# An Affine Invariant $k$-Nearest Neighbor Regression Estimate

Gérard Biau, Luc Devroye, Vida Dujmovic, Adam Krzyzak

# An Affine Invariant $k$-Nearest Neighbor Regression Estimate

**Gérard Biau**[1]
*Université Pierre et Marie Curie*[2] *& Ecole Normale Supérieure*[3]*, France*
gerard.biau@upmc.fr

**Luc Devroye**
*McGill University, Canada*[4]
lucdevroye@gmail.com

**Vida Dujmović**
*Carleton University, Canada*[5]
vida@scs.carleton.ca

**Adam Krzyżak**
*Concordia University, Canada*[6]
krzyzak@cs.concordia.ca

**Abstract**

We design a data-dependent metric in $\mathbb{R}^d$ and use it to define the $k$-nearest neighbors of a given point. Our metric is invariant under all affine transformations. We show that, with this metric, the standard $k$-nearest neighbor regression estimate is asymptotically consistent under the usual conditions on $k$, and minimal requirements on the input data.

*Index Terms* — Nonparametric estimation, Regression function estimation, Affine invariance, Nearest neighbor methods, Mathematical statistics.

*2010 Mathematics Subject Classification*: 62G08, 62G05, 62G20.

[1]Corresponding author.

[2]Research partially supported by the French National Research Agency under grant ANR-09-BLAN-0051-02 "CLARA".

[3]Research carried out within the INRIA project "CLASSIC" hosted by Ecole Normale Supérieure and CNRS.

[4]Research sponsored by NSERC Grant A3456 and FQRNT Grant 90-ER-0291.

[5]Research sponsored by NSERC Grant RGPIN 402438-2011.

[6]Research sponsored by NSERC Grant N00118.

# 1 Introduction

The prediction error of standard nonparametric regression methods may be critically affected by a linear transformation of the coordinate axes. It is typically the case for the popular $k$-nearest neighbor ($k$-NN) predictor (Fix and Hodges [11, 12], Cover and Hart [7], Cover [5, 6]), where a mere rescaling of the coordinate axes has a serious impact on the capabilities of this estimate. This is clearly an undesirable feature, especially in applications where the data measurements represent physically different quantities, such as temperature, blood pressure, cholesterol level, and the age of the patient. In this example, a simple change in, say, the unit measure of the temperature parameter will lead to totally different results, and will thus force the statistician to use a somewhat arbitrary preprocessing step prior to the $k$-NN estimation process. Furthermore, in several practical implementations, one would like, for physical or economical reasons, to supply the freshly collected data to some machine without preprocessing.

In this paper, we discuss a variation of the $k$-NN regression estimate whose definition is not affected by affine transformations of the coordinate axes. Such a modification could save the user a subjective preprocessing step and would save the manufacturer the trouble of adding input specifications.

The data set we have collected can be regarded as a collection of independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$-valued random variables $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, independent of and with the same distribution as a generic pair $(\mathbf{X}, Y)$ satisfying $\mathbb{E}|Y| < \infty$. The space $\mathbb{R}^d$ is equipped with the standard Euclidean norm $\|.\|$. For fixed $\mathbf{x} \in \mathbb{R}^d$, our goal is to estimate the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ using the data $\mathcal{D}_n$. In this context, the usual $k$-NN regression estimate takes the form

$$r_n(\mathbf{x}; \mathcal{D}_n) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i)}(\mathbf{x}),$$

where $(\mathbf{X}_{(1)}(\mathbf{x}), Y_{(1)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n)}(\mathbf{x}), Y_{(n)}(\mathbf{x}))$ is a reordering of the data according to increasing distances $\|\mathbf{X}_i - \mathbf{x}\|$ of the $\mathbf{X}_i$'s to $\mathbf{x}$. (If distance ties occur, a tie-breaking strategy must be defined. For example, if $\|\mathbf{X}_i - \mathbf{x}\| = \|\mathbf{X}_j - \mathbf{x}\|$, $\mathbf{X}_i$ may be declared "closer" if $i < j$, i.e., the tie-breaking is done by indices.) For simplicity, we will suppress $\mathcal{D}_n$ in the notation and write $r_n(\mathbf{x})$ instead of $r_n(\mathbf{x}; \mathcal{D}_n)$. Stone [37] showed that, for all $p \geq 1$, $\mathbb{E}[r_n(\mathbf{X}) - r(\mathbf{X})]^p \to 0$ for all possible distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}|Y|^p < \infty$, whenever $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Thus, the $k$-NN estimate behaves asymptotically well, without exceptions. This property is called $L_p$ universal consistency.

Clearly, any affine transformation of the coordinate axes influences the $k$-NN estimate through the norm $\|.\|$, thereby illuminating an unpleasant face of the procedure. To illustrate this remark, assume that a nontrivial affine transformation $T : \mathbf{z} \mapsto A\mathbf{z}+\mathbf{b}$ (that is, a nonsingular linear transformation $A$ followed by a translation $\mathbf{b}$) is applied to both $\mathbf{x}$ and $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Examples include any number of combinations of rotations, translations, and linear rescalings. Denote by $\mathcal{D}'_n = (T(\mathbf{X}_1), Y_1), \ldots, (T(\mathbf{X}_n), Y_n)$ the transformed sample. Then, for such a function $T$, one has $r_n(\mathbf{x}; \mathcal{D}_n) \neq r_n(T(\mathbf{x}); \mathcal{D}'_n)$ in general, whereas $r(\mathbf{X}) = \mathbb{E}[Y|T(\mathbf{X})]$ since $T$ is bijective. Thus, to continue our discussion, we are looking in essence for a regression estimate $r_n$ with the following property:

$$r_n(\mathbf{x}; \mathcal{D}_n) = r_n(T(\mathbf{x}); \mathcal{D}'_n). \tag{1.1}$$

We call $r_n$ affine invariant. Affine invariance is indeed a very strong but highly desirable property. In $\mathbb{R}^d$, in the context of $k$-NN estimates, it suffices to be able to define an affine invariant distance measure, which is necessarily data-dependent. With this objective in mind, we develop in the next section an estimation procedure featuring (1.1) which in form coincides with the $k$-NN estimate, and establish its consistency in Section 3. Proofs of the most technical results are gathered in Section 4.

It should be stressed that what we are after in this article is an estimate of $r$ which is invariant by an affine transformation of *both* the query point $\mathbf{x}$ and the original regressors $\mathbf{X}_1, \ldots, \mathbf{X}_n$. When the sole regressors are subject to such a transformation, it is then more natural to talk of "affine equivariant" regression estimates rather than of "affine invariant" ones; this is more in line with the terminology used, for example, in Ollila, Hettmansperger, and Oja [28] and Ollila, Oja, and Koivunen [29]. These affine invariance and affine equivariance requirements, however, are strictly equivalent.

There have been many attempts in the nonparametric literature to achieve affine invariance. One of the most natural ones relates to the so-called transformation-retransformation proposed by Chakraborty, Chaudhuri, and Oja [3]. That method and many variants have been discussed in texts such as [9] and [17] for pattern recognition and regression, respectively, but they have also been used in kernel density estimation (see, e.g., Samanta [36]). It is worth noting that, computational issues aside, the transformation step (i.e., premultiplication of the regressors by $\hat{M}_n^{-1}$, where $\hat{M}_n$ is an affine equivariant scatter estimate) may be based on a statistic $\hat{M}_n$ that does not require finiteness of any moment. A typical example is the scatter estimate proposed in Tyler [38] or Hettmansperger and Randles [20]. Rather, our procedure takes

ideas from the classical nonparametric literature using concepts such as multivariate ranks. It is closed in spirit of the approach of Paindaveine and Van Bever [31], who introduce a class of depth-based classification procedures that are of a nearest neighbor nature.

There are also attempts at getting invariance to other transformations. The most important concept here is that of invariance under monotone transformations of the coordinate axes. In particular, any strategy that uses only the coordinatewise ranks of the $\mathbf{X}_i$'s achieves this. The onus, then, is to show consistency of the methods under the most general conditions possible. For example, using an $L_p$ norm on the $d$-vectors of differences between ranks, one can show that the classical $k$-NN regression function estimate is universally consistent in the sense of Stone [37]. This was observed by Olshen [30], and shown by Devroye [8] (see also Gordon and Olshen [15, 16], Devroye and Krzyżak [10], and Biau and Devroye [2] for related works). Rules based upon statistically equivalent blocks (see, e.g., Anderson [1], Quesenberry and Gessaman [34], Gessaman [13], Gessaman and Gessaman [14], and Devroye, Györfi, and Lugosi [9, Section 21.4]) are other important examples of regression methods invariant with respect to monotone transformations of the coordinate axes. These methods and their generalizations partition the space with sets that contain a fixed number of data points each.

It would be interesting to consider in a future paper the possibility of morphing the input space in more general ways than those suggested in the previous few paragraphs of the present article. It should be possible, in principle, to define appropriate metrics to obtain invariance for interesting large classes of nonlinear transformations, and show consistent asymptotic behaviors.

# 2 An affine invariant $k$-NN estimate

The $k$-NN estimate we are discussing is based upon the notion of empirical distance. Throughout, we assume that the distribution of $\mathbf{X}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and that $n \geq d$. Because of this density assumption, any collection $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d}$ $(1 \leq i_1 < i_2 < \ldots < i_d \leq n)$ of $d$ points among $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are in general position with probability 1. Consequently, there exists with probability 1 a unique hyperplane in $\mathbb{R}^d$ containing these $d$ random points, and we denote it by $\mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})$.

With this notation, the empirical distance between $d$-vectors $\mathbf{x}$ and $\mathbf{x}'$ is

defined as

$$\rho_n(\mathbf{x}, \mathbf{x}') = \sum_{1 \le i_1 < \ldots < i_d \le n} \mathbf{1}_{\{\text{segment } (\mathbf{x},\mathbf{x}') \text{ intersects the hyperplane } \mathcal{H}(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_d})\}}.$$

Put differently, $\rho_n(\mathbf{x}, \mathbf{x}')$ just counts the number of hyperplanes in $\mathbb{R}^d$ passing through $d$ out of the points $\mathbf{X}_1, \ldots, \mathbf{X}_n$, that are separating $\mathbf{x}$ and $\mathbf{x}'$. Roughly, "near" points have fewer intersections, see Figure 1 that depicts an example in dimension 2.

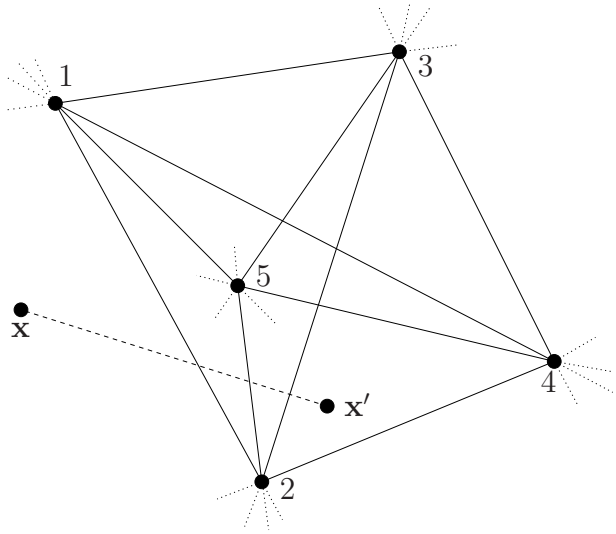


Figure 1: An example in dimension 2. The empirical distance between $\mathbf{x}$ and $\mathbf{x}'$ is 4. (Note that the hyperplane defined by the pair $(3, 5)$ indeed cuts the segment $(\mathbf{x}, \mathbf{x}')$, so that the distance is 4, not 3.)

This hyperplane-based concept of distance is known in the multivariate rank tests literature as the empirical lift-interdirection function (Oja and Paindaveine [27], see also Randles [35], Oja [26], and Hallin and Paindaveine [18] for companion concepts). It was originally mentioned (but not analyzed) in Hettmansperger, Möttönen, and Oja [19], and independently suggested as an affine invariant alternative to ordinary metrics in the monograph of Devroye, Gÿorfi, and Lugosi [9, Section 11.6]. We speak throughout of distance even though, for a fixed sample of size $n$, $\rho_n$ is only defined with probability 1 and is not a distance measure *stricto sensu* (in particular, $\rho_n(\mathbf{x}, \mathbf{x}') = 0$ does not imply that $\mathbf{x} = \mathbf{x}'$). Nevertheless, this empirical distance is invariant under affine transformations $\mathbf{x} \mapsto A\mathbf{x} + \mathbf{b}$, where $A$ is some arbitrary nonsingular linear map and $\mathbf{b}$ any offset vector (see, for instance, Oja and Paindaveine [27, Section 2.4]).

Now, fix $\mathbf{x} \in \mathbb{R}^d$ and let $\rho_n(\mathbf{x}, \mathbf{X}_i)$ be the empirical distance between $\mathbf{x}$ and some observation $\mathbf{X}_i$ in the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$. (That is, the number of hyperplanes in $\mathbb{R}^d$ passing through $d$ out of the observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$, that are cutting the segment $(\mathbf{x}, \mathbf{X}_i)$). In this context, the $k$-NN estimate we are considering still takes the familiar form

$$r_n(\mathbf{x}) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i)}(\mathbf{x}),$$

with the important difference that now the data set $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ is reordered according to increasing values of the empirical distances $\rho_n(\mathbf{x}, \mathbf{X}_i)$, *not* the original Euclidean metric. By construction, the estimate $r_n$ has the desired affine invariance property and, moreover, it coincides with the standard (Euclidean) estimate in dimension $d = 1$. In the next section, we prove the following theorem. The distribution of the random variable $\mathbf{X}$ is denoted by $\mu$.

**Theorem 2.1 (Pointwise $L_p$ consistency)** *Assume that $\mathbf{X}$ has a probability density, that $Y$ is bounded, and that the regression function $r$ is $\mu$-almost surely continuous. Then, for $\mu$-almost all $\mathbf{x} \in \mathbb{R}^d$ and all $p \geq 1$, if $k_n \to \infty$ and $k_n/n \to 0$,*

$$\mathbb{E}\,|r_n(\mathbf{x}) - r(\mathbf{x})|^p \to 0 \quad \textit{as } n \to \infty.$$

The following corollary is a consequence of Theorem 2.1 and the Lebesgue dominated convergence theorem.

**Corollary 2.1 (Global $L_p$ consistency)** *Assume that $\mathbf{X}$ has a probability density, that $Y$ is bounded, and that the regression function $r$ is $\mu$-almost surely continuous. Then, for all $p \geq 1$, if $k_n \to \infty$ and $k_n/n \to 0$,*

$$\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^p \to 0 \quad \textit{as } n \to \infty.$$

The conditions of Stone's universal consistency theorem given in [37] are not fulfilled for our estimate. For the standard nearest neighbor estimate, a key result used in the consistency proof by Stone is that a given data point cannot be the nearest neighbor of more than a constant number (say, $3^d$) other points. Such a universal constant does not exist after our transformation is applied. That means that a single data point can have a large influence on the regression function estimate. While this by itself does not imply that the estimate is not universally consistent, it certainly indicates that any such proof will require new insights. The addition of two smoothness constraints,

namely that $\mathbf{X}$ has a density (without, however, imposing any continuity conditions on the density itself) and that $r$ is $\mu$-almost surely continuous, is sufficient.

The complexity of our procedure in terms of sample size $n$ and dimension $d$ is quite high. There are $\binom{n}{d}$ possible choices of hyperplanes through $d$ points. This collection of hyperplanes defines an arrangement, or partition of $\mathbb{R}^d$ into polytopal regions, also called cells or chambers. Within each region, the distance to each data point is constant, and thus, a preprocessing step might consist of setting up a data structure for determining to which cell a given point $\mathbf{x} \in \mathbb{R}^d$ belongs: This is called the point location problem. Meiser [24] showed that such a data structure exists with the following properties: (1) it takes space $\mathcal{O}(n^{d+\varepsilon})$ for any fixed $\varepsilon > 0$, and (2) point location can be performed in $\mathcal{O}(\log n)$ time. Chazelle's cuttings [4] improve (1) to $\mathcal{O}(n^d)$. Chazelle's processing time for setting up the data structure is $\mathcal{O}(n^d)$. Still in the preprocessing step, one can determine for each cell in the arrangement the distances to all $n$ data points: This can be done by walking across the graph of cells or by brute force. When done naively, the overall set-up complexity is $\mathcal{O}(n^{2d+1})$. For each cell, one might keep a pointer to the $k$ nearest neighbors. Therefore, once set up, the computation of the regression function estimate takes merely $\mathcal{O}(\log n)$ time for point location, and $\mathcal{O}(k)$ time for retrieving the $k$ nearest neighbors.

One could envisage a reduction in the complexity by defining the distances not in terms of all hyperplanes that cut a line segment, but in terms of the number of randomly drawn hyperplanes that make such a cut, where the number of random draws is now a carefully selected number. By the concentration of binomial random variables, such random estimates of the distances are expected to work well, while keeping the complexity reasonable. This idea will be explored elsewhere.

# 3 Proof of the theorem

Recall, since $\mathbf{X}$ has a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$, that any collection $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d}$ $(1 \leq i_1 < i_2 < \ldots < i_d \leq n)$ of $d$ points among $\mathbf{X}_1, \ldots, \mathbf{X}_n$ defines with probability 1 a unique hyperplane $\mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})$ in $\mathbb{R}^d$. Thus, in the sequel, since no confusion is possible, we will freely refer to "the hyperplane $\mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})$ defined by $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d}$" without further explicit mention of the probability 1 event.

Let us first fix some useful notation. The distribution of the random variable $\mathbf{X}$ is denoted by $\mu$ and its density with respect to the Lebesgue measure is

denoted by $f$. For every $\varepsilon > 0$, we let $\mathcal{B}_{\mathbf{x},\varepsilon} = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{x}\| \leq \varepsilon\}$ be the closed Euclidean ball with center at $\mathbf{x}$ and radius $\varepsilon$. We write $A^c$ for the complement of a subset $A$ of $\mathbb{R}^d$. For two random variables $Z_1$ and $Z_2$, the notation

$$Z_1 \leq_{\text{st}} Z_2$$

means that $Z_1$ is stochastically dominated by $Z_2$, that is, for all $t \in \mathbb{R}$,

$$\mathbb{P}\{Z_1 > t\} \leq \mathbb{P}\{Z_2 > t\}.$$

Our first goal is to show that for $\mu$-almost all $\mathbf{x}$, as $k_n/n \to 0$, the quantity $\max_{i=1,\ldots,k_n} \|\mathbf{X}_{(i)}(\mathbf{x}) - \mathbf{x}\|$ converges to 0 in probability, i.e., for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left\{ \max_{i=1,\ldots,k_n} \|\mathbf{X}_{(i)}(\mathbf{x}) - \mathbf{x}\| > \varepsilon \right\} = 0. \tag{3.1}$$

So, fix such a positive $\varepsilon$. Let $\delta$ be a real number in $(0, \varepsilon)$ and $\gamma_n$ be a positive real number (eventually function of $\mathbf{x}$ and $\varepsilon$) to be determined later. To prove identity (3.1), we use the following decomposition, which is valid for all $\mathbf{x} \in \mathbb{R}^d$:

$$\begin{aligned}
&\mathbb{P}\left\{ \max_{i=1,\ldots,k_n} \|\mathbf{X}_{(i)}(\mathbf{x}) - \mathbf{x}\| > \varepsilon \right\} \\
&\quad \leq \mathbb{P}\left\{ \min_{\substack{i=1,\ldots,n \\ \mathbf{X}_i \in \mathcal{B}^c_{\mathbf{x},\varepsilon}}} \rho_n(\mathbf{x}, \mathbf{X}_i) < \gamma_n \right\} \\
&\qquad + \mathbb{P}\left\{ \max_{\substack{i=1,\ldots,n \\ \mathbf{X}_i \in \mathcal{B}_{\mathbf{x},\delta}}} \rho_n(\mathbf{x}, \mathbf{X}_i) \geq \gamma_n \right\} \\
&\qquad + \mathbb{P}\left\{ \text{Card}\left\{ i = 1, \ldots, n : \|\mathbf{X}_i - \mathbf{x}\| \leq \delta \right\} < k_n \right\} \\
&\quad := \mathbf{A} + \mathbf{B} + \mathbf{C}.
\end{aligned} \tag{3.2}$$

The convergence to 0 of each of the three terms above—from which identity (3.1) immediately follows—are separately analyzed in the next three paragraphs.

**Analysis of A.** As for now, taking an affine geometry point of view, we keep $\mathbf{x}$ fixed and see it as the origin of the space. Recall that each point in the Euclidean space $\mathbb{R}^d$ (with the origin at $\mathbf{x}$) may be described by its hyperspherical coordinates (see, e.g., Miller [25, Chapter 1]), which consist of a nonnegative radial coordinate $r$ and $d-1$ angular coordinates $\theta_1, \ldots, \theta_{d-1}$,

where $\theta_{d-1}$ ranges over $[0, 2\pi)$ and the other angles range over $[0, \pi]$ (adaptation of this definition to the cases $d = 1$ and $d = 2$ is clear). For a $(d-1)$-dimensional vector $\Theta = (\theta_1, \ldots, \theta_{d-1})$ of hyperspherical angles, we let $\mathcal{B}_{\mathbf{x},\varepsilon}(\Theta)$ be the unique closed ball anchored at $\mathbf{x}$ in the direction $\Theta$ and with diameter $\varepsilon$ (see Figure 2 which depicts an illustration in dimension 2). We also let $\mathcal{L}_{\mathbf{x}}(\Theta)$ be the axe defined by $\mathbf{x}$ and the direction $\Theta$, and let as well $\mathcal{S}_{\mathbf{x},\varepsilon}(\Theta)$ be the open segment obtained as the intersection of $\mathcal{L}_{\mathbf{x}}(\Theta)$ and the interior of $\mathcal{B}_{\mathbf{x},\varepsilon}(\Theta)$.

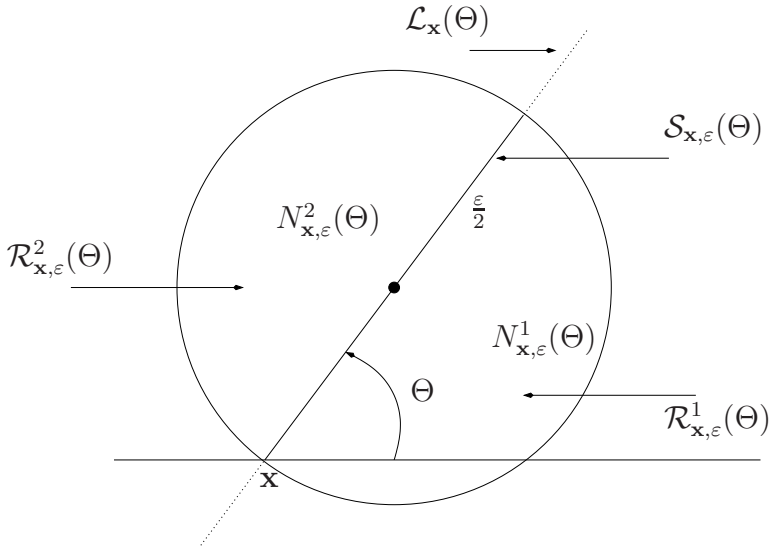


Figure 2: The ball $\mathcal{B}_{\mathbf{x},\varepsilon}(\Theta)$ and related notation. Illustration in dimension 2.

Next, for fixed $\mathbf{x}$, $\varepsilon$ and $\Theta$, we split the ball $\mathcal{B}_{\mathbf{x},\varepsilon}(\Theta)$ into $2^{d-1}$ disjoint regions $\mathcal{R}^1_{\mathbf{x},\varepsilon}(\Theta), \ldots, \mathcal{R}^{2^{d-1}}_{\mathbf{x},\varepsilon}(\Theta)$ as follows. First, the Euclidean space $\mathbb{R}^d$ is sequentially divided into $2^{d-1}$ symmetric quadrants rotating around the axe $\mathcal{L}_{\mathbf{x}}(\Theta)$ (boundary equalities are broken arbitrarily). Next, each region $\mathcal{R}^j_{\mathbf{x},\varepsilon}(\Theta)$ is obtained as the intersection of one of the $2^{d-1}$ quadrants and the ball $\mathcal{B}_{\mathbf{x},\varepsilon}(\Theta)$.

The numbers of sample points falling in each of these regions are denoted hereafter by $N^1_{\mathbf{x},\varepsilon}(\Theta), \ldots, N^{2^{d-1}}_{\mathbf{x},\varepsilon}(\Theta)$ (see Figure 2). Letting finally $V_d$ be the volume of the unit $d$-dimensional Euclidean ball, we are now in a position to control the first term of inequality (3.2).

**Proposition 3.1** *For $\mu$-almost all $\mathbf{x} \in \mathbb{R}^d$ and all $\varepsilon > 0$ small enough,*

$$\mathbb{P}\left\{ \min_{\substack{i=1,\ldots,n \\ \mathbf{X}_i \in \mathcal{B}^c_{\mathbf{x},\varepsilon}}} \rho_n(\mathbf{x}, \mathbf{X}_i) < \gamma_n \right\} \to 0 \quad \textit{as } n \to \infty,$$

*provided*

$$\gamma_n = n^d \left( \frac{V_d}{2^{2d+1}} \varepsilon^d f(\mathbf{x}) \right)^{2^{d-1}} .$$

**Proof of Proposition 3.1** Set

$$p_{\mathbf{x},\varepsilon} = \min_{j=1,\ldots,2^{d-1}} \inf_{\Theta} \mu \left\{ \mathcal{R}^j_{\mathbf{x},\varepsilon}(\Theta) \right\},$$

where the infimum is taken over all possible hyperspherical angles $\Theta$. We know, according to technical Lemma 4.1, that for $\mu$-almost all $\mathbf{x}$ and all $\varepsilon > 0$ small enough,

$$p_{\mathbf{x},\varepsilon} \geq \frac{V_d}{2^{2d}} \varepsilon^d f(\mathbf{x}) > 0. \tag{3.3}$$

Thus, in the rest of the proof, we fix such an $\mathbf{x}$ and assume that $\varepsilon$ is small enough so that the inequalities above are satisfied.

Let $\mathbf{X}^\star$ be defined as the intersection of the line $(\mathbf{x}, \mathbf{X})$ with $\mathcal{B}_{\mathbf{x},\varepsilon}$, and let $\Theta^\star$ be the (random) hyperspherical angle corresponding to $\mathbf{X}^\star$ (see Figure 3 for an example in dimension 2).

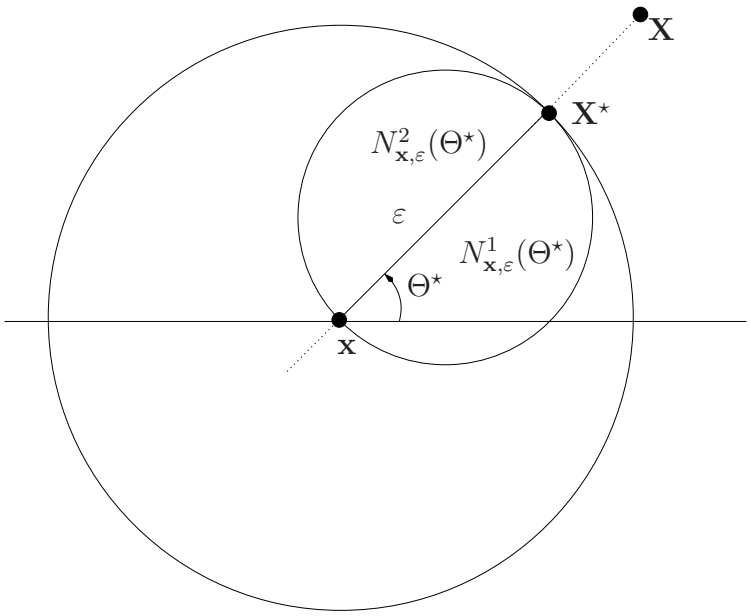


Figure 3: The ball $\mathcal{B}_{\mathbf{x},\varepsilon}(\Theta^\star)$ in dimension 2.

Denote by $N_{\mathbf{x},\varepsilon}(\Theta^\star)$ the number of hyperplanes passing through $d$ out of the

observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and cutting the segment $\mathcal{S}_{\mathbf{x},\varepsilon}(\Theta^\star)$. We have

$$\begin{aligned}
\mathbb{P}\left\{ \min_{\substack{i=1,\ldots,n \\ \mathbf{X}_i \in \mathcal{B}^c_{\mathbf{x},\varepsilon}}} \rho_n(\mathbf{x}, \mathbf{X}_i) < \gamma_n \right\} &\leq n\mathbb{P}\left\{\rho_n(\mathbf{x}, \mathbf{X}^\star) < \gamma_n\right\} \\
&= n\mathbb{P}\left\{N_{\mathbf{x},\varepsilon}(\Theta^\star) < \gamma_n\right\} \\
&\leq n\mathbb{P}\left\{ \frac{N^1_{\mathbf{x},\varepsilon}(\Theta^\star) \ldots N^{2^{d-1}}_{\mathbf{x},\varepsilon}(\Theta^\star)}{n^{2^{d-1}-d}} < \gamma_n \right\},
\end{aligned}$$

where the last inequality follows from technical Lemma 4.2. Thus,

$$\begin{aligned}
\mathbb{P}\left\{ \min_{\substack{i=1,\ldots,n \\ \mathbf{X}_i \in \mathcal{B}^c_{\mathbf{x},\varepsilon}}} \rho_n(\mathbf{x}, \mathbf{X}_i) < \gamma_n \right\} &\leq n \sum_{j=1}^{2^{d-1}} \mathbb{P}\left\{ N^j_{\mathbf{x},\varepsilon}(\Theta^\star) < \left(\gamma_n n^{2^{d-1}-d}\right)^{1/2^{d-1}} \right\} \\
&= n \sum_{j=1}^{2^{d-1}} \mathbb{P}\left\{ N^j_{\mathbf{x},\varepsilon}(\Theta^\star) < \gamma_n^{1/2^{d-1}} n^{1-d/2^{d-1}} \right\}.
\end{aligned}$$

Clearly, conditionally on $\Theta^\star$, each $N^j_{\mathbf{x},\varepsilon}(\Theta^\star)$ satisfies

$$\text{Binomial } (n, p_{\mathbf{x},\varepsilon}) \leq_{\text{st}} N^j_{\mathbf{x},\varepsilon}(\Theta^\star)$$

and consequently, by inequality (3.3),

$$\text{Binomial } \left(n, \frac{V_d}{2^{2d}}\varepsilon^d f(\mathbf{x})\right) \leq_{\text{st}} N^j_{\mathbf{x},\varepsilon}(\Theta^\star).$$

Thus, for each $j = 1, \ldots, 2^{d-1}$, by Hoeffding's inequality for binomial random variables (Hoeffding [21]), we are led to

$$\begin{aligned}
&\mathbb{P}\left\{ N^j_{\mathbf{x},\varepsilon}(\Theta^\star) < \gamma_n^{1/2^{d-1}} n^{1-d/2^{d-1}} \right\} \\
&\quad = \mathbb{E}\left[ \mathbb{P}\left\{ N^j_{\mathbf{x},\varepsilon}(\Theta^\star) < \gamma_n^{1/2^{d-1}} n^{1-d/2^{d-1}} | \Theta^\star \right\} \right] \\
&\quad \leq \exp\left[ -2\left( \gamma_n^{1/2^{d-1}} n^{1-d/2^{d-1}} - n\frac{V_d}{2^{2d}}\varepsilon^d f(\mathbf{x}) \right)^2 / n \right]
\end{aligned}$$

as soon as $\gamma_n^{1/2^{d-1}} n^{1-d/2^{d-1}} < n\frac{V_d}{2^{2d}}\varepsilon^d f(\mathbf{x})$. Therefore, taking

$$\gamma_n = n^d \left( \frac{V_d}{2^{2d+1}}\varepsilon^d f(\mathbf{x}) \right)^{2^{d-1}},$$

we obtain

$$\mathbb{P}\left\{\min_{\substack{i=1,\dots,n\\ \mathbf{X}_i\in\mathcal{B}^c_{\mathbf{x},\varepsilon}}}\rho_n(\mathbf{x},\mathbf{X}_i)<\gamma_n\right\}\leq 2^{d-1}n\exp\left[-n\left(\frac{V_d}{2^{2d}}\varepsilon^d f(\mathbf{x})\right)^2/2\right].$$

The upper bound goes to 0 as $n\to\infty$. ■

**Analysis of B.** Consistency of the second term in inequality (3.2) is established in the following proposition.

**Proposition 3.2** *For $\mu$-almost all $\mathbf{x}\in\mathbb{R}^d$, all $\varepsilon>0$ and all $\delta>0$ small enough,*

$$\mathbb{P}\left\{\max_{\substack{i=1,\dots,n\\ \mathbf{X}_i\in\mathcal{B}_{\mathbf{x},\delta}}}\rho_n(\mathbf{x},\mathbf{X}_i)\geq\gamma_n\right\}\to 0\quad \textit{as } n\to\infty,$$

*provided*

$$\gamma_n=n^d\left(\frac{V_d}{2^{2d+1}}\varepsilon^d f(\mathbf{x})\right)^{2^{d-1}}. \tag{3.4}$$

**Proof of Proposition 3.2** Fix $\mathbf{x}$ in a set of $\mu$-measure 1 such that $f(\mathbf{x})>0$ and denote by $N_{\mathbf{x},\delta}$ the number of hyperplanes that cut the ball $\mathcal{B}_{\mathbf{x},\delta}$. Clearly,

$$\mathbb{P}\left\{\max_{\substack{i=1,\dots,n\\ \mathbf{X}_i\in\mathcal{B}_{\mathbf{x},\delta}}}\rho_n(\mathbf{x},\mathbf{X}_i)\geq\gamma_n\right\}\leq\mathbb{P}\left\{N_{\mathbf{x},\delta}\geq\gamma_n\right\}.$$

Observe that, with probability 1,

$$N_{\mathbf{x},\delta}=\sum_{1\leq i_1<\dots<i_d\leq n}\mathbf{1}_{\{\mathcal{H}(\mathbf{X}_{i_1},\dots,\mathbf{X}_{i_d})\cap\mathcal{B}_{\mathbf{x},\delta}\neq\emptyset\}},$$

whence, since $\mathbf{X}_1,\dots,\mathbf{X}_n$ are identically distributed,

$$\begin{aligned}\mathbb{E}[N_{\mathbf{x},\delta}]&=\binom{n}{d}\mathbb{P}\left\{\mathcal{H}(\mathbf{X}_1,\dots,\mathbf{X}_d)\cap\mathcal{B}_{\mathbf{x},\delta}\neq\emptyset\right\}\\&\leq\frac{n^d}{d!}\mathbb{P}\left\{\mathcal{H}(\mathbf{X}_1,\dots,\mathbf{X}_d)\cap\mathcal{B}_{\mathbf{x},\delta}\neq\emptyset\right\}.\end{aligned}$$

Consequently, given the choice (3.4) for $\gamma_n$ and the result of technical Lemma 4.3, it follows that

$$\mathbb{E}[N_{\mathbf{x},\delta}]<\gamma_n/2$$

for all $\delta$ small enough, independently of $n$. Thus, using the bounded difference inequality (McDiarmid [23]), we obtain, still with the choice

$$\gamma_n = n^d \left( \frac{V_d}{2^{2d+1}} \varepsilon^d f(\mathbf{x}) \right)^{2^{d-1}},$$

$$\begin{aligned}
\mathbb{P}\left\{ N_{\mathbf{x},\delta} \geq \gamma_n \right\} &\leq \mathbb{P}\left\{ N_{\mathbf{x},\delta} - \mathbb{E}[N_{\mathbf{x},\delta}] \geq \gamma_n/2 \right\} \\
&\leq \exp\left( -2 \frac{(\gamma_n/2)^2}{n^{2d-1}} \right) \\
&= \exp\left[ - \left( \frac{V_d}{2^{2d+1}} \varepsilon^d f(\mathbf{x}) \right)^{2^d} n/2 \right].
\end{aligned}$$

This upper bound goes to zero as $n$ tends to infinity, and this concludes the proof of the proposition. ■

**Analysis of C.** To achieve the proof of identity (3.1), it remains to show that the third and last term of (3.2) converges to 0. This is done in the following proposition.

**Proposition 3.3** *Assume that $k_n/n \to 0$ as $n \to \infty$. Then, for $\mu$-almost all $\mathbf{x} \in \mathbb{R}^d$ and all $\delta > 0$,*

$$\mathbb{P}\left\{ \text{Card}\left\{ i = 1, \ldots, n : \|\mathbf{X}_i - \mathbf{x}\| \leq \delta \right\} < k_n \right\} \to 0 \quad \text{as } n \to \infty.$$

**Proof of Proposition 3.3** Recall that the collection of all $\mathbf{x}$ with $\mu(\mathcal{B}_{\mathbf{x},\tau}) > 0$ for all $\tau > 0$ is called the support of $\mu$, and note that it may alternatively be defined as the smallest closed subset of $\mathbb{R}^d$ of $\mu$-measure 1 (Parthasarathy [32, Chapter 2]). Thus, fix $\mathbf{x}$ in the support of $\mu$ and set

$$p_{\mathbf{x},\delta} = \mathbb{P}\{\mathbf{X} \in \mathcal{B}_{\mathbf{x},\delta}\},$$

so that $p_{\mathbf{x},\delta} > 0$. Then the following chain of inequalities is valid:

$$\begin{aligned}
&\mathbb{P}\left\{ \text{Card}\left\{ i = 1, \ldots, n : \|\mathbf{X}_i - \mathbf{x}\| \leq \delta \right\} < k_n \right\} \\
&\quad = \mathbb{P}\left\{ \text{Binomial}\ (n, p_{\mathbf{x},\delta}) < k_n \right\} \\
&\quad \leq \mathbb{P}\left\{ \text{Binomial}\ (n, p_{\mathbf{x},\delta}) \leq n p_{\mathbf{x},\delta}/2 \right\} \\
&\qquad \text{(for all } n \text{ large enough, since } k_n/n \text{ tends to 0)} \\
&\quad \leq \exp(-n p_{\mathbf{x},\delta}^2/2),
\end{aligned}$$

where the last inequality follows from Hoeffding's inequality (Hoeffding [21]). This terminates the proof of Proposition 3.3. ■

We have proved so far that, for $\mu$-almost all $\mathbf{x}$, as $k_n/n \to 0$, the quantity $\max_{i=1,\ldots,k_n} \|\mathbf{X}_{(i)}(\mathbf{x}) - \mathbf{x}\|$ converges to 0 in probability. By the elementary inequality

$$\mathbb{E}\left[\frac{1}{k_n}\sum_{i=1}^{k_n} \mathbf{1}_{\{\|\mathbf{X}_{(i)}(\mathbf{x})-\mathbf{x}\|>\varepsilon\}}\right] \leq \mathbb{P}\left\{\max_{i=1,\ldots,k_n} \|\mathbf{X}_{(i)}(\mathbf{x}) - \mathbf{x}\| > \varepsilon\right\},$$

it immediately follows that, for such an $\mathbf{x}$,

$$\mathbb{E}\left[\frac{1}{k_n}\sum_{i=1}^{k_n} \mathbf{1}_{\{\|\mathbf{X}_{(i)}(\mathbf{x})-\mathbf{x}\|>\varepsilon\}}\right] \to 0 \tag{3.5}$$

provided $k_n/n \to 0$. We are now ready to complete the proof of Theorem 2.1.

Fix $\mathbf{x}$ in a set of $\mu$-measure 1 such that consistency (3.5) holds and $r$ is continuous at $\mathbf{x}$ (this is possible by the assumption on $r$). Because $|a+b|^p \leq 2^{p-1}(|a|^p + |b|^p)$ for $p \geq 1$, we see that

$$\begin{aligned}\mathbb{E}\,|r_n(\mathbf{x}) - r(\mathbf{x})|^p &\leq 2^{p-1}\mathbb{E}\left|\frac{1}{k_n}\sum_{i=1}^{k_n}\left[Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right]\right|^p \\ &\quad + 2^{p-1}\mathbb{E}\left|\frac{1}{k_n}\sum_{i=1}^{k_n}\left[r\left(\mathbf{X}_{(i)}(\mathbf{x})\right) - r(\mathbf{x})\right]\right|^p.\end{aligned}$$

Thus, by Jensen's inequality,

$$\begin{aligned}\mathbb{E}\,|r_n(\mathbf{x}) - r(\mathbf{x})|^p &\leq 2^{p-1}\mathbb{E}\left|\frac{1}{k_n}\sum_{i=1}^{k_n}\left[Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right]\right|^p \\ &\quad + 2^{p-1}\mathbb{E}\left[\frac{1}{k_n}\sum_{i=1}^{k_n}\left|r\left(\mathbf{X}_{(i)}(\mathbf{x})\right) - r(\mathbf{x})\right|^p\right] \\ &:= 2^{p-1}\mathbf{I}_n + 2^{p-1}\mathbf{J}_n.\end{aligned}$$

Firstly, for arbitrary $\varepsilon > 0$, we have

$$\begin{aligned}\mathbf{J}_n &= \mathbb{E}\left[\frac{1}{k_n}\sum_{i=1}^{k_n}\left|r\left(\mathbf{X}_{(i)}(\mathbf{x})\right) - r(\mathbf{x})\right|^p \mathbf{1}_{\{\|\mathbf{X}_{(i)}(\mathbf{x})-\mathbf{x}\|>\varepsilon\}}\right] \\ &\quad + \mathbb{E}\left[\frac{1}{k_n}\sum_{i=1}^{k_n}\left|r\left(\mathbf{X}_{(i)}(\mathbf{x})\right) - r(\mathbf{x})\right|^p \mathbf{1}_{\{\|\mathbf{X}_{(i)}(\mathbf{x})-\mathbf{x}\|\leq\varepsilon\}}\right],\end{aligned}$$

whence

$$\begin{aligned}\mathbf{J}_n &\leq 2^p \zeta^p \, \mathbb{E}\left[\frac{1}{k_n}\sum_{i=1}^{k_n} \mathbf{1}_{\{\|\mathbf{X}_{(i)}(\mathbf{x})-\mathbf{x}\|>\varepsilon\}}\right] \\ &\quad + \left[\sup_{\mathbf{y}\in\mathbb{R}^d:\|\mathbf{y}-\mathbf{x}\|\leq\varepsilon} |r(\mathbf{y})-r(\mathbf{x})|\right]^p \\ &\quad \text{(since } |Y|\leq\zeta\text{)}.\end{aligned}$$

The first term on the right-hand side of the latter inequality tends to 0 by (3.5) as $k_n/n \to 0$, whereas the rightmost one can be made arbitrarily small as $\varepsilon \to 0$ since $r$ is continuous at $\mathbf{x}$. This proves that $\mathbf{J}_n \to 0$ as $n \to \infty$.

Next, by successive applications of inequalities of Marcinkiewicz and Zygmund [22] (see also Petrov [33, pages 59-60]), we have for some positive constant $C_p$ depending only on $p$,

$$\begin{aligned}\mathbf{I}_n &\leq C_p \, \mathbb{E}\left[\frac{1}{k_n^2}\sum_{i=1}^{k_n} \left|Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right|^2\right]^{p/2} \\ &\leq \frac{(2\zeta)^p C_p}{k_n^{p/2}} \\ &\quad \text{(since } |Y|\leq\zeta\text{)}.\end{aligned}$$

Consequently, $\mathbf{I}_n \to 0$ as $k_n \to \infty$, and this concludes the proof of the theorem.

# 4 Some technical lemmas

The notation of this section is identical to that of Section 3. In particular, it is assumed throughout that $\mathbf{X}$ has a probability density $f$ with respect to the Lebesgue measure $\lambda$ on $\mathbb{R}^d$. This requirement implies that any collection $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d}$ $(1 \leq i_1 < i_2 < \ldots < i_d \leq n)$ of $d$ points among $\mathbf{X}_1, \ldots, \mathbf{X}_n$ define with probability 1 a unique hyperplane $\mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})$ in $\mathbb{R}^d$. Recall finally that, for $\mathbf{x} \in \mathbb{R}^d$ and $\varepsilon > 0$, we set

$$p_{\mathbf{x},\varepsilon} = \min_{j=1,\ldots,2^{d-1}} \inf_{\Theta} \mu\left\{\mathcal{R}^j_{\mathbf{x},\varepsilon}(\Theta)\right\},$$

where the infimum is taken over all possible hyperspherical angles $\Theta$, and the regions $\mathcal{R}^j_{\mathbf{x},\varepsilon}(\Theta)$, $j = 1, \ldots, 2^{d-1}$, define a partition of the ball $\mathcal{B}_{\mathbf{x},\varepsilon}(\Theta)$. Recall also that the numbers of sample points falling in each of these regions are denoted by $N^1_{\mathbf{x},\varepsilon}(\Theta), \ldots, N^{2^{d-1}}_{\mathbf{x},\varepsilon}(\Theta)$. For a better understanding of the next lemmas, the reader should refer to Figure 2 and Figure 3.

**Lemma 4.1** *For $\mu$-almost all $\mathbf{x} \in \mathbb{R}^d$ and all $\varepsilon > 0$ small enough,*

$$p_{\mathbf{x},\varepsilon} \geq \frac{V_d}{2^{2d}} \varepsilon^d f(\mathbf{x}) > 0.$$

**Proof of Lemma 4.1** We let $\mathbf{x}$ be a Lebesgue point of $f$, that is, an $\mathbf{x}$ such that for any collection $\mathcal{A}$ of subsets of $\mathcal{B}_{\mathbf{0},1}$ with the property that for all $A \in \mathcal{A}$, $\lambda(A) \geq c\lambda(\mathcal{B}_{\mathbf{0},1})$ for some fixed $c > 0$,

$$\lim_{\varepsilon \to 0} \sup_{A \in \mathcal{A}} \left| \frac{\displaystyle\int_{\mathbf{x}+\varepsilon A} f(\mathbf{y}) \mathrm{d}\mathbf{y}}{\lambda\{\mathbf{x} + \varepsilon A\}} - f(\mathbf{x}) \right| = 0, \tag{4.1}$$

where $\mathbf{x} + \varepsilon A = \{\mathbf{y} \in \mathbb{R}^d : (\mathbf{y} - \mathbf{x})/\varepsilon \in A\}$. As $f$ is a density, we know that $\mu$-almost all $\mathbf{x}$ satisfy this property (see, for instance, Wheeden and Zygmund [39]). Moreover, since $f$ is $\mu$-almost surely positive, we may also assume that $f(\mathbf{x}) > 0$.

Thus, keep such an $\mathbf{x}$ fixed. Fix also $j \in \{1, \ldots, 2^{d-1}\}$, and set

$$p^j_{\mathbf{x},\varepsilon} = \inf_{\Theta} \mu \left\{ \mathcal{R}^j_{\mathbf{x},\varepsilon}(\Theta) \right\}.$$

Taking for $\mathcal{A}$ the collection of regions $\mathcal{R}^j_{\mathbf{0},1}(\Theta)$ when the hyperspherical angle $\Theta$ varies, that is,

$$\mathcal{A} = \left\{ \mathcal{R}^j_{\mathbf{0},1}(\Theta) : \Theta \in [0, \pi]^{d-2} \times [0, 2\pi) \right\},$$

and observing that

$$\lambda \left\{ \mathcal{R}^j_{\mathbf{x},\varepsilon}(\Theta) \right\} = \frac{V_d}{2^{d-1}} \left( \frac{\varepsilon}{2} \right)^d,$$

we may write, for each $j = 1, \ldots, 2^{d-1}$,

$$\begin{aligned}
\left| \frac{2^{d-1} p^j_{\mathbf{x},\varepsilon}}{V_d(\varepsilon/2)^d} - f(\mathbf{x}) \right| &= \left| \inf_{\Theta} \frac{\mu \left\{ \mathcal{R}^j_{\mathbf{x},\varepsilon}(\Theta) \right\}}{\lambda \left\{ \mathcal{R}^j_{\mathbf{x},\varepsilon}(\Theta) \right\}} - f(\mathbf{x}) \right| \\
&= \left| \inf_{A \in \mathcal{A}} \frac{\displaystyle\int_{\mathbf{x}+\varepsilon A} f(\mathbf{y}) \mathrm{d}\mathbf{y}}{\lambda\{\mathbf{x} + \varepsilon A\}} - f(\mathbf{x}) \right| \\
&\leq \sup_{A \in \mathcal{A}} \left| \frac{\displaystyle\int_{\mathbf{x}+\varepsilon A} f(\mathbf{y}) \mathrm{d}\mathbf{y}}{\lambda\{\mathbf{x} + \varepsilon A\}} - f(\mathbf{x}) \right|.
\end{aligned}$$

The conclusion follows from identity (4.1). ■

**Lemma 4.2** *Fix* $\mathbf{x} \in \mathbb{R}^d$, $\varepsilon > 0$ *and* $\Theta \in [0,\pi]^{d-2} \times [0, 2\pi)$. *Let* $N_{\mathbf{x},\varepsilon}(\Theta)$ *be the number of hyperplanes passing through* $d$ *out of the observations* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *and cutting the segment* $\mathcal{S}_{\mathbf{x},\varepsilon}(\Theta)$. *Then, with probability 1,*

$$N_{\mathbf{x},\varepsilon}(\Theta) \geq \frac{N^1_{\mathbf{x},\varepsilon}(\Theta) \ldots N^{2^{d-1}}_{\mathbf{x},\varepsilon}(\Theta)}{n^{2^{d-1}-d}}.$$

**Proof of Lemma 4.2** If one of the $N^j_{\mathbf{x},\varepsilon}(\Theta)$ $(j = 1, \ldots, 2^{d-1})$ is zero, then the result is trivial. Thus, in the rest of the proof, we suppose that each $N^j_{\mathbf{x},\varepsilon}(\Theta)$ is positive and note that this implies $n \geq 2^{d-1}$.

Pick sequentially $2^{d-1}$ observations, say $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{2^{d-1}}}$, in the $2^{d-1}$ regions $\mathcal{R}^1_{\mathbf{x},\varepsilon}(\Theta), \ldots, \mathcal{R}^{2^{d-1}}_{\mathbf{x},\varepsilon}(\Theta)$. By construction, the polytope defined by these $2^{d-1}$ points cuts the axe $\mathcal{L}_{\mathbf{x},\varepsilon}(\Theta)$. Consequently, with probability 1, any hyperplane drawn according to $d$ out of these $2^{d-1}$ points cuts the segment $\mathcal{S}_{\mathbf{x},\varepsilon}(\Theta)$. The result follows by observing that there are exactly $N^1_{\mathbf{x},\varepsilon}(\Theta) \ldots N^{2^{d-1}}_{\mathbf{x},\varepsilon}(\Theta)$ such polytopes. ■

**Lemma 4.3** *For* $1 \leq i_1 < \ldots < i_d \leq n$, *let* $\mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})$ *be the hyperplane passing through* $d$ *out of the observations* $\mathbf{X}_1, \ldots, \mathbf{X}_n$. *Then, for all* $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbb{P}\left\{\mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d}) \cap \mathcal{B}_{\mathbf{x},\delta} \neq \emptyset\right\} \to 0 \quad \text{as } \delta \downarrow 0.$$

**Proof of Lemma 4.3** Given two hyperplanes $\mathcal{H}$ and $\mathcal{H}'$ in $\mathbb{R}^d$, we denote by $\Phi(\mathcal{H}, \mathcal{H}')$ the (dihedral) angle between $\mathcal{H}$ and $\mathcal{H}'$. Recall that $\Phi(\mathcal{H}, \mathcal{H}') \in [0, \pi]$ and that it is defined as the angle between the corresponding normal vectors.

Fix $1 \leq i_1 < \ldots < i_d \leq n$. Let $\mathcal{E}_\delta$ be the event

$$\mathcal{E}_\delta = \left\{\|\mathbf{X}_{i_j} - \mathbf{x}\| > \delta : j = 1, \ldots, d-1\right\},$$

and let $\mathcal{H}(\mathbf{x}, \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}})$ be the hyperplane passing through $\mathbf{x}$ and the $d-1$ points $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}$. Clearly, on $\mathcal{E}_\delta$, the event $\{\mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d}) \cap \mathcal{B}_{\mathbf{x},\delta} \neq \emptyset\}$ is the same as

$$\left\{\Phi\left(\mathcal{H}(\mathbf{x}, \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}), \mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})\right) \leq \Phi_\delta\right\},$$

where $\Phi_\delta$ is the angle formed by $\mathcal{H}(\mathbf{x}, \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}})$ and the hyperplane going trough $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}$ and tangent to $\mathcal{B}_{\mathbf{x},\delta}$ (see Figure 4 for an example in dimension 2).

Thus, with this notation, we may write

$$\begin{aligned}
&\mathbb{P}\left\{\mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d}) \cap \mathcal{B}_{\mathbf{x},\delta} \neq \emptyset\right\} \\
&\quad \leq \mathbb{P}\{\mathcal{E}_n^c\} + \mathbb{P}\left\{\mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d}) \cap \mathcal{B}_{\mathbf{x},\delta} \neq \emptyset, \mathcal{E}_n\right\} \\
&\quad \leq \mathbb{P}\{\mathcal{E}_n^c\} + \mathbb{P}\left\{\Phi\left(\mathcal{H}(\mathbf{x}, \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}), \mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})\right) \leq \Phi_\delta, \mathcal{E}_n\right\}.
\end{aligned}$$

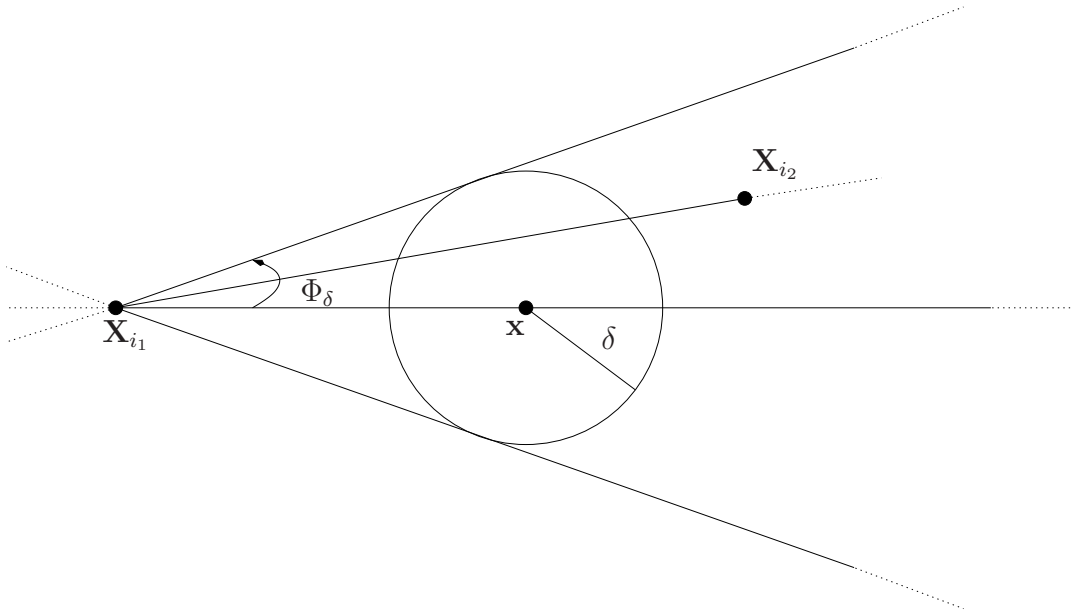


Figure 4: The hyperplanes $\mathcal{H}(\mathbf{x}, \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}})$ and $\mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})$, and the angle $\Phi_\delta$. Illustration in dimension 2.

Since $\mathbf{X}$ has a density, the first of the two terms above tends to zero as $\delta \downarrow 0$. To analyze the second term, first note that, conditionally on $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}$, the angle $\Phi(\mathcal{H}(\mathbf{x}, \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}), \mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d}))$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. This follows from the following two observations: *i*) the random variable $\mathbf{X}_{i_d}$ has a density with respect to the Lebesgue measure on $\mathbb{R}^d$, and *ii*) conditionally on $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}$, $\Phi(\mathcal{H}(\mathbf{x}, \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}), \mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d}))$ is obtained from $\mathbf{X}_{i_d}$ via translations, orthogonal transformations, and the arctan function.

Thus, writing

$$\begin{aligned}
&\mathbb{P}\left\{\Phi\left(\mathcal{H}(\mathbf{x}, \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}), \mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})\right) \le \Phi_\delta, \mathcal{E}_n\right\} \\
&= \mathbb{E}\big[\mathbf{1}_{\mathcal{E}_n} \mathbb{P}\left\{\Phi\left(\mathcal{H}(\mathbf{x}, \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}), \mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})\right) \le \Phi_\delta | \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}\right\}\big]
\end{aligned}$$

and noting that, on the event $\mathcal{E}_n$, for fixed $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}$, $\Phi_\delta \downarrow 0$ as $\delta \downarrow 0$, we conclude by the Lebesgue dominated convergence theorem that

$$\mathbb{P}\left\{\Phi\left(\mathcal{H}(\mathbf{x}, \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{d-1}}), \mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})\right) \le \Phi_\delta\right\} \to 0 \quad \text{as } \delta \downarrow 0.$$

■

**Acknowledgments.** We thank two anonymous referees for valuable comments and insightful suggestions.

# References


[1] T. Anderson. Some nonparametric multivariate procedures based on statistically equivalent blocks. In P. Krishnaiah, editor, *Multivariate Analysis*, pages 5–27, New York, 1966. Academic Press.

[2] G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101:2499–2518, 2010.

[3] B. Chakraborty, P. Chaudhuri, and H. Oja. Operating transformation retransformation on spatial median and angle test. *Statistica Sinica*, 8:767–784, 1998.

[4] B. Chazelle. Cutting hyperplanes for divide-and-conquer. *Discrete Computational Geometry*, 9:145–158, 1993.

[5] T.M. Cover. Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:50–55, 1968.

[6] T.M. Cover. Rates of convergence for nearest neighbor procedures. In *Proceedings of the Hawaii International Conference on Systems Sciences*, pages 413–415, Honolulu, 1968.

[7] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[8] L. Devroye. A universal $k$-nearest neighbor procedure in discrimination. In B.V. Dasarathy, editor, *Nearest Neighbor Pattern Classification Techniques*, pages 101–106, Los Alamos, 1991. IEEE Computer Society Press.

[9] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[10] L. Devroye and A. Krzyżak. New multivariate product density estimators. *Journal of Multivariate Analysis*, 82:88–110, 2002.

[11] E. Fix and J.L. Hodges. *Discriminatory analysis. Nonparametric discrimination: Consistency properties.* Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[12] E. Fix and J.L. Hodges. *Discriminatory analysis: Small sample performance*. Technical Report 11, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1952.

[13] M. Gessaman. A consistent nonparametric multivariate density estimator based on statistically equivalent blocks. *The Annals of Mathematical Statistics*, 41:1344–1346, 1970.

[14] M. Gessaman and P. Gessaman. A comparison of some multivariate discrimination procedures. *Journal of the American Statistical Association*, 67:468–472, 1972.

[15] L. Gordon and R.A. Olshen. Asymptotically efficient solutions to the classification problem. *The Annals of Statistics*, 6:515–533, 1978.

[16] L. Gordon and R.A. Olshen. Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10:611–627, 1980.

[17] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

[18] M. Hallin and D. Paindaveine. Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *The Annals of Statistics*, 30:1103–1133, 2002.

[19] T.P. Hettmansperger, J. Möttönen, and H. Oja. The geometry of the affine invariant multivariate sign and rank methods. *Journal of Nonparametric Statistics*, 11:271–285, 1998.

[20] T.P. Hettmansperger and R.H. Randles. A practical affine equivariant multivariate median. *Biometrika*, 89:851–860, 2002.

[21] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[22] J. Marcinkiewicz and A. Zygmund. Sur les fonctions indépendantes. *Fundamenta Mathematicae*, 29:60–90, 1937.

[23] C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, 1989*, London Mathematical Society Lecture Note Series 141, pages 148–188. Cambridge University Press, 1989.

[24] S. Meiser. Point location in arrangements of hyperplanes. *Information and Computation*, 106:286–303, 1993.

[25] K.S. Miller. *Multidimensional Gaussian Distributions*. Wiley, New York, 1964.

[26] H. Oja. Affine invariant multivariate sign and rank tests and corresponding estimates: A review. *Scandinavian Journal of Statistics*, 26:319–343, 1999.

[27] H. Oja and D. Paindaveine. Optimal signed-rank tests based on hyperplanes. *Journal of Statistical Planning and Inference*, 135:300–323, 2005.

[28] E. Ollila, T.P. Hettmansperger, and H. Oja. Estimates of regression coefficients based on sign covariance matrix. *Journal of the Royal Statistical Society Series B*, 64:447–466, 2002.

[29] E. Ollila, H. Oja, and V. Koivunen. Estimates of regression coefficients based on lift rank covariance matrix. *Journal of the American Statistical Association*, 98:90–98, 2003.

[30] R. Olshen. Comments on a paper by C.J. Stone. *The Annals of Statistics*, 5:632–633, 1977.

[31] D. Paindaveine and G. Van Bever. *Nonparametric consistent depth-based classifiers*. ECARES working paper 2012-014, Brussels, 2012.

[32] K.R. Parthasarathy. *Probability Measures on Metric Spaces*. AMS Chelsea Publishing, Providence, 2005.

[33] V.V. Petrov. *Sums of Independent Random Variables*. Springer-Verlag, Berlin, 1975.

[34] C. Quesenberry and M. Gessaman. Nonparametric discrimination using tolerance regions. *The Annals of Mathematical Statistics*, 39:664–673, 1968.

[35] R.H. Randles. A distribution-free multivariate sign test based on interdirections. *Journal of the American Statistical Association*, 84:1045–1050, 1989.

[36] M. Samanta. A note on uniform strong convergence of bivariate density estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 28:85–88, 1974.

[37] C.J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5:595–645, 1977.

[38] D.E. Tyler. A distribution-free $M$-estimator of multivariate scatter. *The Annals of Statistics*, 15:234–251, 1987.

[39] R.L. Wheeden and A. Zygmund. *Measure and Integral. An Introduction to Real Analysis*. Marcel Dekker, New York, 1977.