ESAIM: PS 28 (2024) 132–160
https://doi.org/10.1051/ps/2024002

ESAIM: Probability and Statistics
www.esaim-ps.org

UNIVERSAL CONSISTENCY OF THE k-NN RULE IN METRIC SPACES AND NAGATA DIMENSION. II

Sushma Kumari[1] and Vladimir G. Pestov[2,3,*]

Abstract. We continue to investigate the k nearest neighbour (k-NN) learning rule in complete separable metric spaces. Thanks to the results of Cérou and Guyader (2006) and Preiss (1983), this rule is known to be universally consistent in every such metric space that is sigma-finite dimensional in the sense of Nagata. Here we show that the rule is strongly universally consistent in such spaces in the absence of ties. Under the tie-breaking strategy applied by Devroye, Györfi, Krzyżak, and Lugosi (1994) in the Euclidean setting, we manage to show the strong universal consistency in non-Archimedian metric spaces (*i.e.*, those of Nagata dimension zero). Combining the theorem of Cérou and Guyader with results of Assouad and Quentin de Gromard (2006), one deduces that the k-NN rule is universally consistent in metric spaces having finite dimension in the sense of de Groot. In particular, the k-NN rule is universally consistent in the Heisenberg group which is not sigma-finite dimensional in the sense of Nagata as follows from an example independently constructed by Korányi and Reimann (1995) and Sawyer and Wheeden (1992).

Mathematics Subject Classification. 62H30, 54F45.

Received July 21, 2023. Accepted February 21, 2024.

1. Introduction

The problem of describing those (separable, complete) metric spaces in which the k nearest neighbour classifier is universally (weakly) consistent still remains open. The same applies to the strong universal consistency under some reasonable tie-breaking strategy. In this paper, we are motivated by those two problems and closely related questions.

The main tool in this direction is the theorem by Cérou and Guyader [1], who have shown that the k-NN classifier is (weakly) consistent under the assumption that the regression function $\eta(x)$ satisfies the weak Lebesgue–Besicovitch differentiation property. While it is unknown if this property actually follows from the consistency of the k-NN classifier, it is now possible to deduce the universal consistency for every metric space having the weak Lebesgue–Besicovitch property for every probability measure. A large class of such metric spaces was previously isolated by Preiss [2]: the so-called sigma-finite dimensional metric spaces in the sense of Nagata [3, 4]. Thus, it follows that in every separable metric space that is sigma-finite dimensional in the sense

Keywords and phrases: k-NN classifier, universal consistency, strong universal consistency, distance ties, Nagata dimension, de Groot dimension, sigma-finite dimensional metric spaces, Heisenberg group, Lebesgue–Besicovitch property.

[1] Defense Institute of Advanced Technology (DIAT), Pune, Maharashtra 411025 India.
[2] Departamento de Matemática, Universidade Federal da Paraíba, João Pessoa, PB, Brazil.
[3] Department of Mathematics and Statistics, University of Ottawa, Ottawa ON K1N 6N5, Canada.

* Corresponding author: vpest283@uottawa.ca

of Nagata the k-NN classifier is universally consistent. In the part I of this work [5], we have given a direct proof of the result in the spirit of the original argument of Stone for Euclidean spaces [6], illustrating the similarities and differences of the argument in this more general setting.

One observation of the present paper is that the conclusion of the result holds for a strictly more general class of metric spaces. Assouad and Quentin de Gromard have shown [7] that the Lebesgue–Besicovitch differentiation property is true for metric spaces that are finite dimensional in the sense of de Groot. In particular, modulo the results of [1], the k-NN classification rule is universally consistent in such spaces. Among the most studied examples of such metric spaces is the Heisenberg group $\mathbb{H}$. It is known that the Heisenberg group has infinite Nagata dimension (this was shown independently by Korányi and Reimann [8] and Sawyer and Wheeden [9]). In fact, their argument also implies that $\mathbb{H}$ is not sigma-finite dimensional in the sense of Nagata. Thus, the k-NN classifier is universally consistent in the Heisenberg group, and the property of being sigma-finite dimensional in the sense of Nagata is not a necessary condition. This observation, the subject of Section 3, refutes the conjecture made by us in part I [5].

It is also noteworthy that the example of the Heisenberg group answers in the negative a question asked by Preiss in 1983 [2]: suppose a metric space Ω satisfies the Lebesgue–Besicovitch differentiation property for every sigma-finite locally finite measure, will it satisfy the strong Lebesgue–Besicovitch differentiation property for every such measure too? While this must be well-known to the experts, we are unaware of this being mentioned explicitly anywhere.

In the remaining part of the article we proceed to the strong universal consistency of the k-NN classifier in metric spaces. In Section 4 we show that in the absence of distance ties, the k-NN rule is strongly universally consistent in every separable sigma-finite dimensional space in the sense of Nagata. The argument follows closely the proof in the Euclidean case belonging originally to Devroye and Györfi [10] and Zhao [11] as presented in the book [12] (Thm. 11.1). Clearly, the key geometric lemma using Nagata dimension is a bit different. Section 4 is a revised version of a part of the PhD thesis of the first-named author [13].

Adopting a specific paradigm of uniform tie-breaking belonging to Devroye, Györfi, Krzyżak, and Lugosi [14] who applied it in the Euclidean case, we show that the k-NN classifier is strongly universally consistent in the non-Archimedian metric spaces, that is, those satisfying the strong triangle inequality: $d(x,z) \le \max\{d(x,y), d(y,z)\}$. The same holds in a slightly more general class of metric spaces of Nagata dimension zero. We were unable to extend the result to all (sigma) finite dimensional metric spaces in the sense of Nagata, but already the non-Archimedian case is, we believe, important, as it is, intuitively, where the distance ties occur most often. It is worth noting that a direct analogue of a crucial technical geometric lemma proved in [14] in the Euclidean case fails in non-Archimedian metric spaces with measure, revealing a rather interesting difference in their underlying geometries. This is the subject of our Section 5.

In the concluding short Section 6, we propose a new version of the conjecture aimed to describe those complete separable metric spaces in which the k-NN classifier is universally consistent.

2. Preliminaries: learning rules

2.1. Learning in a measurable space

Let $\Omega = (\Omega, \mathcal{A})$ be a measurable space, that is, a non-empty set Ω equipped with a sigma-algebra of subsets $\mathcal{A}$. The product $\Omega \times \{0,1\}$ becomes a measurable space in a natural way. The elements $x \in \Omega$ are known as *unlabelled points*, and elements $(x,y) \in \Omega \times \{0,1\}$ are *labelled points.* A finite sequence of labelled points, $\sigma = (x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n) \in \Omega^n \times \{0,1\}^n$, is a *labelled sample.* Here it is probably important to stress that a sample is a sequence and not a subset, as it may have repetitions.

A *classifier* in Ω is a mapping

$$T\colon \Omega \to \{0,1\},$$

assigning a label to every point. The mapping is usually assumed to be measurable (or, more generally, universally measurable, that is, measurable with regard to the intersection of all possible completions of the sigma-algebra). This assumption is necessary in order for things like the misclassification error to be well defined, although some authors are allowing for non-measurable maps, working with the outer measure instead.

Let $\tilde{\mu}$ be a probability measure defined on the measurable space $\Omega \times \{0,1\}$. Denote (X,Y) a random element of $\Omega \times \{0,1\}$ following the law $\tilde{\mu}$. The misclassification error of a classifier T is the quantity

$$\begin{aligned} \mathrm{err}_{\tilde{\mu}}(T) &= \tilde{\mu}\{(x,y) \in \Omega \times \{0,1\} \colon T(x) \neq y\} \\ &= P[T(X) \neq Y]. \end{aligned}$$

The misclassification error cannot be smaller than the *Bayes error,* which is the infimum (in fact, the minimum) of the errors of all the classifiers T defined on Ω:

$$\ell^* = \ell^*_{\tilde{\mu}} = \inf_T \mathrm{err}_{\tilde{\mu}}(T).$$

A (*supervised binary classification*) *learning rule* in $(\Omega, \mathcal{A})$ is a mapping, g, that, when shown a labelled sample, σ, produces a classifier, $g(\sigma)$. In other words, a learning rule determines a label of each point x on the basis of a labelled learning sample σ:

$$g \colon \bigcup_{n=1}^{\infty} \Omega^n \times \{0,1\}^n \times \Omega \ni (\sigma, x) \mapsto g(\sigma)(x) \in \{0,1\}.$$

Again, the map above is usually assumed to be (universally) measurable with regard to the natural sigma-algebra generated by $\mathcal{A}$ through the finite products and then countable unions.

We denote the restriction of g to $\Omega^n \times \{0,1\}^n$ by g_n. This way, one can think of a learning rule g as a sequence of maps and write $g = (g_n)$.

The labelled datapoints are modelled by a sequence of independent, identically distributed random elements $(X_n, Y_n) \in \Omega \times \{0,1\}$ following the law $\tilde{\mu}$. For each n, the *misclassification error* of the rule g restricted to $\Omega^n \times \{0,1\}^n$, that is, g_n, is the random variable

$$\mathrm{err}_{\tilde{\mu}} g_n = \mathrm{err}_{\tilde{\mu}} g_n(D_n),$$

where D_n is a random labelled n-sample, $D_n = (X_1, Y_1, X_2, Y_2, \ldots, X_n, Y_n)$.

Define the measure $\mu = \tilde{\mu} \circ \pi^{-1}$, where π is the first coordinate projection of $\Omega \times \{0,1\}$. This is a probability measure on $(\Omega, \mathcal{A})$. Now define a finite measure μ_1 on Ω by $\mu_1(A) = \tilde{\mu}(A \times \{1\})$. Clearly, μ_1 is absolutely continuous with regard to μ. Define the *regression function*, $\eta \colon \Omega \to [0,1]$, as the corresponding Radon–Nikodým derivative

$$\begin{aligned} \eta(x) &= \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu} \\ &= P[Y = 1 \mid X = x], \end{aligned}$$

that is, the conditional probability for x to be labelled 1. (For the Radon–Nikodým theorem in our abstract setting, see [15], 232E and 232B.)

Notice that since the regression function η, together with the measure μ, allows to fully reconstruct the measure $\tilde{\mu}$, a learning problem in a measurable space $(\Omega, \mathcal{A})$ can be alternatively given either by the measure $\tilde{\mu}$ or by the pair (μ, η). We will sometimes denote the corresponding Bayes error by $\ell^*_{\mu,\eta}$.

Given a classifier $T = \chi_C$, the misclassification error can be written as

$$\operatorname{err}_{\tilde{\mu}}(T) = \int_C (1-\eta)\,\mathrm{d}\mu + \int_{\Omega\setminus C} \eta\,\mathrm{d}\mu. \tag{2.1}$$

Now it is easy to see that the Bayes error $\ell^* = \ell^*_{\mu,\eta}$ is achieved at exactly those classifiers T satisfying

$$T(x) = \begin{cases} 1, & \text{for } \mu\text{-almost all } x \text{ such that } \eta(x) > \frac{1}{2}, \\ 0, & \text{for } \mu\text{-almost all } x \text{ such that } \eta(x) < \frac{1}{2}. \end{cases}$$

(At the points where η equals $1/2$, the value of a Bayes classifier – or any classifier – does not affect the error.) Such classifiers are known as *Bayes classifiers.*

A rule g is *consistent* (or *weakly consistent*) under $\tilde{\mu}$ if

$$\operatorname{err}_{\tilde{\mu}} g_n \xrightarrow{p} \ell^*_{\tilde{\mu}},$$

where the convergence is in probability, and *universally consistent* if g is consistent under every probability measure $\tilde{\mu}$ on $\Omega \times \{0,1\}$. In this paper, *consistency* will be synonymous with *weak consistency.*

In a similar way, one defines the strong consistency. A *labelled sample path* is an infinite sequence of i.i.d. elements of $\Omega \times \{0,1\}$ each one following the law $\tilde{\mu}$. A rule g is *strongly consistent* under $\tilde{\mu}$ if

$$\operatorname{err}_{\tilde{\mu}} g_n(D_n) \to \ell^*_{\tilde{\mu}},$$

where the convergence is along almost every infinite labelled sample path $D_\infty = (X_1, Y_1), (X_2, Y_2), \ldots$, and D_n denotes the initial segment of the path D_∞. A rule is *strongly universally consistent* if it is strongly consistent under every probability measure on the space of labelled points. Clearly, strong consistency implies consistency.

Recall that the *Borel sigma-algebra* (or *Borel structure*) of a topological space Ω is the smallest sigma-algebra containing all open sets. In particular, every metric on a set generates a Borel sigma-algebra. A *standard Borel space* is a set equipped with a sigma-algebra that is the Borel sigma-algebra generated by some complete separable metric. The usual setting for statistical learning is a standard Borel space as Ω. This will be the setting for our paper as well. However, apriori there are no restrictions for studying learning problems in more general measurable spaces.

2.2. The k nearest neighbour classification rule

Let now Ω be a metric space. The k-NN classifier in Ω is a learning rule, defined by selecting the label $g_n(\sigma)(x) \in \{0,1\}$ for a point x on the basis of a labelled n-sample $\sigma = \sigma_n = (x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n)$, $x_i \in \Omega$, $y_i \in \{0,1\}$, by the majority vote among the values of y_i corresponding to the $k = k_n$ nearest neighbours of x in the learning sample σ.

There is an issue of possibly occurring ties, which come in two types. One is the voting tie, when k is even and we may have a split vote. This can be broken, in fact, in any way, without affecting the consistency of the classifier. For instance, in such cases one can always choose the label 1 (as we do below), or just assign the label in a random way. Or else one can only work with odd values of k_n.

It may also be that there are more than k nearest neighbours of x within σ that are at the same distance. This requires a tie-breaking rule. Given k and $n \geq k$, define

$$r^{\sigma_n}_{k\text{-NN}}(x) = \min\{r \geq 0 \colon \sharp\{i = 1, 2, \ldots, n \colon x_i \in \bar{B}(x, r)\} \geq k\}. \tag{2.2}$$

In other words, this is the smallest radius of a closed ball around x containing at least k nearest neighbours of x in the sample σ_n.

A *k nearest neighbour mapping* is a function

$$N_k : \Omega^n \times \Omega \to \Omega^k$$

which, given an unlabelled n-sample σ and a point x, selects a k-subsample $N_k^\sigma(x) \sqsubset \sigma$ so that

1. all elements of $N_k^\sigma(x)$ are at a distance $\leq r_{k\text{-NN}}^{\sigma_n}(x)$ from x, and
2. all points x_i in σ that are at a distance strictly less than $r_{k\text{-NN}}^{\sigma_n}(x)$ to x are in $N_k^\sigma(x)$.

The k nearest neighbour mapping $N_k^\sigma(x)$ (which we will sometimes shorten to $N_k(x)$) can be deterministic or stochastic, in which case it will depend on an additional random variable, independent of the sample path. An example of the kind would be to give the sample σ a random order, under a uniform distribution on the group of n-permutations, and break the distance ties by selecting among the tied neighbours on the sphere the smallest ones under the order selected.

Here is a formal definition of the k-NN learning rule:

$$g_n^{k\text{-NN}}(\sigma)(x) = \theta\left[\frac{1}{k}\sum_{x_i \in N_k^\sigma(x)} y_i - \frac{1}{2}\right].$$

Above, θ is the Heaviside function:

$$\theta(t) = \begin{cases} 1, & \text{if } t \geq 0, \\ 0, & \text{if } t < 0. \end{cases}$$

The k-NN rule was historically the first classification learning rule in a standard Borel space whose universal consistency was established, by Charles J. Stone [6].

Theorem 2.1 (C.J. Stone, 1977)**.** *The k-nearest neighbour classifier is universally consistent in the finite-dimensional Euclidean space whenever $n \to \infty$, $k = k_n \to \infty$, $k/n \to 0$.*

The k-NN classifier is no longer universally consistent in more general separable metric spaces, in fact already in the infinite-dimensional Hilbert space ℓ^2, as noted in [1]. An example of this kind (constructed for the needs of real analysis) belongs to Preiss [16]. (See this example adapted for the k-NN classifier in [5], Sect. 2.) This brings up the question of characterizing those metric spaces in which the k-NN classifier is universally consistent, and so far the problem remains open.

2.3. Strong consistency

Under the — possibly the most natural — randomized method of tie-breaking, the k-NN classifier is *never* strongly universally consistent. Let (Z_n) be a sequence of i.i.d. random variables distributed uniformly in the unit interval $\mathbb{I} = [0, 1]$, and independent on data. In case of distance ties, we choose among the points $x_{n_1}, x_{n_2}, \ldots, x_{n_m}$ at an equal distance to x those points whose corresponding instances z_{n_i} are the smallest. (See for example [1], bottom of p. 341.)

Proposition 2.2. *If a sequence of values of k, (k_n), goes to infinity sufficiently slowly, then the k-NN classifier, under the uniform random tie-breaking using the auxiliary variables $Z_i \in \mathbb{I}$ as above, is not strongly universally consistent in any metric space.*

Proof. Let the underlying probability measure μ on Ω be a Dirac measure concentrated in one point, and let the regression function η take a value $p \in (0, 1)$, $p \neq 1/2$ at the unique point of the measure support. This way, the nature of the metric space becomes totally irrelevant, as everything reduces to a trivial one-point domain, $\Omega = \{*\}$. A sample path in this context is just a Bernoulli sequence (Y_n) of random labels 0 and 1

with probability of success p, together with an i.i.d. sequence $Z_n \in \mathbb{I}$ of tie-breaking values, the two sequences being independent. The Bayes error for our problem equals $\min\{p, 1-p\}$. It is achieved at the Bayes (optimal) classifier, returning the label 1 if $p > 1/2$ and the label 0 if $p < 1/2$. (Here, we need the assumption $p \neq 1/2$: for $p = 1/2$ any prediction would achieve the Bayes error $1/2$.) Strong universal consistency would require that for a.e. Bernoulli sequence (Y_n) and a.e. tie-breaking sequence (Z_n), the k-NN classifier always predicts the Bayes label, starting with some i large enough.

Fix a summable sequence (δ_i), $\delta_i \in (0,1)$. Choose recursively sequences $n_i \uparrow \infty$ and $\epsilon_i \downarrow 0$ in such a way that for every i, if we randomly choose n_i i.i.d. uniform elements of the interval, $Z_1, Z_2, \ldots, Z_{n_i}$, then with confidence $> 1 - \delta_i$

1. at least $\lceil \ln i \rceil$ elements Z_i belong to the interval $[0, \epsilon_i)$, while
2. none of Z_i belong to $[0, \epsilon_{i+1})$.

Now define for each n

$$k_n = \lceil \ln i \rceil, \text{ if } n_i \leq n < n_{i+1}.$$

The first Borel–Cantelli lemma implies that almost surely, for some j occurs the event A_j "a sample path (Z_i) satisfies the conditions (1) and (2) for all $i \geq j$."

Denote Θ the event "the k-NN classifier returns a wrong label infinitely often". We will show that at least for some values of p, it is an almost sure event. We will condition on the tie-breaking path (Z_i). Almost surely, (Z_i) is in A_j for some j. So let us fix j and a path (z_i) belonging to A_j. The properties of A_j imply that, for all i, m, such that $j \leq i < m$, the $k = \lceil \ln i \rceil$ smallest elements among $z_1, z_2, \ldots, z_{n_i}$ belong to the interval $(\epsilon_{i+1}, \epsilon_i)$, while the $k = \lceil \ln m \rceil$ smallest elements among $z_1, z_2, \ldots, z_i, \ldots, z_{n_m}$ belong to the interval $(0, \epsilon_m)$. Since $\epsilon_m \leq \epsilon_{i+1}$, the two intervals are disjoint, so the sets of tie-breaking values at the steps n_i and n_m are disjoint too, and the subsamples $N_{k_{n_i}}^{\sigma_{n_i}}(*)$ and $N_{k_{n_m}}^{\sigma_{n_m}}(*)$ of nearest neighbours selected by the classifier to make a prediction at the steps i and m are disjoint (are indexed with disjoints sets of integers). We conclude: the sets of labels of the k-nearest neighbours chosen at the moments n_i, $i \geq j$ according to our procedure form a sequence of independent random variables with values in $\{0,1\}^{k_{n_i}}$. Consequently, the predictions made at the steps n_i, $i \geq j$ also form an independent sequence.

Denote W_i the event "the k-NN classifier returns the wrong label at the step n_i when using the sequence (z_i) for tie-breaking". According to the above, the sequence of events (W_i), $i \geq j$ is independent. The probability for the k-NN classifier to return the wrong label (that is, 1 if $p < 1/2$ and 0 if $p > 1/2$) at the step n_i, $i \geq j$ is at least $\min\{p, 1-p\}^{k_{n_i}} = \min\{p, 1-p\}^{\lceil \ln i \rceil}$ (this is the probability of the event where all k nearest neighbours have the same label opposite to the Bayes one).

Now let $p = \mathrm{e}^{-1} \approx 0.368\ldots$. We have

$$\begin{aligned}\sum_{i=j}^{\infty} \min\{p, 1-p\}^{\lceil \ln i \rceil} &= \sum \mathrm{e}^{-\lceil \ln i \rceil} \\ &\geq \mathrm{e}^{-1} \sum \mathrm{e}^{-\ln i} \\ &= \mathrm{e}^{-1} \sum_{i=j}^{\infty} \frac{1}{i},\end{aligned}$$

a divergent sequence. The events (W_i), $i \geq j$ are independent, the sequence $p(W_i)$ is divergent. The second Borel–Cantelli lemma implies that, almost surely, W_i occur infinitely often. In other words, if $p = \mathrm{e}^{-1}$ and our (z_i) is used for tie-breaking, the k-NN rule will return the wrong label infinitely often for almost all labelling sequences (Y_i). Since the sequences (Z_i) and (Y_i) are mutually independent, we conclude by the Fubini theorem that our event Θ occurs with probability one (the same holds in fact whenever p belongs to $[\mathrm{e}^{-1}, 1/2) \cup (1/2, 1-\mathrm{e}^{-1}]$). $\square$

In view of this observation, one way to get strong consistency results is to make k grow fast enough. For some results obtained in this direction, see [17]. We do not touch upon this approach in our paper.

Another possibility is to assume that there are no distance ties, that is, there are no atoms and all the spheres have measure zero. This happens in the Euclidean case, for instance, if the underlying distribution has Lebesgue density. Under this assumption, strong consistency for the k-NN classifier in the Euclidean space is a result due to Devroye and Györfi [10] and to Zhao [11]. We will extend the same conclusion to all sigma-finite dimensional metric spaces in the sense of Nagata in Section 4.

Finally, a modified randomized tie-breaking approach to the k-NN classifier was proposed by Devroye, Györfi, Krzyżak, and Lugosi in [14]. As before, the data path is enlarged by adding an independent i.i.d. sequence of tie-breaking variables (Z_n) taking value in $\mathbb{I}$. The difference with the previous approach is that the test data point is also modelled not by a single random variable $X \sim \mu$ but a pair of random variables, (X, Z), where Z is independent of X and of the data and follows the uniform distribution on $\mathbb{I}$. In the case of distance ties, the points X_i, $i \in J$ all at the same distance from X are ordered in accordance with the corresponding values of Z_i, $i \in J$, the closest ones to Z being chosen first. (The previously described approach corresponds to the case of Z taking a constant value zero.)

Under this mode of tie-breaking, the classifier is being built not in Ω proper but rather in the extended domain $\Omega \times \mathbb{I}$, equipped with the product of μ and the uniform measure λ, and whose regression function is the composition of η with the projection on the first coordinate. In the Euclidean case $\Omega = \mathbb{R}^d$ it was shown by Devroye *et al.* [14] that the resulting classifier, which is, strictly speaking, not the k-NN classifier but a modification thereof, converges along almost every sample path to the Bayes classifier on $\Omega \times \mathbb{I}$, obtained by composing the Bayes classifier for Ω with the first coordinate projection. Even if for any fixed value $Z = z$ the same argument as in our Proposition 2.2 shows that the wrong predictions may occur infinitely often, the expected error averaged over $Z \in \mathbb{I}$ converges to zero for almost all sample paths. Thus, if one now wants to obtain a strongly consistent learning rule on Ω proper, one has to average the predictions along every fibre $\{x\} \times \mathbb{I}$, that is, take the majority vote over all values of the auxiliary variable Z. In this approach, essentially, one combines the k-NN with ensemble learning.

In Section 5, we will establish strong consistency within the above approach for non-Archimedian metric spaces, and the proof shows interesting geometric differences from the Euclidean case.

3. Dimension in the sense of de Groot and the Heisenberg group

The aim of this section is to observe that a complete separable metric space in which the k-NN classifier is universally consistent need not be sigma-finite dimensional in the sense of Nagata. We begin by reminding the important result by Cérou and Guyader.

Theorem 3.1 (Cérou and Guyader, [1])**.** *Let Ω be a separable complete metric space equipped with a probability measure μ (the distribution law of data) and a regression function $\eta\colon \Omega \to [0,1]$ (the conditional probability for a point to be labelled 1). Suppose further that the regression function satisfies the weak Lebesgue–Besicovitch differentiation property:*

$$\frac{1}{\mu(B(x,r))} \int_{B(x,r)} \eta(x)\, \mathrm{d}\mu(x) \to \eta(x), \tag{3.1}$$

where the convergence is in measure, that is, for each $\epsilon > 0$,

$$\mu\left\{ x \in \Omega\colon \left| \frac{1}{\mu(B(x,r))} \int_{B(x,r)} \eta(x)\, \mathrm{d}\mu(x) - \eta(x) \right| > \epsilon \right\} \to 0 \text{ when } r \downarrow 0.$$

Then the k-NN classifier is (weakly) consistent for the supervised learning problem (μ, η) in Ω.

Now, some necessary concepts and results related to the Nagata dimension. (For a more detailed presentation with many examples, see Part I of our work [5].) The following definition is Preiss' generalization [2] of Nagata's original concept. Recall that a family γ of subsets of a set Ω has *multiplicity* $\leq n$ if every point of Ω is contained in at most n elements of γ.

Definition 3.2. Let Ω be a metric space and X a metric subspace, let $\delta \in \mathbb{N}$ and $s > 0$. Then X has *Nagata dimension* $\leq \delta$ *on the scale* s *inside of* Ω if every finite family of closed balls in Ω with centres in X and radii $< s$ admits a subfamily having multiplicity $\leq \delta + 1$ in Ω which covers all the centres of the original balls. The Nagata dimension of X within Ω on the scale $s > 0$, denoted $\dim^s_{Nag}(X, \Omega)$ or sometimes simply $\dim_{Nag}(X, \Omega)$, is the smallest δ such that X has Nagata dimension $\leq \delta$ on the scale s inside Ω. We say that a subspace X has a finite Nagata dimension in Ω if X has finite dimension in Ω on some suitable scale $s > 0$.

Here is a reformulation that we will use. A family of balls in a metric space is *disconnected* if the centre of each ball of the family does not belong to any other ball.

Proposition 3.3. *For a subspace X of a metric space Ω, one has*

$$\dim^s_{Nag}(X, \Omega) \leq \delta$$

if and only if every disconnected family of closed balls in Ω of radii $< s$ with centres in X has multiplicity $\leq \delta + 1$.

For a proof, see *e.g.* [5], Proposition 7.2. Here is another important property: the Nagata dimension does not increase when we form the closure of a subspace.

Proposition 3.4 (See [5], Prop. 7.4)**.** *Let X be a subspace of a metric space Ω, satisfying $\dim^s_{Nag}(X, \Omega) \leq \delta$. Then $\dim^s_{Nag}(\bar{X}, \Omega) \leq \delta$, where $\bar{X}$ is the closure of X in Ω.*

Definition 3.5 (Preiss, [2])**.** A metric space Ω is said to be *sigma-finite dimensional in the sense of Nagata* if $\Omega = \cup_{i=1}^{\infty} X_n$, where every subspace X_n has finite Nagata dimension in Ω on some scale $s_n > 0$ (where the scales s_n are possibly all different).

Remark 3.6. Because of Proposition 3.4, we can assume all X_n to be closed. Also, it is easy to see that the union of two subspaces having finite Nagata dimension each also has a finite Nagata dimension (Prop. 7.5 in [5]), so we can in addition assume that X_n form an increasing chain.

Remark 3.7. In view of the preceding remark, the Baire Category argument implies that every complete metric space Ω that is sigma-finite dimensional in the sense of Nagata contains a non-empty open subspace that has finite Nagata dimension in Ω.

Now we can remind the theorem of Preiss.

Theorem 3.8 (Preiss [2])**.** *Let Ω be a complete separable metric space. Then the following two properties are equivalent.*

1. *For every locally finite Borel measure μ on Ω, every $L^1(\mu)$-function $f : \Omega \to \mathbb{R}$ satisies the strong Lebesgue–Besicovitch differentiation property: for μ-a.e. $x \in \Omega$,*

$$\frac{1}{\mu(B(x,r))} \int_{B(x,r)} f(x)\, \mathrm{d}\mu(x) \to f(x) \ \text{ as } r \downarrow 0. \tag{3.2}$$

2. *Ω is sigma-finite dimensional in the sense of Nagata.*

It should be noted that the original note of Preiss [2] only contained a brief sketch of the proof of the implication (1)⇒(2). The implication (2)⇒(1) was worked out in detail by Assouad and Quentin de Gromard

in [7] for the case of finite Nagata dimension (from this, the deduction of the sigma-finite dimensional case is straightforward).

By combining Theorems 3.8 and 3.1, one obtains:

Corollary 3.9. *The k-nearest neighbour classifier is (weakly) universally consistent in every complete separable metric space sigma-finite dimensional in the sense of Nagata.*

In Part I [5] we have given a direct proof of this result along the geometric ideas of the original proof of Stone [6].

Note that Preiss' result asserts a strong version of the Lebesgue–Besicovitch property, while the result of Cérou and Guyader only requires the weak version of it as an assumption. Turns out, there is a class of metric spaces that "fills the gap" between the two. For that, we need to give some more definitions.

Definition 3.10 ([18]; [7], 3.5)**.** Let $\delta \in \mathbb{N}$. A metric space Ω has de Groot dimension $\leq \delta$ if it satisfies the following property. For every closed ball $\bar{B}(a,r)$ in Ω with centre a and radius $r > 0$, if $x_1, \ldots, x_{\delta+1} \in \bar{B}(a,r)$, then there are $i \neq j$ with $d(x_i, x_j) \leq r$.

Proposition 3.11 (Prop. 3.1 in [7])**.** *A metric space Ω has de Groot dimension $\leq \delta$ if and only if every finite family of closed balls having the same radii admits a subfamily covering all the centres of the original balls and having multiplicity $\leq \delta + 1$.*

Proof. Necessity: let $\bar{B}(x_1, r), \ldots, \bar{B}(x_N, r)$ be a finite family of closed balls having the same radius. Take any maximal disconnected subfamily of those balls. It covers all the centres by maximality (here we use the fact that the radii of all the balls are the same). Also, this maximal disconnected subfamily has multiplicity $\leq \delta + 1$ because of our assumption on de Groot dimension: assuming there were x belonging to $\delta + 2$ balls, the r-ball centred at x of radius r would contain $\delta + 2$ points two by two at a distance $> r$ from each other.

Sufficiency: apply the property to the family of balls $\bar{B}(x_i, r)$, $i = 1, 2, \ldots, \delta + 1$, where $x_i \in \bar{B}(x, r)$. All of the above closed balls contain x, so at least one of those balls, say $\bar{B}(x_i, r)$, will be missing from a subfamily containing all the centres; then $x_i \in \bar{B}(x_j, r)$, $j \neq i$, so $d(x_i, x_j) \leq r$. □

Thus, in view of Proposition 3.3, de Groot dimension of a metric space is always bounded by the Nagata dimension on the scale $+\infty$. For the space $\mathbb{R}^n$ equipped with an arbitrary norm, the two dimensions are equal ([7], 4.9). In a more general case, in fact, already in the infinite-dimensional Hilbert space ℓ^2, the distinguishing examples are easy to construct.

Example 3.12. The convergent sequence $2^{-n}e_n$, $n \geq 0$, where e_n are elements of the standard orthonormal basis in the Hilbert space ℓ^2, together with the limit 0, equipped with the induced metric, has infinite Nagata dimension on every scale $s > 0$. Indeed, each closed ball of radius 2^{-n}, centred at $2^{-n}e_n$, contains 0 as the only other element of the space, and so admits no subfamily of finite multiplicity containing all the centres.

At the same time, this sequence has de Groot dimension 2. Call n the *index* of a point $x = 2^{-n}e_n$, and let the index of zero be infinite. Denote the index $i(x)$. Given a closed ball of centre a in this space and three points inside the ball, order them according to the increasing index, x_1, x_2, x_3. If now $i(a) \leq i(x_1)$, then x_2 and x_3 are closer to each other than x_3 is to a. And if $i(x_1) < i(a)$, then the distance between x_2 and x_3 is smaller than between a and x_1. (And notice that de Groot dimension is not equal to one as the example of a ball of radius $1/2$ centred at $x = 2^{-3}e_3$ and containing two points, $x_1 = 2^{-1}e_1$ and $x_2 = 2^{-2}e_2$ shows.)

This space is complete (even compact) and sigma-finite dimensional in the sense of Nagata being the union of countably many singletons: a singleton trivially has Nagata dimension zero in every ambient metric space.

A source of metric spaces of finite de Groot dimension is provided by the doubling metric spaces.

Definition 3.13. A metric space X is *doubling* if there is a constant $C > 0$ such that for every $x \in X$ and $r > 0$, the closed ball $\bar{B}(x, r)$ can be covered with at most C closed balls of radius $r/2$.

The following is a simple exercise. (Cover a closed r-ball with $\leq C$ many $r/2$-balls and notice that among any $C + 1$ points, at least two belong to the same closed $r/2$-ball.)

Proposition 3.14. *Every doubling metric space has finite de Groot dimension (bounded by the constant C from Def. 3.13).*

Metric spaces of finite de Groot dimension satisfy the weak Lebesgue-Besicovitch differentiation property.

Theorem 3.15 (Assouad and Quentin de Gromard, [7], Prop. 3.3.1(b)+Prop. 3.1)**.** *Let a complete separable metric space Ω have finite de Groot dimension. Then for every probability Borel measure μ on Ω, every $L^1(\mu)$-function $f\colon \Omega \to \mathbb{R}$ satisies the weak Lebesgue–Besicovitch differentiation property:*

$$\frac{1}{\mu(B(x,r))}\int_{B(x,r)} f(x)\,\mathrm{d}\mu(x) \to f(x) \tag{3.3}$$

in measure, when $r \downarrow 0$.

Combining this result with that of Cérou and Guyader (Thm. 3.1 above), we arrive at:

Corollary 3.16. *The k-nearest neighbour classifier is universally consistent in every complete separable metric space having finite de Groot dimension.*

It would be certainly interesting to give a direct proof of the result in the spirit of Stone. Moreover, the versions of de Groot dimension on a given scale and of sigma-finite dimensional spaces in the sense of de Groot that exactly parallel the definition of Preiss can be easily stated, so it is natural to ask a number of questions about such spaces. For instance, is it true that a metric space has the weak Lebesgue–Besicovitch property if and only if it is sigma-finite dimensional in the sense of de Groot? See the concluding Section 6 for an exact formulation.

An example of a complete separable metric space of finite de Groot dimension that is not sigma-finite dimensional in the sense of Nagata is provided by the Heisenberg group $\mathbb{H}$ equipped with one of the natural metrics that we now proceed to describe.

Topologically, the Heisenberg group $\mathbb{H}$ is identified with the Euclidean space $\mathbb{R}^3$, and is equipped with the following group multiplication:

$$(x,y,z)\cdot(x',y',z') = (x+x', y+y', z+z'+Cxy'-Cyx')\,. \tag{3.4}$$

Here $x,x',y,y',z,z' \in \mathbb{R}$, and $C \neq 0$ is a real constant. Different choices of C result in algebraically isomorphic groups: a group isomorphism from the above version to the one determined by the constant C' is given by a linear map multiplying each vector by C/C'.

The operation in (3.4) clearly makes $\mathbb{H}$ into a topological group, in fact a Lie group, when equipped with the Euclidean topology.

For all values of C with $|C| \leq 4$ the formula

$$|(x,y,z)|_{\mathbb{H}} = \left((x^2+y^2)^2+z^2\right)^{1/4}$$

defines a *group norm* on $\mathbb{H}$, in the sense that $|p^{-1}|_{\mathbb{H}} = |p|_{\mathbb{H}}$ and

$$|p\cdot q|_{\mathbb{H}} \leq |p|_{\mathbb{H}} + |q|_{\mathbb{H}}.$$

The latter is a consequence of the following particular case of a result of Cygan [19] (using notation and concepts from, and better be looked at jointly with, the article [20]):

$$\left(\left[(x+x')^2+(y+y')^2\right]^2+16(z+z'+xy'-yx')^2\right)^{1/4} \leq \left((x^2+y^2)^2+16z^2\right)^{1/4} + \left((x'^2+y'^2)^2+16z'^2\right)^{1/4}.$$

Given an expression on the right of equation (3.4), denote $\varepsilon = \pm 1$ the product of the signs of $z + z'$ and of $Cxy' - Cyx'$. Assuming $|C| \leq 4$, the norm of the product $(x, y, z) \cdot (x', y', z')$ is less than or equal to

$$\left[\left[(x + x')^2 + (y + y')^2\right]^2 + 16\left(\frac{\varepsilon z}{4} + \frac{\varepsilon z'}{4} + xy' - yx'\right)^2\right]^{1/4},$$

and applying Cygan's inequality, we arrive at the product of norms of (x, y, z) and (x', y', z').

Consequently, a left-invariant metric on $\mathbb{H}$ is defined by

$$d(p, q) = |p^{-1} \cdot q|_{\mathbb{H}},$$

and is clearly compatible with the Euclidean topology. This distance is known as a (*Cygan-*)*Korányi* distance. Thus, it is the unique left-invariant metric such that

$$d(e, p) = |p|_{\mathbb{H}}.$$

It is well-known and readily seen that the group $\mathbb{H}$ equipped with a Cygan–Korányi distance is doubling. In fact, the doubling property holds for any compatible left-invariant metric on $\mathbb{H}$ that is *homogeneous* in the sense that if we apply to the group the transformation $(x, y, z) \mapsto (tx, ty, t^2 z)$ for $t > 0$, then the distance between any pair of points increases by the factor of t. (It can actually be shown that every such metric is automatically compatible with the Euclidean topology, see [21].) In this form, the doubling property is enough to establish for a single ball of radius $r = 1$ say centred at zero, and it follows from local compactness of the Euclidean space. As the Cygan–Korányi metric is both left-invariant and homogeneous (an easy calculation), the statement follows. In particular, we conclude from the result of Assouad and Quentin de Gromard (Thm. 3.15):

Corollary 3.17. *The Heisenberg group $\mathbb{H}$ equipped with a Cygan–Korányi metric satisfies the weak Lebesgue–Besicovitch property for every Borel probability measure μ and every $L^1(\mu)$-function.*

According to the result of Cérou and Guyader (Thm. 3.1), we now have:

Corollary 3.18. *The k-NN learning rule is universally consistent in the Heisenberg group $\mathbb{H}$ equipped with a Cygan–Korányi metric.*

At the same time, the metric space $\mathbb{H}$ with a Cygan–Korányi distance need not be sigma-finite dimensional in the sense of Nagata. For the next result, we choose a version of the group law corresponding to the value $C = -2$ in the multiplication formula (3.4), following [8]. Thus,

$$(x, y, z) \cdot (x', y', z') = (x + x', y + y', z + z' - 2xy' + 2yx'). \tag{3.5}$$

Essentially, by fixing C, we select a version of the Cygan–Korányi metric, because the groups are all isomorphic between themselves for different values of $C \neq 0$.

Lemma 3.19 (Korányi and Reimann, [8], p. 17; Sawyer and Wheeden, [9], Lem. 4.4, p. 863)**.** *Let $C = -2$. There exists a sequence (p_n) of elements of $\mathbb{H}$ with $r_n = |p_n|_{\mathbb{H}} \to 0$ so that the family of balls $\bar{B}(p_n, r_n)$ is disconnected.*

We find it useful to present a proof, following [8] and somewhat expanding the argument.

Proof. By identifying $\mathbb{R}^2$ with the complex plane $\mathbb{C}$, we can write the multiplication law (3.5) in the group $\mathbb{H} = \mathbb{C} \times \mathbb{R}$ as

$$(z, t)(z', t') = (z + z', t + t' + 2\,\mathrm{Im}\, z\bar{z}').$$

The neutral element of the group is $(0,0)$, and the inverse of (z,t) is simply $(-z,-t)$. Consequently, the formula for the left-invariant Cygan–Korányi metric becomes:

$$\begin{aligned} d((z,t),(z',t')) &= \left|(z,t)^{-1}(z',t')\right|_{\mathbb{H}} \\ &= |(-z,-t)(z',t')|_{\mathbb{H}} \\ &= |(-z+z',-t+t'-2\operatorname{Im} z\bar{z}')|_{\mathbb{H}} \\ &= \left(|-z+z'|^4+|-t+t'-2\operatorname{Im} z\bar{z}'|^2\right)^{1/4}. \end{aligned}$$

Let (z,t) and (z',t') be two points on the unit sphere of $\mathbb{H}$ around the neutral element 0 that are different from $(0,0,\pm 1)$ (so that $z,z'\neq 0$). Notice that $\operatorname{Re} z\bar{z}'$ is the inner product of z and z' as vectors of $\mathbb{R}^2$. As $r\downarrow 0$, we have up to the second order terms in r:

$$\begin{aligned} d((rz,r^2t),(z',t'))^4 &= |-rz+z'|^4+\left|-r^2t+t'-2\operatorname{Im} rz\bar{z}'\right|^2 \\ &= \left(r^2|z|^2+|z'|^2-2r\operatorname{Re} z\bar{z}'\right)^2+\left(-r^2t+t'-2\operatorname{Im} rz\bar{z}'\right)^2 \\ &\overset{O(r^2)}{\approx} |z'|^4-4|z'|^2r\operatorname{Re} z\bar{z}'+t'^2-4t'r\operatorname{Im} z\bar{z}' \\ &= 1-4r\left(|z'|^2\operatorname{Re} z\bar{z}'+t'\operatorname{Im} z\bar{z}'\right). \end{aligned}$$

If the bracketed term on the right is strictly negative,

$$|z'|^2\operatorname{Re} z\bar{z}'+t'\operatorname{Im} z\bar{z}'<0, \tag{3.6}$$

then for sufficiently small $r>0$

$$d((rz,r^2t),(z',t'))>1,$$

so for any $\rho>0$, using the homogeneity property of the metric,

$$d((r\rho z,r^2\rho^2t),(\rho z',\rho^2t'))>\rho. \tag{3.7}$$

Since the complex number $t'+|z'|^2i$ has modulus one, it can be written as $\mathrm{e}^{\psi i}$, so the condition in equation (3.6) becomes

$$\operatorname{Im}(\mathrm{e}^{\psi i}z\bar{z}')<0. \tag{3.8}$$

Now we define two sequences of reals

$$\psi_j=-\frac{\pi}{2}\frac{1}{(j+1)^2}+\pi,\quad \theta_j=\frac{\pi}{2}\frac{j-1}{j}$$

and a sequence of points on the unit sphere in $\mathbb{H}$

$$(z_j,t_j)=\left(\mathrm{e}^{\theta_j i}\sqrt{\sin\psi_j},\cos\psi_j\right).$$

Notice that

$$\mathrm{e}^{\psi_j i}=t_j+|z_j|^2i.$$

Since for $n > j$ we have

$$\pi < \theta_{j+1} - \theta_j + \psi_j \leq \theta_n - \theta_j + \psi_j < \frac{3\pi}{2},$$

(there is a small typo in the second displayed formula on p. 18 in [8]), it follows that

$$\mathrm{Im}(\mathrm{e}^{\psi_j i} z_n \bar{z}_j) \leq \mathrm{Im}(\mathrm{e}^{\psi_j i} z_{j+1} \bar{z}_j) < 0.$$

Now the radii $r_j > 0$ are being chosen recursively, using equation (3.7), in such a way that each element $(r_j z_j, r_j^2 t_j)$ is outside the finitely many closed balls already selected. □

Since all of the above closed balls $\bar{B}(p_n, r_n)$ contain zero (the identity of $\mathbb{H}$), the Nagata dimension of $\mathbb{H}$ is infinite by Proposition 3.3, as was noted by Assouad and Quentin de Gromard [7], 4.7(f). But in fact, the construction implies more.

Corollary 3.20. *The group $\mathbb{H}$ equipped with the Cygan–Korányi metric is not sigma-finite dimensional in the sense of Nagata.*

Proof. Assuming $\mathbb{H}$ were sigma-finite dimensional, by our Remark 3.7, it would contain a non-empty open subset U which has finite Nagata dimension in $\mathbb{H}$. Select any $p \in U$. Since the metric is left-invariant and so the left translation $q \mapsto p^{-1} \cdot q$ is an isometry, the set $p^{-1} \cdot U$ also has finite Nagata dimension. Since this set is a neighbourhood of identity, it contains all elements of the sequence (x_n) chosen as in Theorem 3.19, beginning with n large enough. This contradicts the finite dimensionality of the set $p^{-1} \cdot U$ inside $\mathbb{H}$ in the sense of Nagata. □

Thus, the Heisenberg group $\mathbb{H}$ provides an example of a metric space possessing the weak Lebesgue–Besicovitch property — in particular, on which the k-NN classifier is universally (weakly) consistent — and which is not sigma-finite dimensional.

Remark 3.21. The influential 1983 paper by Preiss [2] mentioned that it was unknown whether a complete separable metric space Ω satisfies the weak Lebesgue–Besicovitch differentiation property for every Borel locally finite measure if and only if Ω satisfies the strong Lebesgue–Besicovitch differentiation property for every Borel locally finite measure. The later developments have shown the answer to be negative, in fact the Heisenberg group with the Cygan–Korányi metric provides a distinguishing example in view of Corollary 3.17, Corollary 3.20 and Preiss's Theorem 3.8, (1)⇒(2). This fact must be well known to the specialists, even if we have not found it mentioned explicitly anywhere.

4. Strong consistency in the absence of distance ties

A probability measure μ on a metric space Ω has a zero probability of distance ties if the measure of every sphere $S_r(x)$, $x \in \Omega$, $r \geq 0$ is zero. In particular, such a measure is non-atomic (the case $r = 0$). In this section, we will show that the result by Devroye and Györfi [10] and Zhao [11] about the strong universal consistency of the k-NN classifier in the Euclidean space in the absence of distance ties is valid in all complete separable sigma-finite dimensional metric spaces in the sense of Nagata – again, in the case where distance ties occur with zero probability. We will follow the presentation of the proof of Theorem 11.1 in [12], however, as to be expected, the extension requires certain technical modifications, not all of which concern Lemma 4.6 below.

Theorem 4.1. *Under the zero probability of distance ties, the k-NN learning rule is strongly universally consistent in every complete separable metric space that is sigma-finite dimensional in the sense of Nagata.*

Remark 4.2. The result is certainly of interest in the setting of all finite-dimensional normed spaces (not just the Euclidean ones), because in such a space there are no distance ties whenever the underlying distribution has

density with regard to the Lebesgue measure. It is hard to think of a similar natural condition for sigma-finite dimensional metric spaces beyond the normed spaces case. One of the most interesting classes – and in which the distance-based classifiers are of practical interest [22] – is given by the non-Archimedian metric spaces, satisfying the strong triangle inequality, which are essentially the metric spaces of Nagata dimension zero. It is not difficult to see that a non-Archimedian metric on a separable space only takes a countable number of distinct values. (Indeed, given such a space, Ω, choose a contable dense subset X and apply the strong triangle inequality to deduce that for any $x, y \in \Omega$ there are $a, b \in X$ with $d(x, y) = d(a, b)$.) This means the distance ties will always occur with strictly positive probability. A rather natural example where the ties are overwhelming was worked out by us in Part I [5], Example 6.4.

Recall from Section 2 that strong consistency of a learning rule (g_n) means that along $\tilde{\mu}^\infty$-almost every infinite labelled sample path $\sigma_\infty \in \Omega^\infty \times \{0,1\}^\infty$, the learning error converges to the Bayes error:

$$\mathrm{err}_{\mu,\eta}(g_n(\sigma_n)) \to \ell^*_{\mu,\eta}.$$

Here $\ell^*_{\mu,\eta}$ is the Bayes error of the learning problem (μ, η), and $\mathrm{err}_{\mu,\eta}(g_n(\sigma_n))$ is the error of the classifier given by the learning rule on the sample input σ_n, the initial n-segment of the path σ_∞. The convergence here is that of a sequence of reals.

Getting back to the k-NN learning rule, denote η_n the approximation to the regression function:

$$\eta_n(X) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in N_k(X)\}} Y_i,$$

where the sum is over all k nearest neighbours of X. We have a classical estimate valid in all metric spaces (see [1], Prop. 1.1):

$$\mathrm{err}_{\mu,\eta}(g_n) - \ell^*_{\mu,\eta} \le 2\mathbb{E}_\mu \left\{ |\eta(X) - \eta_n(X)| \middle| D_n \right\}.$$

Therefore, the strong consistency would follow if we could show that along almost every sample path,

$$\mathbb{E}_\mu |\eta(X) - \eta_n(X)| \to 0.$$

A sigma-finite dimensional metric space Ω can be represented as the union of a countable increasing chain of measurable (even closed should we wish, see Rem. 3.6) subspaces (F_m), each having finite Nagata dimension in Ω, in such a way that $\mu(F_m) \to 1$. Thus, the strong consistency would follow if we could prove that for each fixed m, along almost every sample path,

$$\mathbb{E}_\mu \left\{ |\eta(X) - \eta_n(X)| \, X \in F_m \right\} \to 0,$$

where the expectation is conditional, that is, essentially, a normalized integral over F_m. The way to prove this is through the Borel–Cantelli lemma: we want to show that the expected value of the difference $|\eta(X) - \eta_n(X)|$ over F_m normally concentrates in n. We have no control over the rate of convergence of this difference to zero, so it may be very slow, but what matters is that it should be roughly uniform: if for every $\epsilon > 0$, starting with n sufficiently large, the probability of a deviation larger than ϵ is of the order $\exp(-n\epsilon^2)$, we are done: for almost every sample path, beginning with some n, the deviation over F_m will be below ϵ. Thus, the following lemma, modelled on Theorem 11.1 in [12], will settle the proof of Theorem 4.1, and the rest of the section will be just devoted to a proof of lemma.

Lemma 4.3. *Let Ω be a complete separable metric space, and let Q be a Borel subset. Suppose Q has Nagata dimension $\le \beta$ in Ω on a scale s. Let μ be a probability measure on Ω with zero probability of ties, and let $\eta\colon \Omega \to$*

$[0,1]$ *be a regression function. Suppose* $\mu(Q) > 0$. *Let* $\tilde{\mu}$ *be a probability measure on* $\Omega \times \{0,1\}$ *corresponding to* (μ, η). *For* $\varepsilon > 0$, *whenever* $k, n \to \infty$ *and* $k/n \to 0$, *there is a* n_0 *such that for* $n > n_0$,

$$\mathbb{P}\left(\mathbb{E}_{\tilde{\mu}} \left\{ |\eta(X) - \eta_n(X)| \, X \in Q \right\} > \varepsilon \right) \leq 4\mathrm{e}^{-\frac{n\varepsilon^2 \mu(Q)^2}{18(\beta+1)^2}}.$$

Let μ be a Borel probability measure on a complete separable metric space Ω. Let $0 < \alpha \leq 1$. We define

$$r_\alpha(x) = \inf\{r > 0 : \mu(B(x,r)) \geq \alpha\}. \tag{4.1}$$

Lemma 4.4. *Let* μ *be a probability measure with zero probability of ties. Then* $\mu(B(x, r_\alpha(x))) = \alpha$ *for every* x.

Proof. Clearly, $r_\alpha(x) > 0$. The measure of every open ball of radius $< r_\alpha(x)$ is strictly less than α. By the sigma-additivity of μ, the measure of the open ball of radius $r_\alpha(x)$ is $\leq \alpha$, and the measure of the corresponding closed ball $\bar{B}(x, r_\alpha(x))$ is $\geq \alpha$. By our assumption, the sphere is a null set, so the two values are equal. □

Lemma 4.5. *The real-valued function* r_α *defined as in* (4.1) *is 1-Lipschitz continuous and converges to zero as* $\alpha \to 0$ *at each point of the support of the measure.*

Proof. Since $B(y, r_\alpha(y)) \subseteq B(y, \rho(x,y) + r_\alpha(x))$ (the latter ball contains $B(x, r_\alpha(x))$ and so has measure $\geq \alpha$), we have $r_\alpha(y) \leq \rho(x,y) + r_\alpha(x)$. Therefore, r_α is 1-Lipschitz. The second assertion is clear. □

The following technical result is an analogue of Lemma 11.1 in [12].

Lemma 4.6. *Let* Ω *be a complete separable metric space and let* Q *be a Borel subset having Nagata dimension* $\leq \beta$ *in* Ω *on the scale* s. *Assume that* μ *is a probability measure on* Ω *with zero probability of ties. For* $y \in \Omega$, *define*

$$\begin{aligned} D(y,\alpha) &= \{x \in \Omega : y \in B(x, r_\alpha(x))\} \\ &= \{x \in \Omega \colon d(x,y) < r_\alpha(x)\}. \end{aligned}$$

Then $\mu(D(y,\alpha) \cap Q) \leq (\beta+1)\alpha$ *for all* α *small enough.*

Proof. First of all, notice that the set $D(y,\alpha)$ is open, so it makes sense to talk of its measure. Indeed, if $x \in D(y,\alpha)$, then the open ball of radius $\delta = (1/2)\left[r_\alpha(x) - d(x,y)\right] > 0$ around x also belongs to $D(y,\alpha)$: every element x' of such a ball satisfies

$$r_\alpha(x') > r_\alpha(x) - \delta \geq d(x,y) + \delta \geq d(x',y)$$

(the first inequality is due to Lem. 4.5, the rest follow from the triangle inequality).

Now let $\varepsilon > 0$. By Luzin's theorem, there is a compact set $K \subseteq D(y,a) \cap Q \cap \operatorname{supp}\mu$ such that $\mu(D(y,a) \cap Q \setminus K) < \varepsilon$. As $\varepsilon > 0$ is arbitrary, we need to only get the desired upper bound for $\mu(K)$.

It follows from Lemma 4.5 that r_α converges to 0 uniformly on K when α goes to 0. Choose $\alpha_0 > 0$ such that for $0 < \alpha \leq \alpha_0$, we have $r_\alpha(x) < s$ for all $x \in K$.

Every open ball $B(x, r_\alpha(x))$ centered at $x \in K$ contains y, therefore

$$\bar{B}(x, d(x,y)) \subseteq B(x, r_\alpha(x)). \tag{4.2}$$

Let $D = \{a_n : n \in \mathbb{N}\}$ be a countable dense subset of K. Since Q has metric dimension β in Ω on the scale s, (4.2) implies that for every n there exists a set of $\leq \beta + 1$ centers $\{x_1^n, \ldots, x_{\beta+1}^n\} \subseteq \{a_1, \ldots, a_n\}$ such that the closed balls $\bar{B}(x_i^n, d(x_i^n, y))$, $i = 1, 2, \ldots, \beta+1$ cover $\{a_1, \ldots, a_n\}$.

As K is compact, we can recursively select a subset of indices $I \subseteq \mathbb{N}$ so that each sequence of centres x_i^n, $i = 1, 2, \ldots, \beta+1$, $n \in I$ converges to some point $x_i \in K$. We claim that the union of closed balls $\bar{B}(x_i, d(x_i, y)), 1 \leq i \leq \beta+1$ covers K, which will finish the proof in view of the inclusion (4.2).

As closure of the finite union is the union of closures and since the balls are closed, it is enough to show that $D = \{a_m\}_{m\in\mathbb{N}}$ is contained in the union of $\bar{B}(x_i, d(x_i, y)), 1 \leq i \leq \beta+1$. Fix m. There are $i_0 \in \{1, 2, \ldots, \beta+1\}$ and an infinite set of indices $J \subseteq I$ such that a_m belongs to all the balls $\bar{B}(x_{i_0}^n, d(x_{i_0}^n, y))$, $n \in J$. It follows that

$$\begin{aligned} d(a_m, x_0) &= \lim_{n\in J} d(a_m, x_{i_0}^n) \\ &\leq \lim_{n\in J} d(x_{i_0}^n, y) \\ &= d(x_{i_0}, y), \end{aligned}$$

so $a_m \in \bar{B}(x_{i_0}, d(x_{i_0}, y))$. □

Now, to the proof of Lemma 4.3. As in equation (4.1), denote $r_{k/n}(x)$ the unique solution to the equation

$$\mu(B(x, r_{k/n}(x)) = \frac{k}{n}$$

(*cf.* Lem. 4.4). Let η_n^* be another approximation of η,

$$\eta_n^*(X) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{I}_{\{\rho(X_i, X) < r_{k/n}(X)\}} Y_i. \tag{4.3}$$

By the triangle inequality,

$$|\eta(X) - \eta_n(X)| \leq |\eta(X) - \eta_n^*(X)| + |\eta_n^*(X) - \eta_n(X)|. \tag{4.4}$$

Like in equation (2.2), denote $r_{k\text{-NN}}(x)$ the smallest radius of a closed ball around x containing at least k nearest neighbours of x (we suppress the symbol of the sample). In the absence of distance ties, the closed $r_{k\text{-NN}}(x)$-ball a.s. contains exactly k nearest neighbours. Of the two closed balls around x, one of radius $r_{k\text{-NN}}(x)$ and the other of radius $r_{k/n}(x)$, one is necessarily contained in the other, so the symmetric difference, which we tentatively denote $\Delta(x)$, is just the set-theoretic difference of the two balls, though we do not know in which order. With this in mind, we have for the second term on the right-hand side of above equation (4.4),

$$\begin{aligned} |\eta_n^*(X) - \eta_n(X)| &= \frac{1}{k} \left| \sum_{i=1}^{n} \mathbb{I}_{\{\rho(X_i, X) < r_{k/n}(X)\}} Y_i - \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in N_k(X)\}} Y_i \right| \\ &= \frac{1}{k} \left| \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in \Delta(X)\}} Y_i \right| \\ &\leq \frac{1}{k} \left| \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in \Delta(X)\}} \right| \\ &= \frac{1}{k} \left| \sum_{i=1}^{n} \mathbb{I}_{\{\rho(X_i, X) < r_{k/n}(X)\}} - \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in N_k(X)\}} \right| \\ &= \left| \frac{1}{k} \sum_{i=1}^{n} \mathbb{I}_{\{\rho(X_i, X) < r_{k/n}(X)\}} - 1 \right|, \end{aligned} \tag{4.5}$$

because $N_k(X)$ contains exactly k points.

Next we show that the latter term converges to zero. Let $\hat{\eta}_n(X)$ be equal to $\frac{1}{k}\sum_{i=1}^{n}\mathbb{I}_{\{\rho(X_i,X)<r_{k/n}(X)\}}$ and let $\hat{\eta}(X)$ be identically equal to 1. Conditionally on $X=x$, the expected value of the random variable under the absolute sign is zero (LLN), which allows to pass to variance. Using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathbb{E}_{\tilde{\mu}^n}\{\mathbb{E}_{\mu}\{|\eta_n^*(X)-\eta_n(X)|\}\} &\leq \mathbb{E}_{\tilde{\mu}^n}\{\mathbb{E}_{\mu}\{|\hat{\eta}_n(X)-\hat{\eta}(X)|\}\}\\
&\leq \mathbb{E}_{\mu}\left\{\sqrt{\mathbb{E}_{\tilde{\mu}^n}\{|\hat{\eta}_n(X)-\hat{\eta}(X)|^2\}}\right\}\\
&\leq \mathbb{E}_{\mu}\left\{\sqrt{\frac{n}{k^2}Var\{\mathbb{I}_{\{\rho(X_i,X)<r_{k/n}(X)\}}\}}\right\}\\
&\leq \mathbb{E}_{\mu}\left\{\sqrt{\frac{n}{k^2}\mu(B(X,r_{k/n}(X)))}\right\}\\
&= \mathbb{E}_{\mu}\left\{\sqrt{\frac{n}{k^2}\frac{k}{n}}\right\}\\
&= \frac{1}{\sqrt{k}},
\end{aligned}$$

which term goes to zero as $k\to\infty$.

For the first term on the right hand side of equation (4.4),

$$\begin{aligned}
&\mathbb{E}_{\tilde{\mu}^n}\{\mathbb{E}_{\mu}\{|\eta(X)-\eta_n^*(X)|\}\}\\
&\leq \mathbb{E}_{\tilde{\mu}^n}\{\mathbb{E}_{\mu}\{|\eta(X)-\eta_n(X)|\}\}+\mathbb{E}_{\tilde{\mu}^n}\{\mathbb{E}_{\mu}\{|\eta_n(X)-\eta_n^*(X)|\}\}\\
&\leq \mathbb{E}_{\mu}\{\mathbb{E}_{\tilde{\mu}^n}\{|\eta(X)-\eta_n(X)|\}\}+\mathbb{E}_{\tilde{\mu}^n}\{\mathbb{E}_{\mu}\{|\eta_n(X)-\eta_n^*(X)|\}\}\\
&\to 0 \text{ as } n,k\to\infty, k/n\to 0,
\end{aligned}$$

where we used the fact that $\mathbb{E}_{\mu^n}\{|\eta(X)-\eta_n(X)|\}\to 0$ because the k-NN rule in our setting is weakly consistent due to the results of Preiss and Cérou–Guyader (Cor. 3.9).

The random variables $|\eta(X)-\eta_n^*(X)|$ and $|\hat{\eta}_n(X)-\hat{\eta}(X)|$ admit realisations as Borel measurable functions on $\Omega^\infty\times\{0,1\}^\infty\times\Omega$ taking values in $[0,1]$. Thus, the convergence in expectation implies convergence in measure, and consequently their restrictions to $\Omega^\infty\times\{0,1\}^\infty\times Q$, where by our assumption $Q\subseteq\Omega$ has a strictly positive measure, converge to zero as well, in measure and in expectation. So, for a given $\varepsilon>0$ we can choose n,k so large that

$$\mathbb{E}_{\tilde{\mu}^n}\{\mathbb{E}_{\mu}\{|\eta(X)-\eta_n^*(X)|\mid X\in Q\}\}+\mathbb{E}_{\tilde{\mu}^n}\{\mathbb{E}_{\mu}\{|\hat{\eta}_n(X)-\hat{\eta}(X)|\mid X\in Q\}\}<\frac{\varepsilon}{6}. \tag{4.6}$$

Suppose we have random variables X,Y, such that $\mathbb{E}X+\mathbb{E}Y<\epsilon_2$. Let $\epsilon_1>\epsilon_2>0$. The event $X+Y>\epsilon_1$ implies that either $|X-\mathbb{E}X|$ or $|Y-\mathbb{E}Y|$ strictly exceeds $(\epsilon_1-\epsilon_2)/2$. Indeed, assuming otherwise,

$$X+Y\leq|X-\mathbb{E}X|+\mathbb{E}X+|Y-\mathbb{E}Y|+\mathbb{E}Y<\epsilon_2+(\epsilon_1-\epsilon_2)=\epsilon_1.$$

Writing

$$\begin{aligned}
\mathbb{E}_{\mu}\{|\eta(X)-\eta_n(X)|\mid X\in Q\} &\leq \mathbb{E}_{\mu}\{|\eta(X)-\eta_n^*(X)|\mid X\in Q\}+\mathbb{E}_{\mu}\{|\eta_n^*(X)-\eta_n(X)|\mid X\in Q\}\\
&\leq \mathbb{E}_{\mu}\{|\eta(X)-\eta_n^*(X)|\mid X\in Q\}+\mathbb{E}_{\mu}\{|\hat{\eta}_n(X)-\hat{\eta}(X)|\mid X\in Q\}
\end{aligned}$$

(by (4.5)) and applying the above observaton with $\epsilon_1 = \varepsilon/2$ and $\epsilon_2 = \varepsilon/6$, we have

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{E}_\mu\{|\eta(X) - \eta_n(X)| \mid X \in Q\} > \frac{\varepsilon}{2}\right) \\
&\leq \mathbb{P}\left(\mathbb{E}_\mu\{|\eta(X) - \eta_n^*(X)| \mid X \in Q\} - \mathbb{E}_{\tilde{\mu}^n}\{\mathbb{E}_\mu\{|\eta(X) - \eta_n^*(X)| \mid X \in Q\}\} > \frac{\varepsilon}{6}\right) + \\
&\quad \mathbb{P}\left(\mathbb{E}_\mu\{|\hat{\eta}_n(X) - \hat{\eta}(X)| \mid X \in Q\} - \mathbb{E}_{\tilde{\mu}^n}\{\mathbb{E}_\mu\{|\hat{\eta}_n(X) - \hat{\eta}(X)| \mid X \in Q\}\} > \frac{\varepsilon}{6}\right), \qquad (4.7)
\end{aligned}$$

where we used the inequality (4.6). Now we will separately estimate the probability of deviations in the two last terms.

For the first term let θ be a function defined on labeled samples, $\theta : (\Omega \times \{0,1\})^n \to [0,\infty)$ as

$$\theta(\sigma_n) = \mathbb{E}_\mu\{|\eta(X) - \eta_n^*(X)| \mid X \in Q\}.$$

Let a new sample $\sigma_n^{'}$ be formed by replacing (x_i, y_i) with $(\hat{x}_i, \hat{y}_i)$. The difference of values of η_{ni}^* computed at the original sample and the altered one is at most $1/k$. For elements of Q, the value can only change at the points of the set $D(x_i, k/n) \cap Q$. According to Lemma 4.6, the μ-measure of the latter set is bounded by $(\beta+1)k/n$ whenever $r_{k/n}$ is sufficiently small (smaller than the scale s, in fact). Therefore, the normalized (conditional) measure of this set in Q is bounded by $(\beta+1)k/\mu(Q)n$, and

$$\begin{aligned}
|\theta(\sigma_n) - \theta(\sigma_n^{'})| &\leq \frac{1}{k} \cdot (\beta+1)\frac{k}{\mu(Q)n} \\
&= \frac{\beta+1}{\mu(Q)n}. \qquad (4.8)
\end{aligned}$$

Let us remind a classical concentration inequality.

Theorem 4.7 (Azuma, McDiarmid)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables taking values in a space Ω, and let a function $f\colon \Omega^n \to \mathbb{R}$ satisfy the following Lipschitz condition with regard to the Hamming distance: whenever just the i-th coordinate in the argument $(x_1, x_2, \ldots, x_n)$ is changed, the value of the function changes by at most $c_i > 0$. Then the probability of the deviation of the random variable $f(X_1, X_2, \ldots, X_n)$ from the expected value by at least $t > 0$ is bounded by*

$$2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

We conclude that

$$\mathbb{P}\left(\mathbb{E}_\mu\{|\eta(X) - \eta_n^*(X)| \mid X \in Q\} - \mathbb{E}_{\tilde{\mu}^n}\{\mathbb{E}_\mu\{|\eta(X) - \eta_n^*(X)| \mid X \in Q\}\} > \frac{\varepsilon}{6}\right) \leq 2\exp\left(-\frac{\varepsilon^2\mu(Q)^2 n}{18(\beta+1)^2}\right).$$

An identical argument applied to $\hat{\eta}_n$ results in a similar concentration estimate for the second term in equation (4.7), and we are done.

5. Strong consistency in the non-Archimedean case

Here we show that the randomized tie-breaking approach to the k-NN classifier in the presence of distance ties adopted by Devroye, Györfi, Krzyżak, and Lugosi [14] (see our Sect. 2.3) and used by them to prove the

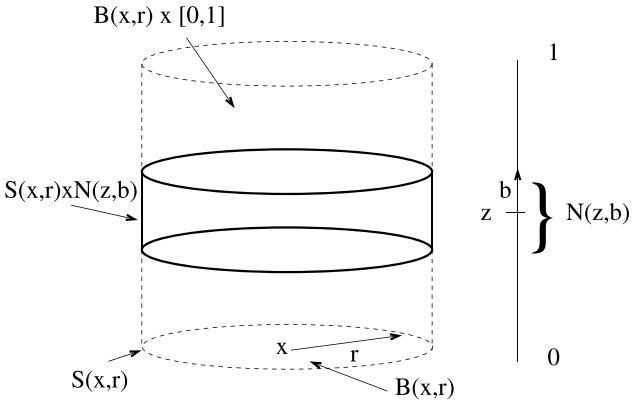

FIGURE 1. The set $B(x,z,r,b) = B(x,r) \times \mathbb{I} \cup S(x,r) \times N(z,b)$.

strong universal consistency of the k-NN classifier in the Euclidean setting works also in the case of metric spaces of non-Archimedian metric spaces: those whose metric satisfies the strong triangle inequality:

$$d(x,y) \leq \max\{d(x,z), d(z,y)\}.$$

However, the proof becomes somewhat trickier, revealing some interesting geometric features of non-Archimedian spaces with measure.

Theorem 5.1. *The k-NN classifier is strongly universally consistent in every complete separable non-Archimedian metric space, under the tie-breaking strategy of Devroye, Györfi, Krzyżak, and Lugosi.*

Remark 5.2. A slightly more general class of metric spaces is formed by those of Nagata dimension zero: a metric space is non-Archimedian if and only if it has Nagata dimension zero on the scale $s = +\infty$, see [5], Example 5.3. Our result above requires a minimal amount of adjustments to be extended to the complete separable metric spaces of Nagata dimension zero on some scale $s > 0$. We decided to avoid technicalities in order to make the argument in the proof of Lemma 5.7 below a little clearer.

We begin with combinatorial preparations. For $z \in \mathbb{I}$ and $b \geq 0$, denote

$$N(z,b) = \{x \in \mathbb{I} \colon |z - x| \leq b\}.$$

Let Ω be a metric space. Given $x \in \Omega$, $z \in \mathbb{I}$, $r, b > 0$, define, just like in [14], the set

$$B(x,z,r,b) = B(x,r) \times \mathbb{I} \cup S(x,r) \times N(z,b) \subseteq \Omega \times [0,1]. \tag{5.1}$$

(See Fig. 1.)

Now let $\alpha > 0$. Given $(x,z) \in \Omega \times \mathbb{I}$, denote $r_\alpha(x)$ as before (Eq. (4.1)), being the infimum of all the radii $r > 0$ such that the open ball of radius r around x has measure $\geq \alpha$. It the presence of atoms, it may happen that $r = 0$; we adopt the convention that $B(x,0)$ is the empty set, and $\bar{B}(x,0) = S(x,0) = \{x\}$.

Lemma 5.3. $\mu(B(x, r_\alpha(x)) \leq \alpha$.

Proof. The statement is trivially true if $r_\alpha(x) = 0$. Otherwise, approximate $r_\alpha(x) > 0$ with a strictly increasing sequence of radii $r_n \uparrow r_\alpha(x)$, and use sigma-additivity. □

Now define

$$b_\alpha(x,z) = \inf\{b>0\colon \mu\otimes\lambda(B(x,z,r_\alpha(x),b)) \geq \alpha\}. \tag{5.2}$$

Lemma 5.4. *The function $b_\alpha\colon \Omega\times\mathbb{I}\to\mathbb{R}$ is Borel measurable.*

Proof. One has

$$b_\alpha(x,z) = \begin{cases} b_\alpha(x,1/2), & \text{if } b_\alpha(x,1/2)\leq z\leq 1-b_\alpha(x,1/2),\\ 2b_\alpha(x,1/2)-\min\{z,1-z\} & \text{otherwise.}\end{cases}$$

Thus, it suffices to prove that $b_\alpha(x,1/2)$ is measurable as a function of $x\in\Omega$. This can be written as

$$b_\alpha(x,1/2) = \begin{cases} 0, & \text{if } \mu(S_{r_\alpha(x)}(x))=0,\\ \frac{1}{2}(\alpha-\mu(B_{r_\alpha(x)}(x))\mu(S_{r_\alpha(x)}(x))^{-1} & \text{if } \mu(\bar B_{r_\alpha(x)}(x))>\alpha,\\ \frac{1}{2}, & \text{if } \mu(\bar B_{r_\alpha(x)}(x))=\alpha>\mu(B_{r_\alpha(x)}(x)).\end{cases}$$

Everything now reduces to proving the measurability of the maps $x\mapsto\mu\left(B_{r_\alpha(x)}(x)\right)$ and $x\mapsto\mu\left(\bar B_{r_\alpha(x)}(x)\right)$.

As the function $x\mapsto r_\alpha(x)$ is continuous (in fact, 1-Lipschitz, see Lem. 4.5), when x is fixed and $x_n\to x$, we have $r_\alpha(x_n)\to r_\alpha(x)$. For every $\epsilon>0$, from the triangle inequality, when n is large enough, the closed ball $\bar B_{r_\alpha(x_n)}(x_n)$ is contained in the ϵ-neighbourhood of the ball $\bar B_{r_\alpha(x)}(x)$. When $\epsilon\downarrow 0$, the measure of this ϵ-neighbourhood converges to the measure of $\bar B_{r_\alpha(x)}(x)$ by sigma-additivity of μ, so we conclude

$$\limsup_{n\to\infty}\mu\left(\bar B_{r_\alpha(x_n)}(x_n)\right)\leq\mu\left(\bar B_{r_\alpha(x)}(x)\right).$$

Thus, the function $x\mapsto\mu\left(\bar B_{r_\alpha(x)}(x)\right)$ is upper semi-continuous, hence Borel measurable. An identical argument works for the sphere in place of the closed ball, and this suffices. □

For $r=r_\alpha(x)$ we have

$$\mu\otimes\lambda(B(x,r)\times\mathbb{I}) = \mu(B(x,r))\leq\alpha\leq\mu(\bar B(x,r)) = \mu\otimes\lambda(\bar B(x,r)\times\mathbb{I}).$$

In case where the two values are different, the function

$$[0,1]\ni b\mapsto\mu(S(x,r))\times\lambda(N(z,b))\in[0,\mu(S(x,r))]$$

is continuous and surjective, so the value α is achieved. We have:

Lemma 5.5. *For every α, $0<\alpha<1$,*

$$\mu\otimes\lambda(B((x,z,r_\alpha(x),b_\alpha(x,z))) = \alpha.$$

Now, given $\alpha\in(0,1]$ and $(x,z)\in\Omega\times\mathbb{I}$, define

$$\begin{aligned} D(x,z,\alpha) &= \{(y,w)\in\Omega\times\mathbb{I}\colon (x,z)\in B(y,w,r_\alpha(y),b_\alpha(y,w)))\}\\ &= \{(y,w)\in\Omega\times\mathbb{I}\colon \text{ either } d(x,y)<r_\alpha(y) \text{ or } d(x,y)=r_\alpha(y) \text{ and } |z-w|\leq b_\alpha(y,w)\}.\end{aligned}$$

The functions $d(x,y)$ and $|z-w|$ with x,z fixed are Lipschitz continuous with constant 1, as is $r_\alpha(y)$ by Lemma 4.5. The function $b_\alpha(y,w)$ is Borel measurable by Lemma 5.4. It follows that the set $D(x,z,\alpha)$, defined by inequalities involving those four functions, is Borel measurable.

The argument by Devroye *et al.* [14] was based on the following technical result (Lem. 3, *loco citato*): in the finite-dimensional Euclidean domain $\Omega = \mathbb{R}^d$, for every $x \in \Omega$, $z \in \mathbb{I}$ and $\alpha > 0$,

$$(\mu \otimes \lambda)(D(x,z,\alpha)) \leq C\alpha, \tag{5.3}$$

where $C = C(d)$ is a constant depending on the dimension of the space.

However, this kind of bound does not hold in more general finite-dimensional spaces in the sense of Nagata. In fact, it already fails in the non-archimedean metric spaces. Here is a counter-example.

Example 5.6. Let Ω be any infinite complete non-archimedean metric space having at least one non-isolated point. For example, one can take any of the classical examples such as the space of p-adic numbers $\mathbb{Q}_p$, or the Cantor space $\{0,1\}^\omega$ with the metric $d(x,y) = 2^{-\min\{i\colon x_i \neq y_i\}}$.

Fix a non-isolated point $x \in \Omega$ and a sequence x_n converging to x, such that $r_n = d(x,x_n)$ is strictly decreasing. Denote $S_n = S_n(x_n, r_n)$ the spheres, $B_n = B(x_n, r_n)$ the open balls, and $\bar{B}_n = \bar{B}(x_n, r_n)$ the closed balls.

For every n we have the following, somewhat counter-intuitive, property, due to the strong triangle inequality:

$$\bar{B}_{n+1} \subseteq S_n. \tag{5.4}$$

Indeed, let $y \in \bar{B}_{n+1}$, that is, $d(x_{n+1}, y) \leq r_{n+1}$. Then $d(x,y) \leq \max\{d(x,x_{n+1}), d(x_{n+1},y)\} = r_{n+1}$, because $x \in S_{n+1}$. We have:

$$d(x_n, y) \leq \max\{d(x_n,x), d(x,y)\} = r_n,$$

and at the same time

$$r_n = d(x_n,x) \leq \max\{d(x_n,y), d(y,x)\}.$$

So we must have $d(x_n,y) = r_n$, proving equation (5.4).

In particular, the open balls B_n are all two-by-two disjoint: if $n < m$, then $B_n \cap S_n = \emptyset$, while $B_m \subseteq S_n$. Also, the spheres S_n form a nested sequence: $S_1 \supseteq S_2 \supseteq \ldots$.

Choose a probability measure μ on Ω so that for $n = 1, 2, \ldots$

$$\mu(B_n) = \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)}.$$

Then $\cup B_n$ has a full measure. Denote

$$s_n = \mu(S_n) = \mu\left(\bigcup_{i=n+1}^{\infty} B_i\right) = \frac{1}{n+1}.$$

Fix a sufficiently small $\alpha > 0$, in the sense to be defined shortly. Denote $\zeta_n = \alpha(n+1)$. When $n \leq \alpha^{-1} - 1$, we have $\zeta_n \in [0,1]$. If $y \in B_n$, then $d(x,y) = \max\{d(x,x_n), d(x_n,y)\} = r_n$. We have $B(y,r_n) = B(x_n,r_n) = B_n$.

Therefore, for every $y \in B_n$ and $w \in [0,1]$,

$$\begin{aligned}(\mu \otimes \lambda)\left(B(y,w,r_n,\zeta_n)\right) &\le \mu(B_n) + 2\zeta_n \mu(S_n) \\ &= \frac{1}{n(n+1)} + 2\alpha \\ &\le 3\alpha,\end{aligned}$$

whenever $n \ge \alpha^{-1/2}$.

Now let x be the same non-isolated point as above, and set $z = 0$. By the above reasoning, the set $D(x,z,3\alpha)$ contains every pair (y,w) with $y \in B_n$ and $w \le \zeta_n$, provided that

$$\alpha^{-1/2} \le n \le \alpha^{-1} - 1.$$

When $\alpha > 0$ is sufficiently small, we have

$$\lceil \alpha^{-1/2} \rceil < \alpha^{-2/3} < \alpha^{-4/5} < \lfloor \alpha^{-1} \rfloor - 1.$$

Since the balls B_n are pairwise disjoint, we have

$$\begin{aligned}(\mu \otimes \lambda)\left(D(x,z,3\alpha)\right) &\ge \sum_{n=\lceil \alpha^{-1/2} \rceil}^{\lfloor \alpha^{-1} \rfloor - 1} \zeta_n \mu(B_n) \\ &= \sum_{n=\lceil \alpha^{-1/2} \rceil}^{\lceil \alpha^{-1} \rceil - 1} \alpha(n+1)\frac{1}{n(n+1)} \\ &= \alpha \sum_{n=\lceil \alpha^{-1/2} \rceil}^{\lceil \alpha^{-1} \rceil - 1} \frac{1}{n} \\ &\ge \alpha \int_{\alpha^{-2/3}}^{\alpha^{-4/5}} \frac{\mathrm{d}x}{x} \\ &= \alpha \left[-\frac{4}{5}\ln\alpha + \frac{2}{3}\ln\alpha \right] \\ &= -\frac{2\alpha \ln \alpha}{15},\end{aligned}$$

which expression is $\omega(\alpha)$ as $\alpha \to 0$. Thus, unlike in the Euclidean case, there is no upper bound on the size of the set $D(x,z,\alpha)$ that is linear in α.

However, this example does not contradict the strong consistency of the k-NN rule. Indeed, the proof of [14] proceeds as follows. The inequality (5.3), taken with $\alpha = k/n$, implies, like in our earlier argument (Eq. (4.8)), that the misclassification error of an auxiliary rule is a Lipschitz function on the cube $\Omega^n \times \{0,1\}^n$, equipped with the normalized Hamming distance, with C being the Lipschitz constant. The Azuma inequality bounds the probability that the error deviates by more than $\varepsilon > 0$ from the expectation by an expression of the form $2\exp(-\epsilon^2 n/C^2)$, and the sequence of such upper bounds is summable, allowing one to use the Borel–Cantelli lemma. If we now assume that the upper bound on the size of $D(x,z,\alpha)$ is of the form $-\alpha \ln \alpha$, and substitute $\alpha = k/n$, the Azuma inequality bounds the probability of a large deviation by something like $2\exp(-\mathrm{e}^2 n/(\ln n)^2)$. This sequence is still summable over n, so the Borel–Cantelli argument applies.

It turns out that in fact the upper bound in the above example is (up to a constant) exact.

Lemma 5.7. *Let Ω be a non-Archimedian metric space equipped with a Borel probability measure μ, let $x \in \Omega$, $z \in [0,1]$, and $\alpha > 0$. Then*

$$(\mu \otimes \lambda)(D(x,z,\alpha)) \leq 4\alpha(-\ln \alpha + 1).$$

Proof. Fix $\alpha > 0$. We will estimate the measure of the set $D(x,z,\alpha) \setminus (\{x\} \times \mathbb{I})$. If x is not an atom, this makes no difference. If $\mu\{x\} > 0$, then for any pair of the form $(x,w) \in D(x,z,\alpha)$ the product measure of the set $\{x\} \times N(z,|w-z|)$ does not exceed α. This means $|z-w| \leq \alpha\mu\{x\}^{-1}$. Consequently, the measure of the set $D(x,z,\alpha) \cap (\{x\} \times \mathbb{I})$ is bounded by 2α, and we will just add this value to our estimate at the end.

It simplifies things to estimate separately the measure of the intersection of the above set, $D(x,z,\alpha) \setminus (\{x\} \times \mathbb{I})$, with $\operatorname{supp}\mu \times (0,z)$ and with $\operatorname{supp}\mu \times (z,1)$; as the arguments are identical, we will only do the former.

By approximating the measurable set $(D(x,z,\alpha) \setminus \{x\} \times \mathbb{I}) \cap \operatorname{supp}\mu \times (0,z)$ with a compact subset K from inside to any given accuracy (Luzin's theorem), we can concentrate on bounding the measure of K.

Denote $\mathcal{V}$ the family of all open subsets of $\Omega \times \mathbb{I}$ of the form $B(y,r) \times (w,z)$, where $(y,w) \in K$, $w < z$, and $r = d(y,x)$. Notice that they cover K. Choose a finite subcover of K with sets of this form, say $B(y_i,r_i) \times (w_i,z)$, $i = 1,2,\ldots,N$. Because the metric is non-Archimedian, we can assume these open balls, $B(y_i,r_i)$, to be disjoint from each other. (Indeed, assume $B(y_i,r_i)$ and $B(y_j,r_j)$ intersect, $i \neq j$. Then one of them is entirely contained in the other, say $B(y_i,r_i) \subseteq B(y_j,r_j)$, and as $r_i = d(y_i,x) = d(y_j,x) = r_j$ by the strong triangle inequality, we have $B(y_i,r_i) = B(y_j,r_j)$. Now out of the two sets $B(y_i,r_i) \times (w_i,z)$ and $B(y_j,r_j) \times (w_j,z)$, one contains the other, depending on whether w_i or w_j is smaller, so one can be discarded.) Also, we can discard all balls of zero μ-measure. Order them in such a way that the radii r_i decrease, and whenever $r_i = r_{i+1}$, we have $w_i \leq w_{i+1}$.

Now, some more non-Archimedian geometry. For every i, $d(y_i,y_{i+1}) = r_i$: it cannot be strictly smaller because the open balls $B(y_i,r_i)$ and $B(y_{i+1},r_{i+1})$ are disjoint, and cannot be strictly larger because both points are at a distance $\leq r_i$ from x. As a consequence, $B(y_{i+1},r_{i+1}) \subseteq S(y_i,r_i)$. Indeed, if $y \in B(y_{i+1},r_{i+1})$, then $d(y_i,y) \leq \max\{d(y_i,y_{i+1}), d(y_{i+1},y) = r_i$, and the strict inequality is again impossible because the open balls do not meet. Notice that it is possible that $r_i = r_{i+1}$, in which case the closed ball $\bar{B}(y_{i+1},r_{i+1})$ will coincide with $\bar{B}(y_i,r_i)$. Write for short $B_i = B(y_i,r_i)$, $S_i = S(y_i,r_i)$. To sum up, the open balls B_i are two-by-two disjoint, and if $i < j$, then $B_j \subseteq S_i$. Also, write $\xi_i = z - w_i$.

Denote $b_i = \mu(B_i)$, $i = 1,2,\ldots$, and $s_i = \sum_{j=i+1}^{N} b_j$, $i = 0,1,2,\ldots$. These b_i and s_i are all strictly positive by the choice of K. Also, $s_i \leq \mu(S_i)$ for $i \geq 1$. For all $i = 1,2,\ldots,N$,

$$\mu(S_i) \times \xi_i \leq \alpha,$$

so in particular

$$\xi_i \leq \frac{\alpha}{\mu(S_i)} \leq \frac{\alpha}{s_i}.$$

Thus,

$$b_i\xi_i = \mu(B_i)\xi_i \leq (s_{i-1} - s_i)\frac{\alpha}{s_i} = \alpha\left(\frac{s_{i-1}}{s_i} - 1\right).$$

Denote

$$\gamma_i = \frac{s_{i-1}}{s_i} - 1 = \frac{s_{i-1} - s_i}{s_i}.$$

We have $\gamma_i > 0$. Let $n \leq N$ be the largest integer satisfying $s_n \geq \alpha$. (If it does not exist, then $\mu(K) \leq \alpha$ and

we are done.) Clearly,

$$\begin{aligned}\mu(K) &\le \sum_{i=1}^{N} \mu \otimes \lambda(B_i \times (w_i, z)) \\ &\le \sum_{i=1}^{n} b_i \xi_i + \mu\left(\cup_{i=n+1}^{N} B_i\right) \\ &\le \alpha \sum_{i=1}^{n} \gamma_i + \alpha. \end{aligned} \tag{5.5}$$

So it is enough to estimate $\sum_{i=1}^{n} \gamma_i$. With this purpose, write

$$\frac{s_i}{s_{i-1}} = 1 - \delta_i,$$

where $\delta_i > 0$. Thus,

$$\delta_i = 1 - \frac{s_i}{s_{i-1}} = \frac{s_{i-1} - s_i}{s_{i-1}} = \gamma_i \frac{s_i}{s_{i-1}},$$

and

$$\gamma_i = \delta_i \frac{s_{i-1}}{s_i}.$$

Also, $s_i = s_{i-1}(1 - \delta_i)$, so

$$s_n = s_0 \prod_{i=1}^{n} (1 - \delta_i) \le \prod_{i=1}^{n} (1 - \delta_i).$$

Notice that if for some $i \le N$ we have $s_{i-1}/s_i > 2$, then

$$s_i < s_{i-1} - s_i = b_i = \mu(B_i) \le \alpha,$$

meaning $i \ge n + 1$. Therefore, $s_{i-1}/s_i \le 2$ for all $i \le n$, so

$$\gamma_i \le 2\delta_i$$

for all $i = 1, \ldots, n$.

As for all $t \in [0, 1)$, $\ln(1 - t) \le -t$, we get:

$$\begin{aligned}\ln \alpha &\le \ln s_n \\ &\le \ln \prod_{i=1}^{n} (1 - \delta_i) \\ &= \sum_{i=1}^{n} \ln(1 - \delta_i) \\ &\le -\sum_{i=1}^{n} \delta_i, \end{aligned}$$

hence

$$\sum_{i=1}^{n} \delta_i \leq -\ln \alpha,$$

and

$$\sum_{i=1}^{n} \gamma_i \leq -2 \ln \alpha,$$

whence we get the estimate using (5.5) and the remark at the start of the proof. □

Remark 5.8. The main result of this section, Theorem 5.1, would be established in the general case of a complete separable metric space sigma-finite dimensional in the sense of Nagata if we could verify the following. Suppose a subspace Q of a complete metric space Ω has Nagata dimension β on a scale $s > 0$ in Ω. Is it true that, for some absolute constant $C > 0$ and all sufficiently small α,

$$\mu(D(x, z, \alpha) \cap Q) \leq -C(\beta + 1)\alpha \ln \alpha?$$

Of course one could think of weaker estimates that will also suffice.

We will model the proof of Theorem 5.1 on the proof of Theorem 1 in [14]. First, we remind that, by Lemma 5.5, for every pair (x, z) there is a unique pair $(r_{k/n}(x), b_{k/n}(x, z))$ defined as in equations (4.1) and (5.2) with $\alpha = k/n$. This leads us to define the regression function approximation

$$\eta_n^*(X, Z) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{I}_{\{(X_i, Z_i) \in B(X, Z, r_{k/n}(X), b_{k/n}(X, Z))\}} Y_i. \tag{5.6}$$

We also have the regression function approximation

$$\eta_n(X, Z) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{I}_{\{(X_i, Z_i) \in N_k(X, Z)\}} Y_i,$$

where the choice of the set $N_k(X, Z)$ of k nearest neighbours of X using the auxiliary random variable Z is made using the same tie-breaking strategy as described at the beginning of the section.

We need to prove that, first, the difference $\eta(X) - \eta_n(X, Z)$ converges to zero in expectation (or in probability), which would mean the (weak) consistency of the algorithm, and second, for every $\epsilon > 0$ the probabilities of an ϵ-deviation of $\eta(X) - \eta_n(X, Z)$ from its expected value taken over all n-samples form a summable sequence. This will allow to apply the first Borel–Cantelli lemma and deduce the strong consistency.

We have, taking the expectation over random samples (*i.e.*, $\mathbb{E}$ stands for $\mathbb{E}_{\sigma \sim \mu^n \otimes \lambda^n}$),

$$\begin{aligned} |\eta(X) - \eta_n(X, Z)| \leq\ & |\eta(X) - \mathbb{E}\eta_n^*(X, Z)| \\ & + |\mathbb{E}\eta_n^*(X, Z) - \eta_n^*(X, Z)| + |\eta_n^*(X, Z) - \eta_n(X, Z)|. \end{aligned} \tag{5.7}$$

We will verify the convergence to zero in expectation for all three terms. As to the deviation bound, the first term does not depend on a random sample, so the ϵ-deviation is improbable. We will deduce a summable bound for the second term, while the conclusion for the third will come for free as a particular case.

Notice that whenever a metric space with measure satisfies the strong Lebesgue–Besicovitch property (Eq. 3.2), that is, for a.e. x,

$$\frac{1}{\mu(B(x,r))}\int_{B(x,r)} f(x)\,\mathrm{d}\mu(x) \to f(x), \tag{5.8}$$

one can use closed balls in place of open balls and the a.e. convergence will still take place. Indeed, as every closed ball is the intersection of a sequence of open balls of the same centre, we have, by sigma-additivity,

$$\frac{1}{\mu(\bar{B}(x,r))}\int_{\bar{B}(x,r)} f(x)\,\mathrm{d}\mu(x) = \lim_{\epsilon\downarrow r}\frac{1}{\mu(B(x,\epsilon))}\int_{B(x,\epsilon)} f(x)\,\mathrm{d}\mu(x),$$

from where the statement follows.

Let now $f\colon \Omega \to \mathbb{R}$ be an $L^1(\mu)$-function. By the main theorem of Preiss from [2] (reproduced above as Thm. 3.8), combined with our observation in the previous paragraph, almost every $x \in \Omega$ has the property: given $\epsilon > 0$, one can select $\rho > 0$ so small that when $0 < r < \rho$, then the average value of f in the r-ball around x, either open or closed, is ϵ-close to $f(x)$. We can also see f as a function on $\Omega \times \mathbb{I}$ which only depends on the first argument, $x \in \Omega$. Now let x, $\epsilon > 0$, and r are as above, and $z, \xi \in [0,1]$. Denote provisionally

$$\overset{t}{B}(x,r) = \{y \in \Omega\colon (y,t) \in B(x,z,r,\xi)\} = \begin{cases} \bar{B}(x,r), & \text{if } |z-t| \le \xi, \\ B(x,r), & \text{otherwise} \end{cases}$$

the "horizontal section" of $B(x,z,r,\xi)$ at the height $t \in [0,1]$. Now the Fubini theorem implies

$$\begin{aligned}\int_{B(x,z,r,\xi)} |f(y)-f(x)|\,\mathrm{d}\mu(y)\mathrm{d}\lambda(t) &= \int_0^1 \mathrm{d}\lambda(t)\int_{\overset{t}{B}(x,r)} |f(y)-f(x)|\,\mathrm{d}\mu(y) \\ &< \int_0^1 \epsilon\mu(\overset{t}{B}(x,r))\,\mathrm{d}\lambda(t) \\ &= \epsilon(\mu\otimes\lambda)B(x,z,r,\xi).\end{aligned}$$

Thus, for μ-a.e. $x \in \Omega$, the average value of f over $B(x,z,r,\xi)$ converges to $f(x)$ when $r \downarrow 0$. In particular, this conclusion applies to the regression function η and its average value over $B(x,z,r_{k/n}(x), b_{k/n}(x,z))$ when $n,k \to \infty$ and $k/n \to 0$.

It follows that the expected value $\mathbb{E}_{\sigma\sim\mu^n\otimes\lambda^n}\eta_n^*(x,z)$ of the approximation $\eta_n^*(x,z)$ taken over all random labelled n-samples converges to $\eta(x)$ for a.e. x,z as $n,k \to \infty$ and $k/n \to 0$:

$$\begin{aligned}\mathbb{E}_{\sigma\sim\mu^n\otimes\lambda^n}\eta_n^*(x,z) &= \frac{1}{\mu\otimes\lambda(B(x,z,r_{k/n}(x),b_{k/n}(x,z)))}\int_{B(x,z,r_{k/n}(x),b_{k/n}(x,z))} \mathbb{E}(Y \mid X=x', Z=z')\,\mathrm{d}\mu(x')\mathrm{d}\lambda(z') \\ &\to \mathbb{E}(Y \mid X=x, Z=z) \\ &= \eta(x).\end{aligned}$$

By the dominated convergence theorem, the integral of the first term in equation (5.7) over our extended domain, $\Omega \times \mathbb{I}$, converges to zero:

$$\mathbb{E}_{\mu\otimes\lambda}|\eta(X) - \mathbb{E}_{\tilde{\mu}^n}\eta_n^*(X,Z)| \to 0.$$

For the second term in equation (5.7) we use the argument already seen in the proof of Lemma 4.3. If the labelled sample is changed in one labelled point, then the value of

$$|\mathbb{E}_{\tilde{\mu}^n}\eta_n^*(x,z) - \eta_n^*(x,z)|$$

may change by at most $1/k$, and only on a set of points (x,z) having measure at most $4(k/n)(\ln n - \ln k + 1)$, thanks to Lemma 5.7. Therefore, the integral

$$\int |\mathbb{E}_{\tilde{\mu}^n}\eta_n^*(x,z) - \eta_n^*(x,z)|\, d(x,z)$$

changes its value by at most $(4/n)(\ln n - \ln k + 1) \leq 4\ln n/n$. The Azuma–McDiarmid inequality (Thm. 4.7) implies that

$$\left| \int |\mathbb{E}\eta_n^*(x,z) - \eta_n^*(x,z)|\, d(x,z) - \mathbb{E}\int |\mathbb{E}\eta_n^*(x,z) - \eta_n^*(x,z)|\, d(x,z) \right| \leq 2\exp\left(-\frac{\mathrm{e}^2 n}{8(\ln n)^2}\right).$$

This is a summable sequence.

For the second term in equation (5.7) it remains to show convergence to zero in expectation. We perform a familiar trick with the Cauchy–Schwarz inequality and the variance:

$$\begin{aligned}
\mathbb{E}_{\tilde{\mu}^n}\int |\mathbb{E}_{\tilde{\mu}^n}\eta_n^*(x,z) - \eta_n^*(x,z)|\, d(x,z) &\leq \int \sqrt{\mathbb{E}_{\tilde{\mu}^n}\,|\mathbb{E}_{\tilde{\mu}^n}\eta_n^*(x,z) - \eta_n^*(x,z)|^2}\, d(x,z) \\
&\leq \int \sqrt{\frac{1}{k^2} n \mathrm{Var}\left(Y \mathbb{I}_{(X,Z)\in B(x,z,r_{k/n}(x),b_{k/n}(x,z))}\right)} d(x,z) \\
&\leq \int \sqrt{\frac{1}{k^2} n\mu\otimes\lambda(B(x,z,r_{k/n}(x),b_{k/n}(x,z))} d(x,z) \\
&\leq \int \sqrt{\frac{1}{k^2} n \frac{k}{n}} d(x,z) \\
&= \sqrt{\frac{1}{k}} \to 0.
\end{aligned}$$

Now the third term in equation (5.7). Let $(X_{(k)}, Z_{(k)})$ denote the k-th nearest neighbour of X in the random sample (in the order defined by the adopted tie-breaking). Denote $R_n = d(X, X_{(k)})$ and $B_n = |Z - Z_{(k)}|$. Then

$$\begin{aligned}
|\eta_n^*(X,Z) - \eta_n(X,Z)| &= \frac{1}{k}\left|\sum_{i=1}^n \mathbb{I}_{\{(X_i,Z_i)\in B(X,Z,r_{k/n}(X),b_{k/n}(X,Z))\}} Y_i - \sum_{i=1}^n \mathbb{I}_{\{(X_i,Z_i)\in B(X,Z,R_n,B_n)\}} Y_i\right| \\
&\leq \frac{1}{k}\sum_{i=1}^n \left|\mathbb{I}_{\{(X_i,Z_i)\in B(X,Z,r_{k/n}(X),b_{k/n}(X,Z))\}} - \mathbb{I}_{\{(X_i,Z_i)\in B(X,Z,R_n,B_n)\}}\right| \\
&= \frac{1}{k}\left|\sum_{i=1}^n \mathbb{I}_{\{(X_i,Z_i)\in B(X,Z,r_{k/n}(X),b_{k/n}(X,Z))\}} - 1\right|. \qquad (5.9)
\end{aligned}$$

We have used the following three observations: the empirical measure of the symmetric difference of the sets $B(X,Z,r_{k/n}(X),b_{k/n}(X,Z)$ and $B(X,Z,R_n,B_n)$ bounds the error, for the latter set this empirical measure is always one, and among the two intersections of the sample with these sets one always contains the other. Now

introduce the regression function $\hat{\eta} \equiv 1$ and the corresponding approximation

$$\hat{\eta}^* = \frac{1}{k} \sum_{i=1}^{n} \mathbb{I}_{\{(X_i,Z_i) \in B(X,Z,r_{k/n}(X),b_{k/n}(X,Z))\}}.$$

The last line of the equation (5.9) becomes $|\hat{\eta}^* - \mathbb{E}\hat{\eta}^*|$, and is therefore just a special case of the second term correspoding to the constant regression function $\hat{\eta} \equiv 1$.

6. The revised conjecture

We propose the following conjecture (a revised version of the conjecture previously stated by us in [5]).

Conjecture 5.1. For a complete separable metric space Ω, the following are equivalent.

1. The k-NN classifier is (weakly) universally consistent in Ω.
2. For every sigma-finite locally finite Borel measure μ on Ω, every $L^1(\mu)$-function $f\colon \Omega \to \mathbb{R}$ satisies the weak Lebesgue–Besicovitch differentiation property:

$$\frac{1}{\mu(B(x,r))} \int_{B(x,r)} f(x)\, \mathrm{d}\mu(x) \to f(x) \tag{6.1}$$

   in probability.
3. The space Ω is sigma-finite dimensional in the sense of de Groot, that is, one can represent Ω as a union of subspaces W_n in such a way that for each n and some $\delta_n \in \mathbb{N}$ and $s_n > 0$, every finite family of closed balls with centres in W_n having the same radii $< s_n$ admits a subfamily covering all the centres of the original balls and having multiplicity $\leq \delta_n + 1$ in Ω.

The implication (3)⇒(2) follows from the results of [7], and the implication (2)⇒(1) was established in [1]. Thus, only (1)⇒(3) needs to be verified.

Acknowledgements. We are most grateful to the anonymous ESAIM:PS referee who has read with utmost care both the original version of the article and the subsequent major revision of it, and whose reports permitted us to improve the paper very considerably, in particular to discover Example 5.6, state the correct version of Lemma 5.7, and clarify the proof of Proposition 2.2. Of course the remaining errors and obscure passages are all authors' own.

Funding. S.K. was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 19K20347 and KIXOXIA encouragement research grant at the earlier stages of this work. V.G.P. was supported by the DCR-A fellowship 300050/2022-4 of the Program of Scientific and Technological Development of the State of Paraíba, Brazil by CNPq and FAPESQ.

References

[1] F. Cérou and A. Guyader, Nearest neighbor classification in infinite dimension.*ESAIM Probab. Stat.* **10** (2006) 340–355.
[2] D. Preiss, Dimension of metrics and differentiation of measures. General topology and its relations to modern analysis and algebra, V (Prague, 1981). Heldermann, Berlin. *Sigma Ser. Pure Math.* **3** (1983) 565–568.
[3] J.I. Nagata, On a special metric and dimension. *Fund. Math.* **55** (1964) 181–194.
[4] P.A. Ostrand, A conjecture of J. Nagata on dimension and metrization. *Bull. Amer. Math. Soc.* **71** (1965) 623–625.
[5] B. Collins, S. Kumari and V.G. Pestov, Universal consistency of the k-NN rule in metric spaces and Nagata dimension. *ESAIM Probab. Stat.* **24** (2020) 914–934.
[6] C. Stone, *Consistent nonparametric regression. Ann. Stat.* **5** (1977) 595–645.
[7] P. Assouad and T. Quentin de Gromard, Recouvrements, derivation des mesures et dimensions.*Rev. Mat. Iberoam.* **22** (2006) 893–953.
[8] A. Korányi and H.M. Reimann, Foundations for the theory of quasiconformal mappings on the Heisenberg group. *Adv. Math.* **111** (1995) 1–87.
[9] E. Sawyer and R.L. Wheeden, Weighted inequalities for fractional integrals on Euclidean and homogeneous spaces. *Amer. J. Math.* **114** (1992) 813–874.

[10] L. Devroye and L. Györfi, Nonparametric Density Estimation. The L_1 View. John Wiley & Sons, New York (1985).
[11] L.C. Zhao, Exponential bounds of mean error for the nearest neighbor estimates of regression functions. *J. Multivariate Anal.* **21** (1987) 168–178.
[12] L. Devroye, L. Györfi and G. Lugosi, A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York (1996).
[13] S. Kumari, *Topics in Random Matrices and Statistical Machine Learning,* Ph.D. thesis, Kyoto University, 2018, 125 pp.
[14] L. Devroye, L. Györfi A. Krzyżak and G. Lugosi, On the strong universal consistency of nearest neighbor regression function estimates. *Ann. Statist.* **22** (1994) 1371–1385.
[15] D.H. Fremlin, Measure Theory. Vol. 2. Broad Foundations, corrected second printing of the 2001 original, Torres Fremlin, Colchester (2003) 563+12 pp. (errata).
[16] D. Preiss, Invalid Vitali theorems, in Abstracta. 7th Winter School on Abstract Analysis. Czechoslovak Academy of Sciences (1979) 58–60.
[17] L. Devroye, Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Z. Wahrsch. Verw. Gebiete* **61** (1982) 467–481.
[18] J. de Groot, On a metric that characterizes dimension. *Can. J. Math.* **9** (1957) 511–514.
[19] J. Cygan, Subadditivity of homogeneous norms on certain nilpotent Lie groups. *Proc. Amer. Math. Soc.* **83** (1981) 69–70.
[20] A. Kaplan, Fundamental solutions for a class of hypoelliptic PDE generated by composition of quadratic forms. *Trans. Amer. Math. Soc.* **258** (1980) 147–153.
[21] E. Le Donne, A primer on Carnot groups: homogenous groups, Carnot-Carathéodory spaces, and regularity of their isometries. *Anal. Geom. Metr. Spaces* **5** (2017) 116–137.
[22] R.A. Martínez Muñoz, *Novas regras de aprendizagem supervisionada utilizando a estrutura dos números p-ádicos* [*New supervised learning rules using the p-adic numbers structure* (in Portuguese)], Ph.D. thesis, Federal University of Santa Catarina, Florianópolis, Brazil, November 2023, 189 pp., https://repositorio.ufsc.br/handle/123456789/103512.