Journal of Machine Learning Research 22 (2021) 1-25 Submitted 10/20; Revised 4/21; Published 7/21

Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces

László Györfi GYORFI@CS.BME.HU
Department of Computer Science and Information Theory
Budapest University of Technology and Economics
1111 Budapest, Hungary

Roi Weiss ROIW@ARIEL.AC.IL
Department of Computer Science
Ariel University
40800 Ariel, Israel

Editor: Samory Kpotufe

Abstract

We study universal consistency and convergence rates of simple nearest-neighbor prototype rules for the problem of multiclass classification in metric spaces. We first show that a novel data-dependent partitioning rule, named Proto-NN, is universally consistent in any metric space that admits a universally consistent rule. Proto-NN is a significant simplification of OptiNet, a recently proposed compression-based algorithm that, to date, was the only algorithm known to be universally consistent in such a general setting. Practically, Proto-NN is simpler to implement and enjoys reduced computational complexity.

We then proceed to study convergence rates of the excess error probability. We first obtain rates for the standard k-NN rule under a margin condition and a new generalized-Lipschitz condition. The latter is an extension of a recently proposed modified-Lipschitz condition from $\mathbb{R}^d$ to metric spaces. Similarly to the modified-Lipschitz condition, the new condition avoids any boundness assumptions on the data distribution. While obtaining rates for Proto-NN is left open, we show that a second prototype rule that hybridizes between k-NN and Proto-NN achieves the same rates as k-NN while enjoying similar computational advantages as Proto-NN. However, as k-NN, this hybrid rule is not consistent in general.

Keywords: universal consistency, rate of convergence, multiclass classification, error probability, k-nearest-neighbor rule, prototype nearest-neighbor rule, metric space

1. Introduction

Let $(\mathcal{X}, \rho)$ be a separable metric space, equipped with its Borel σ-field (Cover and Hart, 1967). Assume that the feature element X takes values in $\mathcal{X}$ and let its label Y take values in $\mathcal{Y} = \{1, \ldots, M\}$. The error probability of an arbitrary decision function $g : \mathcal{X} \to \mathcal{Y}$ is

$$L(g) = \mathbb{P}\{g(X) \neq Y\}.$$

Denote by ν the unknown probability distribution of (X, Y) and let

$$P_j(x) = \mathbb{P}\{Y = j \mid X = x\}, \quad j \in \mathcal{Y}.$$

Then the Bayes decision,

$$g^*(x) = \arg\max_{j \in \mathcal{Y}} P_j(x),$$

minimizes the error probability. This optimal error is denoted by

$$L^* = \mathbb{P}\{g^*(X) \neq Y\}.$$

In the standard model of pattern recognition, we are given labeled samples, $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, which are n independent copies of (X, Y). Based on $\mathcal{D}_n$, one can estimate the regression functions P_j by $P_{j,n}$, $j \in \mathcal{Y}$, and the plug-in classification rule g_n derived from $P_{j,n}$ is

$$g_n(x) = \arg\max_{j \in \mathcal{Y}} P_{j,n}(x).$$

The classifier g_n is *weakly consistent* for the distribution ν if

$$\lim_{n \to \infty} \mathbb{E}\{L(g_n)\} = L^*.$$

It is *strongly consistent* for ν if

$$\mathbb{P}\{\lim_{n \to \infty} L(g_n) = L^*\} = 1.$$

The classifier g_n is *universally consistent* in the metric space $(\mathcal{X}, \rho)$ if it is consistent for *any* distribution ν over the Borel σ-field.

Following the pioneering work of Cover and Hart (1967) and Stone (1977) on nearest-neighbor classification, Zhao (1987); Devroye et al. (1996); Györfi et al. (2006) showed that the k-nearest neighbor rule (k-NN) is universally strongly consistent in the Euclidean space $(\mathbb{R}^d, \|\cdot\|_2)$ provided that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. Kernel-based and various partitioning rules were shown to be universally consistent in $\mathbb{R}^d$ as well (Devroye et al., 1996). For more general (and, in particular, infinite-dimensional) metric spaces, Cérou and Guyader (2006); Forzani et al. (2012) characterized the consistency of the k-NN rule in terms of a Besicovitch-type condition. In particular, Cérou and Guyader (2006) showed that in the binary case, $|\mathcal{Y}| = 2$, a sufficient condition for the k-NN classifier to be weakly consistent for ν is

$$\lim_{r \to 0^+} \mathbb{P}\left\{ \frac{1}{\mu(S_{X,r})} \int_{S_{X,r}} |P_j(z) - P_j(X)| \mu(dz) > \varepsilon \right\} = 0. \tag{1}$$

Here, $S_{x,r} = \{z \in \mathcal{X} : \rho(x, z) \leq r\}$ denotes the closed ball centered at x and having radius r and μ is the marginal distribution of X. It was also shown in Cérou and Guyader (2006) that in the realizable case, where $P_1(x) \in \{0, 1\}$ for all $x \in X$, a violation of (1) implies that k-NN is *inconsistent* (see also Cesari and Colomboni (2021)). Abraham et al. (2006) established the strong consistency of a generalized moving-window rule under the same condition (1).

By Besicovitch's density theorem (Federer, 1969), in $\mathbb{R}^d$—and more generally in any finite-dimensional normed space—condition (1) holds for *all* distributions ν. However, in infinite-dimensional spaces this condition may be violated (Preiss, 1979, 1981). As such, the

k-NN and the moving-window classifiers are not universally consistent in general separable metric spaces. This violation is not an isolated pathology, occurring, for example, in the commonly used separable Gaussian Hilbert spaces (Tišer, 2003). Leveraging the consistency of k-NN in finite dimensions, the filtering technique (i.e., taking the first d_n coordinates in some basis representation for an appropriate d_n) was shown to be universally weakly consistent by Biau et al. (2005). However, that technique is only applicable in Hilbert spaces, as opposed to more general metric spaces.

Up until recently, sufficient and necessary conditions on a metric space under which a universally consistent rule exists in it was an open problem. Building upon the results of Kontorovich and Weiss (2014); Kontorovich et al. (2017) it has been recently shown by Hanneke et al. (2021) that a compression-based prototype classification rule named OptiNet is universally strongly consistent in *any* separable metric space (for example, the space $L^p([a,b])$ is a metric space for any $p \geq 1$ and is separable for $1 \leq p < \infty$ and non-separable for $p = \infty$). Moreover, OptiNet was shown to be universally strongly consistent in any metric space that admits a universally consistent classifier — the first algorithm known to enjoy this property.

Arguably, OptiNet is a rather complex algorithm (see Section 2). The first main result of this paper (Corollary 4) distills the core arguments used to establish OptiNet's consistency and shows that a novel, much simpler classification rule is universally strongly consistent in any metric space that admits a universally consistent classifier. In addition, its consistency proof is much simpler than that of OptiNet and its implementation is trivial. As OptiNet, this algorithm, henceforth termed Proto-NN (and formally presented in Section 2), is a prototype algorithm and has the advantage of reduced computational complexity. Concerning some other prototype nearest-neighbor rules, see Devroye et al. (1996, §19.3).

Another important property of a classification rule is the rate at which the excess error probability $\mathbb{E}\{L(g_n)\} - L^*$ convergences to 0 as $n \to \infty$. For any classification rule g_n, this rate can be arbitrarily slow without further assumptions on the unknown data distribution (Devroye et al., 1996). So to obtain a non-trivial rate of convergence one typically assumes that ν belongs to a large class of distributions meeting some smoothness and tail conditions.

For the case $\mathcal{X} = \mathbb{R}^d$, rates of convergence for several algorithms were obtained under a variety of conditions (Devroye et al., 1996; Györfi et al., 2006). A common assumption is that the regression functions are Hölder-continuous, that is, there are some $C > 0$ and $0 < \beta \leq 1$ such that $\forall x, z \in \mathcal{X}$,

$$|P_j(x) - P_j(z)| \leq C\rho(x,z)^\beta. \tag{2}$$

The case $\beta = 1$ is know as Lipschitz continuity. Denoting the support of μ by

$$\operatorname{supp}(\mu) = \{x \in \mathcal{X} : \mu(S_{x,r}) > 0, \forall r > 0\},$$

it is well known (Györfi et al., 2006) that if $\operatorname{supp}(\mu)$ is bounded and the regression function is Hölder-continuous, then, for $|\mathcal{Y}| = 2$, the k-NN rule with $k = n^{\frac{2\beta}{2\beta+d}}$ achieves the rate

$$\mathbb{E}\{L(g_{k,n})\} - L^* = O(n^{-\frac{\beta}{2\beta+d}}).$$

While this rate holds also for the more general problem of L^1 real-valued bounded regression, Mammen and Tsybakov (1999); Tsybakov (2004); Audibert and Tsybakov (2007) showed

that for binary classification, faster rates can be achieved under an additional margin condition. Recently, Xue and Kpotufe (2018); Puchkin and Spokoiny (2020) generalized the margin condition to the multiclass setting.

Definition 1 *Let $P_{(1)}(x) \geq \cdots \geq P_{(M)}(x)$ be the ordered values of $P_1(x), \ldots, P_M(x)$. Then the **margin condition** means that there are some $\alpha > 0$ and $c^* > 0$ such that*

$$\mathbb{P}\{P_{(1)}(X) - P_{(2)}(X) \leq t\} \leq c^* t^\alpha, \qquad 0 < t \leq 1. \tag{3}$$

Assuming that the regression function is Hölder-continuous, that the margin condition is satisfied, and that the marginal distribution μ of X has a density that satisfies the so called strong density condition, Audibert and Tsybakov (2007); Kohler and Krzyzak (2007); Gadat et al. (2016) showed that for the binary case, the k-NN rule and several other plug-in estimators achieve the rate

$$\mathbb{E}\{L(g_{k,n})\} - L^* = O(n^{-\frac{\beta(1+\alpha)}{2\beta+d}}). \tag{4}$$

Moreover, this rate was shown to be minimax optimal, meaning that, over the class of all distributions meeting the aforementioned conditions, this rate is also a lower bound for *any* classification rule. See Samworth (2012); Blaschzyk and Steinwart (2018) and references therein for even faster rates under stronger conditions on the regression function.

The strong density condition under which the rates in (4) are established requires the density function to be bounded away from zero over the support of μ; a highly restrictive condition that does not hold for many distributions of practical interest. Recently, it has been shown by Döring et al. (2017) that the same rates in (4) are obtained by replacing the strong density condition, together with the Hölder condition (2), with a combined smoothness and tail condition, named modified Lipschitz condition, given by

$$|P_j(x) - P_j(z)| \leq C\mu(S_{x,\rho(x,z)})^{\beta/d}. \tag{5}$$

Chaudhuri and Dasgupta (2014) considered the related condition that there are some $\gamma > 0$ and $C^* > 0$ such that for all x in the support of μ,

$$\left| P_j(x) - \frac{1}{\mu(S_{x,r})} \int_{S_{x,r}} P_j(z) d\mu(z) \right| \leq C^* \mu(S^o_{x,r})^\gamma. \tag{6}$$

Here, $S^o_{x,r} = \{z \in \mathcal{X} : \rho(x,z) < r\}$ denotes the open ball centered at x. They showed that in the binary case, under condition (6) and the margin condition (3), the k-NN rule achieves

$$\mathbb{E}\{L(g_{k,n})\} - L^* = O(n^{-\frac{\gamma(1+\alpha)}{2\gamma+1}}). \tag{7}$$

Evidently, for $\mathcal{X} = \mathbb{R}^d$, the rates in (4) are revisited by setting $\gamma = \beta/d$. More generally, condition (6) can be seen as a uniform Besicovitch condition and is a-priori applicable in any separable metric space. In this paper, we further abstract the modified Lipschitz conditions (5) and (6) and consider the following combined smoothness and tail condition.

Definition 2 *For each $j \in \mathcal{Y}$, the function P_j satisfies the* ***generalized Lipschitz condition*** *if there is a monotonically increasing function $h : [0,1] \to \mathbb{R}^+$ with $h(s) \downarrow 0$ as $s \downarrow 0$ such that for any $x, z \in \mathcal{X}$,*

$$|P_j(x) - P_j(z)| \leq h(\mu(S_{x,\rho(x,z)})). \tag{8}$$

In Section 3, we first obtain rates for the k-NN rule under the generalized Lipschitz condition and the margin condition in terms of the function h (Theorem 6). For the case

$$h(s) = C^* s^\gamma \tag{9}$$

we revisit the same rates as in (7). While obtaining rates for Proto-NN is left open, we proceed to derive rates for a second novel prototype rule that hybridizes between Proto-NN and k-NN (Theorem 7). This rule, which we call Proto-k-NN, allows a reduction in computational complexity by compressing the data into $m = O(n/k)$ prototypes while enjoying the same rates as k-NN; see also Xue and Kpotufe (2018) for results in the same spirit.

On the generalized Lipschitz condition. Note that for condition (8) to be non-trivial, the rate at which $h(s) \to 0$ as $s \to 0$ should appropriately reflect the geometry of $\mathcal{X}$ and the class of distributions μ under consideration. Consider for example two distinct points $x, z \in \mathbb{R}^d$ lying in a region where μ is uniformly distributed. Denoting the Lebesgue measure by λ and abbreviating its measure on any ball by $\lambda(S_{x,r}) = v_d \cdot r^d$, one can verify that for any $0 < r \ll \rho(x,z)$, applying the triangle inequality for $\lceil \rho(x,z)/r \rceil$ times over a path from x to z, the generalized Lipschitz condition (8) implies that for some constant $c > 0$,

$$|P_j(x) - P_j(z)| \leq c \cdot \left(\frac{\rho(x,z)}{r}\right) \cdot h(\lambda(S_{x,r})) = \rho(x,z) \cdot O\left(\frac{h(r^d)}{r}\right) \qquad \text{as } r \to 0.$$

Hence, to allow for non-constant regression functions, one needs $h(\lambda(S_{x,r})) = O(r)$, or equivalently $h(s) = O(s^{1/d})$, as in the modified Lipschitz condition (5). More generally, assuming that μ is absolutely continuous with respect to λ, the Radon-Nikodym theorem (Federer, 1969) asserts that μ has a density, namely, there exists a function $D : \mathbb{R}^d \to \mathbb{R}^+$ such that for λ-almost all $x \in \mathbb{R}^d$,

$$\lim_{r \to 0} \left| \frac{\mu(S_{x,r})}{\lambda(S_{x,r})} - D(x) \right| = 0, \tag{10}$$

such that for any measurable $A \subseteq \mathbb{R}^d$,

$$\mu(A) = \int_A D(x) \lambda(dx).$$

So, in this case, condition (8) with

$$h(s) = C^* s^{\beta/d} \tag{11}$$

essentially becomes

$$|P_j(x) - P_j(z)| \leq C^* D(x)^{\beta/d} \rho(x,z)^\beta.$$

This condition should be compared to the Hölder condition (2); see also Györfi (1981).

Similarly, for a general separable metric space, one may assume that μ is accompanied by an increasing "small-ball probability" function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ with $\lim_{r\to 0} \psi(r) = 0$ and a function $K : \mathcal{X} \to \mathbb{R}^+$ such that for μ-almost all $x \in \mathcal{X}$,

$$\mu(S_{x,r}) \leq K(x)\psi(r) \qquad \text{as } r \to 0. \tag{12}$$

In this case, an appropriate choice would be

$$h(s) = C^* \psi^{-1}(s)^\beta \qquad s \in [0, 1].$$

Condition (8) then becomes

$$|P_j(x) - P_j(z)| \leq C^* \psi^{-1}\big(K(x)\psi(\rho(x, z))\big)^\beta.$$

As an example, consider a doubling measure μ_0 with $\mathrm{supp}(\mu_0) = \mathcal{X}$ (Heinonen, 2012; Rigot, 2018). Such a measure satisfies that there exists $C = C(\mu_0) \geq 1$ such that for any x and any radius $r > 0$,

$$0 < \mu_0(S_{x,2r}) \leq C\mu_0(S_{x,r}). \tag{13}$$

Assuming that μ_0 has no atoms, the doubling property (13) implies that there exists $\tau > 0$ such that

$$\mu_0(S_{x,r}) = \Theta(\psi(r)) = \Theta(r^\tau) \qquad \text{as } r \to 0. \tag{14}$$

Since doubling measures satisfy a differentiation theorem similar to (10), a large class of distributions that satisfy (12) consists of all distributions μ that are absolutely continuous with respect to μ_0. Here, K in (12) is again related to μ's density with respect to μ_0. It is worth mentioning, however, that in infinite-dimensional spaces, doubling measures may not exist and a differentiation theorem may not hold; see Heinonen (2012); Rigot (2018).

In the field of non-parametric functional data analysis (Ferraty and Vieu, 2006; Burba et al., 2009; Baíllo et al., 2011; Ling and Vieu, 2018), where input data items are in the form of random functions, distributions μ accompanied by ψ in the form of (14) are said to have a fractal dimension τ; see Pesin (1993); Bardet (1997) for concrete examples. Similarly to (7), a finite fractal dimension leads to rates of order $n^{-\xi}$ for an appropriate $\xi = \xi(\tau, \alpha, \beta) > 0$. However, in infinite-dimensional spaces, one typically encounters distributions of exponential type, where for some $C > 0$ and $\tau, \tau' > 0$,

$$\psi(r) = \Theta\big(e^{-\frac{1}{r^\tau}\log\left(\frac{1}{r}\right)^{\tau'}}\big) \qquad \text{as } r \to 0. \tag{15}$$

Such distributions include, for example, various diffusion and Gaussian processes; see Ferraty and Vieu (2006, §13) and references therein. In this case, non-parametric estimators suffer from extremely slow convergence rates of order $\log(n)^{-\xi}$, even for estimating the regression function at a *fixed* point $x \in \mathcal{X}$ (Ferraty and Vieu, 2006; Mas, 2012). One may overcome such slow rates by considering a pseudo-metric over $\mathcal{X}$ (where the condition $\rho(x, y) = 0 \Rightarrow x = y$ is removed) instead of a metric, but coming up with an appropriate pseudo-metric can be a challenging task (Ferraty and Vieu, 2006).

Lastly, we note that faster rates can be achieved in infinite-dimensional spaces, uniformly over the support of μ, in terms of covering numbers, assuming the support of μ is totally bounded (Kulkarni and Posner, 1995; Ferraty et al., 2010; Kudraszow and Vieu, 2013; Biau et al., 2010). Here we do not make such an assumption.

2. Universal consistency of Proto-NN

In this section we first show that a simple prototype classification rule, which we call Proto-NN, is universally strongly consistent in any separable metric space. By the recent results of Hanneke et al. (2021), this implies that Proto-NN is in fact universally strongly consistent in any metric space that admits a universally consistent rule; see also Collins et al. (2020). Our consistency result is obtained by first establishing the consistency of Proto-NN for the more general problem of real-valued bounded L^1-regression. Lastly, we show that a slightly modified version of Proto-NN is universally strongly consistent for the general real-valued L^p-regression problem for any $1 \leq p < \infty$ under the (necessary) condition that $\mathbb{E}\{|Y|^p\} < \infty$.

To simplify Proto-NN's analysis, we assume that in addition to the labeled sample $\mathcal{D}_n$, we also have an independent unlabeled sample $\mathbf{X}'_m = \{X'_1, \ldots, X'_m\}$ where the X'_i's are independent copies of X. Introduce the data-driven partition $\mathcal{P}_m$ of $\mathcal{X}$ such that $\mathcal{P}_m$ is a Voronoi partition with the nucleus set $\mathbf{X}'_m$, i.e.,

$$\mathcal{P}_m = \{A_{m,1}, A_{m,2}, \ldots, A_{m,m}\}$$

such that $A_{m,\ell}$ is the Voronoi cell around the nucleus X'_ℓ,

$$A_{m,\ell} = \left\{ x \in \mathcal{X} : \ell = \operatorname*{arg\,min}_{1 \leq i \leq m} \rho(x, X'_i) \right\},$$

where tie breaking is done by indices, i.e., if X'_i and X'_j are equidistant from x, then X'_i is declared "closer" if $i < j$. Proto-NN estimates the regression function P_j over each cell by the piecewise constant function

$$\tilde{P}_{n,j}(x) = \frac{\sum_{i=1}^n \mathbb{I}_{\{Y_i = j, X_i \in A_{m,\ell}\}}}{\sum_{i=1}^n \mathbb{I}_{\{X_i \in A_{m,\ell}\}}}, \quad \text{if } x \in A_{m,\ell}, \tag{16}$$

such that $0/0 = 0$ by definition. Proto-NN is then defined by

$$\tilde{g}_n(x) = \operatorname*{arg\,max}_{j \in \mathcal{Y}} \tilde{P}_{n,j}(x). \tag{17}$$

This rule is just the empirical majority vote over the labeled samples from $\mathcal{D}_n$ that fell into the cell in which x resides, as determined by $\mathbf{X}'_m$. Proto-NN's construction time is $O(mn)$ and a query takes $O(m)$ time.

To establish Proto-NN's universal consistency, we first establish its consistency for the more general problem of L^1 real-valued bounded regression (Györfi et al., 2006). Formally, let the label Y be real-valued, and denote the corresponding regression function by

$$f(x) = \mathbb{E}\{Y \mid X = x\}.$$

Introduce the partitioning regression estimate:

$$f_n(x) = \frac{\sum_{i=1}^n Y_i \mathbb{I}_{\{X_i \in A_{m,\ell}\}}}{\sum_{i=1}^n \mathbb{I}_{\{X_i \in A_{m,\ell}\}}}, \quad \text{if } x \in A_{m,\ell}. \tag{18}$$

Theorem 3 *Let $(\mathcal{X}, \rho)$ be a separable metric space. If Y is bounded and $m = m_n \to \infty$ such that $m_n/n \to 0$, then the estimate f_n is strongly consistent in L^1, that is, for any distribution ν of (X, Y),*

$$\lim_{n\to\infty} \int |f_n(x) - f(x)|\mu(dx) = 0 \qquad a.s.$$

The following corollary establishes the universal consistency of Proto-NN.

Corollary 4 *Let $(\mathcal{X}, \rho)$ be a separable metric space. If $m = m_n \to \infty$ such that $m_n/n \to 0$, then the classification rule $\tilde{g}_n$ is universally strongly consistent, that is, for any distribution of (X, Y),*

$$\lim_{n\to\infty} L(\tilde{g}_n) = L^* \qquad a.s.$$

The last result of this section is an extension of Theorem 3 to real-valued L^1-regression where the assumption of bounded Y is relaxed to the (necessary) condition $\mathbb{E}\{|Y|\} < \infty$. We establish consistency for a modified rule, as similarly done in Györfi et al. (2006, Theorem 23.3). Denoting the index of the cell to which x belongs by $\ell(x) \in \{1, \ldots, m\}$, this modified rule is given by

$$f_n^t(x) = \begin{cases} f_n(x) & \text{if } \sum_{i=1}^n \mathbb{I}_{\{X_i \in A_{m,\ell(x)}\}} \geq \log n, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 5 *Let $(\mathcal{X}, \rho)$ be a separable metric space. If $m = m_n \to \infty$ and $m_n \log n/n \to 0$, then the estimate f_n^t is universally strongly consistent in L^1, that is, for any distribution ν of (X, Y) with $\mathbb{E}\{|Y|\} < \infty$,*

$$\lim_{n\to\infty} \int |f_n^t(x) - f(x)|\mu(dx) = 0 \qquad a.s.$$

By Györfi (1991, Theorem 2), Theorem 5 implies universal strong consistency of f_n^t also for L^p regression with $1 \leq p < \infty$ under the (necessary) condition $\mathbb{E}\{|Y|^p\} < \infty$.

It is interesting to compare Proto-NN to the recently proposed OptiNet classifier of Hanneke et al. (2021), which, to date, was the only algorithm known to be universally consistent in any separable metric space. Denoting the instances in $\mathcal{D}_n$ by $\mathbf{X}_n = \{X_1, \ldots, X_n\}$, OptiNet first constructs several γ-nets of $\mathbf{X}_n$ for different candidate values of $\gamma > 0$ (a γ-net of $\mathbf{X}_n$ is any maximal set $\mathbf{X}(\gamma) \subseteq \mathbf{X}_n$ in which all interpoint distances are at least γ). Each γ-net serves as a nucleus set for a corresponding Voronoi partition of $\mathcal{X}$. For each such partition, a prototype classifier is constructed by taking a majority-vote in each of its cells. Then an optimal γ^* is selected among the different candidates by minimizing a compression-based generalization bound over γ. Alternatively, γ^* can be chosen via a validation procedure using a hold-out dataset. The construction time of OptiNet is $O(n^2)$ and its query time depends on the optimal margin chosen at construction.

Hanneke et al. (2021) observed that a model selection procedure for choosing γ^* cannot be readily avoided, since one cannot choose a-priori a determinstic sequence γ_n for which OptiNet is consistent for all distributions. This is in contrast to Proto-NN, for which any $m = m_n \to \infty$ with $m_n/n \to 0$ will do. Of course, in practice, m should be chosen based on the data, as done with γ^*.

3. Rates of convergence

In this section our focus lies on the rate at which the excess error probability

$$\mathbb{E}\{L(g_n)\} - L^* \to 0.$$

We first derive rates for the k-NN rule under the margin condition (1) and the generalized Lipschitz condition (2) (Theorem 6). Obtaining convergence rates for the universally consistent Proto-NN classifier of Section 2 under these conditions (or any other condition mentioned in Section 1 for that matter) is currently an open research problem. Here, we instead derive rates for a second novel prototype rule, Proto-k-NN, that hybridizes between the Proto-NN and k-NN rules (Theorem 7). As shown below, Proto-k-NN allows a reduction in computational complexity by compressing the data into $m = O(n/k)$ prototypes while enjoying the same rates as the k-NN rule.

We first make an additional simplifying assumption. For a fixed $x \in \mathcal{X}$, let

$$H_x(r) := \mathbb{P}\left(\rho(x, X) \leq r\right), \quad r \geq 0, \tag{19}$$

be the cumulative distribution function of $\rho(x, X)$. In the sequel, we assume that $H_x(\cdot)$ is continuous for each x. This assumption holds, for example, in the case that $\mathcal{X} = \mathbb{R}^d$ and X has a density. If $H_x(\cdot)$ is continuous, then in the definition of nearest neighbors, tie happens with probability zero. In general, one can achieve that $H_x(\cdot)$ is continuous by adding a randomized component to X: take $\tilde{X} = (X, U)$ such that X and U are independent, U is uniformly distributed on $[0, 1]$, and

$$\tilde{\rho}((x, u), (X, U)) := \rho(x, X) + \delta|u - U|$$

for some small $\delta > 0$. One can verify that $\tilde{\rho}$ is indeed a metric.

Rates for the k-NN rule. The k-nearest neighbor rule is defined as follows. We fix $x \in \mathcal{X}$ and reorder the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ according to increasing values of $\rho(x, X_i)$. The reordered data sequence is denoted by

$$(X_{(n,1)}(x), Y_{(n,1)}(x)), \ldots, (X_{(n,n)}(x), Y_{(n,n)}(x)).$$

$X_{(n,k)}(x)$ is the k-th nearest neighbor of x where tie breaking is done by indices. As discussed above, in this paper we assume that tie happens with probability zero. Choose an integer k less than n, then the k-nearest-neighbor estimate of P_j is

$$P_{n,j}(x) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{I}_{\{Y_{(n,i)}(x)=j\}},$$

and the k-nearest-neighbor classification rule is

$$g_{k,n}(x) = \arg\max_{j \in \mathcal{Y}} P_{n,j}(x).$$

Concerning the properties of the k-nearest-neighbor rule and the related literature see Devroye et al. (1996), Györfi et al. (2006), and Biau and Devroye (2015).

In the following theorem we bound the rate of convergence of the excess error probability $\mathbb{E}\{L(g_{k,n})\} - L^*$ for the k-nearest-neighbor classification rule. In this way we extend the results of Kohler and Krzyzak (2007), Gadat et al. (2016), Döring et al. (2017) to the multi-class and to the metric space case. Notice that the paper by Biau et al. (2010) contains rate of convergence results for nearest neighbor regression estimate and Banach space valued features, which has implications for classification, too; cf. Chaudhuri and Dasgupta (2014).

Theorem 6 *Let $(\mathcal{X}, \rho)$ be a separable metric space. Assume that the distribution function $H_x(\cdot)$ is continuous for each x. If the* ***margin condition*** *is satisfied with $0 < \alpha$ and the* ***generalized Lipschitz condition*** *is met, then for $k/\log n \to \infty$,*

$$\mathbb{E}\{L(g_{k,n})\} - L^* = O(1/k^{(1+\alpha)/2}) + O(h(2k/n)^{1+\alpha}).$$

In the case of $\mathcal{X} = \mathbb{R}^d$ and $h(s) = s^{\beta/d}$ as in (11),

$$\mathbb{E}\{L(g_{k,n})\} - L^* = O(1/k^{(1+\alpha)/2}) + O((k/n)^{\beta(1+\alpha)/d})$$

and the choice

$$k_n = \lfloor n^{2\beta/(2\beta+d)} \rfloor \tag{20}$$

yields the order

$$n^{-\frac{\beta(1+\alpha)}{2\beta+d}} \tag{21}$$

as in (4). For the two-class problem, Audibert and Tsybakov (2007, Theorem 3.5) showed that, under the strong density assumption and the margin condition with $\alpha\beta \le d$, (21) is the minimax optimal rate of convergence for the class of β-Hölder-continuous P_j's, that is, the order (21) is a lower bound for *any* classifier.

Rates for the Proto-k-NN rule. We now introduce a second prototype rule, termed Proto-k-NN, that hybridizes between k-NN and Proto-NN. It is essentially a private case of the aggregated denoised 1-NN algorithm introduced by Xue and Kpotufe (2018) and corresponds to the case of using only one subsample. Xue and Kpotufe (2018) established essentially the optimal rates in (4) for this algorithm in $\mathbb{R}^d$ (with finite-sample guarntees) under the Hölder condition (2), the margin condition (3), and the strong density assumption (see Section 1). Here we establish rates under the generalized Lipschitz condition (8) and the margin condition (3) in an arbitrary separable metric space.

Proto-k-NN works as follows. Similarly to Proto-NN (see Section 2), an unlabeled sample $\mathbf{X}'_m$ serves as a nucleus set, inducing a Voronoi partition $\mathcal{P}_m$ of $\mathcal{X}$. Instead of taking a majority vote in each Voronoi cell, Proto-k-NN stores at each nucleus the majority vote among the k-nearest neighbors of that nucleus. Formally, let $X'_{(m,1)}(x)$ be the first nearest neighbor of x among $\mathbf{X}'_m$. We fix $x \in \mathcal{X}$, and let $X_{(n,k)}(X'_{(m,1)}(x))$ be the k-th nearest neighbor of $X'_{(m,1)}(x)$ from the set $\mathbf{X}_n$ and $Y_{(n,k)}(X'_{(m,1)}(x))$ stands for its label. The tie breaking is done by randomization. Again, in this section we assume that tie happens with probability 0. Choose an integer k less than n, then Proto-k-NN estimates $P_j(x)$ by the piecewise constant function

$$\hat{P}_{n,j}(x) = \hat{P}_{n,j}(X'_{(m,1)}(x)) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{I}_{\{Y_{(n,i)}(X'_{(m,1)}(x))=j\}},$$

and the corresponding prototype nearest-neighbor classification rule is

$$\hat{g}_n(x) = \arg\max_{j \in \mathcal{Y}} \hat{P}_{n,j}(x).$$

The construction and query times of Proto-k-NN are $O(mn)$ and $O(m)$ respectively; the same as for Proto-NN.

Theorem 7 *Let $(\mathcal{X}, \rho)$ be a separable metric space. Assume that the distribution function $H_x(\cdot)$ is continuous for each x. If the* ***margin condition*** *is satisfied with $0 < \alpha$ and the* ***generalized Lipschitz condition*** *is met with h that is concave, then for $k/\log n \to \infty$,*

$$\mathbb{E}\{L(\hat{g}_n)\} - L^* = O(1/k^{(1+\alpha)/2}) + O(h(k/n)^{1+\alpha}) + O(h(1/m)^{1+\alpha}).$$

According to this theorem, the proper choice of m is proportional to n/k. For k as in (20), we obtain the same optimal rates as in (21). Note however that, similarly to the k-NN rule, Proto-k-NN is not universally consistent in all separable metric spaces, as established by the same counterexample in Cérou and Guyader (2006, Section 3).

Lastly, we note that in some cases, the distribution μ is concentrated on a finite-dimensional subspace of $\mathcal{X}$ such that on this space the modified Lipschitz condition is satisfied. Then Proto-NN and Proto-k-NN, not knowing this subspace and the intrinsic dimension, automatically achieve good rate of convergence; see Kpotufe and Dasgupta (2012) and references therein for other algorithms that are adaptive to the intrinsic dimension.

4. Proof of universal consistency

Proof of Theorem 3. Put

$$\bar{f}_m(x) = \frac{\int_{A_{m,\ell}} f(z)\mu(dz)}{\mu(A_{m,\ell})}, \quad \text{if } x \in A_{m,\ell}.$$

Then,

$$\int |f_n(x) - \bar{f}_m(x)|\mu(dx) \qquad \text{and} \qquad \int |\bar{f}_m(x) - f(x)|\mu(dx)$$

are called estimation errror and approximation error, respectively. Introduce the notations

$$\vartheta_n(A) = \frac{1}{n}\sum_{i=1}^n Y_i \mathbb{I}_{\{X_i \in A\}}, \qquad \vartheta(A) = \mathbb{E}\{\vartheta_n(A)\}$$

and

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{X_i \in A\}}, \qquad \mu(A) = \mathbb{E}\{\mu_n(A)\}.$$

Concerning the estimation error,

$$\begin{aligned}\int |f_n(x) - \bar{f}_m(x)|\mu(dx) &= \sum_{V \in \mathcal{P}_m} \int_V |f_n(x) - \bar{f}_m(x)|\mu(dx) \\ &= \sum_{V \in \mathcal{P}_m} \left|\frac{\vartheta_n(V)}{\mu_n(V)} - \frac{\vartheta(V)}{\mu(V)}\right| \mu(V).\end{aligned}$$

If L stands for a bound on $|Y|$, then $|\vartheta_n(V)| \le L\mu_n(V)$ and

$$\begin{aligned}
&\int |f_n(x) - \bar{f}_m(x)|\mu(dx) \\
&\le \sum_{V\in\mathcal{P}_m} \left|\frac{\vartheta_n(V)}{\mu_n(V)} - \frac{\vartheta_n(V)}{\mu(V)}\right| \mu(V) + \sum_{V\in\mathcal{P}_m} \left|\frac{\vartheta_n(V)}{\mu(V)} - \frac{\vartheta(V)}{\mu(V)}\right| \mu(V) \\
&\le L \sum_{V\in\mathcal{P}_m} |\mu(V) - \mu_n(V)| + \sum_{V\in\mathcal{P}_m} |\vartheta_n(V) - \vartheta(V)| .
\end{aligned}$$

The first term of the right hand side is a special case of the second one, therefore we bound only the tail distribution of the second term. Let $\sigma(\mathcal{P}_m)$ be the σ-algebra generated by $\mathcal{P}_m$. As in Biau and Györfi (2005), the equality $|a-b| = 2(a-b)^+ + (b-a)$ implies

$$\begin{aligned}
\sum_{V\in\mathcal{P}_m} |\vartheta_n(V) - \vartheta(V)| &= 2 \max_{V\in\sigma(\mathcal{P}_m)} (\vartheta_n(V) - \vartheta(V))^+ + \sum_{V\in\mathcal{P}_m} (\vartheta(V) - \vartheta_n(V)) \\
&= 2 \max_{V\in\sigma(\mathcal{P}_m)} (\vartheta_n(V) - \vartheta(V))^+ + \mathbb{E}\{Y\} - \frac{1}{n}\sum_{i=1}^n Y_i.
\end{aligned}$$

Thus, for any $\varepsilon > 0$, the Hoeffding inequality implies

$$\begin{aligned}
&\mathbb{P}\left\{ \sum_{V\in\mathcal{P}_m} |\vartheta_n(V) - \vartheta(V)| \ge \varepsilon \mid \mathbf{X}'_m \right\} \\
&\le \mathbb{P}\left\{ 2 \max_{V\in\sigma(\mathcal{P}_m)} (\vartheta_n(V) - \vartheta(V)) \ge 2\varepsilon/3 \mid \mathbf{X}'_m \right\} + \mathbb{P}\left\{ \mathbb{E}\{Y\} - \frac{1}{n}\sum_{i=1}^n Y_i \ge \varepsilon/3 \right\} \\
&\le (2^m + 1) e^{-n\varepsilon^2/(18L^2)}.
\end{aligned}$$

Therefore, $m_n/n \to 0$ together with the Borel-Cantelli lemma implies that

$$\sum_{V\in\mathcal{P}_m} |\vartheta_n(V) - \vartheta(V)| \to 0 \qquad a.s.$$

and the consistency of the estimation error is proved.

Concerning the approximation error, refer to Lemma A.1 in Hanneke et al. (2021) such that choose a Lipschitz function f^* with a Lipschitz constant C and with a support contained in a sphere S such that

$$\int |f(x) - f^*(x)|\mu(dx) \le \varepsilon.$$

By examining the proof of Lemma A.1 in Hanneke et al. (2021), we may choose f^* to be bounded by the same bound assumed on $|Y|$. Put

$$\bar{f}^*_m(x) = \frac{\int_{A_{m,\ell}} f^*(z)\mu(dz)}{\mu(A_{m,\ell})}, \quad \text{if } x \in A_{m,\ell}.$$

Then,

$$\begin{aligned}
&\int |\bar{f}_m(x) - f(x)|\mu(dx) \\
&\leq \int |\bar{f}_m(x) - \bar{f}^*_m(x)|\mu(dx) + \int |\bar{f}^*_m(x) - f^*(x)|\mu(dx) + \int |f^*(x) - f(x)|\mu(dx) \\
&\leq \int |\bar{f}^*_m(x) - f^*(x)|\mu(dx) + 2\int |f^*(x) - f(x)|\mu(dx) \\
&\leq \int |\bar{f}^*_m(x) - f^*(x)|\mu(dx) + 2\varepsilon.
\end{aligned}$$

The bound $|Y| \leq L$ implies that $|f^*(x)| \leq L$, therefore

$$\begin{aligned}
&\int |\bar{f}^*_m(x) - f^*(x)|\mu(dx) \\
&= \sum_{V\in\mathcal{P}_m} \int_V |\bar{f}^*_m(x) - f^*(x)|\mu(dx) \\
&= \sum_{V\in\mathcal{P}_m} \frac{1}{\mu(V)} \int_V \left| \int_V f^*(z)\mu(dz) - f^*(x)\mu(V) \right| \mu(dx) \\
&\leq \sum_{V\in\mathcal{P}_m} \frac{1}{\mu(V)} \int_V \int_V |f^*(z) - f^*(x)|\, \mu(dx)\mu(dz) \\
&\leq \sum_{V\in\mathcal{P}_m} \frac{1}{\mu(V)} \int_V \int_V \min\{C\rho(x,z), 2L\}\mu(dx)\mu(dz).
\end{aligned}$$

Recall that $X'_{(m,1)}(x)$ denotes the first nearest neighbor of x among $\mathbf{X}'_m$. For $x, z \in V$,

$$\rho(x,z) \leq \rho(x, X'_{(m,1)}(x)) + \rho(X'_{(m,1)}(x), z) = \rho(x, X'_{(m,1)}(x)) + \rho(z, X'_{(m,1)}(z)),$$

where we applied that V is a Voronoi cell and so for any $x, z \in V$, the nucleuses $X'_{(m,1)}(x)$ and $X'_{(m,1)}(z)$ are identical. Thus,

$$\int |\bar{f}^*_m(x) - f^*(x)|\mu(dx) \leq \int \min\{2C\rho(x, X'_{(m,1)}(x)), 2L\}\mu(dx).$$

Cover and Hart (1967) proved that, for a separable metric space,

$$\rho(x, X'_{(m,1)}(x)) \to 0$$

a.s. as $m \to \infty$, for μ-almost all x, cf. Lemma 6.1 in Györfi et al. (2006). Therefore, the dominated convergence yields

$$\int |\bar{f}^*_m(x) - f^*(x)|\mu(dx) \to 0 \qquad a.s.,$$

concluding the proof of Theorem 3.

Proof of Corollary 4. An extension of Devroye et al. (1996, Theorem 2.2) yields

$$L(\tilde{g}_n) - L^* \leq \sum_{j=1}^{M} \int |P_j(x) - \tilde{P}_{n,j}(x)| \mu(dx).$$

Thus, the corollary is proved if

$$\int |P_j(x) - \tilde{P}_{n,j}(x)| \mu(dx) \to 0$$

a.s., $j = 1, \ldots, M$, which follows from Theorem 3.

Proof of Theorem 5 We write the regression estimator f_n^t as

$$f_n^t(x) = \sum_{i=1}^{n} W_{n,i}(x) Y_i$$

where

$$W_{n,i}(x) = \frac{\mathbb{I}_{\{X_i \in A_{m,\ell(x)}\}}}{\sum_{i=1}^{n} \mathbb{I}_{\{X_i \in A_{m,\ell(x)}\}}} \cdot \mathbb{I}_{\{\sum_{i=1}^{n} \mathbb{I}_{\{X_i \in A_{m,\ell(x)}\}} \geq \log n\}}.$$

Note that the weights are sub-probabilities, namely, for all $x \in \mathcal{X}$,

$$0 \leq \sum_{i=1}^{n} W_{n,i}(x) \leq 1.$$

By Györfi (1991, Theorem 2) (see also Györfi et al. (2006, Lemma 23.3)), it suffices to show:

(i) f_n^t is strongly consistent for L^1 assuming Y is bounded, $|Y| \leq L$;

(ii) there exists $c > 0$ such that for any Y with $\mathbb{E}\{|Y|\} < \infty$,

$$\limsup_n \sum_{i=1}^{n} \int W_{n,i}(x) \mu(dx) |Y_i| \leq c \mathbb{E}\{|Y|\} \quad a.s. \tag{22}$$

To show *(i)*, assume $|Y| \leq L$ and decompose

$$\int |f_n^t(x) - f(x)| \mu(dx) \leq \int |f_n^t(x) - f_n(x)| \mu(dx) + \int |f_n(x) - f(x)| \mu(dx),$$

where f_n is as in (18). By Theorem 3,

$$\lim_{n \to \infty} \int |f_n(x) - f(x)| \mu(dx) = 0 \qquad \text{a.s.}$$

Recall the notation $\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in A\}}$ for $A \subseteq \mathcal{X}$ and let

$$G_n = \{\ell : \mu_n(A_{m,\ell}) \geq \log n / n\} \qquad \text{and} \qquad \mathcal{G}_n = \bigcup_{\ell \in G_n} A_{m,\ell}.$$

Since $f_n^t(x) = f_n(x)$ for $x \in \mathcal{G}_n$ and $f_n^t(x) = 0$ for $x \in \mathcal{G}_n^c$,

$$\int |f_n^t(x) - f_n(x)|\mu(dx) = \int_{\mathcal{G}_n^c} |f_n^t(x) - f_n(x)|\mu(dx) = \int_{\mathcal{G}_n^c} |f_n(x)|\mu(dx) \leq L\mu(\mathcal{G}_n^c).$$

For $c = e^2$ let

$$F_n = \{\ell : \mu(A_{m,\ell}) \geq c \log n/n\} \qquad \text{and} \qquad \mathcal{F}_n = \bigcup_{\ell \in F_n} A_{m,\ell}.$$

Then

$$\mu(\mathcal{G}_n^c) = \mu(\mathcal{G}_n^c \cap \mathcal{F}_n^c) + \mu(\mathcal{G}_n^c \cap \mathcal{F}_n) \leq \mu(\mathcal{F}_n^c) + \mu(\mathcal{G}_n^c \cap \mathcal{F}_n) \leq \frac{cm_n \log n}{n} + \mu(\mathcal{G}_n^c \cap \mathcal{F}_n).$$

The first term convergences to 0 by the Theorem's condition on m_n. For the second term,

$$\mu(\mathcal{G}_n^c \cap \mathcal{F}_n) = \sum_{\ell \in F_n} \mu(A_{m,\ell}) \mathbb{I}_{\{\mu_n(A_{m,\ell}) < \log n/n\}} \leq \sum_{\ell \in F_n} \mu(A_{m,\ell}) \mathbb{I}_{\{\mu_n(A_{m,\ell}) < \mu(A_{m,\ell})/c\}}.$$

Chernoff's bound implies that for any $\ell \in F_n$,

$$\mathbb{P}\left\{\mu_n(A_{m,\ell}) < \mu(A_{m,\ell})/c \mid \mathbf{X}'_m\right\} \leq e^{-n\mu(A_{m,\ell})(1-\frac{1}{c}-\frac{\log c}{c})} \leq e^{-c(1-\frac{1}{c}-\frac{\log c}{c})\log n} \leq n^{-e^2+3}.$$

Thus, for any $0 < \varepsilon < 1$,

$$\begin{aligned}
&\mathbb{P}\left\{\sum_{\ell \in F_n} \mu(A_{m,\ell}) \mathbb{I}_{\{\mu_n(A_{m,\ell}) < \mu(A_{m,\ell})/c\}} > \varepsilon\right\} \\
&= \mathbb{E}\left\{\mathbb{P}\left\{\sum_{\ell \in F_n} \mu(A_{m,\ell}) \mathbb{I}_{\{\mu_n(A_{m,\ell}) < \mu(A_{m,\ell})/c\}} > \varepsilon \mid \mathbf{X}'_m\right\}\right\} \\
&\leq \mathbb{E}\left\{\mathbb{P}\left\{\sum_{\ell \in F_n} \mu(A_{m,\ell}) \mathbb{I}_{\{\mu_n(A_{m,\ell}) < \mu(A_{m,\ell})/c\}} > \varepsilon \sum_{\ell \in F_n} \mu(A_{m,\ell}) \mid \mathbf{X}'_m\right\}\right\} \\
&\leq \mathbb{E}\left\{\sum_{\ell \in F_n} \mathbb{P}\left\{\mathbb{I}_{\{\mu_n(A_{m,\ell}) < \mu(A_{m,\ell})/c\}} > \varepsilon \mid \mathbf{X}'_m\right\}\right\} \\
&= \mathbb{E}\left\{\sum_{\ell \in F_n} \mathbb{P}\left\{\mu_n(A_{m,\ell}) < \mu(A_{m,\ell})/c \mid \mathbf{X}'_m\right\}\right\} \\
&\leq m_n \cdot n^{-e^2+3} \leq n^{-e^2+4},
\end{aligned}$$

which is summable. Hence, by the Borel-Cantelli Lemma,

$$\limsup_n \int |f_n^t(x) - f_n(x)|\mu(dx) = 0 \quad a.s.,$$

concluding the proof of *(i)*.

To show *(ii)*, assume Y satisfies $\mathbb{E}\{|Y|\} < \infty$. We bound

$$\limsup_n \sum_{i=1}^n \int W_{n,i}(x)\,|Y_i|\,\mu(dx) \le \limsup_n \left(\frac{1}{n}\sum_{i=1}^n |Y_i|\right) \cdot \max_i \int nW_{n,i}(x)\mu(dx).$$

By the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n |Y_i| \to \mathbb{E}\{|Y|\} \quad \text{a.s.}$$

Hence, it suffices to show that for $c = e^2$, with probability one,

$$\limsup_n \max_i \int nW_{n,i}(x)\mu(dx) \le c. \tag{23}$$

If $G_n = \emptyset$, then $W_{n,i}(x) = 0$ for all $x \in \mathcal{X}$. Thus,

$$\max_i \int nW_{n,i}(x)\mu(dx) = 0.$$

If $G_n \neq \emptyset$, then since $W_{n,i}(x) = 0$ for all $x \in \mathcal{G}_n^c$,

$$\max_i \int nW_{n,i}(x)\mu(dx) = \max_{i:\ell(X_i)\in G_n} \frac{\mu(A_{m,\ell(X_i)})}{\mu_n(A_{m,\ell(X_i)})}.$$

Then,

$$\begin{aligned}
\mathbb{P}\left\{\max_i \int nW_{n,i}(x)\mu(dx) > c\right\} &= \mathbb{P}\left\{G_n \neq \emptyset, \max_{i:\ell(X_i)\in G_n} \frac{\mu(A_{m,\ell(X_i)})}{\mu_n(A_{m,\ell(X_i)})} > c\right\} \\
&\le n \cdot \mathbb{P}\left\{\ell(X_1) \in G_n, \frac{\mu(A_{m,\ell(X_1)})}{\mu_n(A_{m,\ell(X_1)})} > c\right\} \\
&= n \cdot \mathbb{P}\left\{\ell(X_1) \in G_n \cap F_n, \frac{\mu(A_{m,\ell(X_1)})}{\mu_n(A_{m,\ell(X_1)})} > c\right\} \\
&\le n \cdot \mathbb{E}\left\{\sum_{\ell\in F_n} \mathbb{P}\left\{X_1 \in A_{m,\ell}, \frac{\mu(A_{m,\ell})}{\mu_n(A_{m,\ell})} > c \mid \mathbf{X}'_m\right\}\right\} \\
&= n \cdot \mathbb{E}\left\{\sum_{\ell\in F_n} \mathbb{P}\left\{X_1 \in A_{m,\ell}, \frac{n\mu(A_{m,\ell})}{1+\sum_{i=2}^n \mathbb{I}_{\{X_i\in A_{m,\ell}\}}} > c \mid \mathbf{X}'_m\right\}\right\} \\
&= n \cdot \mathbb{E}\left\{\sum_{\ell\in F_n} \mu(A_{m,\ell}) \cdot \mathbb{P}\left\{\frac{n\mu(A_{m,\ell})}{1+\sum_{i=2}^n \mathbb{I}_{\{X_i\in A_{m,\ell}\}}} > c \mid \mathbf{X}'_m\right\}\right\} \\
&\le n \cdot \mathbb{E}\left\{\sum_{\ell\in F_n} \mu(A_{m,\ell}) \cdot \mathbb{P}\left\{\frac{\mu(A_{m,\ell})}{\mu_n(A_{m,\ell})} > c \mid \mathbf{X}'_m\right\}\right\}.
\end{aligned}$$

Chernoff's bound implies that for any $\ell \in F_n$,

$$\mathbb{P}\left\{\frac{\mu(A_{m,\ell})}{\mu_n(A_{m,\ell})} > c \mid \mathbf{X}'_m\right\} \le e^{-n\mu(A_{m,\ell})(1-\frac{1}{c}-\frac{\log c}{c})} \le e^{-(c-1-\log c)\log n} \le n^{-e^2+3}.$$

Thus,

$$\mathbb{P}\left\{\max_i \int nW_{n,i}(x)\mu(dx) > c\right\} \le n^{-e^2+4} \cdot \mathbb{E}\left\{\sum_{\ell\in F_n} \mu(A_{m,\ell})\right\} \le n^{-e^2+4},$$

which is again summable. Hence, by the Borel-Cantelli Lemma, (23) holds with probability one, concluding the proof of *(ii)* and Theorem 5.

5. Proofs of convergence rates

The rates of convergence for the k-NN and Proto-k-NN classifiers of Section 3 are derived using the following decomposition of the excess error probability.

Lemma 8 *Let g_n be a plug-in rule with estimates $P_{n,j}$ of P_j as in Section 1. Abbreviating*

$$R_l^*(x) = P_{g^*(x)}(x) - P_l(x) \ge 0, \qquad l \in \mathcal{Y},$$

we have

$$\mathbb{E}\{L(g_n)\} - L^* \le \sum_{j=1}^M \sum_{l=1}^M J_{n,j,l}$$

where

$$J_{n,j,l} = \int R_l^*(x)\mathbb{I}_{\{l\neq g^*(x)\}}\mathbb{P}\{|P_{n,j}(x) - P_j(x)| \ge R_l^*(x)/M\}\mu(dx), \qquad j,l \in \mathcal{Y}. \tag{24}$$

Below, we bound $J_{n,j,l}$ for each algorithm separately.

Proof of Lemma 8. For any decision function g,

$$\begin{aligned}
\mathbb{P}\{g(X) \neq Y \mid X\} &= 1 - \mathbb{P}\{g(X) = Y \mid X\} \\
&= 1 - \sum_{j=1}^M \mathbb{P}\{g(X) = Y = j \mid X\} \\
&= 1 - \sum_{j=1}^M \mathbb{I}_{\{g(X)=j\}} P_j(X) \\
&= 1 - P_{g(X)}(X),
\end{aligned}$$

which implies

$$\begin{aligned}\mathbb{E}\{L(g_n)\} - L^* &= \mathbb{E}\{P_{g^*(X)}(X) - P_{g_n(X)}(X)\} \\ &= \int \mathbb{E}\{(P_{g^*(x)}(x) - P_{g_n(x)}(x))\mathbb{I}_{\{g^*(x)\neq g_n(x)\}}\}\mu(dx) \\ &= \int \mathbb{E}\{I_n(x)\}\mu(dx),\end{aligned}$$

where

$$I_n(x) = (P_{g^*(x)}(x) - P_{g_n(x)}(x))\mathbb{I}_{\{P_{g^*(x)}(x) > P_{g_n(x)}(x)\}}\mathbb{I}_{\{P_{n,g_n(x)}(x)\geq P_{n,g^*(x)}(x)\}}.$$

The relation

$$\begin{aligned}&\{P_{n,g_n(x)}(x) - P_{n,g^*(x)}(x) \geq 0\} \\ &= \{P_{n,g_n(x)}(x) - P_{g_n(x)}(x) + P_{g_n(x)}(x) - P_{g^*(x)}(x) + P_{g^*(x)}(x) - P_{n,g^*(x)}(x) \geq 0\} \\ &\subseteq \left\{\sum_{j=1}^{M} |P_{n,j}(x) - P_j(x)| \geq P_{g^*(x)}(x) - P_{g_n(x)}(x)\right\}\end{aligned}$$

yields

$$\mathbb{I}_{\{P_{g^*(x)}(x) > P_{g_n(x)}(x)\}}\mathbb{I}_{\{P_{n,g_n(x)}(x)\geq P_{n,g^*(x)}(x)\}} \leq \mathbb{I}_{\{\sum_{j=1}^{M} |P_{n,j}(x)-P_j(x)|\geq P_{g^*(x)}(x)-P_{g_n(x)}(x)>0\}}.$$

Thus,

$$\begin{aligned}I_n(x) &\leq \sum_{l=1}^{M}(P_{g^*(x)}(x) - P_l(x))\mathbb{I}_{\{\sum_{j=1}^{M} |P_{n,j}(x)-P_j(x)|\geq P_{g^*(x)}(x)-P_l(x)\}}\mathbb{I}_{\{g_n(x)=l\neq g^*(x)\}} \\ &\leq \sum_{l=1}^{M}(P_{g^*(x)}(x) - P_l(x))\mathbb{I}_{\{\sum_{j=1}^{M} |P_{n,j}(x)-P_j(x)|\geq P_{g^*(x)}(x)-P_l(x)\}}\mathbb{I}_{\{l\neq g^*(x)\}} \\ &\leq \sum_{j=1}^{M}\sum_{l=1}^{M}(P_{g^*(x)}(x) - P_l(x))\mathbb{I}_{\{l\neq g^*(x)\}}\mathbb{I}_{\{|P_{n,j}(x)-P_j(x)|\geq (P_{g^*(x)}(x)-P_l(x))/M\}} \\ &= \sum_{j=1}^{M}\sum_{l=1}^{M} R_l^*(x)\mathbb{I}_{\{l\neq g^*(x)\}}\mathbb{I}_{\{|P_{n,j}(x)-P_j(x)|\geq R_l^*(x)/M\}}.\end{aligned}$$

Taking expectation and integrating with respect to μ concludes the proof of Lemma 8.

Proof of Theorem 6. Lemma 8 shows that to bound $\mathbb{E}\{L(g_{k,n})\} - L^*$, it suffices to bound $J_{n,j,l}$ in (24). To this end, let

$$\bar{P}_{n,j}(x) = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\{\mathbb{I}_{\{Y_{(n,i)}(x)=j\}} \mid X_1, \ldots, X_n\} = \frac{1}{k}\sum_{i=1}^{k} P_j(X_{(n,i)}(x)).$$

We bound (24) by

$$J_{n,j,l} \leq J_{n,j,l}^{(1)} + J_{n,j,l}^{(2)},$$

where

$$J_{n,j,l}^{(1)} = \int R_l^*(x)\mathbb{I}_{\{l\neq g^*(x)\}}\mathbb{P}\{|P_{n,j}(x) - \bar{P}_{n,j}(x)| \geq R_l^*(x)/(2M)\}\mu(dx)$$

and

$$J_{n,j,l}^{(2)} = \int R_l^*(x)\mathbb{I}_{\{l\neq g^*(x)\}}\mathbb{P}\{|\bar{P}_{n,j}(x) - P_j(x)| \geq R_l^*(x)/(2M)\}\mu(dx).$$

For each j and l, we show that

$$J_{n,j,l}^{(1)} = O(1/k^{(1+\alpha)/2})$$

and

$$J_{n,j,l}^{(2)} = O(h(2k/n)^{1+\alpha}),$$

from which the theorem follows.

The estimation error $J_{n,j,l}^{(1)}$ can be managed in the same way as in the proof of Lemma 1 in Döring et al. (2017). The Hoeffding inequality implies that

$$\mathbb{P}\{|P_{n,j}(x) - \bar{P}_{n,j}(x)| \geq R_l^*(x)/(2M) \mid X_1, \ldots, X_n\} \leq 2e^{-kR_l^*(x)^2/(2M^2)}.$$

Therefore,

$$J_{n,j,l}^{(1)} \leq 2\mathbb{E}\left\{R_l^*(X)\mathbb{I}_{\{l\neq g^*(X)\}}e^{-kR_l^*(X)^2/(2M^2)}\right\}.$$

The margin condition with parameter α means that for $0 \leq t \leq 1$,

$$G(t) := \mathbb{P}\{R_l^*(X)\mathbb{I}_{\{l\neq g^*(X)\}} \leq t\} \leq \mathbb{P}\{P_{(1)}(X) - P_{(2)}(X) \leq t\} \leq c^* t^\alpha.$$

This implies that

$$\begin{aligned} J_{n,j,l}^{(1)} &\leq 2\int_0^1 se^{-ks^2/(2M^2)}G(ds) \\ &\leq 2c^*\alpha\int_0^1 se^{-ks^2/(2M^2)}s^{\alpha-1}ds \\ &\leq 2c^*\alpha\cdot k^{-(\alpha+1)/2}\cdot\int_0^\infty e^{-u^2/(2M^2)}u^\alpha du \\ &= O(k^{-(\alpha+1)/2}). \end{aligned} \tag{25}$$

Concerning the approximation error $J_{n,j,l}^{(2)}$, we follow the line of proof of Lemma 2 in Döring et al. (2017). The generalized Lipschitz condition implies that

$$\begin{aligned} |\bar{P}_{n,j}(x) - P_j(x)| &\leq \frac{1}{k}\sum_{i=1}^k |P_j(X_{(n,i)}(x)) - P_j(x)| \\ &\leq \frac{1}{k}\sum_{i=1}^k h(\mu(S_{x,\rho(x,X_{(n,i)}(x))})) \\ &\leq h(\mu(S_{x,\rho(x,X_{(n,k)}(x))})). \end{aligned}$$

If the distribution function $H_x(\cdot)$ is continuous for each x, then we apply the probability integral transform (cf. Biau and Devroye (2015), p. 8). As a result, the random variable

$$H_x(\rho(x,X)) = \mu\{S_{x,\rho(x,X)}\} \tag{26}$$

is uniformly distributed on $[0,1]$. For i.i.d. uniformly distributed $U_1,\dots,U_n$, denote by $U_{(1,n)},\dots,U_{(n,n)}$ the corresponding order statistics. Then (26) implies

$$\mu(S_{x,\rho(x,X_{(n,k)}(x))}) \stackrel{\mathcal{D}}{=} U_{(k,n)}.$$

Thus, from the generalized Lipschitz condition one gets

$$\begin{aligned}\mathbb{P}\{R_l^*(x)/(2M) \le |\bar{P}_{n,j}(x) - P_j(x)|\} &\le \mathbb{P}\left\{R_l^*(x)/(2M) \le h(\mu(S_{x,\rho(x,X_{(n,k)}(x))}))\right\} \\ &= \mathbb{P}\left\{R_l^*(x)/(2M) \le h(U_{(k,n)})\right\} \\ &= \mathbb{P}\left\{h^{-1}(R_l^*(x)/(2M)) \le U_{(k,n)}\right\}. \end{aligned} \tag{27}$$

As in the proof of Lemma 3 of Döring et al. (2017), Chernoff's exponential inequality implies

$$\begin{aligned}&\mathbb{P}\{R_l^*(x)/(2M) \le |\bar{P}_{n,j}(x) - P_j(x)|\} \\ &\le \mathbb{P}\left\{\sum_{i=1}^n \mathbb{I}_{\{U_i \le h^{-1}(R_l^*(x)/(2M))\}} < k\right\} \\ &\le e^{-(1-\log 2)k} + \mathbb{I}_{\{h^{-1}(R_l^*(x)/(2M)) < 2k/n\}}. \end{aligned} \tag{28}$$

Applying the margin condition, we get

$$\begin{aligned}J_{n,j,l}^{(2)} &\le \mathbb{E}\Big\{R_l^*(X)\mathbb{I}_{\{l\neq g^*(X)\}}(e^{-(1-\log 2)k} + \mathbb{I}_{\{h^{-1}(R_l^*(X)/(2M))<2k/n\}})\Big\} \\ &\le e^{-(1-\log 2)k} + \int_0^1 s\mathbb{I}_{\{h^{-1}(s/(2M))<2k/n\}}G(ds) \\ &\le e^{-(1-\log 2)k} + c^*\alpha\int_0^1 s^\alpha \mathbb{I}_{\{s<2Mh(2k/n)\}}ds \\ &\le e^{-(1-\log 2)k} + O(h(2k/n)^{1+\alpha}),\end{aligned}$$

concluding the proof of Theorem 6.

Proof of Theorem 7. As in the proof of Theorem 6, we bound (24) by

$$J_{n,j,l} \le J_{n,j,l}^{(1)} + J_{n,j,l}^{(2)},$$

where

$$J_{n,j,l}^{(1)} = \int R_l^*(x)\mathbb{I}_{\{l\neq g^*(x)\}}\mathbb{P}\{|P_{n,j}(X'_{(m,1)}(x)) - \bar{P}_{n,j}(X'_{(m,1)}(x))| > R_l^*(x)/(2M)\}\mu(dx)$$

and

$$J_{n,j,\ell}^{(2)} = \int R_l^*(x)\mathbb{I}_{\{l\neq g^*(x)\}}\mathbb{P}\{|\bar{P}_{n,j}(X'_{(m,1)}(x)) - P_j(x)| > R_l^*(x)/(2M)\}\mu(dx).$$

For each j and l, we show that

$$J^{(1)}_{n,j,l} = O(1/k^{(1+\alpha)/2})$$

and

$$J^{(2)}_{n,j,l} = O(h(k/n)^{\alpha+1}) + O(h(1/m)^{\alpha+1}),$$

from which the theorem follows. The Hoeffding inequality implies that

$$\mathbb{P}\{|P_{n,j}(X'_{(m,1)}(x)) - \bar{P}_{n,j}(X'_{(m,1)}(x))| \geq R^*_l(x)/(2M) \mid \mathbf{X}_n, X'_{(m,1)}(x)\} \leq 2e^{-kR^*_l(x)^2/(2M^2)}.$$

Therefore, similarly to (25), the margin condition implies that

$$J^{(1)}_{n,j,l} \leq 2\mathbb{E}\left\{R^*_l(X)\mathbb{I}_{\{l\neq g^*(X)\}}e^{-kR^*_l(X)^2/(2M^2)}\right\} = O(k^{-(\alpha+1)/2}).$$

The generalized Lipschitz condition implies that

$$\begin{aligned}
&|\bar{P}_{n,j}(X'_{(m,1)}(x)) - P_j(x)| \\
&\leq \frac{1}{k}\sum_{i=1}^{k} |P_j(X_{(n,i)}(X'_{(m,1)}(x))) - P_j(X'_{(m,1)}(x))| + |P_j(X'_{(m,1)}(x)) - P_j(x)| \\
&\leq \frac{1}{k}\sum_{i=1}^{k} h(\mu(S_{X'_{(m,1)}(x),\rho(X'_{(m,1)}(x),X_{(n,i)}(X'_{(m,1)}(x)))})) + h(\mu(S_{x,\rho(x,X'_{(m,1)}(x))})) \\
&\leq h(\mu(S_{X'_{(m,1)}(x),\rho(X'_{(m,1)}(x),X_{(n,k)}(X'_{(m,1)}(x)))})) + h(\mu(S_{x,\rho(x,X'_{(m,1)}(x))})).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{P}\{R^*_l(x)/(2M) \leq |\bar{P}_{n,j}(X'_{(m,1)}(x)) - P_j(x)|\} \\
&\leq \mathbb{P}\left\{R^*_l(x)/(4M) \leq h(\mu(S_{X'_{(m,1)}(x),\rho(X'_{(m,1)}(x),X_{(n,k)}(X'_{(m,1)}(x)))}))\right\} \\
&\quad + \mathbb{P}\left\{R^*_l(x)/(4M) \leq h(\mu(S_{x,\rho(x,X'_{(m,1)}(x))}))\right\}.
\end{aligned}$$

Similarly to (27) and (28) in the proof of Theorem 6, Chernoff's inequality implies

$$\begin{aligned}
&\mathbb{P}\left\{R^*_l(x)/(4M) \leq h(\mu(S_{X'_{(m,1)}(x),\rho(X'_{(m,1)}(x),X_{(n,k)}(X'_{(m,1)}(x)))}))\right\} \\
&= \mathbb{P}\left\{R^*_l(x)/(4M) \leq h(U_{(k,n)})\right\} \\
&= \mathbb{P}\left\{h^{-1}(R^*_l(x)/(4M)) \leq U_{(k,n)}\right\} \\
&\leq \mathbb{P}\left\{\sum_{i=1}^{n}\mathbb{I}_{\{U_i\leq h^{-1}(R^*_l(x)/(4M))\}} < k\right\} \\
&\leq e^{-(1-\log 2)k} + \mathbb{I}_{\{h^{-1}(R^*_l(x)/(4M))<2k/n\}}.
\end{aligned}$$

Applying the margin condition, we get

$$\begin{aligned}
&\mathbb{E}\Big\{R_l^*(X)\mathbb{I}_{\{l\neq g^*(X)\}}\mathbb{P}\{R_l^*(X)/(4M)\leq h(\mu(S_{X'_{(m,1)}(X),\rho(X'_{(m,1)}(X),X_{(n,k)}(X'_{(m,1)}(X)))}))\mid X\}\Big\}\\
&\leq \mathbb{E}\Big\{R_l^*(X)\mathbb{I}_{\{l\neq g^*(X)\}}(e^{-(1-\log 2)k}+\mathbb{I}_{\{h^{-1}(R_l^*(X)/(4M))<2k/n\}})\Big\}\\
&\leq e^{-(1-\log 2)k}+\int_0^1 s\mathbb{I}_{\{h^{-1}(s/(4M))<2k/n\}}G(ds)\\
&\leq e^{-(1-\log 2)k}+c^*\alpha\int_0^1 s^\alpha\mathbb{I}_{\{s<4Mh(2k/n)\}}ds\\
&\leq e^{-(1-\log 2)k}+O(h(2k/n)^{\alpha+1})\\
&= e^{-(1-\log 2)k}+O(h(k/n)^{\alpha+1}),
\end{aligned}$$

where in the last equality we applied the assumption that h is concave. A slight modification of the previous argument yields

$$\mathbb{P}\left\{R_l^*(x)/(4M)<h(\mu(S_{x,\rho(x,X'_{(m,1)}(x))}))\right\}=\mathbb{P}\left\{R_l^*(x)/(4M)<h(U_{(1,m)})\right\},$$

and so

$$\begin{aligned}
&\mathbb{E}\Big\{R_l^*(X)\mathbb{I}_{\{l\neq g^*(X)\}}\mathbb{P}\{R_l^*(X)/(4M)<h(\mu(S_{X,\rho(X,X'_{(m,1)}(X))}))\mid X\}\Big\}\\
&=\mathbb{E}\Big\{R_l^*(X)\mathbb{I}_{\{l\neq g^*(X)\}}\mathbb{P}\left\{R_l^*(X)/(4M)<h(U_{(1,m)})\mid X\right\}\Big\}\\
&\leq \mathbb{E}\left\{\int_0^1 s\mathbb{I}_{\{s<4Mh(U_{(1,m)})\}}G(ds)\right\}\\
&\leq c^*\alpha\mathbb{E}\left\{\int_0^1 s^\alpha\mathbb{I}_{\{s<4Mh(U_{(1,m)})\}}ds\right\}\\
&=O(\mathbb{E}\left\{h(U_{(1,m)})^{\alpha+1}\right\})\\
&=O(h(1/m)^{\alpha+1}),
\end{aligned}$$

where in the last equality we again applied the assumption that h is concave.

Acknowledgments

We thank the editor and referees for carefully reading the manuscript and for the suggested improvements. We also thank Roberto Colomboni for pointing out some inaccuracies in a previous version of the manuscript.

References

Christophe Abraham, Gérard Biau, and Benoît Cadre. On the kernel rule for function classification. *Ann. Inst. Statist. Math.*, 58(3):619–633, 2006.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Amparo Baíllo, Antonio Cuevas, and Ricardo Fraiman. Classification methods for functional data. In *The Oxford Handbook of Functional Data Analysis*. 2011.

Jean Bardet. *Tests d'autosimilarite des processus gaussiens. Dimension fractale et dimension de correlation*. PhD thesis, 1997.

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015.

Gérard Biau and László Györfi. On the asymptotic properties of a nonparametric L_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

Gérard Biau, Florentina Bunea, and Marten H. Wegkamp. Functional classification in Hilbert spaces. *IEEE Trans. Inform. Theory*, 51(6):2163–2172, 2005.

Gérard Biau, Frédéric Cérou, and Arnaud Guyader. Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Trans. Inform. Theory*, 56(4):2034–2040, 2010.

Ingrid Blaschzyk and Ingo Steinwart. Improved classification rates under refined margin conditions. *Electronic Journal of Statistics*, 12(1):793–823, 2018.

Florent Burba, Frédéric Ferraty, and Philippe Vieu. k-nearest neighbour method in functional nonparametric regression. *Journal of Nonparametric Statistics*, 21(4):453–469, 2009.

Frédéric Cérou and Arnaud Guyader. Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics*, 10:340–355, 2006.

Tommaso R Cesari and Roberto Colomboni. A nearest neighbor characterization of lebesgue points in metric measure spaces. *Mathematical Statistics and Learning*, 3(1):71–112, 2021.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014.

Benoît Collins, Sushma Kumari, and Vladimir G Pestov. Universal consistency of the k-nn rule in metric spaces and nagata dimension. *arXiv preprint arXiv:2003.00894*, 2020.

Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer-Verlag New York, Inc., 1996.

Maik Döring, László Györfi, and Harro Walk. Rate of convergence of k-nearest-neighbor classification rule. *The Journal of Machine Learning Research*, 18(1):8485–8500, 2017.

Herbert Federer. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

Frédéric Ferraty, Ali Laksaci, Amel Tadj, and Philippe Vieu. Rate of uniform consistency for nonparametric estimates with functional variables. *Journal of Statistical planning and inference*, 140(2):335–352, 2010.

Liliana Forzani, Ricardo Fraiman, and Pamela Llop. Consistent nonparametric regression for functional data under the Stone–Besicovitch conditions. *IEEE Transactions on Information Theory*, 58(11):6697–6708, 2012.

Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Ann. Statist.*, 44(3):982–1009, 2016.

László Györfi. The rate of convergence of k_n-nn regression estimates and classification rules (corresp.). *IEEE Transactions on Information Theory*, 27(3):362–364, 1981.

László Györfi. Universal consistencies of a regression estimate for unbounded regression functions. In *Nonparametric functional estimation and related topics*, pages 329–338. Springer, 1991.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal Bayes consistency in metric spaces. *To appear in The Annals of Statistics*, 2021.

Juha Heinonen. *Lectures on analysis on metric spaces*. Springer Science & Business Media, 2012.

Michael Kohler and Adam Krzyzak. On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE transactions on information theory*, 53(5):1735–1742, 2007.

Aryeh Kontorovich and Roi Weiss. A Bayes consistent 1-NN classifier. In *Artificial Intelligence and Statistics*, 2014.

Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. In *Advances in Neural Information Processing Systems*, pages 1573–1583, 2017.

Samory Kpotufe and Sanjoy Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, 78(5):1496–1515, 2012.

Nadia L Kudraszow and Philippe Vieu. Uniform consistency of kNN regressors for functional variables. *Statistics & Probability Letters*, 83(8):1863–1870, 2013.

Sanjeev R. Kulkarni and Steven E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inform. Theory*, 41(4):1028–1039, 1995.

Nengxiang Ling and Philippe Vieu. Nonparametric modelling for functional data: selected survey and tracks for future. *Statistics*, 52(4):934–949, 2018.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

André Mas. Lower bound in regression for functional data by representation of small ball probabilities. *Electronic Journal of Statistics*, 6:1745–1778, 2012.

Yakov Pesin. On rigorous mathematical definitions of correlation dimension and generalized spectrum for dimensions. *Journal of statistical physics*, 71(3-4):529–547, 1993.

David Preiss. Invalid Vitali theorems. *Abstracta. 7th Winter School on Abstract Analysis*, pages 58–60, 1979.

David Preiss. Gaussian measures and the density theorem. *Comment. Math. Univ. Carolin.*, 22(1):181–193, 1981.

Nikita Puchkin and Vladimir Spokoiny. An adaptive multiclass nearest neighbor classifier. *ESAIM: Probability and Statistics*, 24:69–99, 2020.

Severine Rigot. Differentiation of measures in metric spaces. *arXiv:1802.02069*, 2018.

Richard J. Samworth. Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, 40(5): 2733–2763, 10 2012.

Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4): 595–620, 1977.

Jaroslav Tišer. Vitali covering theorem in Hilbert space. *Trans. Amer. Math. Soc.*, 355(8): 3277–3289, 2003.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Lirong Xue and Samory Kpotufe. Achieving the time of 1-NN, but the accuracy of k-nn. In *International Conference on Artificial Intelligence and Statistics*, pages 1628–1636. PMLR, 2018.

Lin Cheng Zhao. Exponential bounds of mean error for the nearest neighbor estimates of regression functions. *J. Multivariate Anal.*, 21(1):168–178, 1987.