The Annals of Statistics
2021, Vol. 49, No. 4, 2129–2150
https://doi.org/10.1214/20-AOS2029

UNIVERSAL BAYES CONSISTENCY IN METRIC SPACES

BY STEVE HANNEKE[1], ARYEH KONTOROVICH[2,*], SIVAN SABATO[2,†] AND ROI WEISS[3]

[1]*Toyota Technological Institute at Chicago, steve.hanneke@gmail.com*
[2]*Department of Computer Science, Ben-Gurion University of the Negev, *karyeh@cs.bgu.ac.il; †sabatos@cs.bgu.ac.il*
[3]*Department of Computer Science, Ariel University, roiw@ariel.ac.il*

We extend a recently proposed 1-nearest-neighbor based multiclass learning algorithm and prove that our modification is universally strongly Bayes consistent in all metric spaces admitting *any* such learner, making it an "optimistically universal" Bayes-consistent learner. This is the first learning algorithm known to enjoy this property; by comparison, the k-NN classifier and its variants are not generally universally Bayes consistent, except under additional structural assumptions, such as an inner product, a norm, finite dimension or a Besicovitch-type property.

The metric spaces in which universal Bayes consistency is possible are the "essentially separable" ones—a notion that we define, which is more general than standard separability. The existence of metric spaces that are not essentially separable is widely believed to be independent of the ZFC axioms of set theory. We prove that essential separability exactly characterizes the existence of a universal Bayes-consistent learner for the given metric space. In particular, this yields the first impossibility result for universal Bayes consistency.

Taken together, our results completely characterize strong and weak universal Bayes consistency in metric spaces.

1. Introduction. Since their inception in the 1950s [12]—or, according to some accounts, nearly 1000 years earlier [35]—nearest-neighbor methods have provided an intuitive and reliable suite of techniques for performing classification in metric spaces. For k-NN based methods, it has been generally understood that some notion of finite dimensionality is both necessary and sufficient for the methods to be Bayes consistent under all distributions over the metric space—a property known as *universal Bayes consistency* (UBC). However, a complete characterization of the metric spaces in which *any* nearest neighbor method (or other learners, for that matter) is UBC has been so far unknown. For the problem of multiclass classification, we resolve these questions exhaustively.

To answer these questions, we study a compression-based 1-NN algorithm for multiclass classification, which was proposed in 2017 [28], and shown to be strongly UBC in all metric spaces of bounded diameter and doubling dimension. It was also shown that there exist infinite-dimensional spaces in which this algorithm is strongly Bayes consistent, while classic k-NN based methods provably are not. Left open was the full characterization of metric spaces in which this algorithm is UBC. In this work, we provide this characterization. Moreover, we prove that this algorithm is UBC in any metric space for which a UBC algorithm exists, thus resolving the above fundamental open question about nearest-neighbor methods.

Main results. We design a generalized version of the algorithm used in [28], which we call OptiNet (see Algorithm 1). The contribution of this paper is twofold: (i) We show that OptiNet is universally strongly Bayes consistent in all essentially separable metric spaces.

Received June 2019; revised October 2020.
MSC2020 subject classifications. Primary 54E70, 97K80, 62C12; secondary 03E17, 03E55.
Key words and phrases. Metric space, nearest neighbor, classification, Bayes consistency.

A formal definition of *essential separability*—our broadening of the standard notion of separability—is given in Section 3. Briefly, in an essentially separable metric space, the total mass of every probability measure is contained in some separable subspace. Whether *every* metric space is essentially separable is widely believed to hinge upon set-theoretic axioms that are independent of ZFC, having to do with the existence of certain measurable cardinals (we provide the relevant set-theoretic background in Section 4.1). (ii) We show that in any set-theoretic model that allows the existence of nonessentially separable metric spaces, no (strong or weak) universally Bayes-consistent learner is possible on such spaces. To our knowledge, this is the first construction of a learning setting in which universal Bayes consistency is impossible. In contrast, if one adopts a set-theoretic model in which every metric space is essentially separable, then OptiNet is always universally strongly Bayes consistent. As such, OptiNet is *optimistically* universally Bayes consistent for metric spaces, in a sense analogous to [22]: it succeeds whenever success is possible, and is the first learning algorithm known to enjoy this property. For comparison, k-NN and other existing nearest-neighbor approaches are only universally Bayes consistent under additional structural assumptions, such as an inner product, a norm, a finite dimension or a Besicovitch-type property [3, 4, 6], all of which are significantly stronger assumptions than essential separability.

Taken together, our results completely characterize strong and weak UBC in metric spaces.

Related work. Nearest-neighbor methods were initiated by Fix and Hodges in 1951 [12] and, in the celebrated k-NN formulation, have been placed on a solid theoretical foundation [7, 9, 10, 39, 42]. Following the pioneering work of [9, 39] on nearest-neighbor classification, it was shown by [10, 21, 42] that the k-NN classifier is universally strongly Bayes consistent in $(\mathbb{R}^d, \|\cdot\|_2)$. These results made extensive use of the Euclidean structure of $\mathbb{R}^d$, but in [38] a weak Bayes-consistency result was shown for metric spaces with a bounded diameter and finite doubling dimension, and additional distributional smoothness assumptions.

Consistency of NN-type algorithms in more general (and, in particular, infinite-dimensional) metric spaces was discussed in [1, 3, 4, 6, 13, 33]. Characterizations of Bayes consistency for the standard k-NN [6, 13] and for a generalized "moving window" classification rule [1] were given in terms of a Besicovitch-type condition (see Section 5 for a more detailed discussion). By Besicovitch's density theorem [15], in $(\mathbb{R}^d, \|\cdot\|_2)$, and more generally in finite-dimensional normed spaces, the aforementioned condition holds for all distributions; however, in infinite-dimensional spaces this condition may be violated [36, 37]. The violation of the Besicovitch condition is not an isolated pathology—occurring, for example, in the commonly used Gaussian–Hilbert spaces [40]. Leveraging the consistency of k-NN in finite dimensions, the *filtering* technique (taking the first d coordinates in some basis representation for an appropriate d) was shown to be universally weakly consistent in [3]. However, that technique is only applicable in separable Hilbert spaces, as opposed to more general metric spaces. For compact metric spaces, the SVM algorithm can be made universally Bayes consistent by using an appropriate kernel [8].

Although the classic 1-NN classifier is well known to be inconsistent in general, in recent years a series of papers has presented various ways of learning a *regularized* 1-NN classifier, as an alternative to k-NN. Gottlieb et al. [17] showed that an approximate nearest-neighbor search can act as a regularizer, actually improving generalization performance rather than just injecting noise. This technique was extended to multiclass classification in [30]. In a follow-up work, [31] showed that applying Structural Risk Minimization (SRM) to a margin-regularized data-dependent bound very similar to that in [17] yields a strongly Bayes-consistent 1-NN classifier in doubling spaces with a bounded diameter.

Approaching the problem through the lens of sample compression, a computationally near-optimal nearest-neighbor condensing algorithm was presented in [19] and later extended to

cover semimetric spaces [18]; both were based on constructing γ-nets in spaces with a finite doubling dimension (or its semimetric analogue). As detailed in [31], margin-regularized 1-NN methods enjoy a number of statistical and computational advantages over the traditional k-NN classifier. Salient among these are explicit data-dependent generalization bounds, and considerable runtime and memory savings. Sample compression affords additional advantages, in the form of tighter generalization bounds and increased efficiency in time and space. Recently, [28] provided evidence that this technique has wider applicability than k-NN methods, by exhibiting an infinite-dimensional metric measure space where the compression-based learner is Bayes consistent, while k-NN methods provably fail.

The work of Devroye et al. [10], Theorem 21.2, has implications for 1-NN classifiers in $(\mathbb{R}^d, \|\cdot\|_2)$ that are defined based on data-dependent majority-vote partitions of the space. They showed that a *fixed* mapping from each sample size to a data-dependent partitioning rule, satisfying some regularity conditions, induces a universally strongly Bayes-consistent algorithm. This result requires the partitioning rule to have a VC dimension that grows sublinearly in the sample size, and since this rule must be fixed in advance, the algorithm is not fully adaptive. Theorem 19.3 ibid. proves weak consistency for an inefficient compression-based algorithm, which selects among all the possible compression sets of a certain size, and maintains a certain rate of compression relative to the sample size. The generalizing power of sample compression was independently discovered by [10, 34], and later elaborated upon by [20, 23]. In the context of NN classification, [10] lists various condensing heuristics (which have no known performance guarantees) and leaves open the algorithmic question of how to minimize the empirical risk over all subsets of a given size.

The margin-based technique developed in [17, 30] relied on computing a minimum vertex cover. Thus, it was not possible to make it simultaneously and computationally efficient and Bayes consistent when the number of labels exceeds two, since Vertex Cover on general graphs is an NP-hard problem. Although one could resort to a 2-approximation algorithm for vertex cover, this presents an obstruction to establishing the Bayes consistency of the classifier.

In [27], an active-learning algorithm was presented, which across a broad spectrum of natural noise regimes, reduced the sample complexity roughly quadratically. Along the way, this work circumvented the computational obstacle associated with computing a minimum vertex cover on a general graph: the trick was to construct a γ-net and take the majority label (more accurately, the *plurality*—that is, the most frequent—label; we shall use the more familiar terms "majority label" and "majority vote") in each Voronoi region. The majority is determined by actively querying each region, where the number of calls depends on the density and noise level of the region.

Paper outline. After setting down the definitions in Section 2, we describe in Section 3 the compression-based 1-NN algorithm OptiNet studied in this paper and its consistency on essentially-separable metric spaces is proved. In Section 4, we prove that no universally Bayes-consistent algorithm exists on metric spaces that are not essentially separable. We conclude with a discussion in Section 5. Appendices containing the proofs of several auxiliary lemmas are given in the Supplementary Material [24].

2. Definitions and notation. Our *instance space* is the metric probability space $(\mathcal{X}, \rho, \mu)$, where ρ is a metric and μ is a probability measure. By definition, the Borel σ-algebra $\mathscr{B}$ supporting μ is the smallest σ-algebra containing the open sets of ρ. For any $x \in \mathcal{X}$ and $r > 0$, denote by $B_r(x)$ the open ball of radius r around x under the metric ρ:

$$B_r(x) = \{x' \in \mathcal{X} : \rho(x, x') < r\}.$$

We consider a countable label set $\mathcal{Y}$. The unknown sampling distribution is a probability measure $\bar{\mu}$ over $\mathcal{X} \times \mathcal{Y}$, with marginal μ over $\mathcal{X}$. Denote by $(X, Y) \sim \bar{\mu}$ a pair drawn according to $\bar{\mu}$. The generalization error of a classifier $f : \mathcal{X} \to \mathcal{Y}$ is given by

$$\operatorname{err}(f) := \mathbb{P}_{\bar{\mu}}\big[Y \neq f(X)\big],$$

and its empirical error with respect to a labeled set $S' \subseteq \mathcal{X} \times \mathcal{Y}$ is given by

$$\widehat{\operatorname{err}}\big(f, S'\big) := \frac{1}{|S'|} \sum_{(x,y)\in S'} \mathbf{1}\big[y \neq f(x)\big].$$

The optimal Bayes risk of $\bar{\mu}$ is $R^*_{\bar{\mu}} := \inf \operatorname{err}(f)$, where the infimum is taken over all measurable classifiers $f : \mathcal{X} \to \mathcal{Y}$. We omit the subscript $\bar{\mu}$ when there is no ambiguity and denote the optimal Bayes risk of $\bar{\mu}$ by R^*.

For a labeled sequence $S = (x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ and any $x \in \mathcal{X}$, let $X_{\mathrm{nn}}(x, S)$ be the nearest neighbor of x with respect to S and let $Y_{\mathrm{nn}}(x, S)$ be the nearest-neighbor label of x with respect to S:

$$\big(X_{\mathrm{nn}}(x, S), Y_{\mathrm{nn}}(x, S)\big) := \operatorname*{argmin}_{(x_i,y_i)\in S} \rho(x, x_i),$$

where ties are broken lexicographically, that is, the smallest x_i is chosen, with respect to a fixed total ordering of the space $\mathcal{X}$ (such an ordering can always be chosen to be measurable, see Appendix D). The 1-NN classifier induced by S is defined as $h_S(x) := Y_{\mathrm{nn}}(x, S)$. For any $m \in \mathbb{N}$, any sequence $\boldsymbol{X} = \{x_1, \ldots, x_m\} \in \mathcal{X}^m$ induces a *Voronoi partition* of $\mathcal{X}$, $\mathcal{V}(\boldsymbol{X}) := \{V_1(\boldsymbol{X}), \ldots, V_m(\boldsymbol{X})\}$, where each Voronoi cell is

$$V_i(\boldsymbol{X}) := \Big\{x \in \mathcal{X} : i = \operatorname*{argmin}_{1\le j\le m} \rho(x, x_j)\Big\},$$

again breaking ties lexicographically. In particular, for $\boldsymbol{X} = \{X_i : (X_i, Y_i) \in S\}$, we have $h_S(x) = Y_i$ for all $x \in V_i(\boldsymbol{X})$.

A 1-NN algorithm is a mapping from an i.i.d. labeled sample $S_n \sim \bar{\mu}^n$ to a labeled set $S'_n \subseteq \mathcal{X} \times \mathcal{Y}$, yielding the 1-NN classifier $h_{S'_n}$. While the classic 1-NN algorithm sets $S'_n := S_n$, the algorithm which we analyze chooses S'_n adaptively. More generally, a learning algorithm Alg is a mapping (possibly randomized) from a labeled sequence $S_n = (x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ to $\mathsf{Alg}(S_n) \in \mathcal{Y}^{\mathcal{X}}$, satisfying some natural measurability requirements spelled out in Remark 4.10 below. We say that Alg is *strongly Bayes consistent* under $\bar{\mu}$ if $\operatorname{err}(\mathsf{Alg}(S_n))$ converges to R^* almost surely,

$$\mathbb{P}\Big[\lim_{n\to\infty} \operatorname{err}\big(\mathsf{Alg}(S_n)\big) = R^*\Big] = 1.$$

Similarly, Alg is *weakly Bayes consistent* under $\bar{\mu}$ if $\operatorname{err}(\mathsf{Alg}(S_n))$ converges to R^* in expectation,

$$\lim_{n\to\infty} \mathbb{E}\big[\operatorname{err}\big(\mathsf{Alg}(S_n)\big)\big] = R^*.$$

Obviously, the former implies the latter. We say that Alg is *universally Bayes consistent* on a metric space if Alg is Bayes consistent for every distribution supported on its Borel σ-algebra $\mathscr{B}$. Specializing to $\mathsf{OptiNet}$, we have $\mathsf{Alg}(S_n) = h_{S'_n}$.

For $A \subseteq \mathcal{X}$ and $\gamma > 0$, a *γ-net* of A is any *maximal set* $B \subseteq A$ in which all interpoint distances are at least γ. In separable metric spaces, all γ-nets are at most countable. Denote the diameter of a set $A \subseteq \mathcal{X}$ by $\operatorname{diam}(A) \in [0, \infty]$. For a partition $\mathcal{A}$, $\operatorname{diam}(\mathcal{A})$ denotes the maximum diameter $\operatorname{diam}(A)$ among all cells $A \in \mathcal{A}$.

For $n \in \mathbb{N}$, define $[n] := \{1, \ldots, n\}$. Given a labeled set $S_n = (x_i, y_i)_{i\in[n]}$, $d \in [n]$, and any $\boldsymbol{i} = \{i_1, \ldots, i_d\} \in [n]^d$, denote the subsample of S_n indexed by $\boldsymbol{i}$ by $S_n(\boldsymbol{i}) :=$

TABLE 1
Symbols guide

Symbol	Brief description
$(\mathcal{X}, \rho, \mu)$	metric probability space
$\mathscr{B}$	Borel σ-algebra induced by ρ
$B_r(x)$	open ball of radius r around x
$\mathrm{err}(f)$	generalization error of $f : \mathcal{X} \to \mathcal{Y}$
$\widehat{\mathrm{err}}(f, S')$	empirical error of $f : \mathcal{X} \to \mathcal{Y}$ on S'
R^*	Bayes risk
$S_n = (\boldsymbol{X}_n, \boldsymbol{Y}_n)$	random sample of size n
$S_n(\boldsymbol{i}, \boldsymbol{j})$	subsample $(\boldsymbol{X}_{\boldsymbol{i}}, \boldsymbol{Y}_{\boldsymbol{j}})$ of S_n indexed by $\boldsymbol{i}$ and $\boldsymbol{j}$
$S_n(\boldsymbol{i})$	subsample $S_n(\boldsymbol{i}, \boldsymbol{i})$
$S_n(\boldsymbol{i}, *)$	subsample $(\boldsymbol{X}_{\boldsymbol{i}}, \boldsymbol{Y}^*)$ with true majority vote labels
$\boldsymbol{X}(\gamma)$	γ-net of $\boldsymbol{X}_n$
$S_n(\gamma)$	subsample $(\boldsymbol{X}(\gamma), \boldsymbol{Y}(\gamma))$ with empirical majority votes
$M_n(\gamma) = 2\lvert\boldsymbol{X}(\gamma)\rvert$	size of the compression
h_S	1-NN classifier induced by the labeled set S
$\alpha_n(\gamma)$	empirical error of $h_{S_n(\gamma)}$ on S_n
$\mathrm{UB}_\gamma(A)$	γ-envelope of $A \subseteq \mathcal{X}$
$L_\gamma(A)$	γ-missing mass of $A \subseteq \mathcal{X}$
$\mathcal{V}(\boldsymbol{X})$	Voronoi partition of $\mathcal{X}$ induced by $\boldsymbol{X}$
$\mathsf{Alg}(S) = \hat{h}_S$	classifier obtained by learning algorithm Alg when given sample S

$\{(x_{i_1}, y_{i_1}), \ldots, (x_{i_d}, y_{i_d})\}$. Similarly, for a vector $\boldsymbol{y}' = \{y'_1, \ldots, y'_d\} \in \mathcal{Y}^d$, define $S_n(\boldsymbol{i}, \boldsymbol{y}') := \{(x_{i_1}, y'_1), \ldots, (x_{i_d}, y'_d)\}$, namely the subsample of S_n as determined by $\boldsymbol{i}$ where the labels are replaced with $\boldsymbol{y}'$. Lastly, for $\boldsymbol{i}, \boldsymbol{j} \in [n]^d$, we denote $S_n(\boldsymbol{i}; \boldsymbol{j}) := \{(x_{i_1}, y_{j_1}), \ldots, (x_{i_d}, y_{j_d})\}$.

We use standard order-of-magnitude notation throughout the paper; thus, for $f, g : \mathbb{N} \to [0, \infty)$ we write $f(n) \in O(g(n))$ to mean $\limsup_{n\to\infty} f(n)/g(n) < \infty$ and $f(n) \in o(g(n))$ to mean $\limsup_{n\to\infty} f(n)/g(n) = 0$. Likewise, $f(n) \in \Omega(g(n))$ means that $g(n) \in O(f(n))$. In accordance with common convention, we often use the less precise notation $f(n) = O(g(n))$, etc.

The main notation are summarized in Table 1; some are introduced in later sections.

3. Universal Bayes consistency in separable metric spaces. In this section, we describe a variant of the 1-NN majority-based compression algorithm developed in the series of papers [27–29], adapted to maintain measurability in potentially nonseparable metric spaces. We show that this variant is universally Bayes consistent in all separable metric spaces, and the extension to essential separability is immediate, as will become clear below.

Our variant, OptiNet, is formally presented in Algorithm 1. It operates as follows. The input is the sample S_n; the set of points in the sample is denoted by $\boldsymbol{X}_n = \{X_1, \ldots, X_n\}$. The algorithm defines a set Γ of all scales $\gamma > 0$, which are interpoint distances in $\boldsymbol{X}_n$, and the additional scale $\gamma = \infty$. For each scale in Γ, the algorithm constructs a γ-net of $\boldsymbol{X}_n$; note that any singleton in $\boldsymbol{X}_n$ is an ∞-net. Denote the constructed γ-net by $\boldsymbol{X}(\gamma) := \{X_{i_1}, \ldots, X_{i_{M/2}}\}$, where $M/2 \equiv M_n(\gamma)/2 := \lvert\boldsymbol{X}(\gamma)\rvert$ denotes its size and $\boldsymbol{i} \equiv \boldsymbol{i}(\gamma) := \{i_1, \ldots, i_{M/2}\} \in [n]^{M/2}$ denotes the indices selected from S_n for this γ-net. For each γ-net, OptiNet finds the empirical majority vote labels in the Voronoi cells defined by the partition $\mathcal{V}(\boldsymbol{X}(\gamma)) = \{V_1(\boldsymbol{X}(\gamma)), \ldots, V_{M/2}(\boldsymbol{X}(\gamma))\}$; these labels are denoted by $\boldsymbol{Y}'(\gamma) \in \mathcal{Y}^{M/2}$. Formally, for $i \in [M/2]$,

$$Y'_i(\gamma) := \operatorname*{argmax}_{y\in\mathcal{Y}} \big|\{j \in [n] : X_j \in V_i\big(\boldsymbol{X}(\gamma)\big), Y_j = y\}\big|, \tag{3.1}$$

Algorithm 1 (OptiNet) The 1-NN compression-based algorithm

Input: sample $S_n = (X_i, Y_i)_{i\in[n]}$, confidence $\delta \in (0,1)$
Output: A 1-NN classifier

let $\Gamma := (\{\rho(X_i, X_j) : i, j \in [n]\} \cup \{\infty\}) \setminus \{0\}$
for $\gamma \in \Gamma$ **do**
  let $\boldsymbol{X}(\gamma)$ be a γ-net of $\{X_1, \ldots, X_n\}$
  let $M_n(\gamma) := 2|\boldsymbol{X}(\gamma)|$
  for each $i \in [M_n(\gamma)/2]$, let $Y_i'(\gamma)$ be the most frequent label in $V_i(\boldsymbol{X}(\gamma))$ as in (3.1)
  set $S_n'(\gamma) := (\boldsymbol{X}(\gamma), \boldsymbol{Y}'(\gamma))$
end for
Set $\alpha_n(\gamma) := \widehat{\mathrm{err}}(h_{S_n'(\gamma)}, S_n)$
find $\gamma_n^* \in \operatorname{argmin}_{\gamma\in\Gamma} Q(n, \alpha_n(\gamma), M_n(\gamma), \delta)$, where Q is defined in (3.2)
set $S_n' := S_n'(\gamma_n^*)$
return $h_{S_n'}$

where ties are broken based on a fixed preference order on the countable set $\mathcal{Y}$. The result of the procedure is a labeled set $S_n'(\gamma) := S_n(\boldsymbol{i}(\gamma), \boldsymbol{Y}'(\gamma))$ for every possible scale $\gamma \in \Gamma$. The algorithm then selects one scale $\gamma^* \equiv \gamma_n^*$ from Γ, and outputs the hypothesis that it induces, $h_{S_n'(\gamma^*)}$. The choice of γ^* is based on minimizing a generalization error bound, denoted Q, which upper bounds $\mathrm{err}(h_{S_n'(\gamma)})$ with high probability. The error bound is derived based on a compression-based analysis, as follows.

For an even integer $m \le 2n$, we say that a specific S_n' is an (α, m)-*compression* of S_n if there exist $\boldsymbol{i}, \boldsymbol{j} \in [n]^{m/2}$ such that $S_n' = S_n(\boldsymbol{i}, \boldsymbol{j})$ and $\widehat{\mathrm{err}}(h_{S_n'}, S_n) \le \alpha$. Note that at most m examples from S_n determine $h_{S_n'}$, hence this is a compression scheme of size at most m.

The papers [28] and [29] give a consistency result for the original algorithm of [27], on metric spaces with a finite doubling dimension and a finite diameter, under the following assumptions on the generalization error bound $Q(n, \alpha, m, \delta)$:

Q1. For any $n \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over $S_n \sim \bar{\mu}^n$, for all $\alpha \in [0,1]$ and even $m \in [2n]$: If S_n' is an (α, m)-compression of S_n, then

$$\mathrm{err}(h_{S_n'}) \le Q(n, \alpha, m, \delta).$$

Q2. For any fixed $n \in \mathbb{N}$ and $\delta \in (0,1)$, Q is monotonically increasing in α and in m.

Q3. There is a sequence $\{\delta_n\}_{n=1}^\infty$, $\delta_n \in (0,1)$ such that $\sum_{n=1}^\infty \delta_n < \infty$, and for all m,

$$\lim_{n\to\infty} \sup_{\alpha\in[0,1]} \big(Q(n, \alpha, m, \delta_n) - \alpha\big) = 0.$$

Here, we provide a consistency result that holds for more general metric spaces. We prove that OptiNet is universally strongly Bayes consistent in all *essentially separable* metric spaces. Recall that $(\mathcal{X}, \rho)$ is *separable* if it contains a dense countable set. A metric probability space $(\mathcal{X}, \rho, \mu)$ is separable if there is a measurable $\mathcal{X}' \subseteq \mathcal{X}$ with $\mu(\mathcal{X}') = 1$ such that $(\mathcal{X}', \rho)$ is separable. We will call a metric space $(\mathcal{X}, \rho)$ essentially separable (ES) if, for *every* probability measure μ on $\mathscr{B}$, the metric probability space $(\mathcal{X}, \rho, \mu)$ is separable.

To prove this stronger result, we require a slightly stronger version of property **Q3**.

Q3′. There is a sequence $\{\delta_n\}_{n=1}^\infty$, $\delta_n \in (0,1)$ such that $\sum_{n=1}^\infty \delta_n < \infty$, and for any sequence $m_n \in o(n)$,

$$\lim_{n\to\infty} \sup_{\alpha\in[0,1]} \big(Q(n, \alpha, m_n, \delta_n) - \alpha\big) = 0.$$

Property **Q3′** is slightly stronger than **Q3**, since it allows m to grow as $o(n)$ instead of keeping it as a constant. The compression bound used in [28] does not satisfy this property, since it includes a term of the order $m\log(n)/(n-m)$. Therefore, if $m_n = \Omega(n/\log(n))$, then $m_n = o(n)$, yet this term does not converge to zero for $n \to \infty$, thus precluding consistency of the algorithm in [28] for such cases. We provide here a tighter compression bound, which does satisfy **Q3′**.

LEMMA 3.1. *For $m \le n-2$, define*

$$Q(n,\alpha,m,\delta) := \frac{n}{n-m}\alpha + \sqrt{\frac{8(\frac{n}{n-m})\alpha(m\ln(2en/m)+\ln(2n/\delta))}{n-m}} + \frac{9(m\ln(2en/m)+\ln(2n/\delta))}{n-m}. \tag{3.2}$$

For $m > n-2$, define $Q(n,\alpha,m,\delta) := \max(1, Q(n,\alpha,n-2,\delta))$. Then the function Q satisfies the properties **Q1**, **Q2**, **Q3′**.

The approach to obtaining property **Q3′** is inspired by refinements of compression-based generalization bounds holding for the special case of compression schemes which have a *permutation-invariant* reconstruction function [20]. While $h_{S'_n}$ cannot quite be expressed as a permutation-invariant function of a subset of the (X_i, Y_i) data points, it *can* be expressed as a function that is invariant to permutations of two subsets of (X_i, Y_i) points. This is used in the proof of Lemma 3.1, which is provided in Appendix A.2, to derive the tighter compression bound in (3.2). This bound is derived using Bernstein's inequality over $n-m$ random variables and applying a union bound over all $\binom{n}{m/2}^2$, $1 \le m/2 \le n$, possible compressions.

Our main technical innovation, which allows us to dispose of the finiteness requirements on the dimension and the diameter of the metric space that were assumed in [28], is the sublinear growth of γ-nets. Another straightforward but crucial insight is to approximate functions in $L^1(\mu) := \{f : \int |f|\,d\mu < \infty\}$ by *Lipschitz* ones, rather than by continuous functions with compact support as in [28]. The latter approximation requires local compactness, which essentially amounts to a finite dimensionality condition. Our new approach does not require local compactness or finite dimensionality.

THEOREM 3.2. *Let $(\mathcal{X}, \rho, \mu)$ be a separable metric probability space. Let Q be a generalization bound that satisfies Properties* **Q1**, **Q2**, **Q3′**, *and let δ_n be as stipulated by* **Q3′**. *If the input confidence δ for input size n is set to δ_n, then the 1-NN classifier $h_{S'_n(\gamma_n^*)}$ calculated by* OptiNet *is strongly Bayes consistent on $(\mathcal{X}, \rho, \mu)$:*

$$\mathbb{P}\Big[\lim_{n\to\infty} \operatorname{err}(h_{S'_n(\gamma_n^*)}) = R^*\Big] = 1.$$

REMARK 3.3. OptiNet selects the scale γ based on a compression bound. This creates a close connection between the algorithm and the proof of consistency below. However, it is worth noting that it is possible instead to choose γ based on a hold-out validation set: for instance, using $n/2$ of the n samples to construct the predictor for each possible γ value, and then from among these values γ, one can select the γ whose predictor makes the smallest number of mistakes on the remaining $n/2$ samples. Since the analysis of [28] (see [29]), and its generalization below, show that there exists a choice of γ^* for each n such that OptiNet is Bayes consistent, this alternative technique of selecting γ based on a hold-out sample would only lose an additive $O(\sqrt{\log(n)/n})$ compared to using that γ^*, and hence would also be Bayes consistent.

REMARK 3.4. OptiNet is computationally efficient. Using a farthest first-traversal procedure such as Algorithm 1 in [32], one can construct the γ-nets simultaneously for all γ values, including their corresponding empirical errors, in $O(n^2)$ time, leading to a total runtime of $O(n^2)$.

Given a sample $S_n \sim \bar{\mu}^n$, we abbreviate the optimal empirical error $\alpha_n^* = \alpha(\gamma_n^*)$ and the optimal compression size $M_n^* = M(\gamma_n^*)$ as computed by OptiNet. As discussed above, the labeled set $S'_n(\gamma_n^*)$ computed by OptiNet is a (α_n^*, M_n^*)-compression of the sample S_n. For brevity, we denote

$$Q_n(\alpha, m) := Q(n, \alpha, m, \delta_n).$$

To prove Theorem 3.2, we first follow the standard technique, used also in [29], of decomposing the excess error over the Bayes error into two terms:

$$\begin{aligned}\operatorname{err}(h_{S'_n(\gamma_n^*)}) - R^* &= \big(\operatorname{err}(h_{S'_n(\gamma_n^*)}) - Q_n(\alpha_n^*, M_n^*)\big) + \big(Q_n(\alpha_n^*, M_n^*) - R^*\big) \\ &=: T_{\mathrm{I}}(n) + T_{\mathrm{II}}(n).\end{aligned}$$

We now show that each term decays to zero almost surely. For the first term, $T_{\mathrm{I}}(n)$, we have, similar to [29], that Property **Q1** implies that for any $n > 0$,

$$\mathbb{P}\big[\operatorname{err}(h_{S'_n(\gamma_n^*)}) - Q_n(\alpha_n^*, M_n^*) > 0\big] \le \delta_n. \tag{3.3}$$

Based on the Borel–Cantelli lemma and the fact that $\sum \delta_n < \infty$, we have $\limsup_{n\to\infty} T_{\mathrm{I}}(n) \le 0$ with probability 1.

The main difference from the proof in [29] is in the argument establishing $\limsup_{n\to\infty} T_{\mathrm{II}}(n) \le 0$ almost surely. We now show that the generalization bound $Q_n(\alpha_n^*, M_n^*)$ also approaches the Bayes error R^*, thus proving $\limsup_{n\to\infty} T_{\mathrm{II}}(n) \le 0$ almost surely.

We will show below that there exist $N = N(\varepsilon) > 0$, $\gamma = \gamma(\varepsilon) > 0$, and universal constants $c, C > 0$ such that $\forall n \ge N$,

$$\mathbb{P}\big[Q_n\big(\alpha_n(\gamma), M_n(\gamma)\big) > R^* + \varepsilon\big] \le Cne^{-cn\varepsilon^2} + 1/n^2. \tag{3.4}$$

For any $\gamma > 0$ (even if $\gamma \notin \Gamma$), OptiNet finds γ_n^* such that

$$Q_n(\alpha_n^*, M_n^*) = \min_{\gamma' \in \Gamma} Q_n\big(\alpha_n(\gamma'), M_n(\gamma')\big) \le Q_n\big(\alpha_n(\gamma), M_n(\gamma)\big).$$

The bound in (3.4) thus implies that $\forall n \ge N$,

$$\mathbb{P}\big[Q_n(\alpha_n^*, M_n^*) > R^* + \varepsilon\big] \le Cne^{-cn\varepsilon^2} + 1/n^2. \tag{3.5}$$

By the Borel–Cantelli lemma, this implies that almost surely,

$$\limsup_{n\to\infty} T_{\mathrm{II}}(n) = \limsup_{n\to\infty} \big(Q_n(\alpha_n^*, M_n^*) - R^*\big) \le 0.$$

Since $\forall n$, $T_{\mathrm{I}}(n) + T_{\mathrm{II}}(n) \ge 0$, this implies $\lim_{n\to\infty} T_{\mathrm{II}}(n) = 0$ almost surely, thus completing the proof of Theorem 3.2.

It remains to prove (3.4). We note that a simpler form of (3.4) is proved in [29], where they relied on the finiteness of the dimension and the diameter of the space to upper bound the compression size $M_n(\gamma)$ with probability 1. For $A \subseteq \mathcal{X}$, denote its γ-envelope by $\mathrm{UB}_\gamma(A) := \bigcup_{x\in A} B_\gamma(x)$ and consider the *γ-missing mass* of S_n, defined as the following random variable:

$$L_\gamma(S_n) := \mu\big(\mathcal{X} \setminus \mathrm{UB}_\gamma(S_n)\big). \tag{3.6}$$

We bound the left-hand side of (3.4) using a function $n \mapsto t_\gamma(n)$ of order $o(n)$, used to upper bound the compression size; t_γ will be specified below:

$$\begin{aligned}
&\mathbb{P}\big[Q_n\big(\alpha_n(\gamma), M_n(\gamma)\big) > R^* + \varepsilon\big] \\
&\quad \le \mathbb{P}\bigg[Q_n\big(\alpha_n(\gamma), M_n(\gamma)\big) > R^* + \varepsilon \wedge L_\gamma(S_n) \le \frac{\varepsilon}{10} \wedge M_n(\gamma) \le t_\gamma(n)\bigg] \\
&\qquad + \mathbb{P}\big[L_\gamma(S_n) > \varepsilon/10\big] + \mathbb{P}\big[M_n(\gamma) > t_\gamma(n)\big] \\
&\quad =: P_{\mathrm{I}} + P_{\mathrm{II}} + P_{\mathrm{III}}.
\end{aligned} \tag{3.7}$$

First, we bound P_{I}. By a union bound,

$$\begin{aligned}
&\mathbb{P}\bigg[Q_n\big(\alpha_n(\gamma), M_n(\gamma)\big) > R^* + \varepsilon \wedge L_\gamma(S_n) \le \frac{\varepsilon}{10} \wedge M_n(\gamma) \le t_\gamma(n)\bigg] \\
&\quad \le \sum_{d=1}^{t_\gamma(n)} \mathbb{P}\bigg[Q_n\big(\alpha_n(\gamma), M_n(\gamma)\big) > R^* + \varepsilon \wedge L_\gamma(S_n) \le \frac{\varepsilon}{10} \wedge M_n(\gamma) = d\bigg].
\end{aligned}$$

Thus, it suffices to bound each term in the right-hand sum separately. We do so in the following lemma.

LEMMA 3.5. *There exists a function $\varepsilon \mapsto \gamma(\varepsilon)$ for $\varepsilon > 0$, such that under the conditions of Theorem 3.2, there exists an n_0 such that for all $n \ge n_0$, and for all $d \in [t_\gamma(n)]$, letting $\gamma := \gamma(\varepsilon)$,*

$$p_d := \mathbb{P}\bigg[Q_n\big(\alpha_n(\gamma), M_n(\gamma)\big) > R^* + \varepsilon \wedge L_\gamma(S_n) \le \frac{\varepsilon}{10} \wedge M_n(\gamma) = d\bigg] \le e^{-\frac{n\varepsilon^2}{32}}.$$

Applying Lemma 3.5 and summing over all $1 \le d \le t_\gamma(n)$, we have that, for n large enough so that $t_\gamma(n) \le n$,

$$P_{\mathrm{I}} \le \sum_{d=1}^{t_\gamma(n)} p_d \le t_\gamma(n) e^{-\frac{n\varepsilon^2}{32}} \le n e^{-\frac{n\varepsilon^2}{32}}. \tag{3.8}$$

Lemma 3.5 is a generalization of Lemma 10 in [29]. The main difference is that Lemma 10 holds in doubling spaces and uses the fixed map $t_\gamma(n) = 2\lceil \mathrm{diam}(\mathcal{X})/\gamma \rceil^{\mathrm{ddim}}$ for all $n \in \mathbb{N}$. The proof of Lemma 3.5 is the same as that of Lemma 10 in [29], except for two changes that adapt it for a general metric space. First, where Lemma 10 uses the fact that t_γ is set to a constant function, and thus $\lim_{n\to\infty} t_\gamma(n)/n = 0$, the proof of Lemma 3.5 uses instead the property that $t_\gamma(n) = o(n)$, which again leads to the same limit.

In addition, the proof of Lemma 3.5 employs a new result, Lemma 3.6 given below, instead of Lemma 8 from [29]. Lemma 8 from [29] states that for metric spaces with a finite doubling dimension and diameter, Bayes error R^* can be approached using classifiers defined by the true majority-vote labeling over fine partitions of $\mathcal{X}$. Here, we prove that this holds for general metric spaces. Let $\mathcal{V} = \{V_1, \dots\}$ be a countable partition of $\mathcal{X}$, and define the function $I_{\mathcal{V}} : \mathcal{X} \to \mathcal{V}$ such that $I_{\mathcal{V}}(x)$ is the unique $V \in \mathcal{V}$ for which $x \in V$. For any measurable set $\varnothing \ne E \subseteq \mathcal{X}$, define the true majority-vote label $y^*(E)$ by

$$y^*(E) = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X \in E), \tag{3.9}$$

where ties are broken lexicographically. Given $\mathcal{V}$ and a measurable set $W \subseteq \mathcal{X}$, define the true majority-vote classifier $h^*_{\mathcal{V},W} : \mathcal{X} \to \mathcal{Y}$ given by

$$h^*_{\mathcal{V},W}(x) = y^*\big(I_{\mathcal{V}}(x) \cap W\big). \tag{3.10}$$

The new lemma can now be stated as follows.

LEMMA 3.6. *Let $\bar{\mu}$ be a probability measure on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a metric probability space. For any $\nu > 0$, there exists a diameter $\beta = \beta(\nu) > 0$ such that for any countable measurable partition $\mathcal{V} = \{V_1, \dots\}$ of $\mathcal{X}$ and any measurable set $W \subseteq \mathcal{X}$ satisfying*:

(i) $\mu(\mathcal{X} \setminus W) \le \nu$
(ii) $\operatorname{diam}(\mathcal{V} \cap W) \le \beta$,

the true majority-vote classifier $h^*_{\mathcal{V},W}$ defined in (3.10) *satisfies*

$$\operatorname{err}(h^*_{\mathcal{V},W}) \le R^* + 5\nu.$$

The proof of Lemma 3.6 is identical to that of Lemma 8 from [29], except for the following change: in the proof of Lemma 8 from [29], they use their Lemma 7, which states that on doubling spaces, the set of continuous functions with compact support is dense in $L^1(\mu)$. To remove the requirement of compact support, which restricts the type of spaces for which this lemma holds, we use instead a stronger approximation result, which states that Lipschitz functions are dense in $L^1(\mu)$ for any metric probability space. For completeness, we include a proof of this fact in Lemma A.1 in the Supplementary Material [24], where a complete proof of Lemma 3.6 is also given.

Having established Lemma 3.6, this completes the necessary generalizations to obtain Lemma 3.5, whose proof is given in Appendix A.4 for completeness. This proves the bound on P_{I} claimed in (3.8).

We now turn to constructing the function t_γ, which bounds the compression size (i.e., twice the γ-net size) with high probability, and bounding P_{II} and P_{III}.

LEMMA 3.7. *Let $(\mathcal{X}, \rho, \mu)$ be a separable metric probability space. For $S_n \sim \mu^n$, let $\boldsymbol{X}(\gamma)$ be any γ-net of S_n. Then, for any $\gamma > 0$, there exists a function $t_\gamma : \mathbb{N} \to \mathbb{R}_+$ in $o(n)$ such that*

$$\mathbb{P}\Big[\sup_{\gamma\text{-nets}\boldsymbol{X}(\gamma)} 2\big|\boldsymbol{X}(\gamma)\big| \ge t_\gamma(n)\Big] \le 1/n^2. \tag{3.11}$$

This result can be compared to the case of finite-dimensional and finite-diameter metric spaces, in which one can set $t_\gamma(n) := 2\lceil \frac{\operatorname{diam}(\mathcal{X})}{\gamma} \rceil^{\mathrm{ddim}}$ for all $n \in \mathbb{N}$, where ddim is the (finite) doubling dimension and $\operatorname{diam}(\mathcal{X})$ is the diameter of the space, and get that $\mathbb{P}[M_n(\gamma) \ge t_\gamma(n)] = 0$. The proof of Lemma 3.7 is provided in Appendix A of the Supplementary Material [24].

This lemma implies that $P_{\mathrm{III}} \le 1/n^2$, while a bound on P_{II}, which bounds the γ-missing-mass $L_\gamma(S_n)$, is furnished by the following lemma, whose proof is given in Appendix A of the Supplementary Material [24].

LEMMA 3.8. *Let $(\mathcal{X}, \rho, \mu)$ be a separable metric probability space, $\gamma > 0$ be fixed, and the γ-missing mass L_γ defined as in* (3.6). *Then there exists a function $u_\gamma : \mathbb{N} \to \mathbb{R}_+$ in $o(1)$, such that for $S_n \sim \mu^n$ and all $t > 0$,*

$$\mathbb{P}\big[L_\gamma(S_n) \ge u_\gamma(n) + t\big] \le \exp\big(-nt^2\big). \tag{3.12}$$

Taking n sufficiently large so that $u_\gamma(n)$, as furnished by Lemma 3.8, satisfies $u_\gamma(n) \le \varepsilon/20$, and invoking Lemma 3.8 with $t = \varepsilon/20$, we have

$$P_{\mathrm{II}} = \mathbb{P}\big[L_\gamma(S_n) > \varepsilon/10\big] \le e^{-\frac{n\varepsilon^2}{400}}. \tag{3.13}$$

Plugging (3.8), (3.13) and $P_{\mathrm{III}} \le 1/n^2$ into (3.7), we get that (3.4) holds, which completes the proof of Theorem 3.2.

4. Essential separability is necessary for universal Bayes consistency. Recall that a metric space $(\mathcal{X}, \rho)$ is *essentially separable* (ES) if for every probability measure μ on the Borel σ-algebra $\mathscr{B}$, the metric probability space $(\mathcal{X}, \rho, \mu)$ is separable; namely, there is an $\mathcal{X}' \subseteq \mathcal{X}$ with $\mu(\mathcal{X}') = 1$ such that $(\mathcal{X}', \rho)$ is separable. In Theorem 3.2, we established that OptiNet is indeed universally Bayes consistent (UBC) for all such spaces. As such, essential separability of a metric space is sufficient for the existence of a UBC learning rule in that space. In this section, we show that essential separability is also necessary for such a rule to exist.

The metric spaces one typically encounters in statistics and machine learning are all ES, as reflected by Dudley's remark that "for practical purposes, a probability measure defined on the Borel sets of a metric space is always concentrated in some separable subspace" [11]. The question of whether non-ES metric spaces exist at all turns out to be rather subtle. It is widely believed that the existence of non-ES spaces is *independent of the ZFC axioms of set theory* (see Section 4.1 for further details). In other words, it is believed that assuming that ZFC is consistent, its axioms neither necessitate nor preclude the existence of non-ES metric spaces.

The main contribution of this section is to show that in any non-ES metric space (if one exists), no learning rule is UBC.

THEOREM 4.1. *Let $(\mathcal{X}, \rho)$ be a non-ES metric space equipped with the Borel σ-algebra $\mathscr{B}$. Then no (weak or strong) UBC algorithm exists on $(\mathcal{X}, \rho)$.*

Combining this result with Theorem 3.2, the following result is immediate, revealing that OptiNet is *optimistically* UBC (adopting the terminology of [22]), in the sense that the only required assumption on $(\mathcal{X}, \rho)$ is that UBC learning is *possible*.

COROLLARY 4.2. OptiNet *is UBC in every metric space for which there exists a UBC learning rule.*

REMARK 4.3. Theorem 4.1 is somewhat unusual, in that it identifies a setting in which no universal Bayes-consistent procedure exists. To our knowledge, this is the first such impossibility result. Also unusual, for a statistics paper, is the appearance of esoteric set theory. See [2] for another recent result discussing a setting in which learnability is independent of ZFC.

In the next section, we provide necessary preliminaries. Theorem 4.1 is proved in Section 4.2.

4.1. *Preliminaries.* We collect necessary definitions and known results about non-ES metric spaces. In particular, we connect the existence of non-ES metric spaces with the existence of *real-valued measurable cardinals* (Definition 4.5 below). A thorough treatment of the latter, including most of the material in this subsection, can be found in [26]; a more gentle introduction to the subject can be found in [25]. Throughout the following presentation, we work under the standard Zermelo–Fraenkel set theory together with the Axiom of Choice, commonly abbreviated as ZFC.

Cardinals. We denote the cardinality of a set A by $|A|$. The first infinite (countable) cardinal is denoted by $\aleph_0 = |\omega|$, where ω is the set of all finite cardinals. In particular, $\aleph_0$ is the cardinality of the set of natural numbers, $\aleph_0 = |\mathbb{N}|$. We write $[A]^n$ to denote the family of all subsets of A of size $n \in \mathbb{N}$, and $[A]^{<\omega} := \bigcup_{n\in\mathbb{N}} [A]^n$ is the family of all finite subsets of A.

The smallest uncountable cardinal is denoted by $\aleph_1$. The cardinality of the real numbers, also known as the *continuum*, is $\mathfrak{c} = |\mathbb{R}|$. It is well known that $\mathfrak{c} = 2^{\aleph_0} \geq \aleph_1 > \aleph_0$. The continuum hypothesis states that $\mathfrak{c} = \aleph_1$. It is known that its truth value is independent of ZFC, so that either the continuum hypothesis or its negation can be added as an axiom to ZFC set theory while maintaining its consistency status. In the following, we do not include the continuum hypothesis (or its negation) in our set theory; thus, our discussion includes models of ZFC in which $\mathfrak{c} > \aleph_1$.

Nontrivial probability measures. Let $(\mathcal{X}, \mathscr{B})$ be a measurable space. Recall that a probability measure on $\mathcal{X}$, henceforth called a *measure*, is a function $\mu : \mathscr{B} \to [0, 1]$ satisfying:

(i) $\mu(\varnothing) = 0$ and $\mu(\mathcal{X}) = 1$;
(ii) if $A, B \in \mathscr{B}$ and $A \subseteq B$ then $\mu(A) \leq \mu(B)$;
(iii) if $\{A_i\}_{i=1}^{\infty} \subseteq \mathscr{B}$ are pairwise disjoint then $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

A measure μ is *nontrivial* if it vanishes on singletons: $\mu(\{x\}) = 0$, $\forall x \in \mathcal{X}$. Nontrivial measures play a key role in establishing the impossibility of UBC in non-ES spaces and we will be concerned mainly with such measures.

Another important property of a measure is its additivity. For a cardinal κ, a measure μ is *κ-additive* if for any $\beta < \kappa$ and any pairwise disjoint measurable family $\{A_\alpha \in \mathscr{B} : \alpha < \beta\}$,

$$\mu\Bigg(\bigcup_{\alpha<\beta} A_\alpha\Bigg) = \sum_{\alpha<\beta} \mu(A_\alpha) := \sup_{B\in[\beta]^{<\omega}} \sum_{\alpha\in B} \mu(A_\alpha). \tag{4.1}$$

By definition, any measure is $\aleph_1$-additive, commonly known as *σ-additive*. The following lemma states the main property of nontrivial and κ-additive measures that will be used here. Its proof follows directly from the definitions.

LEMMA 4.4. *Let μ be a nontrivial and κ-additive measure on $\mathscr{B}$. Then any set $A \in \mathscr{B}$ with $|A| < \kappa$ has $\mu(A) = 0$.*

Non-ES metric spaces and real-valued measurable cardinals. Before giving the formal definition of real-valued measurable cardinals and establishing their relation to general non-ES metric spaces, let us first illustrate the main ideas which will be presented below, using a simple example of an uncountable discrete metric space. Consider the metric space $([0, 1], \rho_{\text{dis}})$, where ρ_{dis} is the *discrete metric*, defined as

$$\rho_{\text{dis}}(x, x') := \mathbf{1}[x \neq x'], \quad x, x' \in \mathcal{X}. \tag{4.2}$$

The Borel σ-algebra on $([0, 1], \rho_{\text{dis}})$ is all of $2^{[0,1]}$; thus, all subsets of $[0, 1]$ are measurable. This metric space is clearly nonseparable; the interesting question is whether it is ES. In other words, does there exist a measure on $([0, 1], \rho_{\text{dis}})$ that does not have a separable support?

Note that any nontrivial measure on $([0, 1], \rho_{\text{dis}})$ suffices to prove that it is non-ES. Indeed, for any such measure, Lemma 4.4, together with the fact that all measures are σ-additive, implies that any set of positive measure must have an uncountable cardinality. But any such set is clearly nonseparable, due to the discrete nature of the metric space. Therefore, if a nontrivial measure exists on $([0, 1], \rho_{\text{dis}})$, then it is non-ES. Conversely, if there are no nontrivial measures on $([0, 1], 2^{[0,1]})$, then the discrete metric space admits only trivial measures with countable support. So in this case $([0, 1], \rho_{\text{dis}})$ is ES. Thus, the question of whether $([0, 1], \rho_{\text{dis}})$ is non-ES is equivalent to the question of whether a nontrivial measure exists on this space. It is known that the Lebesgue measure, defined on the Borel σ-algebra generated by open sets on $([0, 1], |\cdot|)$, cannot be extended to the measurable space $([0, 1], 2^{[0,1]})$ while

simultaneously being translation invariant [15]. However, currently, other nontrivial measures on $([0,1], 2^{[0,1]})$ are not ruled out in ZFC.

More generally, given a cardinal κ, let $\mathcal{X}$ be some set of that cardinality, and consider the measurable space $\mathfrak{X}_\kappa := (\mathcal{X}, 2^{\mathcal{X}})$. As above, such a space is induced, for example, by the discrete metric ρ_{dis}. Moreover, whether $(\mathcal{X}, \rho_{\mathrm{dis}})$ is ES depends only on the cardinality κ, and is closely related to the existence of nontrivial measures on $\mathfrak{X}_\kappa$, similarly to the example of $([0,1], \rho_{\mathrm{dis}})$ above. To characterize the cardinalities for which $\mathfrak{X}_\kappa$ is ES, we use the known concept of *real-valued measurable cardinals* (RVMC).

DEFINITION 4.5. A cardinal κ is *real-valued measurable* if there exists a nontrivial and κ-additive measure on $\mathfrak{X}_\kappa$. Any such measure is called a *witnessing measure* for $\mathfrak{X}_\kappa$.

Clearly, any RVMC must be uncountable. We denote by $\kappa_{\min}$ the smallest RVMC; this cardinal exists if some RVMC exists, by the well ordering of the cardinals. The following theorem from [5] characterizes ES metric spaces in terms of $\kappa_{\min}$. Recall that a set $D \subseteq \mathcal{X}$ is *discrete* if for any $x \in D$ there exists some $r_x > 0$ such that $B_{r_x}(x) \cap D = \{x\}$.

THEOREM 4.6 ([5], Appendix III, Theorem 2). *Let $\kappa_{\min}$ be the smallest real-valued measurable cardinal (if one exists). Then a metric space $(\mathcal{X}, \rho)$ is ES if and only if every discrete $D \subseteq \mathcal{X}$ has $|D| < \kappa_{\min}$.*

REMARK 4.7. Theorem 4.6 is stated in [5] only for the case $\kappa_{\min} \leq \mathfrak{c}$. However, one can readily verify that the proof extends essentially verbatim (by replacing "atomless" with "nontrivial") to the case $\kappa_{\min} > \mathfrak{c}$ as well.

It follows that whether a given metric space $(\mathcal{X}, \rho)$ is ES or not depends on whether any RVMC exists and, if one exists, on the cardinality of the smallest such cardinal, $\kappa_{\min}$. Assuming that ZFC is consistent (which cannot be proved in ZFC, by Gödel's second incompleteness theorem), it is well known that one cannot prove in ZFC the existence of real-valued measurable cardinals (RVMC). While it is possible that one can prove in ZFC that RVMCs do not exist, no such proof has been discovered yet. However, quoting Fremlin [14], "at present, almost no-one is seriously searching for a proof in ZFC that real-valued measurable cardinals do not exist." In fact, currently, the vast majority of set-theoreticians believe that the existence of RVMC is *independent* of ZFC, that is, assuming that ZFC is consistent, the existence of an RVMC can neither be proven nor disproven from the axioms of ZFC.

In particular, if one adds to ZFC the axiom that no RVMC exists, then *any* metric space is ES. Alternatively, under some additional properties that are beyond the scope of this paper, one can take $\kappa_{\min}$ to be of cardinality that is arbitrarily large. For more details, see [26], Section 12.

The above relations (and their connection to the results to follow) are further discussed in Section 5 and illustrated in Figure 1 therein.

REMARK 4.8. It is worth mentioning that if one adds to ZFC the continuum hypothesis, $\mathfrak{c} = \aleph_1$, which is well known to be independent of ZFC, then if a RVMC exists, then it must hold that $\kappa_{\min} > \mathfrak{c}$ [26]. In this case, all metric spaces of cardinality $\leq \mathfrak{c}$ are ES. In particular, the metric space $([0,1], \rho_{\mathrm{dis}})$ discussed at the beginning of this section admits only trivial measures with a countable support.

4.2. *UBC is impossible in non-ES metric spaces.* In this section, we prove the following theorem, which readily implies Theorem 4.1.

THEOREM 4.9. *Let $(\mathcal{X}, \rho)$ be a non-ES metric space and let* Alg *be any (possibly random) learning algorithm mapping samples $S \in (\mathcal{X} \times \{0,1\})^{<\omega}$ to classifiers $\mathsf{Alg}(S) \in \{0,1\}^{\mathcal{X}}$. Then, there exist a measure $\bar{\mu}$ on $\mathcal{X} \times \{0,1\}$ (w.r.t. the Borel sets induced by ρ), a measurable classifier $h^* : \mathcal{X} \to \{0,1\}$, and an $\varepsilon > 0$ such that, for $n \in \mathbb{N}$ and $S_n \sim \bar{\mu}^n$,*

$$\limsup_{n\to\infty} \mathbb{E}\big[\mathrm{err}_{\bar{\mu}}\big(\mathsf{Alg}(S_n)\big)\big] \geq \mathrm{err}_{\bar{\mu}}\big(h^*\big) + \varepsilon = R^*_{\bar{\mu}} + \varepsilon, \tag{4.3}$$

where $R^*_{\bar{\mu}}$ is the optimal Bayes error. In particular, no weak or strong UBC algorithm exists for $(\mathcal{X}, \rho)$.

For notational simplicity, in the following we denote $\hat{h}_S := \mathsf{Alg}(S)$ (not to be confused with the 1-NN classifier h_S which we used in previous sections).

REMARK 4.10 (Measurability of Alg). To be strictly clear about definitions here, note that we require that the learning algorithm be *measurable*, in the sense that for every $\bar{\mu}$ and n, for $S \sim \bar{\mu}^n$, $\hat{h}_S$ is a $\mathscr{B}(L^1(\mu))$-measurable random variable, where μ is the marginal of $\bar{\mu}$ on $\mathcal{X}$, and $\mathscr{B}(L^1(\mu))$ is the Borel σ-algebra on the set of all measurable functions $\mathcal{X} \to \{0,1\}$, induced by the $L^1(\mu)$ pseudo-metric. This is a basic criterion, without which the expected risk of $\hat{h}_S$ is not well defined (among other pathologies).

For deterministic algorithms $\hat{h}$, to satisfy the above criterion, it suffices that the function $(s, x) \mapsto \hat{h}_s(x)$ on $(\mathcal{X} \times \{0,1\})^n \times \mathcal{X}$ is a measurable $\{0,1\}$-valued random variable, under the product σ-algebra on $(\mathcal{X} \times \{0,1\})^n \times \mathcal{X}$. To see this, note that for any such function, for $X \sim \mu$ independent of $S \sim \bar{\mu}^n$, for any measurable function $f : \mathcal{X} \to \{0,1\}$, we have that $|\hat{h}_S(X) - f(X)|$ is a measurable random variable; hence the variable $\mathbb{E}[|\hat{h}_S(X) - f(X)| \mid S]$ is well defined and measurable. Therefore, for any $\varepsilon > 0$, the event that $\mathbb{E}[|\hat{h}_S(X) - f(X)| \mid S] \leq \varepsilon$ is measurable. Thus, the inverse images of balls in the $L^1(\mu)$ pseudo-metric are measurable sets, and since these balls generate $\mathscr{B}(L^1(\mu))$, this implies $\hat{h}_S$ is a $\mathscr{B}(L^1(\mu))$-measurable random variable.

In particular, we note that OptiNet satisfies this measurability criterion, since calculating its prediction $\hat{h}_s(x)$ involves only simple operations based on the metric ρ (which are measurable, since by definition, ρ induces the topology generating the Borel σ-algebra), and other basic measurability-preserving operations such as argmin for a finite number of indices indexing measurable quantities. Thus, our requirements of $\hat{h}_S$ in Theorem 4.9 are satisfied by OptiNet.

REMARK 4.11. In [6], Section 2.1, the authors define the metric space $(\mathcal{X}, \rho)$, where $\mathcal{X} = [0,1]$ and

$$\rho(x, x') = \mathbf{1}[x \neq x'] \cdot (1 + \mathbf{1}[xx' \neq 0])$$

and endow it with the distribution μ, which places a mass of $1/2$ on $x = 0$ and spreads the rest of the mass "uniformly" on $(0,1]$. The deterministic labeling $h^*(x) = \mathbf{1}[x > 0]$ is imposed. The authors observe that the optimal Bayes risk is $R^* = 0$ while the (classical) 1-NN classifier achieves an asymptotic expected risk of $1/2$—in contradistinction to the standard result that in finite-dimensional spaces 1-NN is Bayes consistent in the realizable case. The authors then use this example to argue that "[separability] is required even in finite dimension." We find the example somewhat incomplete, because care is not taken to ensure that $(\mathcal{X}, \rho, \mu)$ is a *metric probability space*,that is, that the σ-algebra supporting μ is generated by the open sets

of ρ. Indeed, the Borel σ-algebra generated by ρ is the discrete one, $\mathscr{B} = 2^{[0,1]}$. Endowing the latter with a "uniform" measure implicitly assumes that the Lebesgue measure on the *standard* Borel σ-algebra can be extended to all subsets of $[0, 1]$—a statement known to be equivalent to $\mathfrak{c}$ being larger than or equal to a real-valued measurable cardinal [26]. So the above metric probability space is assumed to be non-ES, as in Theorem 4.9. Another objection is that, under any reasonable notion of *dimension*, the metric space $(\mathcal{X}, \rho)$ would be considered $\mathfrak{c}$-dimensional rather than finite-dimensional.

It is worth mentioning that by Remark 4.8, if one accepts, say, the continuum hypothesis, then the above metric space becomes ES and admits only trivial measures with a countable support (so the standard k-NN, and many other algorithms, are in fact UBC in this space).

To prove Theorem 4.9, we first note that (4.3) indeed implies that no weak or strong UBC algorithm exists for $(\mathcal{X}, \rho)$ by an application of [5], Theorem 5.4. To establish (4.3), note that since $(\mathcal{X}, \rho)$ is non-ES, Theorem 4.6 implies that there exists a discrete set $D \subseteq \mathcal{X}$ with $|D| = \kappa_{\min}$, where $\kappa_{\min}$ is the smallest real-valued measurable cardinal (see Section 4.1). Let $\mathfrak{X}_{|D|} := (D, 2^D)$. By Lemma B.1 in the Supplementary Material [24], $2^D \subseteq \mathscr{B}$. Hence, it suffices to construct the required adversarial measure on $D \times \{0, 1\}$. That being the case, from now on we set without loss of generality $\mathcal{X} := D$ and $\mathscr{B} := 2^D$.

Below, we split the argument for the construction of the required adversarial measure on $\mathcal{X} \times \{0, 1\}$ into two cases:

$$\text{Case (I): } \kappa_{\min} \leq \mathfrak{c} \quad \text{and} \quad \text{Case (II): } \kappa_{\min} > \mathfrak{c}.$$

This is manifested by what is known as Ulam's dichotomy. This dichotomy dictates the nature of nontrivial measures in the two cases. To formally state the dichotomy, we first need some additional definitions.

Let μ be a measure on $\mathcal{X}$. A set $A \subseteq \mathcal{X}$ is an *atom* of μ if $\mu(A) > 0$ and for every measurable $B \subseteq A$ either $\mu(B) = 0$ or $\mu(B) = \mu(A)$. A measure μ is *atomless* if it has no atoms. So in an atomless measure, for any $A \in \mathscr{B}$ with $\mu(A) > 0$ there exists a $B \subset A$ with $0 < \mu(B) < \mu(A)$. Conversely, μ is *purely atomic* if every $A \in \mathscr{B}$ with $\mu(A) > 0$ contains an atom.

Clearly, in a countable space all measures are trivial and purely atomic. However, in uncountable spaces matters are more subtle. While any atomless measure is nontrivial, one might expect that conversely a nontrivial measure cannot contain an atom. However, this is not necessarily the case. In particular, when $\kappa_{\min} > \mathfrak{c}$, measures on $\mathcal{X}$ that are simultaneously nontrivial and purely atomic exist.

Formally, let κ be an RVMC. Recall that a witnessing measure for $\mathfrak{X}_\kappa$ is a nontrivial and κ-additive measure on $\mathfrak{X}_\kappa$, namely, a measure defined over all subsets of $\mathcal{X}$ and that vanishes on any set of cardinality $< \kappa$. We say that κ is *two-valued measurable* if there is a $\{0, 1\}$-valued witnessing measure on $\mathfrak{X}_\kappa$, where a measure is $\{0, 1\}$-valued (or *two-valued*) if $\mu(A) \in \{0, 1\}$ for all $A \in \mathscr{B}$. Clearly, a two-valued measure is purely atomic and satisfies that, for any countable partition $\{P_i\}_{i \in \mathbb{N}} \subseteq \mathscr{B}$ of $\mathcal{X}$, there exists one and only one $j \in \mathbb{N}$ such that $\mu(P_j) = 1$. We say that κ is *atomlessly measurable* if there is an atomless witnessing measure on $\mathfrak{X}_\kappa$. In 1930, Ulam established the following dichotomy (see [15], Section 543).

THEOREM 4.12 (Ulam's dichotomy [41]). *Let κ be a real-valued measurable cardinal. Then*:

(i) *if $\kappa \leq \mathfrak{c}$ then κ is atomlessly measurable and every witnessing measure on $\mathfrak{X}_\kappa$ is atomless*;

(ii) *if $\kappa > \mathfrak{c}$ then κ is two-valued measurable and every witnessing measure on $\mathfrak{X}_\kappa$ is purely atomic*.

In other words, if κ is atomlessly measurable then $\kappa \leq \mathfrak{c}$, while if κ is two-valued measurable then $\kappa > \mathfrak{c}$.

We now proceed to construct the adversarial measures on $\mathcal{X} \times \{0,1\}$ by considering the two cases (I) and (II) above separately.

(*I*) *The case $\kappa_{\min} \leq \mathfrak{c}$.* By Ulam's dichotomy in Theorem 4.12, $|\mathcal{X}| = \kappa_{\min}$ is atomlessly measurable, so there exists an atomless witnessing measure μ on $\mathscr{B}$. Fix such a μ and define the induced set-difference pseudometric

$$\Delta(A, B) = \mu\big(\{A \cup B\} \setminus \{A \cap B\}\big), \quad A, B \in \mathscr{B}.$$

Define the metric space $(\mathscr{U}, \Delta)$, where $\mathscr{U} \subseteq \mathscr{B}$ is the quotient σ-algebra under the equivalence relation $A \sim B \Leftrightarrow \Delta(A, B) = 0$. The measure μ induces the corresponding functional $\tilde{\mu} : \mathscr{U} \to [0,1]$ which agrees with μ on the equivalence classes. The following is proved in Appendix B of the Supplementary Material [24] by an application of Gitik–Shelah theorem [16].

LEMMA 4.13. *Let $\mathcal{X}$ be a set of an atomlessly-measurable cardinality κ and let μ be a witnessing measure on $\mathfrak{X}_\kappa$. Let $(\mathscr{U}, \Delta)$ be as above. Then there exist $\varepsilon > 0$ and $\mathcal{H}_\varepsilon \subseteq \mathscr{U}$ of cardinality $|\mathcal{H}_\varepsilon| = \kappa$ that is ε-separated*:

$$\Delta(U, V) \geq \varepsilon \quad \forall U, V \in \mathcal{H}_\varepsilon, U \neq V.$$

By Lemma 4.13, there exist an $\varepsilon > 0$ and a set $\mathcal{H}_\varepsilon \subseteq \{0,1\}^{\mathcal{X}}$ such that $\forall g, h \in \mathcal{H}_\varepsilon$ with $g \neq h$, $\mu(\{x : g(x) \neq h(x)\}) \geq \varepsilon$ and, furthermore, $\mathcal{H}_\varepsilon$ has cardinality $\kappa_{\min}$: that is, the same cardinality as $\mathcal{X}$. Since $\kappa_{\min}$ is atomlessly-measurable, there exists an atomless witnessing measure π on $(\mathcal{H}_\varepsilon, 2^{\mathcal{H}_\varepsilon})$. We will construct the distribution $\bar{\mu}$ using a random construction, by fixing the marginal μ on $\mathcal{X}$ and setting $\bar{\mu}$ to agree with the classifier h^*, which is π-distributed, independently of the input to the algorithm. This process is described formally below.

First, we introduce a relaxed objective for the learning algorithm Alg. Recall that given a labeled sample $S_n \in (\mathcal{X} \times \{0,1\})^n$, Alg outputs a classifier $\hat{h}_{S_n} \in \{0,1\}^{\mathcal{X}}$. For any sequence $S'_x := \{x'_1, x'_2, \ldots\} \in \mathcal{X}$, $n \in \mathbb{N}$, and $\mathbf{y}' := (y'_1, \ldots, y'_n) \in \{0,1\}^n$, denote $\hat{h}_{S'_x, \mathbf{y}'} := \hat{h}_{\{(x'_1, y'_1), \ldots, (x'_n, y'_n)\}}$ and let $H_{S'_x} := \{\hat{h}_{S'_x, \mathbf{y}'} : n \in \mathbb{N}, \mathbf{y}' \in \{0,1\}^n\}$. This set may be random if the learning algorithm is randomized. Then note that, for any fixed $\bar{\mu}$, denoting by $S := \{(x_1, y_1), (x_2, y_2), \ldots\}$ a countably-infinite sequence of independent $\bar{\mu}$-distributed random variables, and further denoting $S_n := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ and $S_x := \{x_1, x_2, \ldots\}$, we have

$$\inf_n \mathbb{E}\big[\mathrm{err}_{\bar{\mu}}(\hat{h}_{S_n})\big] \geq \mathbb{E}\Big[\inf_{h \in H_{S_x}} \mathrm{err}_{\bar{\mu}}(h)\Big].$$

Now take $\boldsymbol{S}_x = \{x_1, x_2, \ldots\}$ to be an i.i.d. μ-distributed sequence, and let $h^* \sim \pi$ independently of $\boldsymbol{S}_x$. Let $\bar{\mu}$ have marginal μ over $\mathcal{X}$ and define $\bar{\mu}$ such that $\bar{\mu}(\{(x, h^*(x)) : x \in \mathcal{X}\}) = 1$; that is, $\bar{\mu}$ is an h^*-dependent random measure. Note that $\mathrm{err}_{\bar{\mu}}(h^*) = 0$ (a.s.), and hence also that any h has $\mathrm{err}_{\bar{\mu}}(h) = \mu(\{x' : h(x') \neq h^*(x')\})$ (a.s.). Furthermore, by the assumed measurability of the learning algorithm, for each $\mathbf{y}$ we have that $\hat{h}_{\boldsymbol{S}_x, \mathbf{y}}$ is a $\mathscr{B}(L^1(\mu))$-measurable random variable, and h^* is also $\mathscr{B}(L^1(\mu))$-measurable (its distribution is π, which is defined on this σ-algebra). Therefore, $\mu(\{x' : \hat{h}_{\boldsymbol{S}_x, \mathbf{y}}(x') \neq h^*(x')\})$ is a measurable random variable, equal (a.s.) to $\mathrm{err}_{\bar{\mu}}(\hat{h}_{\boldsymbol{S}_x, \mathbf{y}})$.

In particular, this implies that $\mathbb{E}[\inf_{h\in H_{S_x}} \mathrm{err}_{\bar{\mu}}(h)]$ is well defined, and by the law of total expectation,

$$\begin{aligned}\mathbb{E}\Big[\inf_{h\in H_{S_x}} \mathrm{err}_{\bar{\mu}}(h)\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\inf_{h\in H_{S_x}} \mathrm{err}_{\bar{\mu}}(h) \mid H_{S_x}\Big]\Big] \\ &\geq \mathbb{E}\Big[(\varepsilon/2)\mathbb{P}\Big(\inf_{h\in H_{S_x}} \mathrm{err}_{\bar{\mu}}(h) > \varepsilon/2 \mid H_{S_x}\Big)\Big] \\ &= \mathbb{E}\Big[(\varepsilon/2)\pi\Big(h' \in \mathcal{H}_\varepsilon : \inf_{h\in H_{S_x}} \mu(\{x' : h(x') \neq h'(x')\}) > \varepsilon/2\Big)\Big].\end{aligned}$$

Then note that each element of H_{S_x} can be $(\varepsilon/2)$-close to at most one element of $\mathcal{H}_\varepsilon$, and since H_{S_x} is a countable set, this implies that, given H_{S_x}, the set $H^\varepsilon_{S_x} = \{h' \in \mathcal{H}_\varepsilon : \inf_{h\in H_{S_x}} \mu(\{x' : h(x') \neq h'(x')\}) \leq \varepsilon/2\}$ is countable.

But since π vanishes on singletons, we have $\pi(H^\varepsilon_{S_x}) = 0$. Thus, given H_{S_x},

$$\pi\Big(h' \in \mathcal{H}_\varepsilon : \inf_{h\in H_{S_x}} \mu(\{x' : h(x') \neq h'(x')\}) > \varepsilon/2\Big) = 1,$$

so that altogether we have

$$\mathbb{E}\Big[\inf_{h\in H_{S_x}} \mathrm{err}_{\bar{\mu}}(h)\Big] \geq \varepsilon/2.$$

In particular, this also implies there exist fixed choices of h^* for which (4.3) holds. This completes the proof for the case (*I*).

(*II*) *The case* $\kappa_{\min} > \mathfrak{c}$. By Ulam's dichotomy in Theorem 4.12, $|\mathcal{X}| = \kappa_{\min}$ is two-valued measurable, so there exists a two-valued witnessing measure μ on $\mathscr{B}$. As the following lemma shows, μ can be taken to further satisfy a key homogeneity property. The lemma is proved in Appendix C of the Supplementary Material [24], where it is shown to follow by combining Theorems 10.20, 10.22 in [26] and Ulam's theorem 4.12.

LEMMA 4.14. *Let $\mathcal{X}$ be of a two-valued measurable cardinality κ and let $\mathfrak{X}_\kappa = (\mathcal{X}, 2^{\mathcal{X}})$. Then there is a witnessing measure μ on $\mathfrak{X}_\kappa$ such that for any function $f : [\mathcal{X}]^{<\omega} \to \mathbb{R}$, there exists a $U \subseteq \mathcal{X}$ with $\mu(U) = 1$ such that U is* homogeneous *for f, that is, for every $n \in \mathbb{N}$, there exists a $C_n \in \mathbb{R}$ such that $f(W) = C_n$ for all $W \in [U]^n$.*

Let μ be a two-valued witnessing measure on $\mathscr{B} = 2^{\mathcal{X}}$ as furnished by Lemma 4.14. For a label $y \in \{0, 1\}$ and any two-valued witnessing measure ϕ, let $\bar{\phi}_y$ be the measure over $\mathcal{X} \times \{0, 1\}$ with ϕ as its marginal over $\mathcal{X}$ and

$$\bar{\phi}_y(Y = y \mid X = x) = 1 \quad \forall x \in \mathcal{X}.$$

For μ as above, and any other two-valued witnessing measure ϕ, define

$$\lambda_\phi := \frac{2}{3}\bar{\phi}_1 + \frac{1}{3}\bar{\mu}_0. \tag{4.4}$$

We will show that there exists a two-valued witnessing measure $\nu := \nu(\mu, \mathsf{Alg}) \neq \mu$ such that Alg cannot be Bayes-consistent on both λ_μ and λ_ν. To this end, we will use the following properties of the mixture λ_ϕ, proved in Appendix C of the Supplementary Material [24].

LEMMA 4.15. *Let $\nu \neq \mu$ be any two distinct two-valued measures on $(\mathcal{X}, \mathscr{B})$ and let λ_ϕ with $\phi \in \{\mu, \nu\}$ be as in* (4.4).

(i) *Any Bayes-optimal classifier h^* on λ_μ achieves the optimal Bayes-error* $\operatorname{err}_{\lambda_\mu}(h^*) = \frac{1}{3}$ *if and only if* $\mathbb{E}_{X\sim\mu}[h^*(X)] = 1$.

(ii) *Any Bayes-optimal classifier h^* on λ_ν achieves the optimal Bayes-error* $\operatorname{err}_{\lambda_\nu}(h^*) = 0$ *if and only if* $\mathbb{E}_{X\sim\mu}[h^*(X)] = 0$ *and* $\mathbb{E}_{X\sim\nu}[h^*(X)] = 1$.

Let $\mathsf{Alg} : (\mathcal{X}\times\{0,1\})^{<\omega} \to 2^{\mathcal{X}}$ be any (possibly randomized) learning algorithm, and recall that $\hat{h}_S$ denotes the classifier output for data set S; for S and X independent samples from Borel measures on $\mathcal{X}$, we suppose that $\hat{h}_S(X)$ is a measurable random variable (by definition of learning algorithm; see Remark 4.10). Let $\nu \neq \mu$ be a two-valued witnessing measure to be chosen below. Consider the quantity

$$Z_n^\phi := \mathbb{E}_{S_n\sim(\lambda_\phi)^n}\big[\mathbb{E}_{X\sim\mu}\big[\hat{h}_{S_n}(X)\big]\big], \quad \phi \in \{\mu,\nu\}.$$

In the case of a randomized Alg, also add an innermost expectation over the independent randomness of Alg in the above expression. By Lemma 4.15, for Alg to be Bayes consistent on both λ_ν and λ_μ we must have

$$Z_n^\phi \xrightarrow[n\to\infty]{} \delta_{\mu,\phi}, \quad \phi \in \{\mu,\nu\}, \tag{4.5}$$

where $\delta_{\mu,\phi}$ is the Kronecker delta. So to prove the claim it suffices to show that we can choose $\nu := \nu(\mu, \mathsf{Alg})$ such that (4.5) does not hold.

Given a labeled sample $S_n = (\mathbf{X}_n, \mathbf{Y}_n) \sim (\lambda_\phi)^n$ with $\phi \in \{\mu,\nu\}$, let $n_1 := n_1(\mathbf{Y}_n) = \sum_{i=1}^n Y_i$ and $n_0 := n - n_1$ be the random number of samples in S_n with labels 1 and 0, respectively, and let $\mathbf{X}_n^0 \in \mathcal{X}^{n_0}$ and $\mathbf{X}_n^1 \in \mathcal{X}^{n_1}$ be the corresponding instances in $\mathbf{X}_n$. For notational simplicity, we write $\mathbf{X}_n = (\mathbf{X}_n^0, \mathbf{X}_n^1)$ where it is understood that the embedding of $\mathbf{X}_n^0$ and $\mathbf{X}_n^1$ in $\mathbf{X}_n$ is in accordance with $\mathbf{Y}_n$. Note that $\mathbf{Y}_n \sim (\text{Bernoulli}(\frac{2}{3}))^n$ irrespectively of μ and ϕ. In addition, given $\mathbf{Y}_n$ we have that $\mathbf{X}_n^0$ and $\mathbf{X}_n^1$ are independent and $\mathbf{X}_n^0|\mathbf{Y}_n \sim \mu^{n_0}$ and $\mathbf{X}_n^1|\mathbf{Y}_n \sim \phi^{n_1}$. We decompose

$$\begin{aligned} Z_n^\phi &= \mathbb{E}_{S_n\sim(\lambda_\phi)^n}\big[\mathbb{E}_{X\sim\mu}\big[\hat{h}_{S_n}(X)\big]\big] \\ &= \mathbb{E}_{\mathbf{Y}_n}\mathbb{E}_{\mathbf{X}_n|\mathbf{Y}_n}\big[\mathbb{E}_{X\sim\mu}\big[\hat{h}_{(\mathbf{X}_n,\mathbf{Y}_n)}(X)\big]\big] \\ &= \mathbb{E}_{\mathbf{Y}_n}\mathbb{E}_{\mathbf{X}_n^1\sim\phi^{n_1}}\mathbb{E}_{\mathbf{X}_n^0\sim\mu^{n_0}}\big[\mathbb{E}_{X\sim\mu}\big[\hat{h}_{((\mathbf{X}_n^0,\mathbf{X}_n^1),\mathbf{Y}_n)}(X)\big]\big]. \end{aligned} \tag{4.6}$$

Toward applying Lemma 4.14, we first need to translate our reasoning about a random vector $\mathbf{X} = (X_1,\dots,X_k) \sim \phi^k$ with $k \in \mathbb{N}$ into reasoning about the random set of its distinct elements, $W_{\mathbf{X}} := \bigcup_{i=1}^k \{X_i\}$. Since ϕ vanishes on singletons, all instances in $\mathbf{X}$ are distinct with probability one,

$$\mathbb{P}_{\mathbf{X}\sim\phi^k}\big[|W_{\mathbf{X}}| = k\big] = 1. \tag{4.7}$$

Fixing an ordering on $\mathcal{X}$, for any finite set $W = \{w_1,\dots,w_k\} \in [\mathcal{X}]^k$, denote by $\Pi(W)$ the distribution over vectors $\mathbf{X}' = (w_{\pi(1)},\dots,w_{\pi(k)}) \in W^k$ as induced by a random permutation π of the instances in W. Then, by (4.7) and the fact that ϕ^k is a product measure, we have that for any measurable function $f : \mathcal{X}^k \to [0,1]$ the following symmetrization holds:

$$\mathbb{E}_{\mathbf{X}\sim\phi^k}\big[f(\mathbf{X})\big] = \mathbb{E}_{\mathbf{X}\sim\phi^k}\big[\mathbb{E}_{\mathbf{X}'\sim\Pi(W_{\mathbf{X}})}\big[f(\mathbf{X}')\big] \,|\, |W_{\mathbf{X}}| = k\big]. \tag{4.8}$$

For every $\mathbf{Y}_n \in \{0,1\}^n$ define $F_{\mathbf{Y}_n} : [\mathcal{X}]^{n_1} \to \mathbb{R}$ by

$$F_{\mathbf{Y}_n}(W) = \mathbb{E}_{\mathbf{X}^1\sim\Pi(W)}\mathbb{E}_{\mathbf{X}_n^0\sim\mu^{n_0}}\big[\mathbb{E}_{X\sim\mu}\big[\hat{h}_{((\mathbf{X}_n^0,\mathbf{X}^1),\mathbf{Y}_n)}(X)\big]\big], \quad W \in [\mathcal{X}]^{n_1}.$$

In the case of randomized Alg, we also include an innermost conditional expectation over the value of $\hat{h}_{((\mathbf{X}_n^0,\mathbf{X}^1),\mathbf{Y}_n)}(X)$ given $\mathbf{X}_n^0$, $\mathbf{X}^1$, $\mathbf{Y}_n$, X. Putting this in (4.6) while using (4.7) and (4.8),

$$Z_n^\phi = \mathbb{E}_{\mathbf{Y}_n}\mathbb{E}_{\mathbf{X}\sim\phi^{n_1}}\big[F_{\mathbf{Y}_n}(W_{\mathbf{X}}) \,|\, |W_{\mathbf{X}}| = n_1\big].$$

By the choice of μ, Lemma 4.14 implies there exist $C_{\mathbf{Y}_n} \in \mathbb{R}$ and $U_{\mathbf{Y}_n} \subseteq \mathcal{X}$ with $\mu(U_{\mathbf{Y}_n}) = 1$ such that $U_{\mathbf{Y}_n}$ is homogeneous for $F_{\mathbf{Y}_n}$, namely, $F_{\mathbf{Y}_n}(W) = C_{\mathbf{Y}_n}$, $\forall W \in [U_{\mathbf{Y}_n}]^{n_1}$. Let

$$U = \bigcap_{n\in\mathbb{N}} \bigcap_{\mathbf{Y}_n\in\{0,1\}^n} U_{\mathbf{Y}_n}. \tag{4.9}$$

Then U is simultaneously homogeneous for all $\{F_{\mathbf{Y}_n}\}$,

$$F_{\mathbf{Y}_n}(W) = C_{\mathbf{Y}_n} \quad \forall n \in \mathbb{N}, \forall \mathbf{Y}_n \in \{0,1\}^n, \forall W \in [U]^{n_1}. \tag{4.10}$$

In addition, by Lemma C.1 in the Supplementary Material [24], $\mu(U) = 1$.

We are now in position to choose $\nu := \nu(\mu, \mathsf{Alg})$. By Lemma 4.4, we may split U in (4.9) into two disjoint sets B and $U \setminus B$ such that $|B| = |U \setminus B| = |\mathcal{X}|$. Since μ is two-valued, we may assume without loss of generality that $\mu(B) = 0$ (so $\mu(U \setminus B) = 1$). Since $|B|$ is a two-valued measurable cardinal, there exists a two-valued witnessing measure ν' on $(B, 2^B)$ with $\nu'(B) = 1$. Extend ν' to a measure ν over all $\mathscr{B}$ by $\nu(A) = \nu'(A \cap B)$, $\forall A \subseteq \mathcal{X}$. Then $\nu \neq \mu$ and $\nu(U) = \mu(U) = 1$. By the last equality, for $\phi \in \{\mu, \nu\}$ and $\forall k \in \mathbb{N}$, $\Pr_{\mathbf{X}\sim\phi^k}[W_{\mathbf{X}} \in [U]^k \mid |W_{\mathbf{X}}| = k] = 1$. So, for $\phi \in \{\mu, \nu\}$,

$$\begin{aligned} Z_n^\phi &= \mathbb{E}_{\mathbf{Y}_n}\mathbb{E}_{\mathbf{X}\sim\phi^{n_1}}\big[F_{\mathbf{Y}_n}(W_{\mathbf{X}})||W_{\mathbf{X}}| = n_1\big] \\ &= \mathbb{E}_{\mathbf{Y}_n}\mathbb{E}_{\mathbf{X}\sim\phi^{n_1}}\big[F_{\mathbf{Y}_n}(W_{\mathbf{X}})||W_{\mathbf{X}}| = n_1 \wedge W_{\mathbf{X}} \in [U]^{n_1}\big] \\ &= \mathbb{E}_{\mathbf{Y}_n}\mathbb{E}_{\mathbf{X}\sim\phi^{n_1}}\big[C_{\mathbf{Y}_n}||W_{\mathbf{X}}| = n_1 \wedge W_{\mathbf{X}} \in [U]^{n_1}\big] \\ &= \mathbb{E}_{\mathbf{Y}_n}[C_{\mathbf{Y}_n}], \end{aligned}$$

where we used (4.10) and the fact that $C_{\mathbf{Y}_n}$ does not depend on $\mathbf{X}$. Since $\mathbb{E}_{\mathbf{Y}_n}[C_{\mathbf{Y}_n}]$ is independent of ϕ, we conclude that $Z_n^\mu = Z_n^\nu$ for all $n \in \mathbb{N}$. However by (4.5), for Alg to be Bayes consistent on λ_μ and λ_ν we must have $Z_n^\mu \xrightarrow[n\to\infty]{} 1$ and $Z_n^\nu \xrightarrow[n\to\infty]{} 0$. Thus Alg cannot be Bayes consistent on both λ_μ and λ_ν. In particular, (4.3) holds with $\varepsilon = 1/4$.

5. Discussion. We have exhibited a computationally efficient multiclass learning algorithm, OptiNet, that is universally strongly Bayes consistent (UBC) in all essentially separable (ES) metric spaces. In contrast, we showed that in non-ES spaces, no algorithm can be UBC. As such, OptiNet is optimistically universal (in the terminology of [22])—it is universally Bayes consistent in all metric spaces that admit such a learner. We note that in this work, we do not study the rates of decay of the excess risk, leaving this challenging open problem for future study.

By definition, any separable metric space is ES. As discussed in Section 1, consistency of NN-type algorithms in general separable metric spaces was studied in [1, 3, 4, 6, 13, 33]. In particular, in [1, 6, 13], a characterization of the metric spaces in which an algorithm is universally Bayes consistent was given for several such algorithms, in terms of Besicovitch-type conditions. As a notable example, it is shown in [6] that for any separable metric space $\mathcal{X}$, a sufficient condition for the k-NN algorithm (with an appropriate choice of the number of neighbors k) to be Bayes consistent for a distribution $\bar{\mu}$ over $\mathcal{X} \times \{0,1\}$ is that for all $\varepsilon > 0$,

$$\lim_{r\to 0^+} \mathbb{P}\left\{\frac{1}{\mu(B_r(X))}\int_{B_r(X)} |\eta(z) - \eta(X)|\,\mathrm{d}\mu(z) > \varepsilon\right\} = 0, \tag{5.1}$$

where μ is the marginal of $\bar{\mu}$ over $\mathcal{X}$ and $\eta(x) := \mathbb{P}(Y = 1 \mid X = x)$. It is also shown in [6] that in the realizable case, where $\eta(x) \in \{0,1\}$ for all $x \in X$, a violation of (5.1) implies that k-NN is inconsistent. Say that a metric space satisfies the *universal Besicovitch condition* if (5.1) holds *for all* measures $\bar{\mu}$ over the Borel σ-algebra. By Besicovitch's density

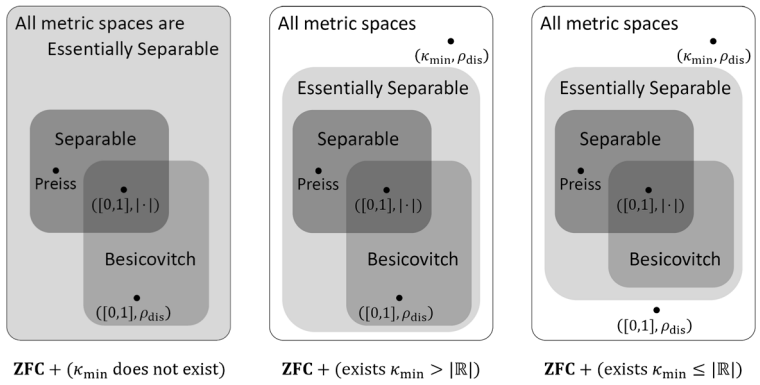

FIG. 1. *The classes of metric spaces discussed in this paper and their inclusion relationships in the three cases where*: *no RVMC exists* (*left*); *minimal RVMC is* $\kappa_{\min} > \mathfrak{c}$ (*middle*); *and minimal RVMC is* $\kappa_{\min} \leq \mathfrak{c}$ (*right*). *All three cases are believed to be valid extensions of ZFC. The metric space* $(\kappa_{\min}, \rho_{dis})$ *corresponds to a discrete one of cardinality* $\kappa_{\min}$; *it is not ES but any discrete metric space of cardinality* $< \kappa_{\min}$ *is ES. The shaded area named "Besicovitch" and the specific metric space "Preiss" are as discussed in the text.*

theorem [15], Section 472, the metric space $(\mathbb{R}^d, \|\cdot\|_2)$—and more generally, any finite-dimensional normed space—satisfies this condition, so k-NN is UBC on such spaces. In contrast, in infinite-dimensional separable spaces, such as ℓ_2, a violation of (5.1) can occur [36, 37, 40]. One such example is the separable metric probability space studied in [28], building upon a construction of Preiss [36]. While the k-NN algorithm is provably not UBC in this space, OptiNet is. As far as we know, OptiNet is the first algorithm known to be UBC (weakly or strongly) in any separable metric space.

As discussed in Section 4.1, the essential separability of nonseparable metric spaces is believed to depend on set-theoretic axioms that are independent of ZFC, and in particular on the cardinality of the minimal RVMC, $\kappa_{\min}$: a metric space is non-ES if and only if it contains a discrete subset of cardinality $\kappa_{\min}$. Figure 1 gives a pictorial illustration of the possible relationships between the following types of metric spaces: separable, (uniform) Besicovitch, ES, and all spaces, depending on the set-theoretic model. If one adopts a model in which no RVMC exist, then any discrete subspace of a metric space admits only trivial, purely-atomic measures. In this case, abbreviated as ZFC + ($\kappa_{\min}$ does not exist) in the left panel of Figure 1, all metric spaces are ES, and OptiNet is UBC on any metric space. Alternatively, if one adopts a set-theoretic model in which an RVMC exists, then discrete subspaces of $\mathcal{X}$ of cardinality $\geq \kappa_{\min}$ admit also nontrivial measures. As shown in Section 4.2, such measures exclude the possibility of a UBC algorithm. The nature of the nontrivial measures, being atomless or purely atomic, depends on whether $\kappa_{\min} > \mathfrak{c}$ or $\kappa_{\min} \leq \mathfrak{c}$, which are illustrated on the middle and right panels of Figure 1, respectively.

Lastly, we note that our argument for the impossibility of UBC in non-ES metric spaces is based solely on the real-valued measurability of the cardinality of discrete subspaces of $\mathcal{X}$. This raises a natural question: Assuming no cardinal is real-valued measurable, are there any topological spaces (which by the results above must be nonmetric) in which no UBC algorithm exists?

To summarize, in this work we provided the first multiclass learning algorithm that is universally Bayes consistent in any metric space where such an algorithm exists. Moreover, we provided a characterization of these metric spaces. The study of learnability in general spaces is fundamental, and provides many open questions for future research.

Acknowledgments. We thank Vladimir Pestov for sharing with us his proof of the existence of a measurable total order. We also thank Robert Furber, Iosif Pinelis, Menachem Kojman, and Roberto Colomboni for helpful discussions.

Funding. Aryeh Kontorovich was supported in part by the Israel Science Foundation (Grant No. 755/15), Paypal and IBM. Sivan Sabato was supported in part by the Israel Science Foundation Grant No. 555/15.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/20-AOS2029SUPP; .pdf). In the Supplementary Material file, we provide the proofs of Lemmas 3.6, 3.7, 3.8, 4.13, 4.14, 4.15, as well as the proofs of the auxiliary Lemmas A.1, B.1, C.1, C.2, D.1.

REFERENCES

[1] ABRAHAM, C., BIAU, G. and CADRE, B. (2006). On the kernel rule for function classification. *Ann. Inst. Statist. Math.* **58** 619–633. MR2327897 https://doi.org/10.1007/s10463-006-0032-1

[2] BEN-DAVID, S., HRUBES, P., MORAN, S., SHPILKA, A. and YEHUDAYOFF, A. (2019). Learnability can be undecidable. *Nat. Mach. Intell.* **1** 44–48. https://doi.org/10.1038/s42256-018-0002-3

[3] BIAU, G., BUNEA, F. and WEGKAMP, M. H. (2005). Functional classification in Hilbert spaces. *IEEE Trans. Inf. Theory* **51** 2163–2172. MR2235289 https://doi.org/10.1109/TIT.2005.847705

[4] BIAU, G., CÉROU, F. and GUYADER, A. (2010). Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Trans. Inf. Theory* **56** 2034–2040. MR2654492 https://doi.org/10.1109/TIT.2010.2040857

[5] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*, 2nd ed. *Wiley Series in Probability and Statistics*: *Probability and Statistics*. Wiley, New York. A Wiley-Interscience Publication. MR1700749 https://doi.org/10.1002/9780470316962

[6] CÉROU, F. and GUYADER, A. (2006). Nearest neighbor classification in infinite dimension. *ESAIM Probab. Stat.* **10** 340–355. MR2247925 https://doi.org/10.1051/ps:2006014

[7] CHAUDHURI, K. and DASGUPTA, S. (2014). Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems* 3437–3445.

[8] CHRISTMANN, A. and STEINWART, I. (2010). Universal kernels on non-standard input spaces. In *Advances in Neural Information Processing Systems* 23: 24*th Annual Conference on Neural Information Processing Systems* 2010 406–414.

[9] COVER, T. M. and HART, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13** 21–27.

[10] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5

[11] DUDLEY, R. M. (1999). *Uniform Central Limit Theorems*. *Cambridge Studies in Advanced Mathematics* **63**. Cambridge Univ. Press, Cambridge. MR1720712 https://doi.org/10.1017/CBO9780511665622

[12] FIX, E. and HODGES J. L., J. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Int. Stat. Rev.* **57** 238–247.

[13] FORZANI, L., FRAIMAN, R. and LLOP, P. (2012). Consistent nonparametric regression for functional data under the Stone–Besicovitch conditions. *IEEE Trans. Inf. Theory* **58** 6697–6708. MR2991802 https://doi.org/10.1109/TIT.2012.2209628

[14] FREMLIN, D. H. (1993). Real-valued-measurable cardinals. In *Set Theory of the Reals* (*Ramat Gan*, 1991). *Israel Math. Conf. Proc.* **6** 151–304. Bar-Ilan Univ., Ramat Gan. MR1234282

[15] FREMLIN, D. H. (2000). *Measure Theory* **1–5**. Torres Fremlin, Colchester.

[16] GITIK, M. and SHELAH, S. (1989). Forcings with ideals and simple forcing notions. *Israel J. Math.* **68** 129–160. MR1035887 https://doi.org/10.1007/BF02772658

[17] GOTTLIEB, L.-A., KONTOROVICH, A. and KRAUTHGAMER, R. (2014). Efficient classification for metric data. *IEEE Trans. Inf. Theory* **60** 5750–5759. MR3252418 https://doi.org/10.1109/TIT.2014.2339840

[18] GOTTLIEB, L.-A., KONTOROVICH, A. and NISNEVITCH, P. (2017). Nearly optimal classification for semimetrics. *J. Mach. Learn. Res.* **18** Paper No. 37, 22. MR3646632

[19] GOTTLIEB, L.-A., KONTOROVICH, A. and NISNEVITCH, P. (2018). Near-optimal sample compression for nearest neighbors. *IEEE Trans. Inf. Theory* **64** 4120–4128. MR3809730 https://doi.org/10.1109/TIT.2018.2822267

[20] GRAEPEL, T., HERBRICH, R. and SHAWE-TAYLOR, J. (2005). PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification. *Mach. Learn.* **59** 55–76.

[21] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. *Springer Series in Statistics*. Springer, New York. MR1920390 https://doi.org/10.1007/b97848

[22] HANNEKE, S. (2017). Learning whenever learning is possible: Universal learning under general stochastic processes. Preprint. Available at arXiv:1706.01418.

[23] HANNEKE, S. and KONTOROVICH, A. (2019). A sharp lower bound for agnostic learning with sample compression schemes. In *Algorithmic Learning Theory* 2019. *Proc. Mach. Learn. Res.* (*PMLR*) **98** 489–505. MR3932856

[24] HANNEKE, S., KONTOROVICH, A., SABATO, S. and WEISS, R. (2021). Supplement to "Universal Bayes consistency in metric spaces." https://doi.org/10.1214/20-AOS2029SUPP

[25] HRBACEK, K. and JECH, T. (1999). *Introduction to Set Theory*, 3rd ed. *Monographs and Textbooks in Pure and Applied Mathematics* **220**. Dekker, New York. MR1697766

[26] JECH, T. (2003). *Set Theory*. *Springer Monographs in Mathematics*. Springer, Berlin. The third millennium edition, revised and expanded. MR1940513

[27] KONTOROVICH, A., SABATO, S. and URNER, R. (2017). Active nearest-neighbor learning in metric spaces. *J. Mach. Learn. Res.* **18** Paper No. 195, 38. MR3827083

[28] KONTOROVICH, A., SABATO, S. and WEISS, R. (2017). Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. In *Advances in Neural Information Processing Systems* 1573–1583.

[29] KONTOROVICH, A., SABATO, S. and WEISS, R. (2017). Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. Preprint. Available at arXiv:1705.08184.

[30] KONTOROVICH, A. and WEISS, R. (2014). Maximum margin multiclass nearest neighbors. In *International Conference on Machine Learning* (*ICML* 2014).

[31] KONTOROVICH, A. and WEISS, R. (2014). A Bayes consistent 1-NN classifier. In *Artificial Intelligence and Statistics* (*AISTATS* 2015).

[32] KPOTUFE, S. and VERMA, N. (2017). Time-accuracy tradeoffs in kernel prediction: Controlling prediction quality. *J. Mach. Learn. Res.* **18** Paper No. 44, 29. MR3655309

[33] KULKARNI, S. R. and POSNER, S. E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inf. Theory* **41** 1028–1039. MR1366756 https://doi.org/10.1109/18.391248

[34] LITTLESTONE, N. and WARMUTH, M. K. (1986). Relating data compression and learnability. Unpublished.

[35] PELILLO, M. (2014). Alhazen and the nearest neighbor rule. *Pattern Recogn. Lett.* **38** 34–37. https://doi.org/10.1016/j.patrec.2013.10.022

[36] PREISS, D. (1979). Invalid Vitali theorems. In *Abstracta. 7th Winter School on Abstract Analysis* 58–60.

[37] PREISS, D. (1981). Gaussian measures and the density theorem. *Comment. Math. Univ. Carolin.* **22** 181–193. MR0609946

[38] SHALEV-SHWARTZ, S. and BEN-DAVID, S. (2014). *Understanding Machine Learning*: *From Theory to Algorithms*. Cambridge Univ. Press, Cambridge.

[39] STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645. MR0443204

[40] TIŠER, J. (2003). Vitali covering theorem in Hilbert space. *Trans. Amer. Math. Soc.* **355** 3277–3289. MR1974687 https://doi.org/10.1090/S0002-9947-03-03296-3

[41] ULAM, S. M. (1930). *Zur Masstheorie in der Allgemeinen Mengenlehre*. Uniwersytet, seminarjum matematyczne.

[42] ZHAO, L. C. (1987). Exponential bounds of mean error for the nearest neighbor estimates of regression functions. *J. Multivariate Anal.* **21** 168–178. MR0877849 https://doi.org/10.1016/0047-259X(87)90105-9