CONVERGENCE OF NEAREST NEIGHBOR SELECTIVE CLASSIFICATION

BY AKASH KUMAR[1,a], SANJOY DASGUPTA[2,b]

[1]*Department of Computer Science and Engineering, UCSD,* [a]*akk002@ucsd.edu*

[2]*Department of Computer Science and Engineering, UCSD,* [b]*dasgupta@eng.ucsd.edu*

An elementary approach to *selective classification* (also known as *classification with a reject option*) is the (k, k')-rule: given a query x, find its k nearest neighbors in the training set and if at least k' of them have the same label, then predict that label; otherwise, abstain. We study this method under minimal assumptions to understand its convergence properties and its tradeoffs between error and abstention rate.

1. Introduction. In safety-critical applications of machine learning such as medical diagnosis, and human-facing applications such as automated assistants, the cost of a wrong judgment can be extremely high. In such scenarios, it makes sense for a predictive model to abstain or defer to an expert on instances on which it is not confident. Allowing a classifier to abstain is called *selective classification* or *classification with a reject option*.

There is an extensive machine learning literature on selective classification, mostly having to do with methods for adding a reject option to standard classifiers such as support vector machines. The goal is to reduce the error rate while keeping the coverage over the instance space as high as possible. This tradeoff, between error and coverage, has been studied in some theoretical work, starting with the formulation of [9, 8] and spreading to a variety of settings such as agnostic learning [41, 38], SVM [17, 22], deep learning [19, 20, 21], and online learning [10, 18]. The focus of our work is selective classification using *nearest neighbor classifiers*.

Among learning rules, nearest neighbor (NN) classifiers are perhaps the simplest non-parametric methods. For a given $k > 0$ and a training set of size n, the k-NN classifier assigns a label according to the majority vote of the k nearest training samples. The asymptotic consistency of these classifiers in standard (no-rejection) learning is fairly well understood [16, 7, 14], and it is of interest to broaden these results to other tasks. The convergence results are primarily in the statistical learning framework, in which all data (past and future) are assumed to come from some unknown underlying distribution P on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the label space. We will take $\mathcal{X}$ to be a separable metric space with distance metric $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and marginal distribution μ. The nearest neighbor literature typically takes the label space $\mathcal{Y}$ to be binary, although we will allow it to be any countable set. The conditional probability distribution of labels given instances is specified by the collection of functions $\eta := \{\eta_i : \mathcal{X} \to [0,1] \,|\, \eta_i(x) = P(Y = i \,|\, X = x)\}_{i \in \mathcal{Y}}$.

For a given $x \in \mathcal{X}$, the most likely (Bayes-optimal) label is $\ell(x) := \arg\max_{i \in \mathcal{Y}} \eta_i(x)$, breaking ties arbitrarily (say, according to a fixed ordering on $\mathcal{Y}$). The asymptotic goal of selective classification [8, 26, 4] is to make this prediction if it is likely—that is, if it has probability greater than some specified threshold—and to abstain otherwise. For *rejection* threshold $\tau > 0$, the target abstaining classifier is the function $g^*_\tau : \mathcal{X} \to \mathcal{Y} \cup \{?\}$ given by

$$g^*_\tau(x) = \begin{cases} \ell(x) & \text{if } \eta_{\ell(x)}(x) \geq \tau \\ ? & \text{otherwise} \end{cases} \tag{1}$$

where "?" is abstention. We can think of this as the Bayes-optimal rule with rejection threshold τ.

Given: training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, rejection threshold: $0 < \tau \leq 1$
To predict at $x \in \mathcal{X}$:
1. Find the smallest ball $B_k(x)$ centered at x containing exactly k training points[1]
2. **If** there exists label $j \in \mathcal{Y}$ such that $|\{i : x_i \in B_k(x), y_i = j\}| \geq k'$:
   return label j
3. **Else**
   return "I don't know" or ?

Algorithm 1: Selective (k, k')-nearest neighbor rule

In this work, we study the selective (k, k')-nearest neighbor rule. Given a training set of size n, this rule $g_n : \mathcal{X} \to \mathcal{Y} \cup \{?\}$ answers any query $x \in \mathcal{X}$ as follows: it finds the k nearest neighbors of x in the training set, and if at least k' of them have the same label, it predicts that label; otherwise it abstains (see Algorithm 1). To achieve a rejection threshold of τ, it is natural to take $k' = \lceil \tau k \rceil$. We will assume $\tau > 1/2$ and thus $k' > k/2$.

We are interested in whether this (k, k')-rule (for k growing as a suitable function of n) converges to g^*_τ, and if so, the rate at which this convergence occurs and the tradeoffs that are realized between error and coverage. To study these theoretical questions, we define three risk functionals. Informally, these capture the probability of: predicting the wrong label; abstaining when we shouldn't; and predicting when we should abstain. More concretely,

$$\textbf{(error rate)}\ \mathcal{R}_e(g) := \Pr_X [g(X) \notin \{?, \ell(X)\}] \tag{2}$$

$$\textbf{(type 1 abstention rate)}\ \mathcal{R}^1_a(g) := \Pr_X \left[\eta_{\ell(X)}(X) \geq \tau,\ g(X) = ?\right] \tag{3}$$

$$\textbf{(type 2 abstention rate)}\ \mathcal{R}^2_a(g) := \Pr_X \left[\eta_{\ell(X)}(X) < \tau,\ g(X) \neq ?\right] \tag{4}$$

These functionals are zero for the target model g^*_τ, and thus $\mathcal{R}_e(g_n)$, $\mathcal{R}^1_a(g_n)$ and $\mathcal{R}^2_a(g_n)$ capture different aspects of the relative performance of g_n to g^*_τ. Our main results and contributions are summarized below:

I. *Two simple NN-based selective classifiers*. We analyze both (a) a fixed-parameter (k, k')-NN rule with rejection threshold τ (taking $k' = \lceil \tau k \rceil$), and (b) an adaptive NN rule that searches over k with calibrated confidence bands, specialized to selective classification (see Section 5).

II. *Universal consistency under weak assumptions*. For the (k, k') rule, we establish almost-sure convergence of the sum $\mathcal{R}_e + \mathcal{R}^1_a + \mathcal{R}^2_a$ to zero in any metric measure space satisfying the Lebesgue differentiation condition, allowing countably many labels (Theorem 3.2, Theorem 3.3). For the adaptive rule, we prove an analogous universal consistency result under the same condition (Theorem 5.2).

III. *Distribution-dependent finite-sample bounds*. We give high-probability, finite-sample bounds for error ($\mathcal{R}_e$), type-1 abstention ($\mathcal{R}^1_a$), and type-2 abstention ($\mathcal{R}^2_a$). For the (k, k') rule these are stated via interior sets defined by mass p and margin Δ (Theorem 4.1), with a VC-free variant (Theorem 4.2). For the adaptive rule, analogous bounds are given in terms of pointwise τ-safeness and τ-saliency (Theorem 5.4).

IV. *Smoothness and margin rates*. Under (α, L)-smoothness of η and a β-margin condition around τ, we obtain rates for all three functionals. For the (k, k') rule, we show that the classification error rate $\mathcal{R}_e = 0$ is achieved for sufficiently large samples, while the

[1] Ties can be broken randomly.

abstention rates $\mathcal{R}_a^1$ and $\mathcal{R}_a^2$ decrease at explicit rates (Theorem 3.4, Corollary 1). For the adaptive rule, we derive rates upper bounded by $(c\log(n/\delta)/n)^{\frac{\alpha\beta}{2\alpha+1}}$ (Theorem 5.3).

V. Lower bounds and countable labels. We prove a general lower bound showing $\mathbb{E}_n\left[\mathcal{R}_a^1 + \mathcal{R}_a^2\right] \gtrsim \mu(\zeta_n)$ (Theorem 4.3), clarifying tightness of our abstention rates.

1.1. *Background and Related Work.*

1.1.0.1. Selective classification. Although there is a substantial machine learning literature on selective classification, only a small portion of it is devoted to rigorous theoretical results.

The early work of [9, 8] investigated the notion of classification with a reject option, and used Bayesian decision theory to derive the ideal estimator g_τ^* mentioned above. Shortly thereafter, [25] showed consistency of the nearest neighbor-based (k, k') rule under continuity assumptions, in the case of binary classification.

In the framework of statistical learning theory, [26] introduced a single risk functional that captures both error and abstention, and gave rates of convergence of this functional for plug-in estimators (of η), under smoothness and margin conditions. They also gave rates for empirical minimizers of the risk. This ERM problem is typically computationally intractable, so the subsequent work of [4] introduced a convex surrogate loss (a hinge loss with two "hinges") and gave rates of convergence as well as calibration results. Other related works include [22, 11].

A PAC-like framework for selective classification was introduced by [15] for the realizable setting—where g_τ^* can be captured perfectly by the hypothesis class under consideration—and was later extended to the more general agnostic setting [41]. These papers are particularly interested in "perfect learners" (originally introduced in [35]) that achieve zero error at the expense of a certain amount of abstention.

This is also a requirement in the KWIK ("knows what it knows") framework of [30], which can be thought of as a variant of the mistake-bound model of online learning. In the KWIK model, data is chosen adversarially and appears one point at a time, indefinitely. At each time t, a new point x_t arrives, the learner either makes a prediction or abstains, and in the latter case, the label y_t is revealed. The key constraints are that the learner is never allowed to make a mistake and can abstain only finitely often. Related models of online learning with abstention appear in [36, 43, 32].

These are some representative results from the theory of selective classification. Other papers of interest include [27, 42, 29, 12, 20, 34, 21, 33, 5, 19].

Our work focuses on the nearest neighbor (k, k')-rule. We show universal consistency—that is, consistency without any conditions on the conditional probability function—in metric measure spaces and countable label spaces. Moreover, rather than combining error and abstention into a single functional, we obtain separate rates of convergence for three functionals that capture different aspects of selective classification.

1.1.0.2. Nearest neighbor classification. In the standard (no-rejection) setting of classifier learning, the consistency of nearest neighbor is well-studied [16, 37, 14, 13, 7, 3] and universal consistency is known to hold in a variety of metric measure spaces. Of course, there are no distribution-free rates of convergence in nonparametric estimation, but rates are available under smoothness (typically Holder conditions on η) and "Tsybakov margin" conditions; see [28, 23, 39, 31, 7]. Moreover, [7] also obtain general rates of convergence that do not require smoothness (or other) assumptions, and are stated in terms of distribution-specific quantities.

Our work follows the methodology of [7] and adapts it to the setting of selective classification. Along the way, we extend the analysis to allow countably many (rather than binary) labels and to accommodate three separate risk functionals that capture different aspects of error and abstention.

2. Preliminaries. Consider the instance space to be a separable metric space $(\mathcal{X}, \rho)$ and map the label set $\mathcal{Y}$ to the set of naturals. All data are assumed to be drawn i.i.d. from a fixed unknown distribution P over $\mathcal{X} \times \mathcal{Y}$. We consider a training sample S_n of size n drawn from P as

$$S_n := \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$$

Let μ denote the marginal distribution on $\mathcal{X}$: if (X, Y) is a random sample from P then

$$\mu(S) = \Pr(X \in S) \tag{5}$$

for any measurable set $S \subseteq \mathcal{X}$. Similarly, for any measurable set S with $\mu(S) > 0$, the conditional expectation of $Y = i$ given $X \in S$ for $i \in \mathcal{Y}$ is

$$\eta_i(S) = \frac{1}{\mu(S)} \int_S \eta_i(x)\, d\mu(x)$$

We study algorithms whose analysis depend heavily on the probability masses and biases of balls in $\mathcal{X}$. For $x \in \mathcal{X}$ and $r \geq 0$, let $B(x, r)$ denote the closed ball of radius r centered at x,

$$B(x, r) := \left\{ x' \in \mathcal{X} \,\middle|\, \rho(x, x') \leq r \right\}$$

For any $0 \leq p \leq 1$, define $r_p(x)$ to be the smallest radius r such that $B(x, r)$ has probability mass at least p,

$$r_p(x) = \inf \{r \geq 0 : \mu(B(x, r)) \geq p\}.$$

It follows that $\mu(B(x, r_p(x))) \geq p$.

2.1. *Empirical estimates.* For any set $S \subseteq \mathcal{X}$, we define its empirical count and probability mass corresponding to training set $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ as

$$\#_n(S) = |\{i : X_i \in S\}|$$

$$\mu_n(S) = \frac{\#_n(S)}{n}$$

Define the empirical bias of set S with respect to to any label $i \in \mathcal{Y}$ as

$$\hat{\eta}_i(S) = \frac{\sum_{j : X_j \in S} 1\{Y_j = i\}}{\#_n(S)}$$

2.2. *Nearest neighbor classifier with a reject option.* In our work, we consider a (k, k')-nearest neighbor rule with a reject option for selective classification. For a given training sample S_n and the fixed scalar $k \leq n$, a (k, k')-nearest neighbor rule, upon given a test point x, picks k points in S_n nearest to x wrt to the distance metric ρ, and assigns a label if a majority vote over one of the labels is at least $\frac{k'}{k}$; otherwise abstains. We denote such a classifier as $g_{k,n}$, interchangeably used as g_n. More formally,

$$g_{k,n}(x) = \begin{cases} i & \text{if } \max_j \hat{\eta}_j(B_k(x)) \geq \frac{k'}{k} \\ ? & \text{o.w.} \end{cases}$$

where $B_k(x)$ is the ball around x with k nearest neighbors in S_n. More recently, [3] studied adaptive nearest neighbor classifiers, denoted as g_n, that adapts to the choice of the parameter k by increasing radius of the balls $B_k(x)$. Similar to [3], we use a statistical confidence parameter δ to bound failure probability of the classifier.

For a given threshold $\tau > 0$, we consider an algorithm that induces a (k, k')- nearest neighbor rule with threshold $k' = \tau \cdot k$ (see Section 3).

3. Universal consistency and margin bounds of selective (k, k')-nearest neighbor rule. In this section, we study the convergence properties of the (k, k')-nearest neighbor rule of Algorithm 1 for the three risk functionals as discussed in Eq. (2)-(4). In particular, we show universal consistency under the mild assumption of Lebesgue differentiation condition in Theorem 3.2 (see Section 3.1). Furthermore, we provide rate of convergence for these risk functionals when the conditional probability distribution satisfies a Holder-type smoothness condition in Theorem 3.4 and Corollary 1 (see Section 3.2).

3.1. *Universal consistency under Lebesgue differentiation condition.* Here, we will study the convergence of the risk functionals to zero as the training size n grows. This property of an estimator is termed as *universal consistency* in a metric measure space $(\mathcal{X}, \rho, \mu)$ if it has the limiting behaviour for any conditional probability distribution η. In our work, we show universal consistency of the (k, k')-nearest neighbor rule of Algorithm 1 if the metric measure space $(\mathcal{X}, \rho, \mu)$ satisfies *Lebesgue differentiation condition*.

First, we provide a definition for Lebesgue differentiation condition for any metric space $(\mathcal{X}, \rho, \mu)$:

DEFINITION 3.1 (Lebesgue differentiation condition). We say a metric measure space $(\mathcal{X}, \rho, \mu)$ satisfies lebesgue differentiation condition if for any bounded measurable $f : \mathcal{X} \to \mathbb{R}$,

$$\lim_{r \downarrow 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} f(x) d\mu = f(x)$$

for almost all (μ-a.e.) $x \in \mathcal{X}$.

A more detailed discussion of the condition is given in [1, 24]. Lebesgue differentiability is known to hold for a variety of settings, e.g in any finite-dimensional normed space or any doubling metric space [24, Chapter 1] or μ is an atomic measure on $\mathcal{X}$.

In the standard learning setting (no-rejection), universal consistency for different versions of nearest neighbor classifiers is well-established in the binary classification setting. [7] achieves consistency of a $(k, \frac{1}{2}k)$-nearest neighbor classifier (for $k/n \to 0$, $k/\log n \to \infty$), whereas [3] showed consistency of an adaptive nearest neighbor classifier for a sequence of confidence parameter (δ_n) such that $\sum_n \delta_n < \infty$ and $\frac{n}{\log(1/\delta_n)} \to \infty$ under Lebesgue continuity. In our work, we establish consistency of the (k, k')-nearest neighbor rule of Algorithm 1 in the infinite label setting.

In the following, we state our result on universal consistency of the (k, k')-nearest neighbor classifiers for selective classification. The proof details are deferred to Appendix C.

THEOREM 3.2 (Universal consistency: Finite VC dimension). *Suppose that the set of balls in $(\mathcal{X}, \rho)$ has finite VC dimension d_0. Assume parameter $1 \geq \tau > \frac{1}{2}$. Let (δ_n) be a sequence in $[0, 1]$ such that $\sum_n \delta_n < \infty$. Pick a sequence of positive integer (k_n) such that (1) $\lim_{n\uparrow\infty} \frac{k_n}{n} = 0$ and (2) $\lim_{n\uparrow\infty} \frac{k_n}{\log(n/\delta_n)} = \infty$. Let the classifier $g_n : \mathcal{X} \to \mathcal{Y} \cup \{?\}$ be the result of applying the $(k_n, \tau k_n)$-nearest neighbor rule of Algorithm 1 with n points chosen i.i.d. from P. Then, the sum $\mathcal{R}_e(g_n) + \mathcal{R}_a^1(g_n) + \mathcal{R}_a^2(g_n) \to 0$ almost surely.*

3.1.0.1. No VC dimension Consistency. We can obtain similar guarantee on consistency even in the case of arbitrary VC dimension of the set of balls in $(\mathcal{X}, \rho)$. We state the result as follows:

THEOREM 3.3 (Universal consistency: General case). *Assume parameter* $1 \geq \tau > \frac{1}{2}$. *Let* (δ_n) *be a sequence in* $[0,1]$ *such that* $\sum_n \delta_n < \infty$. *Pick a sequence of positive integer* (k_n) *such that (1)* $\lim_{n\uparrow\infty} \frac{k_n}{n} = 0$ *and (2)* $\lim_{n\uparrow\infty} \frac{k_n}{\log(n/\delta_n)} = \infty$. *Let the classifier* $g_n : \mathcal{X} \to \mathcal{Y} \cup \{?\}$ *be the result of applying the* $(k_n, \tau k_n)$*-nearest neighbor rule of Algorithm 1 with* n *points chosen i.i.d. from* P. *Then, the sum* $\mathcal{R}_e(g_n) + \mathcal{R}_a^1(g_n) + \mathcal{R}_a^2(g_n) \to 0$ *almost surely.*

3.2. *Margin bounds for smooth distributions.* In the previous subsection, we established the universal consistency for the risk functionals of the (k, k')-nearest neighbor rule of Algorithm 1. As is customary in statistical learning literature, we study *margin* bounds for the rule when the underlying data distribution P satisfies a large margin condition.

In order to establish these bounds, we consider a notion of smoothness with respect to the marginal distribution on instances as discussed in the earlier work of [7] on nearest neighbor classification.

We define this smoothness in the form of a weaker version of Holder's continuity on the conditional distribution η.

DEFINITION 3.1 (weak Holder-continuity). For some $\alpha, L > 0$, we say the conditional probability function η is (α, L)-smooth in metric measure space $(\mathcal{X}, \rho, \mu)$ if for all $x \in \mathcal{X}$, $r > 0$, and any $j \in \mathcal{Y}$,

$$|\eta_j(x) - \eta_j(B(x,r))| \leq L \cdot \mu(B^\circ(x,r))^\alpha \tag{6}$$

In the binary classification setting, [2, 39, 7] studied a margin condition that allows η to be less concentrated around the decision boundary of $1/2$. For our setting, in essence, τ is the decision boundary for prediction or abstention. This has been motivated in the prior work of [26] on selective classification. To capture this notion of margin of the distribution P around the rejection threshold $\tau > 0$, we define for any scalar $\Delta > 0$

$$\mathscr{M}(\Delta) = \mu\left(\left\{x \in \mathcal{X} \,\middle|\, |\eta_{\ell(x)}(x) - \tau| \leq \Delta\right\}\right) \tag{7}$$

Now, we show that under the (α, L)-smoothness of η, we can bound the rate of convergence in terms of $\mathscr{M}(\Delta)$ for a specific choice of Δ. We state the result in Theorem 3.4 with the proof details deferred to Appendix E.

THEOREM 3.4. *Suppose* η *is* (α, L)*-smooth in* $(\mathcal{X}, \rho, \mu)$ *for some* $\alpha, L \geq 0$. *Suppose that the set of balls in* (X, ρ) *has finite VC dimension* d_0. *Assume that* $\tau > \frac{1}{2}$. *Let* $0 < \delta < 1$ *denote the confidence parameter. Pick a positive integer* $k_n \geq c' \log(n/\delta)$ *(for some constant* $c' > 0$*) for training size* n. *Suppose* (k, k')*-nearest neighbor rule (Algorithm 1) with* $k = k_n$ *and* $k' = \tau k_n$ *defines a classifier* g_n *based on* n *training points chosen i.i.d. from* P. *Then, for any choice of training size* $n \geq L^{\frac{1}{\alpha}} \cdot (c_0 d_0 + 1) \log\left(\frac{L^{\frac{1}{\alpha}} \cdot (c_0 d_0 + 1)}{\delta}\right)$, *with probability at least* $(1 - \delta)$ *over the choice of training data, we have*

1. $\mathcal{R}_e(g_n) = 0$
2. $\mathcal{R}_a^1(g_n) \leq \mathscr{M}\left(c_1 \sqrt{\frac{d_0 \log n + \log \frac{1}{\delta}}{k_n}}\right)$
3. $\mathcal{R}_a^2(g_n) \leq \mathscr{M}\left(c_2 \sqrt{\frac{d_0 \log n + \log \frac{1}{\delta}}{k_n}}\right)$

where $c_0, c_1, c_2 > 0$ *are universal constants.*

We can give explicit rate of convergence for the three risk functionals if $\mathscr{M}(\Delta)$ can be bounded in terms of Δ. Similar to [7], we define β-margin condition to bound $\mathscr{M}(\Delta)$ as follows:

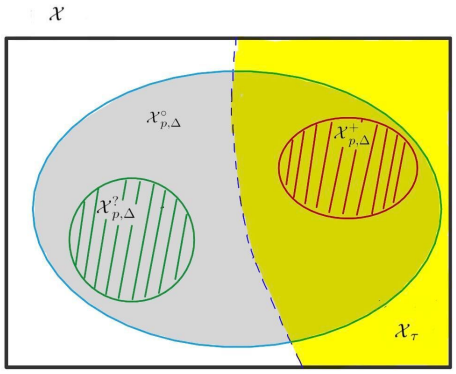

Fig 1: Schematic diagram to represent the set construction.

DEFINITION 3.2 (β-margin). For any $\beta \geq 0$ and threshold $\tau > 0$, we say P satisfies the β-margin condition around τ if there exists a constant $C > 0$ such that

$$\mu\left(\left\{x \in \mathcal{X} \,\middle|\, |\eta_{\ell(x)}(x) - \tau| \leq \sigma\right\}\right) \leq C \cdot \sigma^{\beta}$$

We now obtain margin bounds for the risk functionals of (k, k')-NN rule under (α, L)-smoothness and β-margin conditions. The proof is deferred to Appendix E.

COROLLARY 1. Suppose η is (α, L)-smooth in $(\mathcal{X}, \rho, \mu)$ for some $\alpha, L \geq 0$ and satisfies the β-margin condition with constant $C > 0$. Suppose that the set of balls in (X, ρ) has finite VC dimension d_0. Assume that $\tau > \frac{1}{2}$. Let $0 < \delta < 1$ denote the confidence parameter. Pick a positive integer $k_n \geq c' \log(n/\delta)$ (for some constant $c' > 0$) for training size n. Suppose (k, k')-nearest neighbor rule (Algorithm 1) with $k = k_n$ and $k' = \tau k_n$ defines a classifier g_n based on n training points chosen i.i.d. from P. Then, for any choice of training size $n \geq L^{\frac{1}{\alpha}} \cdot (c_0 d_0 + 1) \log\left(\frac{L^{\frac{1}{\alpha}} \cdot (c_0 d_0 + 1)}{\delta}\right)$, with probability at least $(1 - \delta)$ over the choice of training data, we have

1. $\mathcal{R}_e(g_n) = 0$
2. $\mathcal{R}_a^1(g_n) \leq c_1' \left(\sqrt{\frac{d_0 \log n + \log \frac{1}{\delta}}{k_n}}\right)^{\beta}$
3. $\mathcal{R}_a^2(g_n) \leq c_2' \left(\sqrt{\frac{d_0 \log n + \log \frac{1}{\delta}}{k_n}}\right)^{\beta}$

where $c_0, c_1', c_2' > 0$ are universal constants.

4. Finite-sample bounds on risk functionals. In this section, we provide finite-sample upper bounds for the three risk functionals in Theorem 4.1 without any assumptions on the data distribution P. The bounds are expressed in terms of distribution-specific quantities.

Prior work of [7] studied similar bounds for nearest-neighbor classifiers for general data distributions. We extend this for selective classification with (k, k')-nearest neighbor rule of Algorithm 1.

The result crucially requires defining distribution-specific quantities. We would need the notion of *support* of the marginal distribution μ, i.e. defined as

$$\textbf{supp}(\mu) = \{x \in \mathcal{X} : \mu(B(x, r)) > 0 \text{ for all } r > 0\}.$$

For separable metric spaces $(\mathcal{X}, \rho)$ one can show that $\mu(\textbf{supp}(\mu)) = 1$ [13].

4.1. *The effective interiors for prediction and abstention..* It is well-known that for a training size of n the k nearest samples around any test input x lie in a ball $B_k(x)$ of probability mass $\approx \frac{k}{n}$, whereas the average over the k labels has a standard deviation of $\approx \frac{1}{\sqrt{k}}$. We use this intuition to define regions (interiors) around the rejection threshold $\tau > 0$ that captures the behaviour of a (k, k')-nearest neighbor rule.

Consider the set $\mathcal{X}^{+}_{p,\Delta}$ defined for any probability mass $p > 0$ and threshold $\Delta > 0$ as

$$\mathcal{X}^{+}_{p,\Delta} := \left\{ x \in \textbf{supp}(\mu) \,\middle|\, \eta_{\ell(x)}(B(x,r)) \geq \tau + \Delta, \text{ for all } r \leq r_p(x) \right\}$$

The key to this definition is we can show that the (k, k')-nearest neighbor rule of Algorithm 1 accurately predicts labels on this set. Similarly, we define $\mathcal{X}^{\circ}_{p,\Delta}$ for any probability mass $p > 0$ and threshold $\Delta > 0$,

$$\mathcal{X}^{\circ}_{p,\Delta} := \left\{ x \in \textbf{supp}(\mu) \,\middle|\, \max_{j \neq \ell(x)} \eta_j(B(x,r)) < \tau - 2\Delta, \text{ for all } r \leq r_p(x) \right\}$$

On $\mathcal{X}^{\circ}_{p,\Delta}$ we can guarantee that the (k, k')-nearest neighbor rule of Algorithm 1 at least accurately predicts labels or abstains. This motivates a definition of a set of inputs where we can guarantee abstention. We define $\mathcal{X}^{?}_{p,\Delta}$ for any probability mass $p > 0$ and threshold $\Delta > 0$,

$$\mathcal{X}^{?}_{p,\Delta} := \left\{ x \in \textbf{supp}(\mu) \,\middle|\, \max_{j \in \mathcal{Y}} \eta_j(B(x,r)) < \tau - 2\Delta, \text{ for all } r \leq r_p(x) \right\}$$

A key property of the interior sets defined above is that under Lebesgue differentiation condition (see Definition 3.1), as $p, \Delta \to 0$: $\mu(\mathcal{X}_\tau \setminus \mathcal{X}^{+}_{p,\Delta}) \to 0$, where $\mathcal{X}_\tau := \left\{ x \in \mathcal{X} \,|\, \eta_{\ell(x)}(x) \geq \tau \right\}$, $\mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p,\Delta}) \to 0$, and $\mu((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^{?}_{p,\Delta}) \to 0$ (using Lemma C.1, Appendix C).

[3] used a confidence interval Δ that depends on the training size n, nearest neighbors k, and confidence threshold δ for the average label in the region around the query input. More concretely,

$$\Delta(n, k, \delta) := c_1 \sqrt{\frac{d_0 \log n + \log(1/\delta)}{k}} \tag{8}$$

for some universal constant $c_1 > 0$. We show that roughly the same margin around τ is sufficient to control the risk functionals of the (k, k')-nearest neighbor rule where $k = \omega(\log \frac{n}{\delta})$.

In the following, we state the main result of this section in Theorem 4.1 that provides distribution-dependent bounds on the risk functionals. The detailed proof is deferred to Appendix A.

THEOREM 4.1. *Suppose that the set of balls in $(\mathcal{X}, \rho)$ has finite VC dimension d_0. Assume parameter $\tau > \frac{1}{2}$. Assume that $\tau > \frac{1}{2}$. Let $0 < \delta < 1$ denote the confidence parameter. Pick a positive integer $k_n \geq c' \log(n/\delta)$ (for some constant $c' > 0$) for training size n. Suppose (k, k')-nearest neighbor rule (Algorithm 1) with $k = k_n$ and $k' = \tau k_n$ defines a classifier g_n based on n training points chosen i.i.d. from P. With probability at least $(1 - \delta)$ over the choice of training data,*

1. $\Pr_X \left[g_n(X) \neq ?, g_n(X) \neq \ell(X) \right] \leq \mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p,\Delta})$
2. $\Pr_X \left[X \in \mathcal{X}_\tau, \, g_n(X) = ? \right] \leq \mu(\mathcal{X}_\tau \setminus \mathcal{X}^{+}_{p,\Delta})$
3. $\Pr_X \left[x \in \mathcal{X} \setminus \mathcal{X}_\tau, \, g_n(X) \neq ? \right] \leq \mu\left((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^{?}_{p,\Delta} \right)$

where $\mathcal{X}_\tau = \left\{ x \in \textbf{supp}(\mu) \,|\, \eta_{\ell(x)} \geq \tau \right\}$, $p = \frac{ck_n}{n}$, $\Delta = \Delta(n, k_n, \delta)$ for some constant $c > 1$.

Proof Outline: The key to the bounds above is showing that the $(k_n, \tau k_n)$-nearest neighbor rule of Algorithm 1: doesn't give a wrong label on $\mathcal{X}^{\circ}_{p,\Delta}$, correctly predicts on $\mathcal{X}^{+}_{p,\Delta}$, and abstains on $\mathcal{X}^{?}_{p,\Delta}$ for the specific choices of p and Δ.

In order to show these properties, we use a careful choice of total deviation bounds to control the empirical estimates $\mu_n(B)$ and $\hat{\eta}_i(B)$ (see Section 2 for definitions) for any measurable ball B and any label $i \in \mathcal{Y}$. We note that for a large fraction of all possible training sample of size n: empirical mass $\mu_n(B)$ is at least $\frac{k_n}{n}$ for measurable balls B of mass $\approx \frac{k_n}{n}$ and empirical bias $\hat{\eta}_i(B)$ is bounded within $\Delta(n, \#_n(B), \delta)$ to $\eta_i(B)$ for any measurable ball B and any label $i \in \mathcal{Y}$ (see Appendix A.1). More concretely, we state the observations as

$$\forall x \in \text{supp}(\mu)\, \mu_n(B(x, r_{\frac{ck_n}{n}}(x))) \geq \frac{k_n}{n}$$

$$\forall \text{ measurable ball } B, \forall i \in \mathcal{Y}\, |\eta_i(B) - \hat{\eta}_i(B)| \leq \Delta(n, \#_n(B), \delta)$$

4.1.0.1. VC dimension-free finite-sample bounds.

THEOREM 4.2. *Assume parameter $\tau > \frac{1}{2}$. Assume that $\tau > \frac{1}{2}$. Let $0 < \delta < 1$ denote the confidence parameter. Pick a positive integer $k_n \geq c' \log(n/\delta)$ (for some constant $c' > 0$) for training size n. Suppose (k, k')-nearest neighbor rule (Algorithm 1) with $k = k_n$ and $k' = \tau k_n$ defines a classifier g_n based on n training points chosen i.i.d. from P. With probability at least $(1 - \delta)$ over the choice of training data,*

1. $\Pr_X [g_n(X) \neq ?, g_n(X) \neq \ell(X)] \leq \delta + \mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p,\Delta})$
2. $\Pr_X [X \in \mathcal{X}_\tau, g_n(X) = ?] \leq \delta + \mu(\mathcal{X}_\tau \setminus \mathcal{X}^{+}_{p,\Delta})$
3. $\Pr_X [x \in \mathcal{X} \setminus \mathcal{X}_\tau, g_n(X) \neq ?] \leq \delta + \mu\left((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^{?}_{p,\Delta}\right)$

where $\mathcal{X}_\tau = \left\{x \in \textbf{\textit{supp}}(\mu) \,|\, \eta_{\ell(x)} \geq \tau\right\}$, $p = \frac{ck_n}{n}$, $\Delta = \Delta(n, k_n, \delta)$ for some constant $c > 1$.

4.2. *Lower bound on risk functionals.* In this section, we provide a general lower bound on the expected abstention rate for the (k, k')-nearest neighbor g_n induced by Algorithm 1 in the continuous setting (i.e, input space $\mathcal{X}$ is continuous). We define the sets $\zeta^{?}_n$ and ζ^{-}_n and show that g_n abstains on $\zeta^{?}_n$, whereas it gives a label on ζ^{-}_n. We define the sets as follows:

$$\zeta^{?}_n := \left\{ x \in \textbf{supp}(\mu) \,\middle|\, \eta_{\ell(x)} \geq \tau,\ \frac{1}{2} \leq \eta_{\ell(x)}(B(x,r)) \leq \tau + \Delta,\ \text{ for all } r_{k_n/n}(x) \leq r \leq r_{(k_n+\sqrt{k_n}+1)/n}(x) \right\}$$

$$\zeta^{-}_n := \left\{ x \in \textbf{supp}(\mu) \,\middle|\, \eta_{\ell(x)} < \tau,\ \eta_{\ell(x)}(B(x,r)) \geq \tau - \Delta,\ \text{ for all } r_{k_n/n}(x) \leq r \leq r_{(k_n+\sqrt{k_n}+1)/n}(x) \right\}$$

In the following, we show a lower bound on the sum $\mathcal{R}^1_a(g_n) + \mathcal{R}^2_a(g_n)$ in terms of the set $\zeta_n = \zeta^{?}_n \cup \zeta^{-}_n$.

THEOREM 4.3 (lower bound). *Suppose that the set of balls in $(\mathcal{X}, \rho)$ has finite VC dimension d_0. Assume parameter $1 \geq \tau > \frac{1}{2}$. Suppose (k, k')-nearest neighbor rule (Algorithm 1) with $k = k_n$ and $k' = \tau k_n$ defines a classifier g_n based on n training points chosen i.i.d. from P. Then, there exisits a constant $c' > 0$ such that*

$$\mathop{\mathbb{E}}_{n}\left[\mathcal{R}^1_a(g_n) + \mathcal{R}^2_a(g_n)\right] \geq c' \cdot \mu(\zeta_n)$$

Using Theorem 4.1, we note that $\mathcal{R}^1_a(g_n) + \mathcal{R}^2_a(g_n) \leq \mu(\mathcal{X}_\tau \setminus \mathcal{X}^{+}_{p,\Delta}) + \mu\left((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^{?}_{p,\Delta}\right)$. Under the smoothness condition (as considered in Section 3.2), using Theorem 3.4 we show that

$$\mathcal{R}^1_a(g_n) + \mathcal{R}^2_a(g_n) \leq \mathscr{M}\left(c_1 \Delta(n, k_n, \delta)\right) + \mathscr{M}\left(c_2 \Delta(n, k_n, \delta)\right)$$

Using the result on lower bound above we can show that

LEMMA 4.4. *Suppose η is (α, L)-smooth in $(\mathcal{X}, \rho, \mu)$ for some $\alpha, L \geq 0$, then for any n we can show that*

$$\zeta_n \supset \left\{ x \in \textbf{\textit{supp}}(\mathcal{X}) \,\middle|\, |\eta_{\ell(x)} - \tau| \leq \Delta(n, k_n, \delta) - L \left(\frac{k_n + \sqrt{k_n} + 1}{n} \right)^{\alpha} \right\}$$

5. Accuracy of an adaptive nearest neighbor classifier with a reject option. In this section, we study bounds on risk functionals (see (2)-Eq. (4)) of an adaptive version of nearest neighbor classifiers [3].

Similar to [3], we define the notion of advantage around an input x so to avoid misclassification.

5.1. *Pointwise saliency and safeness:.* We define a notion of *safeness* from misclassification of an input, i.e. the classifier is safe not to give a wrong label. For $p, \Delta > 0$, we say a point x is (p, Δ)-safe if the following holds:

$$\forall r \in [0, r_p(x)], \max_{j \neq \ell(x)} \eta_j(B(x, r)) \leq \tau - \Delta, \text{ and } |\eta_{\ell(x)}(B(x, r_p(x))) - \tau| \geq \Delta$$

The τ-safeness of x is defined as the largest value of $p\Delta^2$ over all such (p, Δ) pairs:

$$s_\tau(x) = \sup \left\{ p\Delta^2 \mid x \text{ is } (p, \Delta)\text{-safe} \right\}$$

As convention, we assign $s_\tau(x) := 0$ if the rhs set is empty.

Similarly, we define a notion of saliency of classification or abstention. For $p, \Delta > 0$, we say a point x is (p, Δ)-salient if the following holds:

- if $\eta_{\ell(x)}(x) > \tau$, then $\forall r \in [0, r_p(x)]$, $\eta_{\ell(x)}(B(x, r)) \geq \tau + \Delta$
- if $\eta_{\ell(x)}(x) < \tau$, then $\forall r \in [0, r_p(x)]$, $\max_j \eta_j(B(x, r)) \leq \tau - \Delta$

The τ-saliency of x is defined as the largest value of $p\Delta^2$ over all such (p, Δ) pairs:

$$m_\tau(x) = \begin{cases} \sup \left\{ p\Delta^2 \mid x \text{ is } (p, \Delta)\text{-salient} \right\} & \text{if } \eta_i(x) \neq \tau \\ 0 & \text{if } \eta_i(x) = \tau \end{cases}$$

Note that saliency implies safeness but vice versa need not be true. Since the definition of safeness and saliency require some form of uniformity in behavior of the conditional probability distribution η. Thus, intuitively, we can't hope to have non-zero safeness in the general case. But, under mild condition on the underlying metric measure space, we can show that almost all input in the support have non-zero safeness and saliency.

Given: training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, rejection threshold: $0 < \tau \leq 1$, confidence parameter: $0 < \delta < 1$

To predict at $x \in \mathcal{X}$:

1. **for** iteration $k = 1, 2, \ldots, n$
2. **do** find the smallest ball $B_k(x)$ centered at x containing exactly k training points
3. **if** $\hat{\eta}_j(B_k(x)) \geq \tau + \Delta(n, k, \delta)$
4. **return** $i \in \arg\max_j \hat{\eta}_j(B_k(x))$
5. **else if** $\max_j \hat{\eta}_j(B_k(x)) < \tau - \Delta(n, k, \delta)$
6. **return** "I don't know" or ?
7. **else** continue
8. **return** "I don't know" or ?

Algorithm 2: An adaptive nearest neighbor selective classifier

5.2. *Universal consistency under Lebesgue differentiation condition.* This condition allows non-zero m_τ over every point x such that $\max_i \eta_i(x) \neq \tau$ as stated below:

LEMMA 5.1 (Measure of safeness and saliency). *Suppose metric measure space $(\mathcal{X}, d, \mu)$ satisfy Lebesgue differentiation condition as defined in Definition 3.1. Then, for any conditional probability function $\eta := \{\eta_i\}_{i\in\mathcal{Y}}$, the sets of points:*

$$\left\{x \in \mathcal{X} \,\middle|\, \max_i \eta_i(x) \neq \tau,\, m_\tau(x) = 0\right\}$$

$$\{x \in \mathcal{X} \mid s_\tau(x) = 0\}$$

has measure zero wrt to μ.

Thus, almost surely any input is safe and salient. This is useful because we can hope that if we increase the training samples, then significant number of them would fall in the neighborhood of saliency and safeness; thereby guaranteeing a correct label prediction or abstention.

We provide a formal statement in Theorem 5.2 that states the strong convergence of $\mathcal{R}_e(g_n), \mathcal{R}_a^1(g_n)$ and $\mathcal{R}_a^2(g_n)$ to 0.

We defer the proof in the Appendix of the supplementary. Now, we state our result on universal consistency as follows:

THEOREM 5.2 (Universal consistency). *Suppose the metric measure space (X, ρ, μ) satisfies the lebesgue differentiation condition. Let (δ_n) be a sequence in $[0, 1]$ with (1) $\sum_n \delta_n < \infty$ and (2) $\lim_{n\to\infty}(\log(1/\delta_n))/n = 0$. Let the classifier $g_{n,\delta_n} : \mathcal{X} \to \mathcal{Y} \cup \{?\}$ be the result of applying the adaptive NN procedure of Algorithm 2 with n training points chosen i.i.d. from P. Then, the sum $\mathcal{R}_e(g_{n,\delta_n}) + \mathcal{R}_a^1(g_{n,\delta_n}) + \mathcal{R}_a^2(g_{n,\delta_n}) \to 0$ almost surely.*

In Section 3.2, we studied margin bounds under some well-known smoothness assumptions on the separable metric space $(\mathcal{X}, \rho, \mu)$. In the following section, we consider holder continuity Eq. (30) with respect to the marginal distribution μ (see Definition 3.1) and β-margin condition (see Definition 3.2) with respect to the rejection threshold τ, and show margin bounds with rates depending on $\left(\frac{\log \frac{n}{\delta}}{n}\right)$.

5.3. *Margin bounds for smooth measures.* We now obtain bounds for the abstention and error rates under smoothness and margin conditions with detailed proofs deferred to the Appendix of the supplementary.

THEOREM 5.3. *Suppose η is (α, L)-smooth in $(\mathcal{X}, \rho, \mu)$ and satisfies the β-margin condition (with constant C), for some $\alpha, \beta, L, C \geq 0$. Suppose that the set of balls in (X, ρ, μ) has finite VC dimension d_0. Let $g_n : \mathcal{X} \to \mathcal{Y} \cup \{?\}$ be the result of applying the adaptive NN procedure of Algorithm 2 with n training points chosen i.i.d. from P and with confidence threshold $0 < \delta < 1$. Then, we have the two following statements:*

where $\Delta_\tau = 2\tau - 1$ and C_0 is a positive constant.

(a) *With probability at least $(1 - \delta)$ over the training sample,*

$$\Pr_X\left[g_n(X) \neq ?,\, g_n(X) \neq \ell(X)\right] \leq D' \cdot \left(\frac{c \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}}$$

$$\Pr_X\left[X \in \mathcal{X}_\tau,\, g_n(X) = ?\right] \leq D' \cdot \left(\frac{c \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}}$$

$$\Pr_X\left[x \in (\mathcal{X} \setminus \mathcal{X}_\tau),\, g_n(X) \neq ?\right] \leq D' \cdot \left(\frac{c \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}}$$

(b) $\mathbb{E}_{S_n}[\mathcal{R}_e(g_n)] \leq \delta + D' \cdot \left(\frac{C' \cdot \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}}$

$\mathbb{E}_{S_n}\left[\mathcal{R}_a^1(g_n)\right] \leq \delta + D' \cdot \left(\frac{C' \cdot \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}}$

$\mathbb{E}_{S_n}\left[\mathcal{R}_a^2(g_n)\right] \leq \delta + D' \cdot \left(\frac{C' \cdot \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}}$

for constants $c, C', D' > 0$.

The upper bounds on the three risk functionals for adaptive classifier of Algorithm 2 have convergence rates of the order $\left(\frac{\log n}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}}$ that is comparatively better than the rates we achieve for the (k, k')-nearest neighbor rule of Algorithm 1 that is of the order $\left(\frac{\log n}{k_n}\right)^{\alpha/2}$. This is at the cost of zero error on classification, i.e. $\mathcal{R}_e(g_n) = 0$ for the former case.

5.4. *General upper bounds on classification and abstention.* We will show a general upper bound on risk for any marginal distribution μ and label distribution η. To show that, we analyze effective regions where prediction or abstention could go wrong.

5.4.0.1. The effective interiors of classification and abstention.. Consider the set $\mathcal{X}_z^+$ defined for any scalar $z > 0$, e.g.

$$\mathcal{X}_z^+ := \left\{ x \in \mathbf{supp}(\mu) \,\middle|\, m_\tau(x) \geq z,\, \eta_{\ell(x)}(x) > \tau \right\}$$

The key to this definition is that we can show that Algorithm 2 accurately predicts labels on this set. Similarly, we define $\mathcal{X}_z^\circ$ for any probability mass p and threshold $\Delta > 0$,

$$\mathcal{X}_z^\circ := \{ x \in \mathbf{supp}(\mu) \,|\, s_\tau(x) \geq z \}$$

On $\mathcal{X}_z^\circ$ we can guarantee that the *kNN* at least accurately predicts labels or abstains. This motivates a definition of a set of inputs where we can guarantee abstention. We define $\mathcal{X}_{p,\Delta}^?$ for any probability mass $p > 0$ and threshold $\Delta > 0$,

$$\mathcal{X}_z^? := \left\{ x \in \mathbf{supp}(\mu) \,\middle|\, m_\tau(x) \geq z,\, \eta_{\ell(x)}(x) < \tau \right\}$$

THEOREM 5.4. *Suppose that the set of balls in* $(\mathcal{X}, \rho)$ *has finite VC dimension* d_0. *Assume parameter* $\tau > \frac{1}{2}$. *Let* $0 < \delta < 1$ *denote the confidence parameter in the adaptive NN rule of Algorithm 2, and suppose the algorithm is used to define a classifier* g_n *based on* n *training points chosen i.i.d. from* P. *Then, with probability at least* $(1 - \delta)$,

1. $\Pr_X [g_n(X) \neq ?,\, g_n(X) \neq \ell(X)] \leq \mu(\mathcal{X} \setminus \mathcal{X}_z^\circ)$,
2. $\Pr_X [X \in \mathcal{X}_\tau,\, g_n(X) = ?] \leq \mu(\mathcal{X}_\tau \setminus \mathcal{X}_z^+)$,
3. $\Pr_X [x \in \mathcal{X} \setminus \mathcal{X}_\tau,\, g_n(X) \neq ?] \leq \mu\left((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}_z^?\right)$.

where $z = \left(\frac{c \cdot \log \frac{n}{\delta}}{n}\right)$ *for some constant* $c > 0$.

5.4.0.2. VC dimension-free finite-sample bounds. We can state a VC dimension-free result that is similar to the above except an additional cost of δ in the bounds.

THEOREM 5.5. *Assume parameter* $\tau > \frac{1}{2}$. *Let* $0 < \delta < 1$ *denote the confidence parameter in the adaptive NN rule of Algorithm 2, and suppose the algorithm is used to define a classifier* g_n *based on* n *training points chosen i.i.d. from* P. *Then, with probability at least* $(1 - \delta)$,

1. $\Pr_X [g_n(X) \neq ?,\, g_n(X) \neq \ell(X)] \leq \delta + \mu(\mathcal{X} \setminus \mathcal{X}_z^\circ)$,

2. $\Pr_X\left[X \in \mathcal{X}_\tau,\, g_n(X) = ?\right] \leq \delta + \mu(\mathcal{X}_\tau \setminus \mathcal{X}_z^+)$,
3. $\Pr_X\left[x \in \mathcal{X} \setminus \mathcal{X}_\tau,\, g_n(X) \neq ?\right] \leq \delta + \mu\left((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}_z^?\right)$.

where $z = \left(\frac{c \cdot \log \frac{n}{\delta}}{n}\right)$ *for some constant* $c > 0$.

List of Appendices. Now we list the appendices that contain full proofs and additional details for our results.

- Appendix A proves the distribution-dependent finite-sample bounds for the (k, k') rule in Theorem 4.1 and the VC-free variant in Theorem 4.2. It also includes the universal consistency proofs for the (k, k') rule (Theorem 3.2, Theorem 3.3).
- Appendix B gives the lower bound in Theorem 4.3 and the smoothness/margin proofs for the (k, k') rule (Theorem 3.4, Corollary 1).
- Appendix C establishes the adaptive NN distribution-dependent bounds stated in Theorem 5.4.
- Appendix D contains the universal consistency proof for the adaptive NN rule (Theorem 5.2).
- Appendix E contains the smoothness and margin proofs for the adaptive NN rule (Theorem 5.3).

APPENDIX A: FINITE-SAMPLE BOUNDS ON RISK FUNCTIONALS

In this appendix, we provide the formal proof of Theorem 4.1. First, we provide some useful results to bound the accuracy of empirical estimates of $\mu_n(\cdot)$ and $\hat{\eta}_i(\cdot)$. This is discussed in Appendix A.1. Then, using these total deviation bounds we provide the proof of Theorem 4.1 in Appendix A.2.

A.1. Total deviation bounds. For a given training sample of size n, one could expect that as n grows the empirical estimate of $\mu_n(B)$ of any measurable subset B of $\mathcal{X}$ gets closer to the probability mass $\mu(B)$. [6] established this intuition more concretely in the following lemma:

LEMMA A.1 (Lemma 7, [6]). *There is a universal constant c_0 such that the following holds. Let $\mathcal{C}$ be any class of measurable subsets of $\mathcal{X}$ of VC dimension d_0. Pick any $0 < \delta < 1$. Then with probability at least $1 - \delta^2/2$ over the choice of $(x_1, y_1), \ldots, (x_n, y_n)$, for all $B \in \mathcal{C}$ and for any integer k, we have*

$$\mu(B) \geq \frac{k}{n} + \frac{c_0}{n} \max\left(k, d_0 \log \frac{n}{\delta}\right) \implies \mu_n(B) \geq \frac{k}{n}$$

On the other hand, [3] showed a strong bound on empirical estimate of a conditional probability of events from any collection of events $\mathcal{A}$ and $\mathcal{B}$. They only need finite VC dimension of $\mathcal{A}$ and $\mathcal{B}$ to establish this result. We state it as follows:

THEOREM A.2 (Theorem 5, [3]). *Let P be a probability distribution over $\mathcal{X}$, and let $\mathcal{A}, \mathcal{B}$ be two families of measurable subsets of $\mathcal{X}$ such that VC($\mathcal{A}$),VC($\mathcal{B}$) $\leq d_0$. Let $n \in \mathbb{N}$, and let $x_1, \ldots, x_n$ be n i.i.d samples from P . The, the following event occurs with probability at least $1 - \delta$:*

$$\forall A \in \mathcal{A}, \forall B \in \mathcal{B} : |P(A|B) - P_n(A|B)| \leq \sqrt{\frac{k_0}{\#_n(B)}}$$

where $k_0 = 1000(d_0 \log(8n) + \log(4/\delta))$ and $P_n(A|B) = \frac{\sum_i 1\{x_i A \cap B\}}{\sum_i 1\{x_i \in B\}}$.

Although [3] studied Theorem A.2 for binary classification, it could be used to bound the empirical estimate $\hat{\eta}$ in the countably infinite label setting. Consider $\mathcal{A} = \{\mathcal{X} \times \mathcal{Y}\}, \mathcal{B} = \{C \times \mathcal{Y} : C \in \mathcal{C}\}$. Note that $\mathcal{A}, \mathcal{B}$ have finite VC dimension because of the finiteness of VC dimension of $\mathcal{Y}$ and $\mathcal{C}$ (i.e. product of two VC classes [40]). Using this observation, we state the following useful lemma for countably infinite label set $\mathcal{Y}$.

LEMMA A.3. *There is a universal constant $c_1 \geq 1$ for which the following holds. Let $\mathcal{C}$ be a class of subsets of $\mathcal{X}$ with VC dimension d_0. Pick any $0 < \delta < 1$. Then with probability at least $1 - \delta^2/2$ over the choice of $(x_1, y_1), \ldots, (x_n, y_n)$, for all $B \in \mathcal{C}$ and $i \in \mathcal{Y}$,*

$$|\eta_i(B) - \hat{\eta}_i(B)| \leq \Delta(n, \#_n(B), \delta)$$

where $\#_n(B) = |\{j : x_j \in B\}|$ is the number of points in B and

$$\Delta(n, \#_n(B), \delta) = c_1 \sqrt{\frac{d_0 \log n + \log(1/\delta)}{k}}$$

A.2. Proof of Theorem 4.1. In Algorithm 1, we fix $k_n = \omega(\log \frac{n}{\delta})$. Consider the constant $c = c_0 d_0 + 1$, where c_0 is as stated in Lemma A.1. Consider an arbitrary measurable ball $B \in \mathcal{C}$ such that $\mu(B) = \frac{ck_n}{n}$. Then

$$\mu(B) = \frac{ck_n}{n} = \frac{k_n}{n} + \frac{c_0 d_0 k_n}{n} \geq \frac{k_n}{n} + \frac{c_0}{n} \max\left(k_n, d_0 \log \frac{n}{\delta}\right) \tag{9}$$

This implies that

$$\mu_n(B) \geq \frac{k_n}{n} \tag{10}$$

Thus, for an input $x \in \textbf{supp}(\mu)$, $\mu(B_{k_n}(x)) \leq p_n$, where $B_{k_n}(x)$ is a ball around x such that k_n inputs are picked (see Algorithm 1).

Before, we proof the first part of the theorem, we state the following useful result:

LEMMA A.4. *Suppose that the set of balls in $(\mathcal{X}, \rho)$ has finite VC dimension d_0. Assume parameter $\tau > \frac{1}{2}$. Let $0 < \delta < 1$ denote the confidence parameter. Pick a positive integer k_n for training size n such that (1) $\lim_{n\uparrow\infty} \frac{k_n}{n} = 0$ and (2) $\lim_{n\uparrow\infty} \frac{k_n}{\log(n/\delta)} = \infty$. Suppose (k, k')-nearest neighbor rule (Algorithm 1) with $k = k_n$ and $k' = \tau k_n$ defines a classifier g_n based on n training points chosen i.i.d. from P. With probability at least $(1 - \delta)$ over the choice of training data, for any input $x \in \mathcal{X}^{\circ}_{p,\Delta}$, $g_n(x)$ is either $\ell(x)$ or ?, where $p = \frac{ck_n}{n}$, and $\Delta = \Delta(n, k_n, \delta)$ for some constant $c > 1$.*

PROOF. Fix the $(1 - \delta)$ fraction of all possible training sets of size n such that Lemma A.1, and Lemma A.3 hold. Furthermore, c is used as shown in Eq. (9).

For the sake of contradiction assume that there exists $x \in \mathcal{X}^{\circ}_{p,\Delta}$ such that $g_n(x) = j \notin \{\ell(x), ?\}$. Thus, the following condition must have been satisfied in Algorithm 1: $\hat{\eta}_j(B_{k_n}(x)) \geq \tau$. But using Lemma A.3

$$\eta_j(B_{k_n}(x)) \geq \hat{\eta}_j(B_{k_n}(x)) - \Delta(n, k_n, \delta) \geq \tau - \Delta(n, k_n, \delta)$$

Since $\Delta(n, k_n, \delta) > 0$ we have shown that for a ball $B_{k_n}(x)$ around x of probability mass at most p, $\eta_j(B_{k_n}(x)) > \tau - 2\Delta(n, k_n, \delta)$. This contradicts the assumption that $x \in \mathcal{X}^{\circ}_{p,\Delta}$. Thus, the claim in the lemma is proven. □

Using Lemma A.4, we know that Algorithm 1 doesn't predict a wrong label on $\mathcal{X}^{\circ}_{p,\Delta}$ for p, Δ as stated in Theorem 4.1 on at the least $(1 - \delta)$-fraction of all possible training samples. Thus, for any given $\delta > 0$

$$\Pr_n\left[\Pr_X[g_n(X) = \ell(X) \text{ or } g_n(X) = ?] \geq \mu(\mathcal{X}^{\circ}_{p,\Delta})\right] \geq 1 - \delta$$

This gives

$$\Pr_n\left[\Pr_X[g_n(X) \neq \ell(X), g_n(X) \neq ?] \leq \mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p,\Delta})\right] \geq 1 - \delta$$

Hence, the first part of the theorem is proven.

Now, we would proof the second part of Theorem 4.1 that follows from the following lemma:

LEMMA A.5. *Suppose that the set of balls in $(\mathcal{X}, \rho)$ has finite VC dimension d_0. Assume parameter $\tau > \frac{1}{2}$. Let $0 < \delta < 1$ denote the confidence parameter. Pick a positive integer k_n for training size n such that (1) $\lim_{n\uparrow\infty} \frac{k_n}{n} = 0$ and (2) $\lim_{n\uparrow\infty} \frac{k_n}{\log(n/\delta)} = \infty$. Suppose (k, k')-nearest neighbor rule (Algorithm 1) with $k = k_n$ and $k' = \tau k_n$ defines a classifier g_n based on n training points chosen i.i.d. from P. With probability at least $1 - \delta$ over the choice of training data, for any input $x \in \mathcal{X}^+_{p,\Delta}$, $g_n(x) = \ell(x)$, where $p = \frac{ck_n}{n}$, and $\Delta = \Delta(n, k_n, \delta)$ for some constant $c > 1$.*

PROOF. First, we fix the $(1 - \delta)$-fraction of all possible training sets of size n such that Lemma A.1, and Lemma A.3 hold. Furthermore, c is used as shown in Eq. (9).

We note that, for any $x \in \mathcal{X}^+_{p,\Delta}$,

$$\hat{\eta}_{\ell(x)}(B_{k_n}(x)) \geq \eta_{\ell(x)}(B_{k_n}(x)) - \Delta(n, k_n, \delta) \tag{11}$$

$$\geq (\tau + \Delta(n, k_n, \delta)) - \Delta(n, k_n, \delta) \tag{12}$$

$$= \tau$$

Eq. (11) follows using Lemma A.3. In Eq. (12), we use the observation from Eq. (10) that $B_{k_n}(x) \subseteq B(x, r_p(x))$, thus $\eta_{\ell(x)}(B_{k_n}(x)) \geq \tau + \Delta$.

Since $\tau > \frac{1}{2}$ thus no label other than $\ell(x)$ could satisfy the condition in the Algorithm 1. Hence, $g_n(x) = \ell(x)$. This proves the claim in the lemma. □

Using Lemma A.5, we know that Algorithm 1 predicts a correct label on $\mathcal{X}^+_{p,\Delta}$ for p, Δ as stated in Theorem 4.1 for $1 - \delta$ fraction of all possible training samples. Thus, for any $\delta > 0$

$$\Pr_n \left[\Pr_X [g_n(X) = \ell(X)] \geq \mu(\mathcal{X}^+_{p,\Delta}) \right] \geq 1 - \delta$$

This gives

$$\Pr_n \left[\Pr_X [X \in \mathcal{X}_\tau, g_n = ?] \leq \mu(\mathcal{X}_\tau \setminus \mathcal{X}^+_{p,\Delta}) \right] \geq 1 - \delta$$

LEMMA A.6. *Suppose that the set of balls in $(\mathcal{X}, \rho)$ has finite VC dimension d_0. Assume parameter $\tau > \frac{1}{2}$. Let $0 < \delta < 1$ denote the confidence parameter. Pick a positive integer k_n for training size n such that (1) $\lim_{n\uparrow\infty} \frac{k_n}{n} = 0$ and (2) $\lim_{n\uparrow\infty} \frac{k_n}{\log(n/\delta)} = \infty$. Suppose (k, k')-nearest neighbor rule (Algorithm 1) with $k = k_n$ and $k' = \tau k_n$ defines a classifier g_n based on n training points chosen i.i.d. from P. With probability at least $1 - \delta$ over the choice of training data, for any input $x \in \mathcal{X}^?_{p,\Delta}$, $g_n(x)$ is ?, where $p = \frac{ck_n}{n}$, and $\Delta = \Delta(n, k_n, \delta)$ for some constant $c > 1$.*

The proof follows similar to one given for Lemma A.4. We assume the contradiction and show that if Algorithm 1,for any $x \in \mathcal{X}^?_{p,\Delta}$, gives a label $g_n(x) = j \neq ?$ then

$$\eta_j(B_{k_n}(x)) \geq \hat{\eta}_j(B_{k_n}(x)) - \Delta(n, k_n, \delta) \geq \tau - \Delta(n, k_n, \delta).$$

Since $\mu(B_{k_n}(x)) \leq p$ the asssertion above violates the definition of $\mathcal{X}^?_{p,\Delta}$. Hence, the lemma follows. This yields the third part of the theorem.

APPENDIX B: VC DIMENSION-FREE BOUNDS

In this section, we consider the general case where we don't fix any assumption of finiteness of VC dimension of the set of balls in $(\mathcal{X}, \rho)$.

Now, we state two corollaries to Lemma A.1 and Lemma A.3 where we state the results for any fixed arbitrary input $x \in \mathcal{X}$.

COROLLARY 2. Consider an arbitrary input $x \in \mathcal{X}$. There is a universal constant c_0 such that the following holds. Pick any $0 < \delta < 1$. Then, with probability at least $1 - \delta^2/2$ over the choice of $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ for any ball $B := \mathcal{B}(x, r)$ for some $r > 0$, and for any integer k, we have

$$\mu(B) \leq \frac{k}{n} + \frac{c_0}{n} \max\left(k, \log \frac{n}{\delta}\right) \implies \mu_n(B) \geq \frac{k}{n}$$

Furthermore, we make a remark that the constant c_0 is independent of the choice of the arbitrary input $x \in \mathcal{X}$. This follows from the discussion for Lemma 16 [6].

COROLLARY 3. Consider an arbitrary input $x \in \mathcal{X}$. Pick any $0 < \delta < 1$. Then, with probability at least $1 - \delta^2/2$ over the choice of $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ for any ball $B := \mathcal{B}(x, r)$ where $r > 0$, and for any $i \in \mathcal{Y}$, we have

$$|\eta_i(B) - \hat{\eta}_i(B)| \leq \Delta(n, \#_n(B), \delta)$$

Note that $d_0 = 1$ in the definition of $\Delta(n, \#_n(B), \delta)$. Using Corollary 2 and Corollary 3, we could prove Lemma A.4, Lemma A.5, and Lemma A.6 for any input x that belongs to the sets $\mathcal{X}^{\circ}_{p,\Delta}$, $\mathcal{X}^{+}_{p,\Delta}$, and $\mathcal{X}^{?}_{p,\Delta}$ respectively without any assumption on the VC dimension of the set of balls in $(\mathcal{X}, \rho)$. We restate the result on finite-sample bounds that is similar to Theorem 4.1 but has an extra dependence on the δ parameter in the upper bounds.

THEOREM B.1. *Assume parameter $\tau > \frac{1}{2}$. Assume that $\tau > \frac{1}{2}$. Let $0 < \delta < 1$ denote the confidence parameter. Pick a positive integer $k_n \geq c' \log(n/\delta)$ (for some constant $c' > 0$) for training size n. Suppose (k, k')-nearest neighbor rule (Algorithm 1) with $k = k_n$ and $k' = \tau k_n$ defines a classifier g_n based on n training points chosen i.i.d. from P. With probability at least $(1 - \delta)$ over the choice of training data,*

1. $\Pr_X [g_n(X) \neq ?, g_n(X) \neq \ell(X)] \leq \delta + \mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p,\Delta})$
2. $\Pr_X [X \in \mathcal{X}_\tau, g_n(X) = ?] \leq \delta + \mu(\mathcal{X}_\tau \setminus \mathcal{X}^{+}_{p,\Delta})$
3. $\Pr_X [x \in \mathcal{X} \setminus \mathcal{X}_\tau, g_n(X) \neq ?] \leq \delta + \mu\left((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^{?}_{p,\Delta}\right)$

where $\mathcal{X}_\tau = \{x \in \textbf{\textit{supp}}(\mu) \,|\, \eta_{\ell(x)} \geq \tau\}$, $p = \frac{ck_n}{n}$, $\Delta = \Delta(n, k_n, \delta)$ for some constant $c > 1$.

Here, the constant c is same as the one shown in Eq. (9).

We outline the proof of the first part of the Theorem 4.2. Other parts follow using similar arguments.

For a fixed input $x \in \mathcal{X}^{\circ}_{p,\Delta}$,

$$\Pr_n [g_n(x) \neq \ell(x), g_n(x) \neq ?] \leq \delta^2$$

Now, taking expectation with respect to the sample size and sample variable X, we get

$$\mathop{\mathbb{E}}_{n}\left[\mathop{\mathbb{E}}_{X}\left[\mathbf{1}\left(g_n(X) \neq \ell(X), g_n(X) \neq ?\right) | X \in \mathcal{X}^{\circ}_{p,\Delta}\right]\right] \leq \delta^2$$

Using Markov's inequality we get the bound

$$\Pr_n\left[\Pr_X\left[g_n(X)\neq\ell(X), g_n(X)\neq?|X\in\mathcal{X}^{\circ}_{p,\Delta}\right]\geq\delta\right]\leq\delta$$

Thus, with probability at least $(1-\delta)$ over the choice of training sample

$$\Pr_X\left[g_n(X)\neq\ell(X), g_n(X)\neq?|X\in\mathcal{X}^{\circ}_{p,\Delta}\right]\leq\delta$$

Now, using a basic conditional probability inequality

$$\begin{aligned}\Pr_X\left[g_n(X)\neq\ell(X), g_n(X)\neq?\right]&\leq\Pr_X\left[g_n(X)\neq\ell(X), g_n(X)\neq?|X\in\mathcal{X}^{\circ}_{p,\Delta}\right]+\Pr_X\left[X\in\mathcal{X}^{\circ}_{p,\Delta}\right]\\&\leq\delta+\mu(\mathcal{X}^{\circ}_{p,\Delta})\end{aligned}$$

APPENDIX C: UNIVERSAL CONSISTENCY UNDER LEBESGUE DIFFERENTIATION CONDITION

In this appendix, we provide the proof for Theorem 3.2. This section is organized as, first we state and proof a useful result in Lemma C.1 and then sequentially provide proofs for each consistency result in Theorem 3.2.

In Section 4, for any $p,\Delta>0$, we defined sets $\mathcal{X}^{\circ}_{p,\Delta}$, $\mathcal{X}^{+}_{p,\Delta}$, and $\mathcal{X}^{?}_{p,\Delta}$. In the following, we state a useful result wich talks about the containment of any input $x\in\mathcal{X}$ that has Lebesgue continuity, in one of these sets. First, we define the boundary set $\mathcal{B}_\tau$ as follows:

$$\mathcal{B}_\tau:=\left\{x\in\mathcal{X}\,|\,\eta_{\ell(x)}(x)=\tau\right\}$$

Note that $\mu(\mathcal{B}_\tau)=0$. Using this, we state the lemma:

LEMMA C.1. *Suppose the metric space $(\mathcal{X},\rho,\mu)$ satisfy the lebesgue differentiation condition. Then, there exists a set $\mathcal{X}'\subset\mathcal{X}$ such that $\mu(\mathcal{X}')=0$. Moreover, we have the following items:*

- *For any $x\in\mathcal{X}\setminus\mathcal{X}'$ lies in $\mathcal{X}^{\circ}_{p,\Delta}$ for some $p,\Delta>0$.*
- *For any $x\in\mathcal{X}_\tau\setminus\mathcal{X}'$ lies in $\mathcal{X}^{+}_{p,\Delta}$ for some $p,\Delta>0$.*
- *For any $x\in\mathcal{X}\setminus(\mathcal{X}_\tau\cup\mathcal{X}')$ lies in $\mathcal{X}^{?}_{p,\Delta}$ for some $p,\Delta>0$.*

PROOF. Consider a set $\mathcal{X}'\subset\mathcal{X}$ defined as

$$\mathcal{X}'=\{x\in\mathcal{X}\,|\,\textit{Lebesgue continuity doesn't hold on } x\}\cup\mathcal{B}_\tau$$

Note that $\mu(\mathcal{X}')=0$. Now, we would show the rest of the assertion in the lemma for inputs in $\mathcal{X}\setminus\mathcal{X}'$.

Note that for any $j\neq\ell(x)$, $\tau>\frac{1}{2}\geq\eta_j(x)$. Using lebesgue differentiation condition, we have

$$\lim_{r\downarrow0}\frac{1}{\mu(B(x,r))}\int_{B(x,r)}\eta_j(x)d\mu=\eta_j(x)$$

Fix $\Delta_0:=\frac{\tau-\frac{1}{2}}{2}$. There exists r_0 such that $\forall r'\in[0,r_0]$ $\eta_j(B(x,r'(x)))\leq\eta_j(x)+\Delta_0\leq\frac{1}{2}+\Delta_0=\tau-\frac{\Delta_0}{2}$. Thus, for $p=\mu(B(x,r_0(x)))$, and $\Delta=\frac{\Delta_0}{8}$, $x\in\mathcal{X}^{\circ}_{p,\Delta}$. Hence, the first part of the lemma is proven.

Now, consider an input $x\in\mathcal{X}_\tau\setminus\mathcal{X}'$. Thus, $\eta_{\ell(x)}(x)>\tau$. We can write $\eta_{\ell(x)}(x)=\tau+\epsilon$ for some $\epsilon>0$. Now, fix $\Delta_0=\epsilon-\Delta$ for $\epsilon>\Delta>0$. Using the continuity condition, we know that there exists r_0 such that $\forall r'\in[0,r_0]$ $\eta_j(B(x,r'(x)))\geq\eta_j(x)-\Delta_0=(\tau+\epsilon)-(\epsilon-\Delta)=\tau+\Delta$. Thus, for $p=\mu(B(x,r_0(x)))$, and $\Delta=\frac{\Delta_0}{3}$, $x\in\mathcal{X}^{+}_{p,\Delta}$. This gives the second part of the lemma.

Similarly, we can argue for the third part of the lemma. □

First, we will show that $\mathcal{R}_e(g_n)$ converges to 0 almost surely. Assume p_n, Δ_n are chosen as in Theorem 4.1. Using Lemma A.5, with high probability we have

$$\Pr_X[g_n(X) \neq ?,\ g_n(X) \neq \ell(X)] \leq \mu\left(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p_n,\Delta_n}\right) \tag{13}$$

We state the observation in the following corollary.

COROLLARY 4. Let (δ_n) be any sequence of positive reals, and (k_n) any sequence of positive integers. For each n, define (p_n) and (Δ_n) as in Theorem 4.1. Then, the following holds:

$$\Pr_n\left[\mathcal{R}_e(g_n) > \mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p_n,\Delta_n})\right] \leq \delta_n$$

Using Lemma C.1, we note that as $p_n, \Delta_n \downarrow 0$

$$\lim_{n\to\infty} \mu((\mathcal{X} \setminus \mathcal{X}') \setminus \mathcal{X}^{\circ}_{p_n,\Delta_n}) = 0$$

But then

$$\lim_{n\to\infty} \mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p_n,\Delta_n}) \leq \lim_{n\to\infty} \mu((\mathcal{X} \setminus \mathcal{X}') \setminus \mathcal{X}^{\circ}_{p_n,\Delta_n}) + \mu(\mathcal{X}') = 0 \tag{14}$$

Before we prove the strong convergence we prove this straight forward result on the limiting behaviour of $\mathcal{R}_e(g_n)$,

LEMMA C.2. *For (p_n) and (Δ_n) as stated in Theorem 4.1,*

$$\lim_{n\to\infty} \Pr_n\left[\mathcal{R}_e(g_n) > \mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p_n,\Delta_n})\right] = 0$$

PROOF. Let (δ_n) be an arbitrary sequence in $(0,1)$ converging to zero. Using Corollary 4, we know that

$$\Pr_n\left[\mathcal{R}_e(g_n) > \mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p_n,\Delta_n})\right] \leq \delta_n$$
$$\implies \Pr_n\left[\mathcal{R}_e(g_n) > \mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p_n,\Delta_n})\right] \underset{\delta_n\downarrow 0}{\longrightarrow} 0$$

Since (δ_n) is an arbitrary converging sequence thus the claim holds. □

LEMMA C.3. *Let (δ_n) be a sequence in $[0,1]$ such that $\sum_n \delta_n < \infty$. Pick a sequence of positive integers (k_n) such that (1) $\lim_{n\uparrow\infty} \frac{k_n}{n} = 0$ and (2) $\lim_{n\uparrow\infty} \frac{k_n}{\log(n/\delta_n)} = \infty$. Then, $\mathcal{R}_e(g_n)$ converges to 0 almost surely.*

PROOF. Fix sequences (p_n) and (Δ_n) as stated in Theorem 4.1. Given the choice of (δ_n) and (k_n), note that $(\delta_n), (p_n)$ and (Δ_n) converge to zero. Consider an infinite training set $(X_1, Y_1), (X_2, Y_2), \ldots, .$, and denote this by θ. Pick an arbitrary $\epsilon > 0$. Let N be such that $\sum_{n\geq N} \delta_n \leq \epsilon$. Now, using Corollary 4, we have

$$\Pr\left[\theta \,|\, \exists n \geq N, \text{s.t. } \mathcal{R}_e(g_n)(\theta) > \mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p_n,\Delta_n})\right] \leq \sum_{n\geq N} \delta_n \leq \epsilon.$$

This implies that

$$\Pr\left[\theta \,|\, \forall n \geq N, \text{s.t. } \mathcal{R}_e(g_n)(\theta) \leq \mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p_n,\Delta_n})\right] \geq (1-\epsilon)$$

Using Eq. (14), we note that as $n \uparrow \infty$, $\mu(\mathcal{X} \setminus \mathcal{X}^{\circ}_{p_n,\Delta_n}) \downarrow 0$ since $p_n, \Delta_n \downarrow 0$. Thus,

$$\mathcal{R}_e(g_n)(\theta) \underset{n\uparrow\infty}{\longrightarrow} 0$$

Since ϵ is arbitrarily picked thus $\mathcal{R}_e(g_n)$ converges to 0 almost surely. □

Now, we could show the convergence of $\mathcal{R}_a^1(g_n)$ to 0 almost surely for the (k, k') nearest neighbor classifier of Algorithm 1. Using Theorem 4.1, for fixed scalars $p, \Delta > 0$ as stated, we have

$$\Pr_X[X \in \mathcal{X}_\tau,\, g_n(X) = ?] \leq \mu(\mathcal{X}_\tau \setminus \mathcal{X}^+_{p,\Delta})$$

Similar to Corollary 4, we state the following key corollary:

COROLLARY 5. Let (δ_n) be any sequence of positive reals, and (k_n) any sequence of positive integers. For each n, define (p_n) and (Δ_n) as in Theorem 4.1. Then, the following holds:

$$\Pr_n\left[\mathcal{R}_a^1(g_n) > \mu(\mathcal{X}_\tau \setminus \mathcal{X}^+_{p_n,\Delta_n})\right] \leq \delta_n$$

As shown in Eq. (14), using Lemma C.1 we note that

$$\lim_{n\uparrow\infty} \mu(\mathcal{X}_\tau \setminus \mathcal{X}^+_{p_n,\Delta_n}) = 0$$

This holds because $\mathcal{X}^+_{p_n,\Delta_n} \subseteq \mathcal{X}^\circ_{p_n,\Delta_n}$ (for n sufficiently large). Then, Lemma C.1 implies that for any input $x \in \mathcal{X}_\tau \setminus \mathcal{X}'$ there exists some choice of p_n, Δ_n such that $x \in \mathcal{X}^+_{p_n,\Delta_n}$.

Rest of the argument for the convergence of $\mathcal{R}_a^1(g_n)$ follows as shown of convergence for $\mathcal{R}_e(g_n)$.

Finally, in order to show the convergence of $\mathcal{R}_a^2(g_n)$ to 0 almost surely we use the corollary:

COROLLARY 6. Let (δ_n) be any sequence of positive reals, and (k_n) any sequence of positive integers. For each n, define (p_n) and (Δ_n) as in Theorem 4.1. Then, the following holds:

$$\Pr_n\left[\mathcal{R}_a^2(g_n) > \mu((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^?_{p_n,\Delta_n})\right] \leq \delta_n$$

Furthermore, Lemma C.1 implies that

$$\lim_{n\uparrow\infty} \mu((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^?_{p_n,\Delta_n}) = 0$$

APPENDIX D: LOWER BOUND

In this appendix, we provide the proof of Theorem 4.3 that gives lower bound on the **type 1** and **type 2** risk functionals– $\mathcal{R}_a^1$ and $\mathcal{R}_a^2$.

Similar to [7], we will consider binomial distributions to show the lower bounds. Let $\mathsf{bin}(n,p)$ denote the binomial distribution of the sum of n independent Bernoulli(p) random variables. Denote by $\mathsf{bin}(n,p;\geq k)$ the probability that the sum is at least k; similarly $\mathsf{bin}(n,p;\geq k)$ otherwise.

We will show that the $(k_n,\tau k_n)$-nearest neighbor rule of Algorithm 1 fails on $\zeta_n^?$ by abstaining on them, and fails on ζ_n^- by giving a label.

Consider any arbitrary input $x' \in \zeta_n^?$ (proof for ζ_n^- follows using similar steps). Pick n inputs denoted as $X_1, X_2, \ldots, X_n$ as follows:

I. First pick an input $X_1 \in \mathcal{X}$ according to the marginal distribution of (k_n+1)st nearest neighbor of x'.
II. Now, pick k_n inputs uniformly at random from the distribution μ restricted to the ball $B' := B(x', \rho(x', X_1))$.
III. Then, pick $(n-k_n-1)$ inputs uniformly at random from the distribution μ restricted to the ball $\mathcal{X} \setminus B'$.
IV. Randomly permute the n points obtained in this way.

Since we assume that the marginal distribution μ is continuous thus the (k_n+1)st nearest neighbor of x' is the first input chosen; denote it as $X_{(k_n+1)}(x')$. As shown in [7], $X_{(k_n+1)}(x')$ lies within a ball of probability mass $(k+\sqrt{k}+1)/n$ but not within a ball of probability mass k_n/n. Denote this event as $\mathcal{G}_1$:

$$\mathcal{G}_1 : \; r_{k_n/n}(x') \leq \rho(x', X_{(k_n+1)}(x')) \leq r_{(k+\sqrt{k}+1)/n}(x')$$

Now, using Lemma 17 in [7], we know that there exists an absolute constant $c_1 > 0$ such that $\Pr[\mathcal{G}_1] \geq c_1$.

Now, we wish to lower bound the probability that the k_n inputs sampled in B' lead to abstention given $\mathcal{G}_1$ holds. In other words, we consider the event: $\max_{i\in\mathcal{Y}} \hat{\eta}_i(B') < \tau$. More formally,

$$\mathcal{G}_2 : \text{for all } i \in \mathcal{Y}\, \hat{\eta}_i(B') < \tau$$

In the following, we would provide the lower bound on the the conditional probability of the event: $\mathcal{G}_1 \mid \mathcal{G}_2$.

LEMMA D.1. *There is an absolute constant c_2 such that* $\Pr[\mathcal{G}_2 \mid \mathcal{G}_1] > c_2$.

PROOF. Notice that if $\mathcal{G}_1$ holds then B' is a ball of radius $\rho(x', X_1)$ where X_1 was chosen in the first step of the procedure, and thus $r_{k_n/n}(x') \leq \rho(x', X_1) \leq r_{(k_n+\sqrt{k_n}+1)/n}(x')$. Since $x' \in \zeta_n^?$ we have

$$\frac{1}{2} \leq \eta_{\ell(x')}(B') \leq \tau + \Delta.$$

Denote $\eta_{\ell(x')}(B')$ by p. Now, we can bound $\Pr_n[\mathcal{G}_2 \mid \mathcal{G}_1]$ as follows:

$$\Pr_n[\mathcal{G}_2 \mid \mathcal{G}_1] \geq \Pr_n\left[(1-\tau)k_n < \hat{\eta}_{\ell(x')}(B') < \tau k_n \mid \mathcal{G}_1\right] \tag{15}$$

$$\geq 1 - \mathsf{bin}(k_n, p; \geq \tau k_n) - \mathsf{bin}(k_n, p; \leq (1-\tau)k_n) \tag{16}$$

$$= \mathsf{bin}(k_n, p; > (1-\tau)k_n) - \mathsf{bin}(k_n, p; \geq \tau k_n) \tag{17}$$

$$> 0 \tag{18}$$

In Eq. (15), we note that the event: $(1-\tau)k_n < \hat{\eta}_{\ell(x')}(B') < \tau k_n$ is a subset of the event $\mathcal{G}_2$. Since we sample k_n inputs in B' in step 2 of the procedure we can bound $\Pr_n\left[(1-\tau)k_n < \hat{\eta}_{\ell(x')}(B') < \tau k_n \mid \mathcal{G}_1\right]$ using binomial distribution. We just need to eliminate all the events where one of the inequalities for the event $(1-\tau)k_n < \hat{\eta}_{\ell(x')}(B') < \tau k_n$ doesn't hold. This includes all events (1) when $\hat{\eta}_{\ell(x')}(B') \leq (1-\tau)k_n$, and (2) when $\hat{\eta}_{\ell(x')}(B') \geq \tau k_n$. But the probability of these two events is $\mathsf{bin}(k_n, p; \geq \tau k_n) + \mathsf{bin}(k_n, p; \leq (1-\tau)k_n)$ giving Eq. (16). Eq. (17) and Eq. (18) follow using the definitions.

This gives the claim of the lemma. □

Thus,

$$\Pr_n\left[x' \in \mathcal{X}_\tau, g_n(x') = ?\right] \geq \Pr_n\left[\mathcal{G}_1 \wedge \mathcal{G}_2\right] \geq c_1 c_2$$

Taking expectation with respect of x' we get

$$\mathop{\mathbb{E}}_{n}\left[\Pr_X\left[X \in \mathcal{X}, g_n(X) = ?\right]\right] \geq c_1 c_2 \mu(\zeta_n^?)$$

Similarly, we can show that

$$\mathop{\mathbb{E}}_{n}\left[\Pr_X\left[X \in \mathcal{X} \setminus \mathcal{X}_\tau, g_n(X) \in \mathcal{Y}\right]\right] \geq c_1 c_2 \mu(\zeta_n^-)$$

Summing the two expected abstention rates gives

$$\mathop{\mathbb{E}}_{n}\left[\mathcal{R}_a^1(g_n) + \mathcal{R}_a^2(g_n)\right] \geq c_1 c_2 \mu(\zeta_n)$$

Thus, we have proven the claim of the lemma.

APPENDIX E: PROOF OF THEOREM 3.4

In this appendix, we provide the proof of Theorem 3.4 that states margin bounds on the three risk functionals under (α, L)-smooth distribution (where $\alpha, L \geq 0$)) and β-margin condition on the marginal distribution μ with a constant $C > 0$. The proof is organized in the following manner: we state and prove Lemma E.1 and Lemma E.2 that provide characterizations for subsets $\mathcal{X}^{\circ}_{p,\Delta}$, $\mathcal{X}^{+}_{p,\Delta}$, and $\mathcal{X}^{?}_{p,\Delta}$ for some specific choices of $p, \Delta > 0$. Using these results we prove each item of Theorem 3.4 for some $N > 0$, and subsequently Corollary 1. Finally, Lemma E.3 provides a concrete choice of N in terms of $\alpha, L, C, d_0, \delta$ where d_0 is the VC dimension of all measurable balls of $(\mathcal{X}, \rho)$.

LEMMA E.1. *Suppose η is (α, L)-smooth in $(\mathcal{X}, \rho, \mu)$ for some $\alpha, L \geq 0$. Then, there exists $p, \Delta > 0$ such that the following holds:*

$$\mathcal{X}^{\circ}_{p,\Delta} = \mathcal{X}$$

PROOF. Consider an arbitrary input $x \in \mathcal{X}$. Notice that for $j \neq \ell(x)$, $\eta_j(x) \leq \frac{1}{2}$. Let $r(x)$ be the radius such that $\eta_{\ell(x)}(B(x, r(x))) \geq \tau - 2\Delta$. Since η is (α, L)-smooth, we have:

$$\begin{aligned} & L \cdot \mu(B^{\circ}(x, r(x)))^{\alpha} \geq |\eta_{\ell(x)}(x) - \eta_{\ell(x)}(B(x, r(x)))| \\ \implies & L \cdot \mu(B^{\circ}(x, r(x)))^{\alpha} \geq |\frac{1}{2} - (\tau - 2\Delta)| \\ \implies & \mu(B^{\circ}(x, r(x))) \geq \left(\frac{\tau - \frac{1}{2} - 2\Delta}{L}\right)^{\frac{1}{\alpha}} \end{aligned}$$

Thus, $\eta_j(B(x, r(x))) < \tau + \Delta$ unless the probability mass of $B^{\circ}(x, r(x))$ is at least $\left(\frac{\tau - \frac{1}{2} - 2\Delta}{L}\right)^{\frac{1}{\alpha}}$. Since x is an arbitrary input in $\mathcal{X}$ we define this radius as $r^{\circ}(\Delta)$ for a given Δ. Then, for the probability mass $p^{\circ}(\Delta) := \left(\frac{\tau - \frac{1}{2} - 2\Delta}{L}\right)^{\frac{1}{\alpha}}$ and any threshold $\Delta > 0$ we have the claim. □

LEMMA E.2. *Suppose η is (α, L)-smooth in $(\mathcal{X}, \rho, \mu)$ for some $\alpha, L \geq 0$. Then, there exists $p, \Delta > 0$ such that the following holds:*

$$\begin{aligned} \left\{x \in \mathcal{X} \,|\, \eta_{\ell(x)} \geq \tau + 2\Delta\right\} &\subseteq \mathcal{X}^{+}_{p,\Delta} \\ \left\{x \in \mathcal{X} \,|\, \eta_{\ell(x)} < \tau - 3\Delta\right\} &\subseteq \mathcal{X}^{?}_{p,\Delta} \end{aligned}$$

PROOF. Fix an arbitrary threshold $\Delta > 0$. Consider an input $x \in \mathcal{X}_{\tau}$ such that $\eta_{\ell(x)} \geq \tau + 2\Delta$. Let $r(x)$ be the radius such that $\eta_{\ell(x)}(B(x, r(x))) \leq \tau + \Delta$. Since η is (α, L)-smooth, we have:

$$\begin{aligned} & L \cdot \mu(B^{\circ}(x, r(x)))^{\alpha} \geq |(\tau + 2\Delta) - (\tau + \Delta)| \\ \implies & \mu(B^{\circ}(x, r(x))) \geq \left(\frac{\Delta}{L}\right)^{\frac{1}{\alpha}} \end{aligned}$$

Thus, $\eta_{\ell(x)}(B(x, r(x))) \geq \tau + \Delta$ unless the probability mass of $B^{\circ}(x, r(x))$ is at least $\left(\frac{\Delta}{L}\right)^{\frac{1}{\alpha}}$.

Similarly, assume that for any $x \in \mathcal{X}$ $\eta_{\ell(x)}(x) < \tau - 3\Delta$. Now, if a radius $r(x)$ that leads to $\eta_{\ell(x)}(B(x, r(x)))$ change by at least Δ, then we have

$$\Delta < |\eta_{\ell(x)}(x) - \eta_{\ell(x)}(B(x, r(x)))| \leq L \cdot \mu(B^{\circ}(x, r(x)))^{\alpha} \implies \mu(B^{\circ}(x, r(x))) \geq \left(\frac{\Delta}{L}\right)^{\frac{1}{\alpha}}$$

Same argument can be applied for any other label $i \neq \ell(x)$. Thus, $\max_i \eta_i(B(x, r(x))) < \tau - 2\Delta$ unless the probability mass of $B^\circ(x, r(x))$ is at least $\left(\frac{\Delta}{L}\right)^{\frac{1}{\alpha}}$.

Since x is an arbitrary input in both the cases above we define the radius of interest as $r^+(\Delta)$ for a given $\Delta > 0$. Then, for the probability mass $p^+(\Delta) := \left(\frac{\Delta}{L}\right)^{\frac{1}{\alpha}}$ and any threshold $\Delta > 0$ we achieve the stated claim. □

In Lemma E.1 and Lemma E.2 we obtained two probability masses of interest: p° and p^+. We implicitly assume that $\Delta(n, k_n, \delta) \leq \frac{\tau - \frac{1}{2}}{3}$ as $k_n = \omega(\log \frac{n}{\delta})$, multiplying an appropriate constant to k_n yields the desired inequality. This implies that $p^\circ \geq p^+$. Set $p(n) := p^+(\Delta(n, k_n, \delta))$. Thus, by definition $p(n) = \left(\frac{\Delta(n,k_n,\delta)}{L}\right)^{\frac{1}{\alpha}}$. Since L, α are constants $p(n)$ decays slower than $\frac{ck_n}{n}$ (c as stated in Theorem 4.1), i.e. there exists $N > 0$ such that

$$\forall n \geq N \, p(n) \geq \frac{ck_n}{n} \tag{19}$$

This implies that Lemma E.1 and Lemma E.2 holds for the same rate of p, Δ as stated in Theorem 4.1. Using Eq. (19) and Lemma E.1 we know that $\forall n \geq N$

$$\mathcal{X} \subseteq \mathcal{X}^\circ_{\frac{ck_n}{n}, \Delta(n,k_n,\delta)} \implies \mu(\mathcal{X} \setminus \mathcal{X}^\circ_{\frac{ck_n}{n}, \Delta(n,k_n,\delta)}) = 0$$

Using Theorem 4.1 $\forall n \geq N$ we have,

$$\mathcal{R}_e(g_n) \leq \mu(\mathcal{X} \setminus \mathcal{X}^\circ_{\frac{ck_n}{n}, \Delta(n,k_n,\delta)}) = 0$$

Since $\mathcal{R}_e(g_n) \geq 0$ by definition this gives the first part of Theorem 3.4.

Now, we would bound $\mu(\mathcal{X}_\tau \setminus \mathcal{X}^+_{\frac{ck_n}{n}, \Delta(n,k_n,\delta)})$. Notice that if $x \in \mathcal{X}_\tau$ then $\eta_{\ell(x)}(x) \geq \tau$. Using Lemma E.2 we have $\forall n \geq N$

$$\mathcal{X}_\tau \setminus \mathcal{X}^+_{\frac{ck_n}{n}, \Delta(n,k_n,\delta)} \subseteq \left\{x \in \mathcal{X} \,|\, \tau \leq \eta_{\ell(x)} < \tau + 2\Delta(n, k_n, \Delta)\right\}$$

Thus,

$$x \in \mathcal{X}_\tau \setminus \mathcal{X}^+_{\frac{ck_n}{n}, \Delta(n,k_n,\delta)}, |\eta_{\ell(x)}(x) - \tau| \leq 2\Delta(n, k_n, \delta)$$

This implies that

$$\begin{aligned}
\mu(\mathcal{X}_\tau \setminus \mathcal{X}^+_{\frac{ck_n}{n}, \Delta(n,k_n,\delta)}) &\leq \mu(\left\{x \in \mathcal{X} \,|\, |\eta_{\ell(x)}(x) - \tau| \leq 2\Delta(n, k_n, \delta)\right\}) \\
&\leq \mathscr{M}(2\Delta(n, k_n, \delta)) \\
&= \mathscr{M}\left(2c_1\sqrt{\frac{d_0 \log n + \log \frac{1}{\delta}}{k_n}}\right)
\end{aligned}$$

Using Theorem 4.1 $\forall n \geq N$ we have,

$$\mathcal{R}^1_a(g_n) \leq \mu(\mathcal{X}_\tau \setminus \mathcal{X}^+_{\frac{ck_n}{n}, \Delta(n,k_n,\delta)}) \leq \mathscr{M}\left(2c_1\sqrt{\frac{d_0 \log n + \log \frac{1}{\delta}}{k_n}}\right)$$

This gives the second part of Theorem 3.4.

Now, we would bound $\mu((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^?_{\frac{ck_n}{n}, \Delta(n,k_n,\delta)})$. First, note that for any $x \in (\mathcal{X} \setminus \mathcal{X}_\tau)$ $\eta_{\ell(x)}(x) < \tau$. Using Lemma E.2 we have

$$(\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^?_{\frac{ck_n}{n}, \Delta(n,k_n,\delta)} \subseteq \left\{x \in \mathcal{X} \,|\, \tau - 3\Delta(n, k_n, \delta) \leq \eta_{\ell(x)} < \tau\right\}$$

This gives

$$x \in (\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^?_{\frac{ck_n}{n},\Delta(n,k_n,\delta)}, |\eta_{\ell(x)}(x) - \tau| \leq 3\Delta(n,k_n,\delta)$$

This implies that

$$\begin{aligned}\mu((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^?_{\frac{ck_n}{n},\Delta(n,k_n,\delta)}) &\leq \mu(\{x \in \mathcal{X} \,|\, |\eta_{\ell(x)}(x) - \tau| \leq 3\Delta(n,k_n,\delta)\}) \\ &\leq \mathscr{M}(3\Delta(n,k_n,\delta)) \\ &= \mathscr{M}\left(3c_1\sqrt{\frac{d_0 \log n + \log\frac{1}{\delta}}{k_n}}\right)\end{aligned}$$

Using Theorem 4.1 $\forall n \geq N$ we have,

$$\mathcal{R}^2_a(g_n) \leq \mu((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}^?_{\frac{ck_n}{n},\Delta(n,k_n,\delta)}) \leq \mathscr{M}\left(3c_1\sqrt{\frac{d_0 \log n + \log\frac{1}{\delta}}{k_n}}\right)$$

This gives the third part of Theorem 3.4.

Now, we would prove Corollary 1. The proof is a direct consequence of Theorem 3.4 where we use β-margin condition (see Definition 3.2) to note that

$$\mathscr{M}(\Delta(n,k_n,\delta)) = C \cdot \left(c_1\sqrt{\frac{d_0 \log n + \log\frac{1}{\delta}}{k_n}}\right)^\beta = Cc_1^\beta\left(\sqrt{\frac{d_0 \log n + \log\frac{1}{\delta}}{k_n}}\right)^\beta.$$

Appropriately plugging the constant multiple of Cc_1^β gives the constants c_1' and c_2' in Corollary 1.

Now, we provide give an explicit form for N as shown in the following lemma:

LEMMA E.3. *Suppose η is (α, L)-smooth in $(\mathcal{X}, \rho, \mu)$ for some $\alpha, L \geq 0$. Then, Eq. (19) holds for all*

$$n \geq L^{\frac{1}{\alpha}} \cdot (c_0 d_0 + 1) \log\left(\frac{L^{\frac{1}{\alpha}} \cdot (c_0 d_0 + 1)}{\delta}\right)$$

PROOF. We know that $(1-\delta)$-fraction of all training samples as stated in Theorem 4.1 satisfy Lemma A.1, i.e. for all measurable ball $B \in \mathcal{C}$ (see Lemma A.1) and for any integer k, we have

$$\mu(B) \geq \frac{k}{n} + \frac{c_0}{n}\max\left(k, d_0 \log\frac{n}{\delta}\right) \implies \mu_n(B) \geq \frac{k}{n}$$

In Eq. (19), we require any ball B of mass to have $\left(\frac{\Delta(n,k_n,\delta)}{L}\right)^{\frac{1}{\alpha}} \mu_n(B) \geq \frac{k_n}{n}$. Thus we want

$$\begin{aligned}&\left(\frac{\Delta(n,k_n,\delta)}{L}\right)^{\frac{1}{\alpha}} \geq \frac{k_n}{n} + \frac{c_0}{n}\max\left(k_n, d_0 \log\frac{n}{\delta}\right) \\ \implies &\left(\frac{\Delta(n,k_n,\delta)}{L}\right)^{\frac{1}{\alpha}} \geq \frac{(1 + c_0 d_0)k_n}{n} \\ \text{(expand } \Delta(n,k_n,\delta)) \quad \implies &\frac{n^{2\alpha}c_1^2(d_0 \log n + \log(1/\delta))}{k_n^{2\alpha+1}} \geq L^2 \cdot (c_0 d_0 + 1)^{2\alpha} \\ \implies &\frac{c_1^2 d_0^2 n^{2\alpha}}{k_n^{2\alpha}} \geq L^2 \cdot (c_0 d_0 + 1)^{2\alpha}\end{aligned}$$

Since $c_1, d_0 \geq 1$ we can simplify the condition as

$$\frac{n}{k_n} \geq L^{\frac{1}{\alpha}} \cdot (c_0 d_0 + 1)$$

(we use condition $k_n \geq \log \frac{n}{\delta}$) $$\implies \frac{n}{\log \frac{n}{\delta}} \geq L^{\frac{1}{\alpha}} \cdot (c_0 d_0 + 1)$$

(taking $\log$ over previous step) $$\implies \log \frac{n}{\delta} > \log \left(\frac{L^{\frac{1}{\alpha}} \cdot (c_0 d_0 + 1)}{\delta} \right)$$

(multiplying previous steps) $$\implies n \geq L^{\frac{1}{\alpha}} (c_0 d_0 + 1) \log \left(\frac{L^{\frac{1}{\alpha}} \cdot (c_0 d_0 + 1)}{\delta} \right)$$

□

APPENDIX F: GENERAL BOUNDS FOR ADAPTIVE NEAREST NEIGHBOR CLASSIFIER

In this appendix, we establish the claims of Theorem 5.4 for the adaptive nearest neighbor claasier induced by Algorithm 2. In Section 5.4, we defined the interior sets $\mathcal{X}_z^+$, $\mathcal{X}_z^?$, and $\mathcal{X}_z^\circ$ for any scalar $z \geq 0$. In terms of the measure of these sets, we stated upper bounds on the risk functionals–$\mathcal{R}_e, \mathcal{R}_a^1$, and $\mathcal{R}_a^2$ (see Eq. (2)-Eq. (4)).

We would denote the NN classifier induced by Algorithm 2 as g_n. In order to prove the theorem, we state two keys lemmas to characterize the behavior of the classifier g_n on inputs with high τ-saliency and τ-safeness. In Lemma F.1, we show that if an input has a certain τ-saliency of the order of $\frac{\log \frac{n}{\delta}}{n}$ (where n is the training sample size) then g_n is *very likely* to classify it correctly i.e., same as g_τ^*. In Lemma F.2, we show that if an input has τ-safeness approximately $\frac{\log \frac{n}{\delta}}{n}$ then g_n *very likely* doesn't give a wrong label.

Here, we state Lemma F.1 and complete its proof.

LEMMA F.1 (convergence of salient points). *Pick any $x \in \textbf{\textit{supp}}(\mu)$ with $m_\tau(x) > 0$. Fix any $0 < \delta < 1$. There exists a constant $c_3 > 0$ such that if the number of training points satisfies*

$$n \geq \frac{c_3}{m_\tau(x)} \max\left(\log \frac{1}{m_\tau(x)}, \log \frac{1}{\delta}\right),$$

then with probability at least $1 - \delta^2$ over the choice of training data, adaptive nearest neighbor classifier g_n as shown in Algorithm 2 predicts same as g_τ^*, that is, $g_n(x) = g_\tau^*(x)$.

PROOF OF LEMMA F.1. Fix $c_2 = c_1 \cdot \sqrt{1 + c_0}$ where c_0, c_1 are the universal constants from Lemma A.1 and Lemma A.3. Finally, we pick $c_3 = 4c_2^2$.

First, assume that $\eta_{\ell(x)}(x) > \tau$. Now, we would show Algorithm 2 predicts the maximizing label $\ell(x)$ for the input x. Consider all balls $B \subseteq \mathcal{X}$ centered at x. It is easy to note that the VC dimension of these balls $d_0 = 1$. Now, we would fix a $(1 - \delta^2)$ fraction of all possible choices of the training data of size n. The idea is to pick the useful fraction on which Lemma A.1 and Lemma A.3 hold. Thus, we consider the $(1 - \delta^2)$ faction on which the following properties hold:

1. For any integer k, we have $\#_n(B) \geq k$ whenever $n\mu(B) \geq k + c_0 \max(k, \log\left(\frac{n}{\delta}\right))$.
2. $\forall i \in \mathcal{Y}$, $|\hat{\eta}_i(B) - \eta_i(B)| \leq \Delta(n, \#_n(B), \delta)$

Note that x is (p, Δ)-salient such that $m_\tau(x) := p\Delta^2$. Using the lower bound on n, we can write that:

$$n \cdot m_\tau(x) \geq 4c_2^2 \left(\log n + \log \frac{1}{\delta}\right) \implies \Delta \geq 2c_2 \sqrt{\frac{\log\left(\frac{n}{\delta}\right)}{np}} \tag{20}$$

Fix $k = \frac{np}{1+c_0}$. Using Eq. (20), we get $np \geq 16c_2^2 \log \frac{n}{\delta}$ which implies $k \geq \log \frac{n}{\delta}$. Thus, we have the following:

$$np = k \cdot (1 + c_0) \geq k + c_0 k \geq k + c_0 \max\left(k, \log \frac{n}{\delta}\right) \tag{21}$$

Using this observation, we know that the ball $B = B(x, r_p)$ satisfies the condition that $\#_n(B) \geq k$ which implies that $\Delta(n, k, \delta) \geq \Delta(n, \#_n(B), \delta)$. Now, using Lemma A.3, for the ball B we have:

$$\hat{\eta}_{\ell(x)}(B) \geq \eta_{\ell(x)}(B) - \Delta(n, \#_n(B), \delta) \tag{22}$$

$$\geq \tau + \Delta - \Delta(n, k, \delta) \tag{23}$$

$$= \tau + 2c_2\sqrt{\frac{\log \frac{n}{\delta}}{np}} - \Delta(n, k, \delta) \tag{24}$$

$$\geq \tau + 2c_1\sqrt{\frac{\log \frac{n}{\delta}}{k}} - \Delta(n, k, \delta) \tag{25}$$

$$\geq \tau + \Delta(n, \#_n(B), \delta) \tag{26}$$

Using the definition of advantage, we know that $\eta_{\ell(x)}(B) \geq \tau + \Delta$ which gives Eq. (23). Eq. (24) is a direct consequence of Eq. (20). In Eq. (25), we note the specific choice of the constant c_2. Note that Eq. (26) gives the satisfying condition for Algorithm 2 to provide the label that is $\ell(x)$ as $\tau > \frac{1}{2}$.

Now, we would argue that kNN of Algorithm 2 doesn't flip the label for any ball $B' := B(x, r')$ with $r' < r_p(x)$. Assume the contrary that for B' and label $j \neq \ell(x)$, we have

$$\hat{\eta}_j(B') \geq \tau + \Delta(n, \#_n(B'), \delta)$$

Using Lemma A.3, we get:

$$\eta_j(B') \geq \hat{\eta}_j(B') - \Delta(n, \#_n(B'), \delta) \geq \tau + \Delta(n, \#_n(B'), \delta) - \Delta(n, \#_n(B'), \delta) \geq \tau$$

But we know that $\eta_{\ell(x)}(B') \geq \tau$ and this gives a contradiction. Thus, we are guaranteed that label i is not flipped. Furthermore, the algorithm didn't abstain either on B' as

$$\hat{\eta}_i(B') \geq \eta_i(B') - \Delta(n, \#_n(B'), \delta) > \tau - \Delta(n, \#_n(B'), \delta)$$

where we use $\eta_i(B') > \tau$. This ensures that Algorithm 2 reliably returns the label i.

For the case when $\eta_i(x) < \tau$, similar to Eq. (22), using Eq. (20) we obtain

$$\begin{aligned}
\max_j \hat{\eta}_j(B) &\leq \max_j \eta_j(B) + \Delta(n, \#_n(B), \delta) \\
&\leq \tau - \Delta + \Delta(n, k, \delta) \\
&\leq \tau - 2 \cdot \Delta(n, k, \delta) + \Delta(n, k, \delta) \\
&\leq \tau - \Delta(n, \#_n(B), \delta).
\end{aligned}$$

Thus, Algorithm 2 abstains on the ball B. We can show that it there is no label prediction for a ball $B' = B(x, r')$ with radius $r' < r_p(x)$ as $\forall B'\ \eta_i(B') < \tau$ that gives

$$\max_j \hat{\eta}_j(B') \leq \max_j \eta_j(B') + \Delta(n, \#_n(B'), \delta) < \tau + \Delta(n, \#_n(B'), \delta).$$

Thus, we have proven the stated lemma. □

Now, we state Lemma F.2 and provide its proof subsequently.

LEMMA F.2 (convergence of safe points). *Pick any $x \in \text{supp}(\mu)$ with $s_\tau(x) > 0$. Fix any $0 < \delta < 1$. There exists a constant $c_3 > 0$ such that if the number of training points satisfies*

$$n \geq \frac{c_3}{s_\tau(x)} \max\left(\log \frac{c_3}{s_\tau(x)}, \log \frac{1}{\delta}\right),$$

then with probability at least $1 - \delta^2$ over the choice of training data, adaptive nearest neighbor classifier g_n as shown in Algorithm 2 doesn't predict a wrong label, that is, $g_n(x) \in \{\ell(x), ?\}$.

PROOF. We follow similar steps as for the proof of Lemma F.1. Consider any input x that satisfies the condition in the lemma. Fix c_2 as above. Since x is (p, Δ)-safe such that $s_\tau(x) := p\Delta^2$, as noted in Eq. (20), we get

$$n \cdot s_\tau(x) \geq 4c_2^2 \left(\log n + \log \frac{1}{\delta} \right) \implies \Delta \geq 2c_2 \sqrt{\frac{\log\left(\frac{n}{\delta}\right)}{np}} \tag{27}$$

We fix the ball B centered at x with radius r_p such that $\#_n(B) \geq k$ (for at least $(1-\delta)$-fraction of all n-samples). Notice that $|\eta_{\ell(x)}(x) - \tau| \geq \Delta$. Wlog we assume that $\eta_{\ell(x)}(x) \geq \tau + \Delta$ (other case would be dealt similarly). As shown in Eq. (22)-26, we have

$$\hat{\eta}_{\ell(x)}(B) \geq \tau + \Delta(n, \#_n(B), \delta)$$

Thus, for the ball B Algorithm 2 predicts the maximizing label $\ell(x)$. Now, we would show that for any radius $r' \leq r_p(x)$, kNN doesn't predict a wrong label. Notice that

$$\begin{aligned}
\max_{i \neq \ell(x)} \hat{\eta}_i(B(x, r')) &\leq \max_{i \neq \ell(x)} \eta_i(B(x, r')) + \Delta(n, \#_n(B(x, r')), \delta) \\
&\leq \tau - \Delta + \Delta(n, \#_n(B(x, r')), \delta) \\
&\leq \tau - \Delta(n, \#_n(B(x, r')), \delta) + \Delta(n, \#_n(B(x, r')), \delta) \\
&\leq \tau
\end{aligned}$$

The second last inequality above follows using Eq. (21) and Eq. (27). Thus, kNN wouldn't have given a wrong label for the ball $B(x, r')$ for any $r' \leq r_p(x)$.

Hence, we have shown the claim of the lemma. □

Using Lemma F.1, we can easily show uniform convergence over every point $x \in \textbf{supp}(\mu)$ if the conditions in Lemma F.1 is satisfied. This follows using similar arguments as used in Lemma F.1 where the difference would be in applying the key lemmas: Lemma A.1, and Lemma A.3, for the class of balls $\mathcal{C}$ rather than just for balls around an input.

Now, we proof Theorem 5.4 in the following.

PROOF OF THEOREM 5.4. Consider the $(1-\delta)$-fraction of all possible training samples such that uniform convergence occurs as stated in Lemma F.1 and Lemma F.2.

Now, we consider the first part of Theorem 5.4. Note that for any $x \in \textbf{supp}(\mu)$, if

$$n \geq \frac{c}{s_\tau(x)} \max\left(\log \frac{1}{s_\tau(x)}, \log \frac{1}{\delta} \right),$$

for some constant $c > 0$ (as shown in Lemma F.2),

$$g_n(x) \in \{?, \ell(x)\}.$$

Thus,

$$\Pr_n \left[\Pr_X [g_n(X) \neq ?, g_n(X) \neq \ell(X)] > \mu(\mathcal{X}_z^\circ) \right] \geq 1 - \delta,$$

where $\frac{c \cdot \log \frac{n}{\delta}}{n}$. Thus, g_n could predict a wrong label only on the set $\mathcal{X} \setminus \mathcal{X}_z^\circ$. This implies

$$\Pr_n \left[\Pr_X [g_n(X) \neq ?, g_n(X) \neq \ell(X)] \leq \mu(\mathcal{X} \setminus \mathcal{X}_z^\circ) \right] \geq 1 - \delta$$

Now, we consider the second part of Theorem 5.4 (third part could be proven in a similar manner). Note that for any $x \in \mathbf{supp}(\mu)$, if

$$n \geq \frac{c}{m_\tau(x)} \max\left(\log \frac{1}{m_\tau(x)}, \log \frac{1}{\delta}\right),$$

for some constant $c > 0$ (as shown in Lemma F.1),

$$g_n(x) = g^*_\tau(x).$$

Thus,

$$\Pr_n\left[\Pr_X[X \in \mathcal{X}_\tau,\, g_n(X) = \ell(X)] \geq \mu(\mathcal{X}_z^+)\right] \geq 1 - \delta,$$

where $z = \frac{c \cdot \log \frac{n}{\delta}}{n}$. This gives us

$$\Pr_n\left[\Pr_X[X \in \mathcal{X}_\tau,\, g_n(X) \neq \ell(x)] \leq \mu(\mathcal{X}_\tau \setminus \mathcal{X}_z^+)\right] \geq 1 - \delta$$

$$\implies \Pr_n\left[\Pr_X[X \in \mathcal{X}_\tau,\, g_n(X) = ?] \leq \mu(\mathcal{X}_\tau \setminus \mathcal{X}_z^+)\right] \geq 1 - \delta$$

Using similar arguments the third part of Theorem 5.4 follows. □

APPENDIX G: UNIVERSAL CONSISTENCY OF ADAPTIVE NEAREST NEIGHBOR CLASSIFIERS

In this appendix, we would provide the proof of Theorem 5.2, and thus show the universal consistency of the kNN rule of Algorithm 2. In order to establish that we prove Lemma 5.1.

For the sake of clarity, we restate Lemma 5.1 here.

LEMMA G.1 (Measure of safeness and saliency). *Suppose metric measure space $(\mathcal{X}, \rho, \mu)$ satisfy Lebesgue differentiation condition as defined in Definition 3.1. Then, for any conditional probability function $\eta := \{\eta_i\}_{i \in \mathcal{Y}}$, the sets of points:*

$$\left\{x \in \mathcal{X} \,\middle|\, \eta_{\ell(x)}(x) \neq \tau,\, m_\tau(x) = 0\right\}$$

$$\left\{x \in \mathcal{X} \,\middle|\, s_\tau(x) = 0\right\}$$

has measure zero wrt to μ.

PROOF. First, we prove the result on saliency measure $m_\tau(\cdot)$. Consider an input $x \in \mathcal{X}$ such that $\eta_{\ell(x)} > \tau$. Thus, we could write $\eta_{\ell(x)}(x) = \tau + \epsilon$ for some $\epsilon > 0$. Using lebesgue differentiation condition (assuming it holds on x with $f := \eta_{\ell(x)}$), we have

$$\lim_{r \downarrow 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} \eta_{\ell(x)}(x) d\mu = \eta_{\ell(x)}(x)$$

Thus, there exists a constant $r > 0$ such that $\eta_{\ell(x)}(B(x,r')) \geq \tau + \frac{\epsilon}{2}$ for all $r' \leq r$. Thus, x is $(\mu(B(x,r)), \frac{\epsilon}{2})$-salient, and thus have a non-zero τ-saliency.

Now, consider the case where $\eta_{\ell(x)}(x) < \tau$. We'll break this case into two parts: (1) when $\frac{1}{2} \leq \eta_{\ell(x)}(x) < \tau$, and (2) when $\eta_{\ell(x)}(x) < \frac{1}{2}$.

Part (1): We can write $\eta_{\ell(x)}(x) = \tau - \epsilon$ for some $\epsilon > 0$. Using the continuity condition on the function $\eta_{\ell(x)}$, we get a constant $r > 0$ such that

$$\forall r_0 \leq r,\ \eta_{\ell(x)}(B(x,r_0)) \in [1/2 - \Delta, 1/2 + \Delta]$$

for some $\Delta > (\tau - 1/2) - \epsilon$. But then for any $j \neq \ell(x)$ and $r' \leq r$, $\eta_{\ell(x)}(B(x,r')) \leq 1/2 + \Delta$. Thus, x is $(\mu(B(x,r)), (\tau - 1/2) - \Delta)$-salient in this case.

Part (2): We will use the idea from part (1) over a sum of biases over finite labels. Consider a descending order π of biases $\{\eta_{\pi(i)}(x)\}$. Since $\sum_{\pi(i):i\in\mathcal{Y}} \eta_{\pi(i)}(x) = 1$ there exists $k \geq 1$ such that $\sum_{\pi(i):i\in\{1,2,\ldots,k\}} \eta_{\pi(i)}(x) > \frac{1}{2}$ for the set $\{\pi(1), \pi(2), \ldots, \pi(k)\}$. Denote this finite sum as $\partial_k(x)$. Note that $\partial_k(\cdot)$ as a sum of k measurable functions, i.e. $\eta_{\pi(i)}$'s is a measurable. Now, applying the continuity condition on $\partial_k(\cdot)$, we get constant $r_\partial, r_1, \ldots, r_k > 0$ such that

$$\forall i \forall r_0 \leq r_i,\ \eta_{\pi(i)}(B(x,r_0)) \leq \left(\frac{1}{2} + \frac{\tau - \frac{1}{2}}{2}\right)$$

$$\forall r_0 \leq r_\partial,\ \partial_k(B(x,r_0)) \in \left[\frac{1}{2}, \frac{1}{2} + \Delta_\partial\right]$$

for some $\Delta_\partial > 0$. We can find the appropriate constants because each $\eta_{\pi(i)}(x) < 1/2$ and $\partial_k(x) > 1/2$. Now, pick $r_{\min} = \min\{r_1, r_2, \ldots, r_k, r_\partial\}$ and $\Delta_{\min} = \min\left\{\frac{\tau - \frac{1}{2}}{2}, \Delta_\partial\right\}$. For a ball contained in $B(x, r_{\min})$ for any label j η_j is bounded within $\Delta_{\min}$. Hence, x is $(\mu(B, x, r_{\min}), \Delta_{\min})$-salient in this case.

Thus, we have shown that, if x has the lebesgue continuity condition, $m_\tau(x)$ is non-zero. Since the measure of inputs with maximizing confidence τ or inputs that don't satisfy lebesgue continuity condition is zero, we get the first part of the lemma.

Now, we show the result on safeness measure $s_\tau(\cdot)$. Consider an input $x \in \mathcal{X}$ that satisfy the lebesgue continuity condition. We just need to note that every salient input is also *safe*. If $\eta_{\ell(x)}(x) \neq \tau$ then we note that almost surely $m_\tau(x) > 0$. But then we have the following:

- if $\eta_{\ell(x)}(x) > \tau$, $\exists r, \Delta > 0$, forall $r' \leq r$ $\eta_{\ell(x)}(B(x,r')) \geq \tau + \Delta$
- if $\eta_{\ell(x)}(x) < \tau$, $\exists r, \Delta > 0$, forall $r' \leq r$ $\max_i \eta_i(B(x,r')) \leq \tau - \Delta$.

Using (r, Δ) we can show non-zero safeness of the input x. Thus, x satisfy the safeness condition. Since the measure of inputs x' with $\eta_{\ell(x')}(x') = \tau$ is zero, we show that almost every input in $\mathcal{X}$ is safe.

□

Now, we would show the proof of Theorem 5.2. First, we show the consistency of the kNN rule g_{n,δ_n} defined by Algorithm 2 for error rate $\mathcal{R}_e(g_{n,\delta_n})$.

Using Theorem 5.4, we note the following key observation: for fixed scalars $\delta > 0$ and $z \geq \frac{c \cdot \log \frac{n}{\delta}}{n}$ (for constant $c > 0$), with probability at least $(1 - \delta)$

$$\Pr_X[g_{n,\delta}(X) \neq ?,\ g_{n,\delta}(X) \neq \ell(X)] \leq \mu(\mathcal{X} \setminus \mathcal{X}_z^\circ)$$

This holds because for any $x \in \mathcal{X}_z^\circ$ has τ-safeness $s_\tau(x) \geq z$ that implies $g_{n,\delta}(x) \in \{?, \ell(x)\}$ (as shown in Lemma F.2) on $(1 - \delta)$ fraction of training samples. Now, consider a sequence of confidence parameter (δ_n). Define a sequence of τ-safeness parameter (z_n) as follows:

$$z_n = \frac{c \cdot \log\left(\frac{n}{\delta_n}\right)}{n} \tag{28}$$

We note that as $n \uparrow \infty$, $z_n \downarrow 0$. We state the following useful corollary.

COROLLARY 7. Let (δ_n) be any sequence of positive reals, and (z_n) be a sequence of positive reals as shown in Eq. (28). Then,

$$\Pr_n\left[\Pr_X[g_{n,\delta_n}(X) \neq ?,\ g_{n,\delta_n}(X) \neq \ell(X)] \leq \mu(\mathcal{X} \setminus \mathcal{X}_z^\circ)\right] \geq 1 - \delta_n$$

Now, we note that using Lemma 5.1, we have the following:

$$\lim_{n\uparrow\infty} \mu(\mathcal{X} \setminus \mathcal{X}_{z_n}^\circ) = 0 \tag{29}$$

This holds because $\mathcal{X}_{z_n}^\circ \to \mathcal{X}$ almost surely as almost every $x \in \textbf{supp}(\mathcal{X})$ has non-zero τ-safeness.

Before we prove the strong convergence we prove this straight forward result on the limiting behaviour of $\mathcal{R}_e(g_{n,\delta_n})$.

LEMMA G.2. *Let (δ_n) be any sequence of positive reals, and (z_n) be a sequence of positive reals as shown in Eq. (28). Then,*

$$\lim_{n\uparrow\infty} \Pr_n\left[\mathcal{R}_e(g_{n,\delta_n}) \leq \mu(\mathcal{X} \setminus \mathcal{X}_{z_n}^\circ)\right] = 0$$

PROOF. Fix a sequence of confidence parameter (δ_n) converging to zero. Using Corollary 7, we have

$$\begin{aligned} &\Pr_n\left[\mathcal{R}_e(g_{n,\delta_n}) > \mu(\mathcal{X} \setminus \mathcal{X}_{z_n}^\circ)\right] \leq \delta_n \\ \implies &\Pr_n\left[\mathcal{R}_e(g_{n,\delta_n}) > \mu(\mathcal{X} \setminus \mathcal{X}_{z_n}^\circ)\right] \underset{\delta_n\downarrow 0}{\longrightarrow} 0 \end{aligned}$$

Since (δ_n) is an arbitrary converging sequence thus the claim holds. □

Similar to the proof of Lemma C.3, we show that $\mathcal{R}_e(g_{n,\delta_n})$ converges to 0 almost surely.

LEMMA G.3. *Let (δ_n) be a sequence in $[0,1]$ with (1) $\sum_n \delta_n < \infty$ and (2) $\lim_{n\to\infty}(\log(1/\delta_n))/n = 0$. Let the classifier $g_{n,\delta_n} : \mathcal{X} \to \mathcal{Y} \cup \{?\}$ be the result of applying the adaptive NN procedure of Algorithm 2 with n training points chosen i.i.d. from P. Then, $\mathcal{R}_e(g_{n,\delta_n})$ converges to 0 almost surely.*

PROOF. Pick (z_n) as stated in Eq. (28) using (δ_n). Given the choice of (δ_n), note that (δ_n) and (z_n) converge to zero. Consider an infinite training set $(X_1, Y_1), (X_2, Y_2), \ldots$, and denote this by ω. Pick an arbitrary $\epsilon > 0$. Let N be such that $\sum_{n\geq N} \delta_n \leq \epsilon$. Now, using Corollary 7, we have

$$\Pr\left[\omega \,|\, \exists n \geq N, \text{s.t. } \mathcal{R}_e(g_{n,\delta_n})(\omega) > \mu(\mathcal{X} \setminus \mathcal{X}_{z_n}^\circ)\right] \leq \sum_{n\geq N} \delta_n \leq \epsilon.$$

This implies that

$$\Pr\left[\omega \,|\, \forall n \geq N, \text{s.t. } \mathcal{R}_e(g_{n,\delta_n})(\omega) \leq \mu(\mathcal{X} \setminus \mathcal{X}_{z_n}^\circ)\right] \geq (1 - \epsilon)$$

But note that as $n \uparrow \infty$, using Eq. (28) and Eq. (29), we have $\mu(\mathcal{X} \setminus \mathcal{X}_{z_n}^\circ) \downarrow 0$ because $z_n \downarrow 0$. This gives $\mathcal{R}_e(g_{n,\delta_n})(\omega) \to 0$. Since ϵ is arbitrarily picked, $\mathcal{R}_e(g_{n,\delta_n})$ converges to 0 almost surely. □

Now, we could show the convergence of $\mathcal{R}_a^1(g_{n,\delta_n})$ to 0 almost surely for the adaptive nearest neighbor classifier of Algorithm 2. But then using Lemma F.1 and Theorem 5.4, we note the following key observation: for fixed scalars $\delta > 0$ and $z \geq \frac{c \cdot \log \frac{n}{\delta}}{n}$ (for constant $c > 0$), with probability at least $(1 - \delta)$

$$\Pr_X [X \in \mathcal{X}_\tau,\, g_{n,\delta}(X) = ?] \leq \mu(\mathcal{X}_\tau \setminus \mathcal{X}_z^+)$$

This holds because for any $x \in \mathcal{X}_z^+$ has τ-saliency $m_\tau(x) \geq z$ that implies $g_n(x) = g_\tau^*(x)$ (as shown in Lemma F.1) on $(1 - \delta)$ fraction of training samples.

Similar to Corollary 7, we note the following useful corollary:

COROLLARY 8. Let (δ_n) be any sequence of positive reals, and (z_n) be a sequence of positive reals as shown in Eq. (28). Then,

$$\Pr_n \left[\Pr_X [X \in \mathcal{X}_\tau,\, g_{n,\delta}(X) = ?] \leq \mu(\mathcal{X}_\tau \setminus \mathcal{X}_{z_n}^+) \right] \geq 1 - \delta_n$$

Rest of the proof follows similar steps as above once we note that

$$\lim_{n \uparrow \infty} \mathcal{X}_{z_n}^+ \to \mathcal{X}_\tau \text{ (almost surely)}$$

This holds because every $x \in \textbf{supp}(\mathcal{X})$ has non-zero τ-advantage (see Lemma 5.1).

Finally, in order to show the convergence of $\mathcal{R}_a^2(g_n)$ to 0 almost surely we use the corollary:

COROLLARY 9. Let (δ_n) be any sequence of positive reals, and (z_n) be a sequence of positive reals as shown in Eq. (28). Then,

$$\Pr_n \left[\mathcal{R}_a^2(g_{n,\delta_n}) > \mu((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}_{z_n}^?) \right] \leq \delta_n$$

Furthermore, Lemma 5.1 implies that

$$\lim_{n \uparrow \infty} \mu((\mathcal{X} \setminus \mathcal{X}_\tau) \setminus \mathcal{X}_{z_n}^?) = 0$$

APPENDIX H: BOUNDS FOR ADAPTIVE NEAREST NEIGHBOR CLASSIFIERS UNDER SMOOTH MEASURES

In this appendix, we establish the upper bounds as stated in Theorem 5.3 on the risk functionals–$\mathcal{R}_e$, $\mathcal{R}_a^1$, and $\mathcal{R}_a^2$, when the underlying label distribution η satisfies the (α, L)-smoothness condition (Definition 3.1) and data distribution μ satisfies the β-margin condition (Definition 3.2).

H.1. Useful results. Here, we state some useful results.

LEMMA H.1 (point-wise convergence). *Suppose η is (α, L)-smooth in $(\mathcal{X}, \rho, \mu)$ and satisfies the β-margin condition (with constant C), for some $\alpha, \beta, L, C \geq 0$. Pick any $x \in$ supp(μ) with $m_\tau(x) > 0$. Fix any $0 < \delta < 1$. There exists a constant $c_3 > 0$ such that if the number of training points satisfies*

$$n \geq \frac{c_3}{m_\tau(x)} \max\left(\log \frac{1}{m_\tau(x)}, \log \frac{1}{\delta}\right),$$

then with probability at least $1 - \delta^2$ over the choice of training data, adaptive nearest neighbor classifier g_n as shown in Algorithm 2 predicts same as g_τ^*, that is, $g_n(x) = g_\tau^*(x)$.

The proof for Lemma H.1 follows using similar arguments as used for Lemma F.1.

LEMMA H.2 (convergence of safe points). *Suppose η is (α, L)-smooth in $(\mathcal{X}, \rho, \mu)$ and satisfies the β-margin condition (with constant C), for some $\alpha, \beta, L, C \geq 0$. Pick any $x \in$ supp(μ) with $s_\tau(x) > 0$. Fix any $0 < \delta < 1$. There exists a constant $c_3 > 0$ such that if the number of training points satisfies*

$$n \geq \frac{c_3}{s_\tau(x)} \max\left(\log \frac{1}{s_\tau(x)}, \log \frac{1}{\delta}\right),$$

then with probability at least $1 - \delta^2$ over the choice of training data, adaptive nearest neighbor classifier g_n as shown in Algorithm 2 doesn't predict a wrong label, that is, $g_n(x) \in \{\ell(x), ?\}$.

The proof for Lemma H.2 follows using similar arguments as used for Lemma F.2.

We can easily extend the result to uniform convergence for the class of balls $\mathcal{C}$ rather then just for balls around an input using Lemma H.1.

DEFINITION H.1 (strong Holder-continuity). For some $\alpha, L > 0$, we say the conditional probability function η is (α, L)-smooth in metric measure space $(\mathcal{X}, \rho, \mu)$ if for all $x, x' \in \mathcal{X}$, and any $j \in \mathcal{Y}$,

$$|\eta_j(x) - \eta_j(x')| \leq L \cdot \mu(B^\circ(x, \rho(x, x')))^\alpha \tag{30}$$

Then it is easy to show that for all $j \in \mathcal{Y}$ and $x \in \mathcal{X}$, $\eta_j(B(x, r))$ is continuous as a function of r.

For the following lemma, we would assume that the saliency and safeness condition satisfy the following:

- For τ-saliency, if $x \in$ **supp**(μ) is (p, Δ)-salient then either $\eta_{\ell(x)}(B(x, r_p)) = \tau + \Delta$ (for $\eta_{\ell(x)}(x) > \tau$) or $\eta_{\ell(x)}(B(x, r_p)) = \tau - \Delta$ (for $\eta_{\ell(x)}(x) < \tau$).
- For τ-safeness, if $x \in$ **supp**(μ) is (p, Δ)-safe then $|\eta_{\ell(x)}(x) - \eta_{\ell(x)}(B(x, r_p(x)))| = \Delta$

H.1.0.1. Remark. Consider an input $x \in \mathcal{X}$. Note that if for all $r > 0$ $\eta_{\ell(x)}(B(x, r(x))) > \eta_{\ell(x)}(x)$ then we define $m_\tau(x) := 1 \cdot (\eta_{\ell(x)}(x) - \tau)^2$ where rather 0 (following the convention above).

LEMMA H.3. *Suppose η is (α, L)-smooth in $(\mathcal{X}, \rho, \mu)$ for constants $\alpha, L \geq 0$. For a given confidence parameter τ and constant $a > 0$, we have the following:*

$$\left\{x \in \textbf{\textit{supp}}(\mu) : m_\tau(x) \leq a, \eta_{\ell(x)}(x) \geq \tau\right\} \subseteq \left\{x \in \textbf{\textit{supp}}(\mu) : \tau \leq \eta_{\ell(x)}(x) \leq \tau + \Delta_0 + L \cdot p_0^\alpha\right\},$$

$$\left\{x \in \textbf{\textit{supp}}(\mu) : m_\tau(x) \leq a, \eta_{\ell(x)}(x) < \tau\right\} \subseteq \left\{x \in \textbf{\textit{supp}}(\mu) : \tau - \Delta_0 - L \cdot p_0^\alpha \leq \eta_{\ell(x)}(x) < \tau\right\},$$

$$\left\{x \in \textbf{\textit{supp}}(\mu) : s_\tau(x) \leq a\right\} \subseteq \left\{x \in \textbf{\textit{supp}}(\mu) : \tau - \Delta_0 - L \cdot p_0^\alpha \leq \eta_{\ell(x)}(x) \leq \tau + \Delta_0 + L \cdot p_0^\alpha\right\},$$

where $p_0 = (L\alpha)^{\frac{-2}{2\alpha+1}} \cdot a^{\frac{1}{2\alpha+1}}$ *and* $\Delta_0 = (L\alpha)^{\frac{1}{2\alpha+1}} \cdot a^{\frac{\alpha}{2\alpha+1}}$.

PROOF. We'll proof the first part of the claim of the lemma (others follow using similar arguments).

First, we note that, using Eq. (30) (see Definition 3.1)

$$\eta_{\ell(x)}(x) \leq \eta_{\ell(x)}(B(x, r_p(x))) + L \cdot \mu(B^\circ(x, r_p(x)))^\alpha \leq \tau + \Delta + Lp^\alpha \tag{31}$$

for any pair (p, Δ) such that $p\Delta^2 \leq a$. In order to show the required containment, we need to find the maximum of the rhs in Eq. (31). Thus, we have the following objective[2]:

$$\max_{p,\Delta} (\tau + \Delta + L \cdot p^\alpha)$$

$$\textit{s.t. } p\Delta^2 \leq a$$

We solve this objective using Langrange multipliers. We denote the lagrangian as

$$\mathcal{L}(p, \Delta, s) := \tau + \Delta + L \cdot p^\alpha - s(p\Delta^2 - a).$$

Using Karush–Kuhn–Tucker conditions, we solve the following partial derivatives:

$$\frac{\partial \mathcal{L}}{\partial \Delta} = 1 - 2sp\Delta = 0$$

$$\frac{\partial \mathcal{L}}{\partial p} = L\alpha \cdot p^{\alpha-1} - s\Delta^2 = 0$$

These equations yields the following solutions: $p_0 = (2L\alpha)^{\frac{-2}{2\alpha+1}} \cdot a^{\frac{1}{2\alpha+1}}$ and $\Delta_0 = (2L\alpha)^{\frac{1}{2\alpha+1}} \cdot a^{\frac{\alpha}{2\alpha+1}}$.

Thus, for any $x \in \textbf{supp}(\mu)$ such that $m_\tau(x) \leq a$ and $\eta_{\ell(x)}(x) > \tau$ we get

$$\tau < \eta_{\ell(x)}(x) \leq \tau + \Delta_0 + L \cdot p^\alpha \tag{32}$$

This proofs the first part as claimed in the lemma.

For the second part we realize that

$$\tau - \Delta - Lp^\alpha \leq \eta_{\ell(x)}(B(x, r_p(x))) - L \cdot \mu(B^\circ(x, r_p(x)))^\alpha \leq \eta_{\ell(x)}(x) < \tau$$

if $x \in \textbf{supp}(\mu)$ is (p, Δ)-salient. Solving a constrained objective as above yields the p, Δ as obtained before. Thus, we get the second part as claimed in the lemma.

Third part is a straight-forward implication of the first and second parts of the lemma. □

[2]For this maximization, we omit the constraints on p and γ. The unique solution satisfies the constraints.

H.2. Margin bounds. In this subsection, we provide the proof of Theorem 5.3. Now, we prove the first part of Theorem 5.3.

PROOF OF THEOREM 5.3(a). Using Theorem 5.4, we know that with probability at $(1-\delta)$ over the training sample

$$\Pr_X[g_n(X) \neq ?, g_n(X) \neq \ell(X)] \leq \mu(\mathcal{X} \setminus \mathcal{X}_a^\circ)$$

for $a = \left(\frac{c \cdot \log \frac{n}{\delta}}{n}\right)$.

Thus, we need to bound $\mu(\mathcal{X} \setminus \mathcal{X}_a^\circ)$. But we can, alternately, write $(\mathcal{X} \setminus \mathcal{X}_a^\circ)$ as $\{x \in \textbf{supp}(\mu) : s_\tau(x) \leq a\}$ using the definition of $\mathcal{X}_a^\circ$. Using Lemma H.3, we have

$$\{x \in \textbf{supp}(\mu) : s_\tau(x) \leq a\} \subseteq \left\{x \in \textbf{supp}(\mu) : \tau - \Delta_0 - L \cdot p_0^\alpha \leq \eta_{\ell(x)}(x) \leq \tau + \Delta_0 + L \cdot p_0^\alpha\right\},$$

where $p_0 = (2L\alpha)^{\frac{-2}{2\alpha+1}} \cdot a^{\frac{1}{2\alpha+1}}$ and $\Delta_0 = (2L\alpha)^{\frac{1}{2\alpha+1}} \cdot a^{\frac{\alpha}{2\alpha+1}}$. Thus[3],

$$\begin{aligned}
\mu(\mathcal{X}_0 \setminus \mathcal{X}_a^\circ) &\leq \mu\left(\left\{x \in \textbf{supp}(\mu) : \tau - \Delta_0 - L \cdot p_0^\alpha \leq \eta_{\ell(x)}(x) \leq \tau + \Delta_0 + L \cdot p_0^\alpha\right\}\right) \\
&= \mu\left(\left\{x \in \textbf{supp}(\mu) : |\eta_{\ell(x)}(x) - \tau| \leq \Delta_0 + L \cdot p_0^\alpha\right\}\right) \\
&\leq C \cdot (\Delta_0 + L \cdot p_0^\alpha)^\beta \qquad (33) \\
&= CC_0 \cdot a^{\frac{\alpha\beta}{2\alpha+1}} \qquad (34)
\end{aligned}$$

In Eq. (33), we apply the condition of β-margin on the measure μ. In Eq. (**??**), we use $C_0 = \frac{L^\beta(\alpha+1)^\beta}{(L^2\alpha^2)^{\frac{\alpha\beta}{2\alpha+1}}}$. Plugging in $a = \frac{c \log \frac{n}{\delta}}{n}$ in Eq. (34), with probability at least $1-\delta$, we have

$$\Pr_X[g_n(X) \neq ?, g_n(X) \neq \ell(X)] \leq D' \cdot \left(\frac{c \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}}$$

where we use $D' = CC_0$. This gives the first part of Theorem 5.3(a).

Now, we consider the second part of the theorem. Using Theorem 5.4, we know that with probability at $(1-\delta)$ over the training sample

$$\Pr_X[X \in \mathcal{X}_\tau, g_n(X) = ?] \leq \mu(\mathcal{X} \setminus \mathcal{X}_a^+)$$

for $a = \left(\frac{c \cdot \log \frac{n}{\delta}}{n}\right)$. Using the definition of the set $\mathcal{X}_a^+$, we can rewrite $(\mathcal{X} \setminus \mathcal{X}_a^+)$ as $\left\{x \in \textbf{supp}(\mu) : m_\tau(x) \leq a, \eta_{\ell(x)}(x) \geq \tau\right\}$. Now, using Lemma H.3, we

$$\begin{aligned}
\mu\left(\mathcal{X} \setminus \mathcal{X}_a^+\right) &= \mu\left(\left\{x \in \textbf{supp}(\mu) : m_\tau(x) \leq a, \eta_{\ell(x)}(x) \geq \tau\right\}\right) \\
&\leq \mu\left(\left\{x \in \textbf{supp}(\mu) : \tau \leq \eta_{\ell(x)}(x) \leq \tau + \Delta_0 + L \cdot p_0^\alpha\right\}\right) \\
&\leq \mu\left(\left\{x \in \textbf{supp}(\mu) : |\eta_{\ell(x)}(x) - \tau| \leq \Delta_0 + L \cdot p_0^\alpha\right\}\right) \\
&\leq D' \cdot a^{\frac{\alpha\beta}{2\alpha+1}} \qquad (35)
\end{aligned}$$

where D' is same as above. Now, plugging $a = \frac{c \log \frac{n}{\delta}}{n}$ in Eq. (34), with probability at least $1-\delta$, we have

$$\Pr_X[X \in \mathcal{X}_\tau, g_n(X) = ?] \leq D' \cdot \left(\frac{c \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}}$$

[3]We note that $\Delta_0 = 2L\alpha p_0^\alpha$.

This gives the first part of Theorem 5.3(a). Similarly, we get the following bound

$$\Pr_X[x \in \mathcal{X} \setminus \mathcal{X}_\tau,\, g_n(X) \neq ?] \leq D' \cdot \left(\frac{c \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}}$$

Thus, we have have proven all the bounds as claimed in the theorem.

□

PROOF OF THEOREM 5.3(b). Before we prove the main theorem, we state the following important observation:

LEMMA H.4. *Let the classifier $g_n : \mathcal{X} \to \mathcal{Y} \cup \{?\}$ be the result of applying the adaptive NN procedure of Algorithm 2 with n points chosen i.i.d. from P and with confidence parameter $\delta > 0$ and rejection threshold $\tau > \frac{1}{2}$. Pick an arbitrary input $x \in \mathcal{X}$. Now, consider the following propositions:*

p_1: For $x \in \mathcal{X}$, $\frac{C'}{s_\tau(x)} \max\left(\log \frac{1}{s_{\tau(x)}}, \log \frac{1}{\delta}\right) > n$.

p_1': For $x \in \mathcal{X}$, $\frac{C'}{m_\tau(x)} \max\left(\log \frac{1}{m_{\tau(x)}}, \log \frac{1}{\delta}\right) > n$.

p_2: Let $\mathcal{B}$ be any class of measurable subsets of $\mathcal{X}$ of VC dimension d_0. There exists $B \in \mathcal{B}$ and an integer k such that

$$\mu(B) \geq \frac{k}{n} + \frac{c_0}{n} \max\left(k, d_0 \log \frac{n}{\delta}\right) \implies \mu_n(B) \geq \frac{k}{n}$$

p_3: Let $\mathcal{C}$ be a class of subsets of $\mathcal{X}$ with VC dimension d_0. There exists $C \in \mathcal{C}$ and $i \in \mathcal{Y}$, such that

$$|\hat{\eta}_i(C) - \eta_i(C)| > \Delta(n, \#_n(C), \delta)$$

Then, the following holds:

$$\mathbf{1}(g_n(x) \neq ?,\, g_n(x) \neq \ell(x)) \leq \mathbf{1}(\mathsf{p}_1 \text{ holds}) + \mathbf{1}(\mathsf{p}_2 \text{ holds}) + \mathbf{1}(\mathsf{p}_3 \text{ holds}), \tag{36}$$

$$\mathbf{1}(x \in \mathcal{X}_\tau,\, g_n(x) = ?) \leq \mathbf{1}(\mathsf{p}_1' \text{ holds})\, \mathbf{1}(x \in \mathcal{X}_\tau) + \mathbf{1}(\mathsf{p}_2 \text{ holds}) + \mathbf{1}(\mathsf{p}_3 \text{ holds}), \tag{37}$$

$$\mathbf{1}(x \in \mathcal{X} \setminus \mathcal{X}_\tau,\, g_n(x) \neq ?) \leq \mathbf{1}(\mathsf{p}_1' \text{ holds})\, \mathbf{1}(x \in \mathcal{X} \setminus \mathcal{X}_\tau) + \mathbf{1}(\mathsf{p}_2 \text{ holds}) + \mathbf{1}(\mathsf{p}_3 \text{ holds}). \tag{38}$$

PROOF. We would prove the first claim and argue that the second and third claims follow with similar arguments. First, note that if $g_n(x) = ?$ or $g_n(x) = \ell(x)$, then the upper bound holds trivially.

Now, consider the case when $g_n(x) \neq ?$ and $g_n(x) \neq g_\tau^*(x)$. Then, $\mathbf{1}(g_n(x) \neq g^*(x)) = 1$. If any of the three propositions– $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ hold then Eq. (36) holds. If none of the propositions hold, then we know that the claims of Lemma A.1 and Lemma A.3 hold in addition to the condition for Lemma H.2. As shown in Lemma H.2, either $g_n(x) = ?$ or $g_n(x) = \ell(x)$. But this contradicts the assumed case. Thus, at least one of the three propositions have to be negated. But then,

$$\mathbf{1}(\mathsf{p}_1 \text{ holds}) + \mathbf{1}(\mathsf{p}_2 \text{ holds}) + \mathbf{1}(\mathsf{p}_3 \text{ holds}) \geq 1$$

which gives the first claim.

For the second claim we note that if $\mathcal{X} \in \mathcal{X}_\tau$ and the propositions–p_1', p_2 and p_3 fail, then using Lemma H.1 we have $g_n(x) = \ell(x)$. Rest of the argument follows the proof of the first claim.

For the third claim we note that negation of p_1', p_2, and p_3 are the exact set of premises for Lemma H.1 to hold, and thus $g_n(x) = ?$ when $x \in (\mathcal{X} \setminus \mathcal{X}_\tau)$. This could be sufficiently used to proof the claim in the third part of the lemma. □

Now, we would proof the main claim of Theorem 5.3(b).

First, we define the point-wise error of a classifier g_n on an input $x \in \mathcal{X}$ as

$$\mathcal{R}_e(g(x)) = \mathbf{1}(g_n(x) \neq ?,\, g_n(x) \neq \ell(x))$$

Taking expectation over the training data S_n gives

$$\mathop{\mathbb{E}}_{S_n}[\mathcal{R}_e(g_n(x))] \leq \mathop{\mathbb{E}}_{S_n}[\mathbf{1}\,(g_n(x) \neq ?,\, g_n(x) \neq \ell(x))] \tag{39}$$

Now we invoke Lemma A.1 and Lemma A.3 to conclude

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S_n}[\mathbf{1}\,(g_n(x) \neq \ell(x),\, g_n(x) \neq ?)] \\
&\leq \mathop{\mathbb{E}}_{S_n}\left[\mathbf{1}\left(s_\tau(x) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right)\right] + \mathop{\mathbb{E}}_{S_n}[\mathbf{1}\,(\mathsf{p}_2 \text{ holds})] + \mathop{\mathbb{E}}_{S_n}[\mathbf{1}\,(\mathsf{p}_3 \text{ holds})] && (40)\\
&\leq \mathbf{1}\left(s_\tau(x) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right) + \Pr_{S_n}[\mathsf{p}_2 \text{ holds}] + \Pr_{S_n}[\mathsf{p}_3 \text{ holds}] && (41)\\
&\leq \mathbf{1}\left(s_\tau(x) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right) + \delta^2 && (42)
\end{aligned}$$

We can bound $\mathbb{E}_{S_n}[\mathbf{1}\,(g_n(x) \neq \ell(x),\, g_n(x) \neq ?)]$ in a straight-forward way using Lemma H.4. In Eq. (40), we use a simplified version of proposition p_1. In Eq. (41), we note that $\mathbf{1}\left(s_\tau(x) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right)$ is a constant wrt the expectation over the choice of training data S_n and for $i = 2, 3$, $\mathbb{E}_{S_n}[\mathbf{1}\,(\mathsf{p}_i \text{ holds})] = \Pr_{S_n}[\mathsf{p}_i \text{ holds}]$ by definition. In Eq. (42), we note that for all $i = 2, 3$, $\Pr_{S_n}[\neg\, \mathsf{p}_i \text{ holds}] \geq 1 - \frac{\delta^2}{2}$.

Now, combining Eq. (39) and Eq. (42) and taking the expectation with respect to the random variable $X \sim \mu$, we obtain

$$\mathop{\mathbb{E}}_{X}\left[\mathop{\mathbb{E}}_{S_n}[\mathcal{R}_e(g_n(X))]\right] \leq \mathop{\mathbb{E}}_{X}\left[\mathbf{1}\left(s_\tau(X) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right)\right] + \delta^2. \tag{43}$$

Note that

$$\begin{aligned}
\mathop{\mathbb{E}}_{X}\left[\mathop{\mathbb{E}}_{S_n}[\mathcal{R}_e(g_n(X))]\right] &= \mathop{\mathbb{E}}_{S_n}\left[\mathop{\mathbb{E}}_{X}[\mathbf{1}\,(g_n(X) \neq \ell(x),\, g_n(X) \neq ?)]\right] \\
&= \mathop{\mathbb{E}}_{S_n}\left[\Pr_{X}[g_n(X) \neq ?, g_n(X) \neq \ell(X)]\right] \\
&= \mathop{\mathbb{E}}_{S_n}[\mathcal{R}_e(g_n)]
\end{aligned}$$

Thus, we can rewrite Eq. (43) as

$$\begin{aligned}
\mathop{\mathbb{E}}_{S_n}[\mathcal{R}_e(g_n)] &\leq \Pr_{X}\left[s_\tau(X) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right] + \delta^2 \\
&= \mu\left(s_\tau(X) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right) + \delta^2
\end{aligned}$$

By definition, we know that $\Pr_X\left[s_\tau(X) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right] = \mu\left(s_\tau(X) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right)$, and thus the upper bound follows above.

Using Lemma H.3 and Eq. (34), we get:

$$\mathop{\mathbb{E}}_{S_n}[\mathcal{R}_e(g_n)] \leq D' \cdot \left(\frac{C' \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}} + \delta^2$$

Now, we would proof the second part of Theorem 5.3(b). The result bounds the **type 1** abstention rate $\mathcal{R}_a^1$ of a classifier g_n.

We define the point-wise **type 1** abstention by a classifier g on an input $x \in \mathcal{X}_\tau$ as

$$\mathcal{R}_a^1(g(x)) = \mathbf{1}(x \in \mathcal{X}_\tau,\, g(x) = ?)$$

Taking the expectation over the training sample S_n gives

$$\mathop{\mathbb{E}}_{S_n}\left[\mathcal{R}_a^1(g_n(x))\right] \leq \mathop{\mathbb{E}}_{S_n}[\mathbf{1}(x \in \mathcal{X}_\tau,\, g(x) = ?)] \tag{44}$$

We can bound the RHS in Eq. (H.2) as follows

$$\mathop{\mathbb{E}}_{S_n}[\mathbf{1}(x \in \mathcal{X}_\tau,\, g(x) = ?)]$$

$$\leq \mathop{\mathbb{E}}_{S_n}[\mathbf{1}\left(\mathsf{p}_1 \text{ holds}\right)\mathbf{1}\left(x \in \mathcal{X}_\tau\right)] + \mathop{\mathbb{E}}_{S_n}[\mathbf{1}\left(\mathsf{p}_2 \text{ holds}\right)] + \mathop{\mathbb{E}}_{S_n}[\mathbf{1}\left(\mathsf{p}_3 \text{ holds}\right)] \tag{45}$$

$$\leq \mathbf{1}\left(m_\tau(x) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right)\mathbf{1}\left(x \in \mathcal{X}_\tau\right) + \mathop{\mathbb{E}}_{S_n}[\mathbf{1}\left(\mathsf{p}_2 \text{ holds}\right)] + \mathop{\mathbb{E}}_{S_n}[\mathbf{1}\left(\mathsf{p}_3 \text{ holds}\right)] \tag{46}$$

$$\leq \mathbf{1}\left(m_\tau(x) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right)\mathbf{1}\left(x \in \mathcal{X}_\tau\right) + \delta^2 \tag{47}$$

Eq. (45) follows using Lemma H.4. We use the simplified way to write p_1 in Eq. (46). Eq. (47) follows as derived in Eq. (42).

Now, combining and Eq. (47) and taking the expectation with respect to the random variable $X \sim \mu$ restricted to the set $\mathcal{X}_\tau$, we obtain

$$\begin{aligned}
\mathop{\mathbb{E}}_{X}\left[\mathop{\mathbb{E}}_{S_n}\left[\mathcal{R}_a^1(g_n)\right]\right] &\leq \mathop{\mathbb{E}}_{X}\left[\mathbf{1}\left(m_\tau(X) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right)\mathbf{1}\left(X \in \mathcal{X}_\tau\right)\right] + \delta^2 \\
&\leq \mathop{\mathbb{E}}_{X}\left[\mathbf{1}\left(\eta_{\ell(X)}(X) \geq \tau, m_\tau(X) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right)\right] + \delta^2 \\
&\leq \Pr_{X}\left[\eta_{\ell(X)}(X) \geq \tau, m_\tau(X) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right] + \delta^2 \\
&= \mu\left(\eta_{\ell(X)}(X) \geq \tau, m_\tau(X) < \frac{C' \cdot \log \frac{n}{\delta}}{n}\right) + \delta^2
\end{aligned}$$

Now, using Lemma H.3 and Eq. (35) we get the desired bound

$$\mathop{\mathbb{E}}_{S_n}\left[\mathcal{R}_a^1(g_n)\right] \leq D' \cdot \left(\frac{C' \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}} + \delta^2.$$

For the third part of Theorem 5.3(b) that bounds the **type 2** abstention rate $\mathcal{R}_a^2$ of a classifier g_n, we define the point-wise **type 2** abstention by a classifier on an input $x \in \mathcal{X} \setminus \mathcal{X}_\tau$ as

$$\mathcal{R}_a^2(g(x)) = \mathbf{1}\left(x \in \mathcal{X} \setminus \mathcal{X}_\tau, g(x) \neq ?\right)$$

Rest of the proof follows similar arguments as shown for the proof of $\mathcal{R}_a^1$, and thus we get

$$\mathop{\mathbb{E}}_{S_n}\left[\mathcal{R}_a^2(g_n)\right] \leq D' \cdot \left(\frac{C' \log \frac{n}{\delta}}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}} + \delta^2.$$

This completes the proof of Theorem 5.3(b).

□

REFERENCES

[1] ALBERTI, G. S. (2005). Geometric Measure Theory.

[2] AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Annals of Statistics* **35** 608-633.

[3] BALSUBRAMANI, A., DASGUPTA, S., FREUND, Y. and MORAN, S. (2019). An adaptive nearest neighbor rule for classification. In *NeurIPS*.

[4] BARTLETT, P. L. and WEGKAMP, M. H. (2008). Classification with a Reject Option using a Hinge Loss. *J. Mach. Learn. Res.* **9** 1823-1840.

[5] CHAROENPHAKDEE, N., CUI, Z., ZHANG, Y. and SUGIYAMA, M. (2021). Classification with Rejection Based on Cost-sensitive Classification. *ArXiv* **abs/2010.11748**.

[6] CHAUDHURI, K. and DASGUPTA, S. (2010). Rates of convergence for the cluster tree. In *NIPS*.

[7] CHAUDHURI, K. and DASGUPTA, S. (2014). Rates of Convergence for Nearest Neighbor Classification. In *NIPS*.

[8] CHOW, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* **16** 41-46. https://doi.org/10.1109/TIT.1970.1054406

[9] CHOW, C. K. (1957). An optimum character recognition system using decision functions. *IRE Trans. Electron. Comput.* **6** 247-254.

[10] CORTES, C., DESALVO, G., GENTILE, C., MOHRI, M. and YANG, S. (2018). Online Learning with Abstention. In *ICML*.

[11] CORTES, C., DESALVO, G. and MOHRI, M. (2016). Learning with Rejection. In *ALT*.

[12] CORTES, C., DESALVO, G. and MOHRI, M. (2016). Boosting with Abstention. In *NIPS*.

[13] COVER, T. and HART, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13** 21-27. https://doi.org/10.1109/TIT.1967.1053964

[14] DEVROYE, L., GYÖRFI, L., KRZYŻAK, A. and LUGOSI, G. (1994). On the Strong Universal Consistency of Nearest Neighbor Regression Function Estimates. *Annals of Statistics* **22** 1371-1385.

[15] EL-YANIV, R. and WIENER, Y. (2010). On the Foundations of Noise-free Selective Classification. *J. Mach. Learn. Res.* **11** 1605-1641.

[16] FIX, E. and HODGES, J. L. (1989). Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties. *International Statistical Review* **57** 238.

[17] FUMERA, G. and ROLI, F. (2002). Support Vector Machines with Embedded Reject Option. In *SVM*.

[18] GANGRADE, A., KAG, A., CUTKOSKY, A. and SALIGRAMA, V. (2021). Online Selective Classification with Limited Feedback. In *NeurIPS*.

[19] GANGRADE, A., KAG, A. and SALIGRAMA, V. (2021). Selective Classification via One-Sided Prediction. In *AISTATS*.

[20] GEIFMAN, Y. and EL-YANIV, R. (2017). Selective Classification for Deep Neural Networks. In *NIPS*.

[21] GEIFMAN, Y. and EL-YANIV, R. (2019). SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *ICML*.

[22] GRANDVALET, Y., RAKOTOMAMONJY, A., KESHET, J. and CANU, S. (2008). Support Vector Machines with a Reject Option. In *NIPS*.

[23] GYÖRFI, L. (1981). The rate of convergence of kn-NN regression estimates and classification rules. *IEEE Trans. Inf. Theory* **27** 362-364.

[24] HEINONEN, J. (2001). *Lectures on Analysis on Metric Spaces*. https://doi.org/10.1007/978-1-4613-0131-8

[25] HELLMAN, M. E. (1970). The Nearest Neighbor Classification Rule with a Reject Option. *IEEE Trans. Syst. Sci. Cybern.* **6** 179-185.

[26] HERBEI, R. and WEGKAMP, M. H. (2006). Classification with reject option. *Canadian Journal of Statistics* **34** 709-721. https://doi.org/10.1002/cjs.5550340410

[27] KALAI, A. T., KANADE, V. and MANSOUR, Y. (2009). Reliable Agnostic Learning. *J. Comput. Syst. Sci.* **78** 1481-1495.

[28] 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.

[29] LEI, J. (2014). Classification with confidence. *Biometrika* **101** 755-769.

[30] LI, L., LITTMAN, M. L., WALSH, T. J. and STREHL, A. L. (2008). Knows what it knows: a framework for self-aware learning. *Machine Learning* **82** 399-443.

[31] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth Discrimination Analysis. *Annals of Statistics* **27** 1808-1829.

[32] NEU, G. and ZHIVOTOVSKIY, N. (2020). Fast Rates for Online Prediction with Abstention. In *COLT*.

[33] NI, C., CHAROENPHAKDEE, N., HONDA, J. and SUGIYAMA, M. (2019). On the Calibration of Multiclass Classification with Rejection. In *NeurIPS*.

[34] RAMASWAMY, H. G., TEWARI, A. and AGARWAL, S. (2018). Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics* **12** 530-554.
[35] RIVEST, R. L. and SLOAN, R. (1988). Learning Complicated Concepts Reliably and Usefully (Extended Abstract). In *Proceedings of the Seventh AAAI National Conference on Artificial Intelligence*. *AAAI'88* 635–640. AAAI Press.
[36] SAYEDI-ROSHKHAR, A. S., ZADIMOGHADDAM, M. and BLUM, A. (2010). Trading off Mistakes and Don't-Know Predictions. In *NIPS*.
[37] STONE, C. J. (1977). Consistent Nonparametric Regression. *Annals of Statistics* **5** 595-620.
[38] SZITA, I. and SZEPESVARI, C. (2011). Agnostic KWIK learning and efficient approximate reinforcement learning. In *COLT*.
[39] TSYBAKOV, A. B. (2003). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics* **32** 135-166.
[40] VAN DER VAART, A. and WELLNER, J. A. (2009). A note on bounds for VC dimensions. *Institute of Mathematical Statistics collections* **5** 103-107.
[41] WIENER, Y. and EL-YANIV, R. (2011). Agnostic Selective Classification. In *NIPS*.
[42] YUAN, M. and WEGKAMP, M. H. (2010). Classification Methods with Reject Option Based on Convex Risk Minimization. *J. Mach. Learn. Res.* **11** 111-130.
[43] ZHANG, C. and CHAUDHURI, K. (2016). The Extended Littlestone's Dimension for Learning with Mistakes and Abstentions. In *COLT*.