. *Nature Machine Intelligence*, 3(5):378–386, May 2021.
- [17] Martin Anthony and Peter L Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
- [18] Luc Devroye, Laszlo Györfi, and Adam Krzyżak. The hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227, 1998.
- [19] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.
- [20] GS Watson. Smooth regression analysis. *Sankhya A*: 26: 359-372, (50), 1964.
- [21] Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524. ACM, 1968.
- [22] Reinhard Farwig. Rate of convergence of shepard’s global interpolation formula. *Mathematics of Computation*, 46(174):577–590, 1986.