NEAREST NEIGHBOR CLASSIFICATION IN INFINITE DIMENSION

Frédéric Cérou[1] and Arnaud Guyader[2]

Abstract. Let X be a random element in a metric space $(\mathcal{F}, d)$, and let Y be a random variable with value 0 or 1. Y is called the class, or the label, of X. Let $(X_i, Y_i)_{1 \leq i \leq n}$ be an observed i.i.d. sample having the same law as (X, Y). The problem of classification is to predict the label of a new random element X. The k-nearest neighbor classifier is the simple following rule: look at the k nearest neighbors of X in the trial sample and choose 0 or 1 for its label according to the majority vote. When $(\mathcal{F}, d) = (\mathbb{R}^d, ||.||)$, Stone (1977) proved in 1977 the universal consistency of this classifier: its probability of error converges to the Bayes error, whatever the distribution of (X, Y). We show in this paper that this result is no longer valid in general metric spaces. However, if $(\mathcal{F}, d)$ is separable and if some regularity condition is assumed, then the k-nearest neighbor classifier is weakly consistent.

1. General definitions and results about classification

Let X be a random element with values in a metric space $(\mathcal{F}, d)$, and let Y be a random variable with values 0 or 1. The distribution of the pair (X, Y) is defined by:

- the probability measure μ of X:

$$\mu(B) = \mathbb{P}(X \in B) \text{ for all Borel sets } B \text{ in } \mathcal{F};$$

- and the regression function η of Y in X:

$$\eta(x) = \mathbb{P}(Y = 1 | X = x) \text{ for all } x \in \mathcal{F}.$$

Assume we observe n independent and identically distributed copies $(X_i, Y_i)_{1 \leq i \leq n}$: they are called the training data, and briefly denoted by D_n. Now we would like to guess the label Y of a new random element X, with $X \sim \mu$ independent of the training data. To this aim, one has to construct a function $g_n : \mathcal{F} \to \{0, 1\}$, called a classifier. This classifier is usually obtained from an approximation η_n of η using the simple decision rule:

$$g_n(x) = \left\{ \begin{array}{l} 1 \text{ if } \eta_n(x) > \frac{1}{2} \\ 0 \text{ otherwise.} \end{array} \right.$$
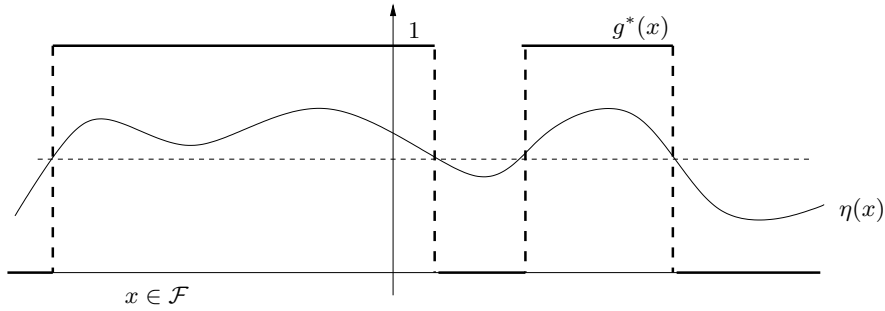
FIGURE 1. The regression function η and the associated Bayes classifier g^*.

It is easy to prove that the best possible solution is the Bayes classifier (see Fig. 1):

$$g^*(x) \triangleq \mathbb{1}_{\{\eta(X) \geq 1/2\}}.$$

More precisely, we have the following proposition (see [7] for a proof):

Proposition 1.1 (optimality of the Bayes classifier)**.** *The quantity* $L^* = \mathbb{P}(g^*(X) \neq Y)$ *is called the Bayes error, or the Bayes risk. For every classifier* $g_n : \mathcal{F} \to \{0, 1\}$

$$\mathbb{P}(g_n(X) \neq Y) \geq L^*.$$

More precisely, if $g_n(x) = \mathbb{1}_{\{\eta_n(x) \geq 1/2\}}$, *then:*

$$0 \leq \mathbb{P}(g_n(X) \neq Y) - L^* = 2 \int_{\mathcal{F}} \left| \eta(x) - \frac{1}{2} \right| \mathbb{1}_{\{g_n(x) \neq g^*(x)\}} \mu(\,\mathrm{d}x) \leq 2\mathbb{E}|(\eta - \eta_n)(X)|.$$

Remark 1.2. If the label Y is a deterministic function of X, then $L^* = 0$.

Of course, in general, one does not know the regression function η, nor the Bayes classifier g^*. From now on, we focus our attention on the k-nearest neighbor classifier, sometimes simply called "nearest neighbor classifier" from now on. Let us define the approximate regression function by:

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

where "$X_i \in k(X)$" means "X_i is one of the k nearest neighbors of X" and the notation $(X_{(i)}, Y_{(i)})$ (we should write $(X_{(i)}(X), Y_{(i)}(X))$ to be completely rigorous) means that the pairs $(X_i, Y_i)_{1 \leq i \leq n}$ have been re-indexed so that:

$$d(X, X_{(1)}) \leq d(X, X_{(2)}) \leq \cdots \leq d(X, X_{(n)}).$$

In case of equality, ties are broken by comparing auxiliary i.i.d. variables $\beta_1, \ldots, \beta_n$, independent of all the other random objects, and uniformly distributed on $(0, 1)$. More precisely, assume we observe $d(X, X_i) = d(X, X_j)$: if $\beta_i < \beta_j$, we decide $d(X_i, X) < d(X, X_j)$, otherwise we decide $d(X_i, X) > d(X, X_j)$. This rule has the interesting feature of making all the $n!$ orderings have the same probability of occurrence.

The associated decision function is

$$g_n(X) \triangleq 1_{\{\eta_n(x) \geq 1/2\}}.$$

The probability of error conditional on D_n is defined by:

$$L_n \triangleq \mathbb{P}(g_n(X) \neq Y | D_n).$$

L_n is a random variable, and its expectation $\mathbb{E}[L_n] = \mathbb{P}(g_n(X) \neq Y)$ is a real number depending on the parameters (k, n). We are interested in the asymptotic behavior, when: $n \to \infty$, $k \to \infty$ and $\frac{k}{n} \to 0$. By convention, from now on, we will simply write " $n \to \infty$" for these asymptotics.

Definition 1.3 (universal consistency). The k-nearest neighbor classifier is:
- universally weakly consistent if: $\lim_{n \to \infty} \mathbb{E}[L_n] = L^*$;
- universally strongly consistent if: $\lim_{n \to \infty} L_n = L^*$ almost surely.

The term "universally" means that the result is independent of the distribution μ and of the regression function η. In the following, we are only interested in weak consistency. The main result is due to Stone.

Theorem 1.4 (Stone (1977)). *With $(\mathcal{F}, d) = (\mathbb{R}^d, ||.||)$, the k-nearest neighbor classifier is universally weakly consistent.*

For the proof, we refer the reader to [7] or [13]. It is based on a geometrical result, known as Stone's Lemma. This powerful and elegant argument can unfortunately not be generalized to infinite dimension.

The notation $(\mathcal{F}, d) = (\mathbb{R}^d, ||.||)$ means that the metric d derives from a vector norm on $\mathbb{R}^d$. As we will see in the next section, this point is essential for the validity of the result. Universal strong consistency in $(\mathbb{R}^d, ||.||)$ has been proved by Devroye *et al.* [6].

2. Consistency in general metric spaces

2.1. Separability of the metric space

To generalize Stone's result, the first natural assumption is separability of the metric space $(\mathcal{F}, d)$. The following example shows that this condition is required even in finite dimension.

Example: a pathological distance on $[0, 1]$

Let us define a distance d on $[0, 1]$ as follows:

$$d(x, x') = \begin{cases} 0 & \text{if } x = x' \\ 1 & \text{if } xx' = 0 \text{ and } x \neq x' \\ 2 & \text{if } xx' \neq 0 \text{ and } x \neq x'. \end{cases}$$

Since the triangle inequality holds, d is a distance on $[0, 1]$. But $([0, 1], d)$ is clearly not separable.

Let μ be the simple distribution defined as follows: with probability one half, one picks the origin 0 ; with probability one half, one picks a point uniformly in $[0, 1]$. Mathematically speaking, if $\lambda_{[0,1]}$ denotes the Lebesgue measure on $[0, 1]$ and δ_0 the Dirac measure at the origin:

$$\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\lambda_{[0,1]}.$$

The way to attribute a label y to a point x in $[0, 1]$ is deterministic: if $x = 0$ then $y = 0$; if $0 < x \leq 1$ then $y = 1$. As Y is a deterministic function of X, the Bayes risk L^* is equal to zero. Nevertheless, it is intuitively

clear that the asymptotic probability of error with the nearest neighbors rule does not converge to 0:

$$\lim_{n\to\infty} \mathbb{E}[L_n] = \frac{1}{2} > L^* = 0.$$

So the nearest neighbors classifier is not weakly consistent in this context, although we are in finite dimension.

In general metric spaces, the separability assumption is sufficient to have convergence of the nearest neighbor to the point of interest. This is what Cover and Hart noticed in 1967 [3]. From now on we will assume that $(\mathcal{F}, d)$ is a separable metric space.

Proposition 2.1 (Cover and Hart (1967)). *If x is in the support of μ, then $\lim_{n\to\infty} d(X_k(x), x) = 0$ with probability one. If X is independent of the data and has probability measure μ, then with probability one*

$$\lim_{n\to\infty} d(X_k(X), X) = 0.$$

We refer to [7] for the proof[1].

2.2. **The Besicovitch condition**

As we will see later, separability of the metric space is not a sufficient assumption for consistency of the nearest neighbor classifier. It is necessary to put an assumption on the regularity of the regression function η with respect to the measure μ. More precisely, we will require a differentiation hypothesis that will be called "the Besicovitch condition". In what follows, we will use the symbol $B_{x,\delta}$ for the closed ball of radius δ centered at x.

Hypothesis (($\mathcal{H}$): the Besicovitch condition). *For every $\varepsilon > 0$*

$$\lim_{\delta\to 0} \mu\{x \in \mathcal{F} : \frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |\eta(z) - \eta(x)| \, \mathrm{d}\mu(z) > \varepsilon\} = 0.$$

Another formulation is the following convergence in probability:

$$\frac{1}{\mu(B_{X,\delta})} \int_{B_{X,\delta}} |\eta - \eta(X)| \, \mathrm{d}\mu \xrightarrow[\delta\to 0]{\mathbb{P}} 0$$

We will discuss this condition in the final section. Let us now give the main result of this paper.

Theorem 2.2 (consistency of the nearest neighbor classifier). *If $(\mathcal{F}, d)$ is separable and if the Besicovitch condition $\mathcal{H}$ is fulfilled, then the nearest neighbor classifier is weakly consistent*

$$\mathbb{E}[L_n] \xrightarrow[n\to\infty]{} L^*.$$

Proof. Thanks to Proposition 1.1, it is sufficient to show convergence in L^1:

$$\lim_{n\to\infty} \mathbb{E}|(\eta - \eta_n)(X)| = 0.$$

Let us introduce another approximation $\tilde{\eta}_n$ of the regression function:

$$\tilde{\eta}_n(x) = \frac{1}{k} \sum_{i=1}^{k} \eta(X_{(i)}(x)).$$

---

[1]The proof there is written in $\mathbb{R}^d$ with its usual norm, but the argument still works in any separable metric space.

Then the triangle inequality gives:

$$\mathbb{E}\,|(\eta - \eta_n)(X)| \le \mathbb{E}|(\eta - \tilde{\eta}_n)(X)| + \mathbb{E}|(\tilde{\eta}_n - \eta_n)(X)|.$$

• $\mathbb{E}|(\tilde{\eta}_n - \eta_n)(X)|$

This step is very classical. Cauchy-Schwarz inequality implies

$$
\begin{aligned}
\mathbb{E}|(\eta_n - \tilde{\eta}_n)(X)| &\le (\mathbb{E}[(\eta_n - \tilde{\eta}_n)^2(X)])^{1/2} \\
&= \left\{ \mathbb{E}\left[\left(\sum_{i=1}^{k} \frac{1}{k}(Y_{(i)} - \eta(X_{(i)})) \right)^2 \right] \right\}^{1/2},
\end{aligned}
$$

which gives

$$\mathbb{E}|(\eta_n - \tilde{\eta}_n)(X)| \le \left\{ \frac{1}{k^2} \sum_{1 \le i,j \le k} \mathbb{E}\left[(Y_{(i)} - \eta(X_{(i)})) \cdot (Y_{(j)} - \eta(X_{(j)})) \right] \right\}^{1/2}.$$

We use the conditioning trick

$$\mathbb{E}\left[(Y_{(i)} - \eta(X_{(i)})) \cdot (Y_{(j)} - \eta(X_{(j)})) \right] = \mathbb{E}\left[\mathbb{E}\left[(Y_{(i)} - \eta(X_{(i)})) \cdot (Y_{(j)} - \eta(X_{(j)}))|X_1, \ldots, X_n, \beta_1, \ldots, \beta_n, X \right] \right],$$

and remark that, conditionally on $X_1, \ldots, X_n, \beta_1, \ldots, \beta_n, X$, we have no problem with ties, the ordering is then fixed. Moreover, conditionally on X_i, the label Y_i has Bernoulli distribution with expectation $\eta(X_i)$ and is independent of all the other random objects. Thus, given $X_1, \ldots, X_n, \beta_1, \ldots, \beta_n, X$, the quantities $(Y_{(i)} - \eta(X_{(i)}))$ are independent centered random variables (see also [7] p.100 l+4). So if $i \ne j$

$$\mathbb{E}\left[(Y_{(i)} - \eta(X_{(i)})) \cdot (Y_{(j)} - \eta(X_{(j)})) \right] = 0.$$

Thus there only remains

$$\mathbb{E}|(\eta_n - \tilde{\eta}_n)(X)| \le \left\{ \frac{1}{k^2} \sum_{i=1}^{k} \mathbb{E}[(Y_{(i)} - \eta(X_{(i)}))^2] \right\}^{1/2},$$

and since $|Y_{(i)} - \eta(X_{(i)})| \le 1$, we have finally

$$\mathbb{E}|(\eta_n - \tilde{\eta}_n)(X)| \le \frac{1}{\sqrt{k}},$$

which proves that $\lim_{n \to \infty} \mathbb{E}|(\eta_n - \tilde{\eta}_n)(X)| = 0$.

• $\mathbb{E}|(\eta - \tilde{\eta}_n)(X)|$

Let $\mathcal{F}_0$ denote the support of μ. Then

$$\mathbb{E}|(\eta - \tilde{\eta}_n)(X)| = \int_{\mathcal{F}_0} \mathbb{E}|(\tilde{\eta}_n - \eta)(x)|\,\mathrm{d}\mu(x).$$

From now on we keep x fixed. We use the conditioning trick again

$$
\begin{aligned}
\mathbb{E}|(\tilde{\eta}_n - \eta)(x)| &= \mathbb{E}\left|\frac{1}{k}\sum_{i=1}^{k}(\eta(X_{(i)}) - \eta(x))\right| \\
&= \mathbb{E}\left[\mathbb{E}\left[\left|\frac{1}{k}\sum_{i=1}^{k}(\eta(X_{(i)}) - \eta(x))\right| \middle| d(x, X_{(k+1)}(x))\right]\right].
\end{aligned}
$$

To simplify notation, let us denote $d_{(j)} = d_{(j)}(x) = d(x, X_{(j)})$, for $1 \leq j \leq n$. Then, using Lemma **??** given in the appendix, we get:

$$
\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{k}\sum_{i=1}^{k}(\eta(X_{(i)}) - \eta(x))\right| \middle| d_{(k+1)}\right] &\leq \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left|(\eta(X_{(i)}) - \eta(x))\right| \middle| d_{(k+1)}\right] \\
&= \left(\tilde{\mu}(B_{x,d_{(k+1)}})\right)^{-1}\int_{B_{x,d_{(k+1)}}}|(\eta(x') - \eta(x))|\,\mathrm{d}\tilde{\mu}(x'),
\end{aligned}
$$

with $\tilde{\mu} = (\mathbb{1}_{U_{x,d_{(k+1)}}} + \frac{1}{2}\mathbb{1}_{S_{x,d_{(k+1)}}})\mu$, where we denote respectively $U_{x,r} = \{y \in \mathcal{F},\ d(x,y) < r\}$ and $S_{x,r} = \{y \in \mathcal{F},\ d(x,y) = r\}$ the open ball centered at x and of radius r, and its boundary. Now it is clear that for any measurable positive function φ, we have

$$
-\frac{1}{2}\int_{B_{x,d_{(k+1)}}}\varphi\,\mathrm{d}\mu \leq \int_{B_{x,d_{(k+1)}}}\varphi\,\mathrm{d}\tilde{\mu} - \int_{B_{x,d_{(k+1)}}}\varphi\,\mathrm{d}\mu \leq 0,
$$

so that

$$
\mathbb{E}\left[\left|\frac{1}{k}\sum_{i=1}^{k}(\eta(X_{(i)}) - \eta(x))\right| \middle| d_{(k+1)}\right] \leq 2\left(\mu(B_{x,d_{(k+1)}})\right)^{-1}\int_{B_{x,d_{(k+1)}}}|(\eta - \eta(x))|\,\mathrm{d}\mu.
$$

Thus we have:

$$
\mathbb{E}|(\eta - \tilde{\eta}_n)(X)| \leq 2\mathbb{E}\left[\left(\mu(B_{X,d_{(k+1)}})\right)^{-1}\int_{B_{X,d_{(k+1)}}}|(\eta - \eta(X))|\,\mathrm{d}\mu\right].
$$

Since the random variables $\left(\mu(B_{X,d_{(k+1)}})\right)^{-1}\int_{B_{X,d_{(k+1)}}}|(\eta - \eta(X))|\,\mathrm{d}\mu$ are all smaller than 1, the proof will be complete if we show that they converge in probability to 0. To see this, fix $\varepsilon > 0$, then for every $\delta_0 > 0$:

$$
\mathbb{P}\left(\left(\mu(B_{X,d_{(k+1)}})\right)^{-1}\int_{B_{X,d_{(k+1)}}}|(\eta - \eta(X))|\,\mathrm{d}\mu > \varepsilon\right) \leq \mathbb{P}(d_{(k+1)(X)} \geq \delta_0)
$$

$$
+ \sup_{0 \leq \delta \leq \delta_0}\mathbb{P}\left(\left(\mu(B_{X,\delta})\right)^{-1}\int_{B_{X,\delta}}|(\eta - \eta(X))|\,\mathrm{d}\mu > \varepsilon\right).
$$

Now, the first term goes to 0 thanks to Cover and Hart's result and the second one also thanks to the Besicovitch assumption $\mathcal{H}$. $\qquad\square$

Remark 2.3. In finite dimension, Devroye [5] has already mentioned that the Besicovitch condition is the cornerstone for nearest neighbor estimates as well as for kernel estimates.
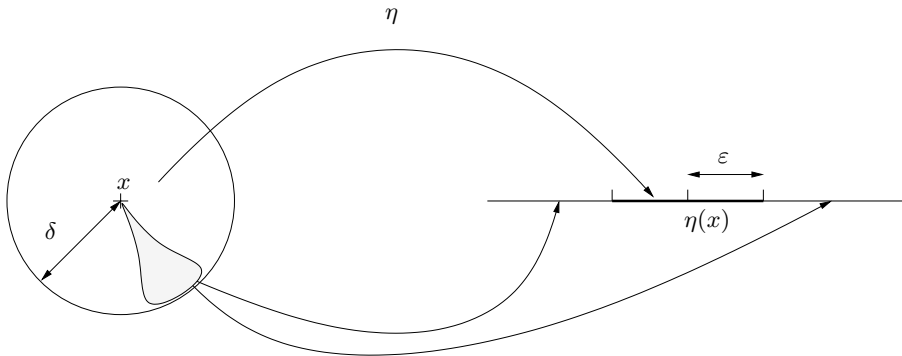
FIGURE 2. $\mu-$continuity: another way to see the Besicovitch condition.

3. DISCUSSION

3.1. Continuity of the regression function

It is clear that if η is continuous on $(\mathcal{F}, d)$, then the Besicovitch condition is automatically fulfilled. Nevertheless, intuitively, continuity is not necessary, since the principle of the nearest neighbor classifier is the following: to guess the label Y of a new point X, just average the labels Y_i for points X_i around X. The continuous version which ensures the validity of this averaging method has an integral form: this is exactly the Besicovitch condition.

As an illustration, we will give an example where the continuity condition on η is not fulfilled, but where the k-nearest neighbor classifier is still consistent anyway. Before that, we formulate a stronger but more tractable assumption than the Besicovitch condition.

Hypothesis (($\mathcal{H}'$): μ-continuity). *For every $\varepsilon > 0$, for μ almost every $x \in \mathcal{F}$*

$$\lim_{\delta \to 0} \frac{\mu \left\{ z \in \mathcal{F} : |\eta(z) - \eta(x)| > \varepsilon \cap d(x,z) < \delta \right\}}{\mu \left\{ z \in \mathcal{F} : d(x,z) < \delta \right\}} = 0.$$

This is a kind of continuity of η with respect to the measure μ, whence the name μ-continuity (see figure 2). Another equivalent definition is the following almost sure convergence:

$$\frac{1}{\mu(B_{X,\delta})} \int_{B_{X,\delta}} \mathbb{1}_{\{\eta - \eta(X) > \varepsilon\}} \, \mathrm{d}\mu \xrightarrow[\delta \to 0]{} 0 \qquad a.s.$$

Proposition 3.1 (μ-continuity $\Rightarrow$ the Besicovitch condition). *If the regression function η is μ-continuous, then the Besicovitch condition is fulfilled.*

Proof. For μ almost every $x \in \mathcal{F}$, take any $\varepsilon > 0$. Let us consider the following decomposition:

$$\begin{aligned}
\frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |\eta(z) - \eta(x)| \, \mathrm{d}\mu &= \frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |\eta(z) - \eta(x)| \mathbb{1}_{\{\eta(z) - \eta(x) \leq \varepsilon\}} \, \mathrm{d}\mu \\
&+ \frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |\eta(z) - \eta(x)| \mathbb{1}_{\{\eta(z) - \eta(x) > \varepsilon\}} \, \mathrm{d}\mu.
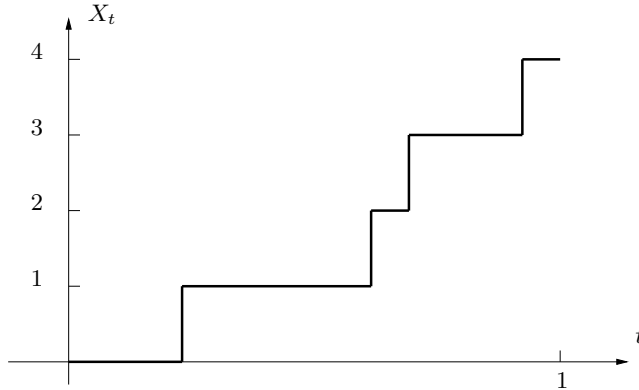\end{aligned}$$

FIGURE 3. A trajectory $(x_t)_{0 \leq t \leq 1}$ of a Poisson process.

Now we use the fact that for all x and z in $\mathcal{F}$, $|\eta(z) - \eta(x)| \leq 1$. This gives:

$$\frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |\eta(z) - \eta(x)| \, \mathrm{d}\mu \leq \varepsilon + \mu \left\{ z \in \mathcal{F} : |\eta(z) - \eta(x)| > \varepsilon | d(x,z) < \delta \right\}$$

So we have

$$\overline{\lim_{\delta \to 0}} \frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |\eta(z) - \eta(x)| \, \mathrm{d}\mu \leq \varepsilon.$$

Since ε is arbitrary, that gives

$$\lim_{\delta \to 0} \frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |\eta(z) - \eta(x)| \, \mathrm{d}\mu = 0.$$

Clearly, this almost sure convergence implies the Besicovitch condition $\mathcal{H}$. $\qquad \square$

More precisely, one can easily see that μ-continuity is equivalent to almost sure convergence in the Besicovitch condition:

$$\frac{1}{\mu(B_{X,\delta})} \int_{B_{X,\delta}} |\eta - \eta(X)| \, \mathrm{d}\mu \xrightarrow[\delta \to 0]{} 0 \qquad a.s.$$

As we see in the following example, the $\mu-$continuity may be easier to check than the Besicovitch condition.

Now for the example: $\mathcal{F}$ is the space of all possible realizations of a Poisson process of fixed intensity 1 between initial time 0 and final time 1. Its elements are denoted $x = (x_t)_{0 \leq t \leq 1}$ or $z = (z_t)_{0 \leq t \leq 1}$. The distance on $\mathcal{F}$ is derived from the L_1 norm:

$$d(x,z) = \|x - z\|_1 = \int_0^1 |x_t - z_t| \, \mathrm{d}t$$

It is clear that $(\mathcal{F}, \|.\|_1)$ is separable: consider for example the processes that jump at rational times between time 0 and time 1. This is a countable set and for every $\delta > 0$ and every $x \in \mathcal{F}$, there exists such a process in the ball $B_{x,\delta}$.

The label of a process x is deterministic and depends only on the final point of the process: if x_1 is even, then $y = 0$. If x_1 is odd, then $y = 1$. Since the label is deterministic, the Bayes risk L^* is null. Moreover, it is not difficult to see that η is nowhere continuous. Indeed, let us fix $x \in \mathcal{F}$, $\delta \in (0,1)$ and consider $z \in \mathcal{F}$ defined as follows (see Fig. 4):

$$z(t) = \begin{cases} x(t) & \text{if } 0 \leq t \leq 1 - \delta \\ x(t) + 1 & \text{if } 1 - \delta < t \leq 1. \end{cases}$$
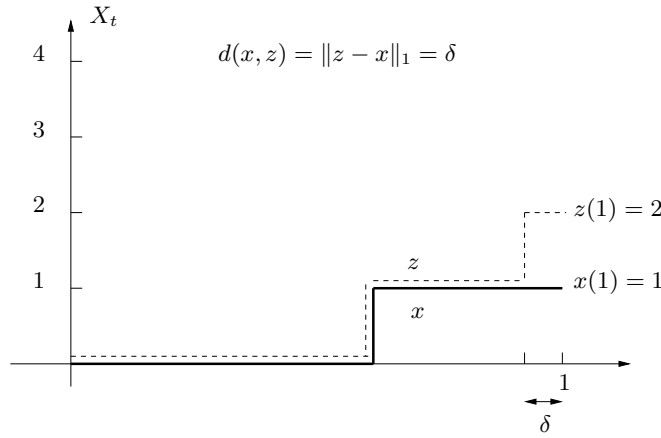
FIGURE 4. The curves x and x' do not have the same label.

So z is at distance δ from x but has not the same label as x: since δ is arbitrary, this proves that η is not continuous at point x. Since x is arbitrary, this proves that η is nowhere continuous. Nevertheless, we prove now that the nearest neighbor rule is consistent, since the Besicovitch condition is fulfilled. For this, we use the more tractable formulation $\mathcal{H}'$. Let us fix $\varepsilon > 0$ and $x \in \mathcal{F}$. The aim is to show that

$$\lim_{\delta \to 0} \frac{\mu\left(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta}\right)}{\mu(B_{x,\delta})} = 0. \tag{1}$$

We can estimate this quantity.

Lemma 3.2 (the ratio of small balls).

$$\lim_{\delta \to 0} \frac{\mu(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta})}{\mu(B_{x,\delta})} = 0.$$

Proof. In a first time, let us suppose that the number M of jumps of the process x is strictly positive and denote by $t_1, \ldots, t_M$ the jumping times, and $t_0 = 0$. In the denominator, a process z is in $B_{x,\delta}$ if it jumps at times $t'_1, \ldots, t'_M$ which are close enough to $t_1, \ldots, t_M$:

$$\forall i \in \{1, \ldots, M\} \qquad |t'_i - t_i| < \frac{\delta}{M}.$$

This is implied by:

$$\forall i \in \{1, \ldots, M\} \qquad t_i - t_{i-1} - \frac{\delta}{M^2} < t'_i - t'_{i-1} < t_i - t_{i-1} + \frac{\delta}{M^2}.$$

Thus

$$\mu\left(z : \forall i \in \{1, \ldots, M\}, \ |t'_i - t_i| < \frac{\delta}{M}\right)$$

$$\geq \ \mu\left(z : \forall i \in \{1, \ldots, M\}, \ t_i - t_{i-1} - \frac{\delta}{M^2} < t'_i - t'_{i-1} < t_i - t_{i-1} + \frac{\delta}{M^2}\right)$$

$$= \ \prod_{i=1}^{M} \int_{-\frac{\delta}{M^2}}^{+\frac{\delta}{M^2}} \exp[-(t_i - t_{i-1}) - s] \, \mathrm{d}s$$

$$\geq \ \left[2\sinh\left(\frac{\delta}{M^2}\right)\right]^M \exp\left[-M \max_{i \in \{1, \ldots, M\}}(t_i - t_{i-1})\right].$$

But doing this, we have not enumerated all processes $z \in B_{x,\delta}$, so we can only conclude that

$$\mu(B_{x,\delta}) \geq C_1(x)\left[\sinh\left(\frac{\delta}{M^2}\right)\right]^M,$$

for some constant $C_1(x)$ which does not depend on δ. For the numerator, since η takes values 0 and 1, we have

$$\mu\left(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta}\right) = \mu\left(\{\eta(z) \neq \eta(x)\} \cap B_{x,\delta}\right).$$

Thus, if we consider the processes z with M jumps at times $t'_1, \ldots, t'_M$ such that

$$\forall i \in \{1, \ldots, M\} \qquad |t'_i - t_i| < \delta,$$

and other jumps at times $t'_{M+1}, \ldots$ such that

$$\forall i > M \qquad |t'_i - 1| < \delta,$$

we get a set S of Poisson processes which is bigger than the one of interest. Briefly speaking

$$(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta}) \subset S.$$

Now, being an element of S implies

$$\forall i \in \{1, \ldots, M\} \qquad t_i - t_{i-1} - \delta < t'_i - t'_{i-1} < t_i - t_{i-1} + \delta,$$

and

$$1 - t_M - 2\delta < t'_{M+1} - t'_M < 1 - t_M + \delta < 1 - t_M + 2\delta.$$

Which gives:

$$\mu\left(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta}\right) = \mu\left(\{\eta(z) \neq \eta(x)\} \cap B_{x,\delta}\right)$$

$$\leq \ \prod_{i=1}^{M} \int_{-\delta}^{+\delta} \exp[-(t_i - t_{i-1}) - s] \, \mathrm{d}s \int_{-2\delta}^{2\delta} \exp[-(1 - t_M) - s] \, \mathrm{d}s$$

$$\leq \ [\sinh(\delta)]^M . \sinh(2\delta). \exp[-M \min_{i \in \{1, \ldots, M\}}(t_i - t_{i-1}) - (1 - t_M)]$$
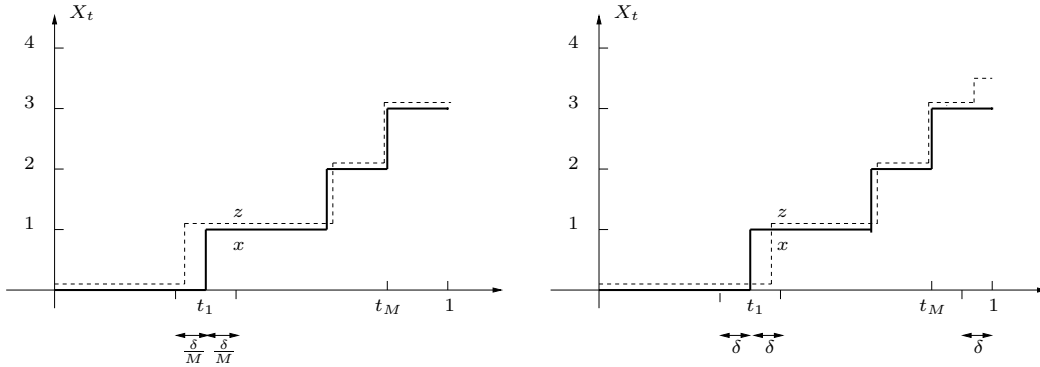
$$= \ C_2(x).[\sinh(\delta)]^M . \sinh(2\delta),$$

FIGURE 5. Estimation of the numerator and denominator of equation (1).

for some constant $C_2(x)$ which does not depend on δ. The ratio of small balls is now

$$\frac{\mu(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta})}{\mu(B_{x,\delta})} \leq \frac{C_2(x)}{C_1(x)} \left[\frac{\sinh(\delta)}{\sinh(\frac{\delta}{M^2})}\right]^M \sinh(2\delta) \xrightarrow[\delta \to 0]{} 0.$$

If x has no jump, the reasoning is the same. □

The result of the Lemma ensures that for every $x \in \mathcal{F}$ and for every $\varepsilon > 0$

$$\lim_{\delta \to 0} \mu\left\{z \in \mathcal{F} : |\eta(z) - \eta(x)| > \varepsilon | d(x, z) < \delta\right\} = 0.$$

As we have seen before, this implies the Besicovitch condition. So the nearest neighbor rule is consistent, although η is nowhere continuous.

3.2. **The Besicovitch assumption in infinite dimension**

In this section, we discuss the the Besicovitch condition. If we restrict ourselves to finite dimension and suppose that the distance d comes from a vector norm, the essential result is the following one (see for instance [8], Chapter 1.7, pp 43-44).

Theorem 3.3 (Lebesgue-Besicovitch differentiation theorem). *Let μ be a Radon measure on $\mathbb{R}^d$ and $f \in L^p_{\text{loc}}(\mathbb{R}^d)$, then*

$$\lim_{\delta \to 0} \frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |f - f(x)|^p \, \mathrm{d}\mu = 0$$

for μ almost every x.

In our case, μ is a probability measure on $\mathbb{R}^d$ and η is bounded by 1, so this result can be directly applied. Devroye [5] has already noted that this is another way to prove Stone's theorem.

Corollary 3.4 (Stone's theorem). *In $(\mathbb{R}^d, ||.||)$, the k-nearest neighbor classifier is universally weakly consistent.*

Remark 3.5. If we wish to estimate the regression function η, our reasoning shows, moreover, that with the nearest neighbor method

$$\lim_{n \to \infty} \mathbb{E}\left[|\eta_n(x) - \eta(x)|^p\right] = 0$$

for μ almost every x.

The Besicovitch condition $\mathcal{H}$ appears also in recent papers on connected problems: Abraham, Biau and Cadre [1] use it for function classification with the kernel rule. Dabo-Niang and Rhomari [4] use it for nonparametric regression estimation in general metric spaces.

Now the question is: what about the Besicovitch density condition in general metric spaces[2]?

This question has been studied by several authors in geometric measure theory, see for example [9, 11, 12]. Unlike the situation in $\mathbb{R}^d$, this condition is no longer automatically fulfilled in infinite dimension. In [11], Preiss introduces a rather technical notion, called the σ-finite dimensionality of a metric on a space. He shows that it is the *sine qua non* condition for the Besicovitch property for all measures on a metric space. Without delving into the details of this notion, let us just mention that it is related to the σ-finite dimensionality of the space. In fact, to return to the Poisson processes above, we were in precisely this situation.

Example. Fix $M \geq 0$ and denote $\mathcal{F}_M$ all possible realizations of the Poisson process that have exactly M jumps. A process that has M jumps can be summarized by an M-dimensional vector of jump times. Then it is obvious that the metric space $(\mathcal{F}_M, \|.\|_1)$ is isometric to $([0, 1]^M, \|.\|_1)$. So that we have the correspondence

$$(\mathcal{F}, \|.\|_1) = \bigcup_{M=0}^{+\infty} (\mathcal{F}_M, \|.\|_1) \sim \bigcup_{M=0}^{+\infty} ([0, 1]^M, \|.\|_1),$$

and the σ-finite dimensionality is clear.

Let us focus now on the classical situation where $(\mathcal{F}, d)$ is a separable Hilbert space and μ a Gaussian measure. Let ν denote the centered and normalized Gaussian measure on $\mathbb{R}$, let (c_n) be a non-increasing sequence of positive numbers such that $\sum_{n=0}^{+\infty} c_n < +\infty$ and let $\ell_2(c)$ be the set of all sequences $x = (x_n)$ such that

$$|x|^2 = \sum_{n=0}^{+\infty} c_n x_n^2 < +\infty.$$

Then the measure $\mu = \nu^{\otimes \mathbb{N}}$ is a σ-additive measure in a Hilbert space, $\ell_2(c)$. More precisely, each Gaussian measure can be represented in this way.

Even in this rather comfortable context, one has to put conditions on the sequence (c_n) to get the Besicovitch property. To be precise, Preiss and Tišer [12] have shown the following result: if there exists $q < 1$ such that

$$\forall n \in \mathbb{N} \qquad \frac{c_{n+1}}{c_n} < q,$$

then the Besicovitch property is true for every function $f \in L^1(\mu)$. Roughly speaking, if we see (c_n) as the sequence of variances of μ along the direction of base vectors, it means that these variances have to decay exponentially fast: this is a very strong condition.

Now let us see an example which shows that if the Besicovitch condition is not fulfilled, there is not much hope for classification with the nearest neighbor rule. This example is due to Preiss [10].

Example: a problematic case for nearest neighbor classification
In this paper, Preiss constructs a Gaussian measure μ in a separable Hilbert space $\mathcal{F}$ and a Borel set $M \subset F$ with $\mu(M) < 1$ such that
$$\lim_{\delta \to 0} \frac{\mu(M \cap B_{x,\delta})}{\mu(B_{x,\delta})} = 1$$
for μ almost every $x \in \mathcal{F}$.

---

[2]Of course, we still suppose that the metric space is separable.

Now suppose that X is distributed with respect to μ and its label Y is deterministic

$$Y = \mathbb{1}_M(X).$$

As usual the Bayes risk is equal to 0. Nevertheless, if we try to apply the nearest neighbor rule to this example, it is intuitively clear that we have some problems in classifying elements $x \in \overline{M}$. Indeed, one can easily prove that

$$\varlimsup_{n\to\infty} L_n^* \geq \frac{1}{2}\mu(\overline{M}) > L^* = 0.$$

It is worth mentioning that this result is not in contradiction with the one of Biau *et al.* [2]. In this paper, they consider a random variable X taking values in a separable Hilbert space $\mathcal{X}$, with label $Y \in \{0,1\}$. They establish the universal weak consistency of a neighbor-**type** classifier, but not of the **classical** nearest neighbor classifier. More precisely, they reduce the infinite dimension of $\mathcal{X}$ by considering only the first d coefficients of the Fourier series expansion of each X_i, and then perform nearest neighbor classification in $\mathbb{R}^d$. In fact, their result and the example above suggest that in infinite dimension, the classical nearest neighbor classification is not the right way to proceed.

Appendix A. Technical lemma

In this section we use the notation of the proof of Theorem 2.2. And for all x in $\mathcal{F}$ and $r \geq 0$ we denote respectively by $B_{x,r}$, $U_{x,r}$ and $S_{x,r}$ the closed ball and the open ball centered at x and of radius r, and their boundary. We recall that in case of equality, the ties are broken by comparing auxiliary i.i.d. variables $\beta_1, \ldots, \beta_n$, independent of all the other random objects, and uniformly distributed in $(0,1)$.

Lemma A.1. Let F be a μ-integrable real function on $\mathcal{F}$. For all x in the support of μ,

$$\mathbb{E}[\frac{1}{k}\sum_{j=1}^{k} F(X_{(j)})| d_{(k+1)}] = C \int_{B_{x,d_{(k+1)}}} F(x')\, \mathrm{d}\tilde{\mu}(x'),$$

where $\tilde{\mu} = (\mathbb{1}_{U_{x,d_{(k+1)}}} + \frac{1}{2}\mathbb{1}_{S_{x,d_{(k+1)}}})\mu$, and C is a normalizing $d_{(k+1)}$-measurable constant

$$C = \left(\tilde{\mu}(B_{x,d_{(k+1)}})\right)^{-1}.$$

Proof. Let $\mathcal{Q}(n)$ be the set of all n-permutations, Q denote the random permutation given by the ordering of the nearest neighbor, and $\tilde{\mathcal{Q}}(n,k)$ all the subsets of k elements in $\{1,\ldots,n\}$. C' will denote a deterministic constant. For all $x \in \mathcal{F}$ and all $j \in \{1,\ldots,n\}$ we will also use the notation $d_j = d(x,X_j)$. We have

$$
\begin{aligned}
\mathbb{E}\left[\sum_{j=1}^{k} F(X_{(j)})| d_{(k+1)}\right] &= \mathbb{E}\left[\sum_{j=1}^{k} F(X_{Q(j)}) \sum_{q\in\mathcal{Q}(n)} \mathbb{1}_{Q=q}| d_{Q(k+1)}\right] \\
&= \sum_{q\in\mathcal{Q}(n)} \mathbb{E}\left[\sum_{j=1}^{k} F(X_{q(j)})\mathbb{1}_{Q=q}| d_{Q(k+1)}\right] \\
&= \sum_{\{q(1),\ldots,q(k)\}\in\tilde{\mathcal{Q}}(n,k)} \mathbb{E}\left[\sum_{j=1}^{k} F(X_{(j)})\mathbb{1}_{\{Q(1),\ldots,Q(k)\}=\{q(1),\ldots,q(k)\}}| d_{Q(k+1)}\right] \\
&= C'\, \mathbb{E}\left[\sum_{j=1}^{k} F(X_j)\mathbb{1}_{\{Q(1),\ldots,Q(k)\}=\{1,\ldots,k\}}| d_{Q(k+1)}\right].
\end{aligned}
$$

The last two equalities come from reordering the terms in the summation, and the fact that all the orderings have the same probability. Then we decompose the event:

$$\mathbb{1}_{\{Q(1),\ldots,Q(k)\}=\{1,\ldots,k\}} \quad = \quad \sum_{\ell=k+1}^{n}\prod_{j=1}^{k}(\mathbb{1}_{d_j<d_\ell}+\mathbb{1}_{\beta_j<\beta_\ell}\mathbb{1}_{d_j=d_\ell})\times\prod_{\substack{h=k+1\\h\neq\ell}}^{n}(\mathbb{1}_{d_h>d_\ell}+\mathbb{1}_{\beta_h\geq\beta_\ell}\mathbb{1}_{d_h=d_\ell}).$$

Then we have

$$\mathbb{E}\left[\sum_{j=1}^{k}F(X_j)\mathbb{1}_{\{Q(1),\ldots,Q(k)\}=\{1,\ldots,k\}}\big|\,d_{Q(k+1)}\right]=$$

$$\sum_{\ell=k+1}^{n}\mathbb{E}\left[\sum_{j=1}^{k}F(X_j)\prod_{j=1}^{k}(\mathbb{1}_{d_j<d_\ell}+\mathbb{1}_{\beta_j<\beta_\ell}\mathbb{1}_{d_j=d_\ell})\prod_{\substack{h=k+1\\h\neq\ell}}^{n}(\mathbb{1}_{d_h>d_\ell}+\mathbb{1}_{\beta_h\geq\beta_\ell}\mathbb{1}_{d_h=d_\ell})\big|\,d_\ell\right],$$

using the fact that the integrand can be non-zero only when $d_\ell = d_{Q(k+1)}$. Then we use the fact that the couples (X_j, d_j) for $j \neq \ell$ are i.i.d., and we get:

$$\mathbb{E}\left[\sum_{j=1}^{k}F(X_j)\prod_{j=1}^{k}(\mathbb{1}_{d_j<d_\ell}+\mathbb{1}_{\beta_j<\beta_\ell}\mathbb{1}_{d_j=d_\ell})\prod_{\substack{h=k+1\\h\neq\ell}}^{n}(\mathbb{1}_{d_h>d_\ell}+\mathbb{1}_{\beta_h\geq\beta_\ell}\mathbb{1}_{d_h=d_\ell})\big|\,d_\ell\right]$$

$$=\quad \sum_{j=1}^{k}\mathbb{E}\left[F(X_j)(\mathbb{1}_{d_j<d_\ell}+\mathbb{1}_{\beta_j<\beta_\ell}\mathbb{1}_{d_j=d_\ell})|d_\ell\right]\times\mathbb{E}\left[\prod_{\substack{m=1\\m\neq j}}^{k}(\mathbb{1}_{d_m>d_\ell}+\mathbb{1}_{\beta_m\geq\beta_\ell}\mathbb{1}_{d_m=d_\ell})|\,d_\ell\right]$$

$$\times\mathbb{E}\left[\prod_{\substack{h=k+1\\h\neq\ell}}^{n}(\mathbb{1}_{d_h>d_\ell}+\mathbb{1}_{\beta_h\geq\beta_\ell}\mathbb{1}_{d_h=d_\ell})|\,d_\ell\right]$$

$$=\quad k\mathbb{E}\left[F(X_1)(\mathbb{1}_{d_1<d_\ell}+\mathbb{1}_{\beta_1<\beta_\ell}\mathbb{1}_{d_1=d_\ell})|d_\ell\right]\times\mathbb{E}\left[\prod_{m=2}^{k}(\mathbb{1}_{d_m>d_\ell}+\mathbb{1}_{\beta_m\geq\beta_\ell}\mathbb{1}_{d_m=d_\ell})|\,d_\ell\right]$$

$$\times\mathbb{E}\left[\prod_{\substack{h=k+1\\h\neq\ell}}^{n}(\mathbb{1}_{d_h>d_\ell}+\mathbb{1}_{\beta_h\geq\beta_\ell}\mathbb{1}_{d_h=d_\ell})|\,d_\ell\right]$$

$$=\quad \frac{k}{\tilde{\mu}(B_{x,d_\ell})}\int_{B_{x,d_\ell}}F\,\mathrm{d}\tilde{\mu}\times\mathbb{E}\left[\prod_{m=1}^{k}(\mathbb{1}_{d_m>d_\ell}+\mathbb{1}_{\beta_m\geq\beta_\ell}\mathbb{1}_{d_m=d_\ell})|\,d_\ell\right]$$

$$\times\mathbb{E}\left[\prod_{\substack{h=k+1\\h\neq\ell}}^{n}(\mathbb{1}_{d_h>d_\ell}+\mathbb{1}_{\beta_h\geq\beta_\ell}\mathbb{1}_{d_h=d_\ell})|\,d_\ell\right],$$

having noticed that

$$\mathbb{E}[F(X_1)(\mathbb{1}_{d_1<d_\ell} + \mathbb{1}_{\beta_1<\beta_\ell}\mathbb{1}_{d_1=d_\ell})|d_\ell] = \int_{B_{x,d_\ell}} F \, d\tilde\mu$$

and

$$\mathbb{E}[\mathbb{1}_{d_1<d_\ell} + \mathbb{1}_{\beta_1<\beta_\ell}\mathbb{1}_{d_1=d_\ell}|d_\ell] = \tilde\mu(B_{x,d_\ell}).$$

Then using the same arguments, we use the reverse transformations:

$$\sum_{\ell=k+1}^n \mathbb{E}\left[\sum_{j=1}^k \frac{1}{k} F(X_j) \prod_{j=1}^k (\mathbb{1}_{d_j<d_\ell} + \mathbb{1}_{\beta_j<\beta_\ell}\mathbb{1}_{d_j=d_\ell}) \prod_{\substack{h=k+1\\h\neq\ell}}^n (\mathbb{1}_{d_h>d_\ell} + \mathbb{1}_{\beta_h\geq\beta_\ell}\mathbb{1}_{d_h=d_\ell})| \, d_\ell\right]$$

$$= \sum_{\ell=k+1}^n \mathbb{E}\left[\frac{1}{\tilde\mu(B_{x,d_\ell})} \int_{B_{x,d_\ell}} F \, d\tilde\mu\right.$$

$$\times \prod_{m=1}^k (\mathbb{1}_{d_m>d_\ell} + \mathbb{1}_{\beta_m\geq\beta_\ell}\mathbb{1}_{d_m=d_\ell}) \prod_{\substack{h=k+1\\h\neq\ell}}^n (\mathbb{1}_{d_h>d_\ell} + \mathbb{1}_{\beta_h\geq\beta_\ell}\mathbb{1}_{d_h=d_\ell})| \, d_\ell\bigg]$$

$$= \mathbb{E}\left[\frac{1}{\tilde\mu(B_{x,d_{Q_{(k+1)}}})} \left(\int_{B_{x,d_{Q_{(k+1)}}}} F \, d\tilde\mu\right) \mathbb{1}_{\{Q(1),\ldots,Q(k)\}=\{q(1),\ldots,q(k)\}}| \, d_{Q_{(k+1)}}\right]$$

$$= \mathbb{E}\left[\mathbb{1}_{\{Q(1),\ldots,Q(k)\}=\{q(1),\ldots,q(k)\}} \frac{1}{\tilde\mu(B_{x,d_{Q_{(k+1)}}})} \int_{B_{x,d_{Q_{(k+1)}}}} F \, d\tilde\mu| \, d_{Q_{(k+1)}}\right].$$

We conclude the proof by reordering all the permutations again:

$$\mathbb{E}\left[\sum_{j=1}^k F(X_{(j)})| \, d_{(k+1)}\right] = \sum_{\{q(1),\ldots,q(k)\}\in\tilde{\mathcal{Q}}(n,k)} \mathbb{E}\left[\sum_{j=1}^k \mathbb{1}_{\{Q(1),\ldots,Q(k)\}=\{q(1),\ldots,q(k)\}}\right.$$

$$\times \frac{1}{\tilde\mu(B_{x,d_{Q_{(k+1)}}})} \int_{B_{x,d_{Q_{(k+1)}}}} F \, d\tilde\mu| \, d_{Q(k+1)}\bigg]$$

$$= \frac{1}{\tilde\mu(B_{x,d_{(k+1)}})} \int_{B_{x,d_{(k+1)}}} F \, d\tilde\mu. \qquad \square$$

Remark A.2. If the probability μ does not put mass on ball surfaces, the proof is much simpler and the result of the Lemma is merely

$$\mathbb{E}\left[\frac{1}{k}\sum_{j=1}^k F(X_{(j)})| \, d_{(k+1)}\right] = \frac{1}{\mu(B_{x,d_{(k+1)}})} \int_{B_{x,d_{(k+1)}}} F(x') \, d\mu(x').$$

This can be seen as a particular case of the following general decorrelation result: if $\varphi : \mathbb{R}^k \to \mathbb{R}$ is a test function, symmetric in all its entries, then

$$\mathbb{E}[\varphi(X_{(1)}, \ldots, X_{(k)}) | d_{(k+1)}] = \mathbb{E}[\varphi(Z_1, \ldots, Z_k) | d_{(k+1)}]$$

with the $(Z_i)_{1 \leq i \leq k}$ i.i.d. random variables distributed according to the restriction of μ to the ball $B_{x,d_{(k+1)}}$.

References

[1] C. Abraham, G. Biau and B. Cadre, *On the kernel rule for function classification.* submitted (2003).

[2] G. Biau, F. Bunea and M.H. Wegkamp, On the kernel rule for function classification. *IEEE Trans. Inform. Theory*, to appear (2005).

[3] T.M. Cover and P.E. Hart, Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **IT-13** (1967) 21–27.

[4] S. Dabo-Niang and N. Rhomari, Nonparametric regression estimation when the regressor takes its values in a metric space, submitted (2001).

[5] L. Devroye, On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.* **9** (1981) 1310–1319.

[6] 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.

[7] L. Devroye, L. Györfi and G. Lugosi, *A probabilistic theory of pattern recognition* **31**, *Applications of Mathematics (New York).* Springer-Verlag, New York (1996).

[8] L.C. Evans and R.F. Gariepy, *Measure theory and fine properties of functions.* Studies in Advanced Mathematics. CRC Press, Boca Raton, FL (1992).

[9] H. Federer, *Geometric measure theory.* Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York (1969).

[10] D. Preiss, Gaussian measures and the density theorem. *Comment. Math. Univ. Carolin.* **22** (1981) 181–193.

[11] D. Preiss, Dimension of metrics and differentiation of measures, in *General topology and its relations to modern analysis and algebra, V (Prague, 1981), Sigma Ser. Pure Math.*, Heldermann, Berlin **3** (1983) 565–568.

[12] D. Preiss and J. Tišer, Differentiation of measures on Hilbert spaces, in *Measure theory, Oberwolfach 1981 (Oberwolfach, 1981)*, Springer, Berlin. *Lect. Notes Math.* **945** (1982) 194–207.

[13] C.J. Stone, Consistent nonparametric regression. *Ann. Statist.* **5** (1977) 595–645. With discussion and a reply by the author.