A Two-Stage Active Learning Algorithm for k-Nearest Neighbors

Nick Rittler [1] **Kamalika Chaudhuri** [1]

Abstract

k-nearest neighbor classification is a popular non-parametric method because of desirable properties like automatic adaption to distributional scale changes. Unfortunately, it has thus far proved difficult to design active learning strategies for the training of local voting-based classifiers that naturally retain these desirable properties, and hence active learning strategies for k-nearest neighbor classification have been conspicuously missing from the literature. In this work, we introduce a simple and intuitive active learning algorithm for the training of k-nearest neighbor classifiers, the first in the literature which retains the concept of the k-nearest neighbor vote at prediction time. We provide consistency guarantees for a modified k-nearest neighbors classifier trained on samples acquired via our scheme, and show that when the conditional probability function $\mathbb{P}(Y = y|X = x)$ is sufficiently smooth and the Tsybakov noise condition holds, our actively trained classifiers converge to the Bayes optimal classifier at a faster asymptotic rate than passively trained k-nearest neighbor classifiers.

1. Introduction

In active learning, the learner has access to unlabeled training data from the target distribution, and can ask an annotator to selectively label a subset of this data. The learner's goal is to develop an accurate classifier while querying as few labels as possible. A significant body of work has demonstrated the power of active learning to lower label complexities versus passive approaches, wherein every point from the underlying distribution is labeled (Balcan et al., 2006; Hanneke, 2007; Beygelzimer et al., 2008; Zhang & Chaudhuri, 2014). In cases where human annotation is a driving cost in supervised training, such guarantees are particularly relevant. This work considers active learning when the underlying classifier is the k-nearest neighbor (k-NN) classifier.

The previous work on active learning for nearest neighbors algorithms is surprisingly sparse. (Dasgupta, 2012) gives asymptotic consistency guarantees for k-NN under a wide class of active strategies, but does not provide any finite sample guarantees. (Kontorovich et al., 2018) uses a compression approach to give guarantees for a 1-nearest neighbor classifier trained on a subset of an actively acquired training sample; while their algorithm does not use many label queries, it does not come with consistency guarantees ensuring convergence to the Bayes-risk in the large sample limit, and also relies on a somewhat computationally intensive algorithm to determine which labels are to be queried. (Njike & Siebert, 2021) gives an algorithm that, similar to (Kontorovich et al., 2018), outputs the prediction of a 1-NN classifier trained on sample points which have been actively determined to be informative. That said, the schemes of both (Kontorovich et al., 2018) and (Njike & Siebert, 2021) do not retain the natural adaptivity to scale changes in the distribution and local de-noising properties of the k-NN vote – which, for one, allow for local estimation of conditional probabilities – instead resorting to various complex algorithmic subroutines to ensure good performance. Our main contribution is thus the exposition of the only active learning scheme that retains the concept of the k-NN vote at prediction time, resulting in a much more direct approach that admits significant convergent rate speed-ups.

Active learning usually relies on some inductive bias or distributional assumption that can exploited to reduce the need for labeled data. The canonical example is that of learning a linear classifier in the realizable case, wherein the a priori knowledge allows the learner to avoid querying examples when one of the possible labelings would be inconsistent with any possible linear classifier and the labels queried so far. The main challenge in actively learning non-parametric classifiers such as nearest neighbors is that this inductive bias is harder to characterize. To facilitate our investigation, we operate under one of the most common distributional assumptions in the non-parametric active learning literature – a smoothness condition on the conditional probability function $\mathbb{P}(Y = y|X = x)$ (Castro & Nowak, 2008; Minsker, 2012). Specifically, we assume that the conditional proba-

[1]Department of Computer Science and Engineering, University of California - San Diego, San Diego, CA, USA. Correspondence to: Nick Rittler <nrittler@ucsd.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

bility function has a property that is a necessary condition for Hölder continuity under mild regularity assumptions.

A natural algorithm in this regime is to look at the regions of space where multiple labels can be found, and then sample those regions more closely to get better nearest neighbor estimates. Based on this principle, we propose an intuitive and computationally efficient approach to determine which labels are to be actively queried.

We show that our active learning algorithm comes with natural consistency guarantees – in the sense that the expected risk of the learned classifier converges to the Bayes-risk as the number of label queries grows to infinity. Additionally, under the Hölder continuity of the conditional probability function $\eta(x) := \mathbb{P}(Y = 1|X = x)$, we provide finite sample guarantees for a modified k-NN classifier trained on samples acquired via our scheme. We discuss general conditions under which these guarantees are tighter than the corresponding passive guarantees, a discussion which quite notably for the non-parametric active learning literature does not rely on low-noise assumptions.

We continue by showing that under the Tsybakov noise condition $\Pr_X(|\eta(X) - \frac{1}{2}| \leq t) \leq Ct^\beta$ (Audibert & Tsybakov, 2005), our actively trained classifiers converge at a significantly faster asymptotic rate than passively trained k-NN classifiers. Roughly speaking, assuming η is α-Hölder continuous on a d dimensional instance space, previous work has shown the fraction of examples on which the learnt classifier disagrees with the Bayes optimal is $\tilde{O}\left((k^{-1/2} + (k/n)^{\alpha/d})^\beta\right)$ for passively trained k-NN after n samples; this improves to $\tilde{O}\left((k^{-1/2} + (f_{k,n} \cdot k/n)^{\alpha/d})^\beta\right)$ for $f_{k,n} \in \tilde{O}\left((k^{-1/2} + (k/n + n^{-1/2})^{\alpha/d})^\beta\right)$ for our actively trained classifier. Our results are a consequence of a new style of guarantee – controlling the disagreement with the Bayes optimal classifier – that is novel in the non-parametric active learning literature, and different from the standard work on rates of convergence of expected risk (Castro & Nowak, 2008; Locatelli et al., 2017). We finish with some discussion of the optimality of the scheme.

2. Related Work

The study of learning in settings where the learner possesses some power over which samples it uses in training has long been of interest to the learning theory community. Though alternative label query models have been explored (Angluin, 1987; Awasthi & Kanade, 2012), most attention has been given to the standard active learning model, wherein the learner is assumed to have functionally unlimited access to unlabeled samples from the target distribution, but must choose which of these examples are to be labeled in the course of training a classifier. Much of this work was originally done in the parametric setting, whereby the learner was tasked with choosing an approximately optimal classifier from a hypothesis class (Balcan et al., 2006; Dasgupta et al., 2007; Beygelzimer et al., 2008). Because it has long been known that there are settings in which active strategies cannot beat passive ones in terms of label complexity (Kääriäinen, 2006), investigation of specific conditions under which active strategies can beat passive ones has been the norm (Balcan & Urner, 2015). The most general description of settings in which parametric active strategies improve over passive strategies is due to (Hanneke & Yang, 2014).

The study of active learning for the training of non-parametric classifiers has lead to a well-developed theory under smoothness of the conditional probability function and low-noise conditions, which are the common surrogates for inductive bias in the non-parametric setting. Perhaps the most notable work in this litany is (Castro & Nowak, 2008). Their work establishes minimax lower bounds on the expected excess risk (over the Bayes-risk) of an actively trained non-parametric classifier, and provides an active scheme for the training of local polynomial classifiers that achieves this lower bound up to log factors. The approximate optimality of their algorithm however assumed access to underlying noise parameters, such as Hölder and Tsbakov parameters. (Locatelli et al., 2017) resolved this issue with an adaptive algorithm that achieves the same guarantees as (Castro & Nowak, 2008) without a priori access to noise parameters. (Minsker, 2012) gives a non-parametric active learning scheme for the training of classifiers based around estimates of the conditional probability function that are locally constant on hypercubes partitioning the instance space. This work is not directly comparable given that k-NN estimates of the conditional probability function are locally constant on a more complex geometric structure that is data-dependent.

For the study of active NN algorithms in particular, the authors are not aware of significant contributions other than the aforementioned (Kontorovich et al., 2018), (Dasgupta, 2012) and (Njike & Siebert, 2021). (Balsubramani et al., 2019) gives guarantees for a nearest neighbor algorithm that adaptively sets k across instance space depending on the local agreement of nearest neighbor votes, but this algorithm is passive with respect to data collection.

3. Preliminaries

3.1. Setting

We study a binary classification setting where instances come from a metric space $(\mathcal{X}, \rho)$, and the label space is $\mathcal{Y} = \{0, 1\}$. We assume that we have access to labeled data from some data generating measure $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$. $\mathbb{P}$ is the

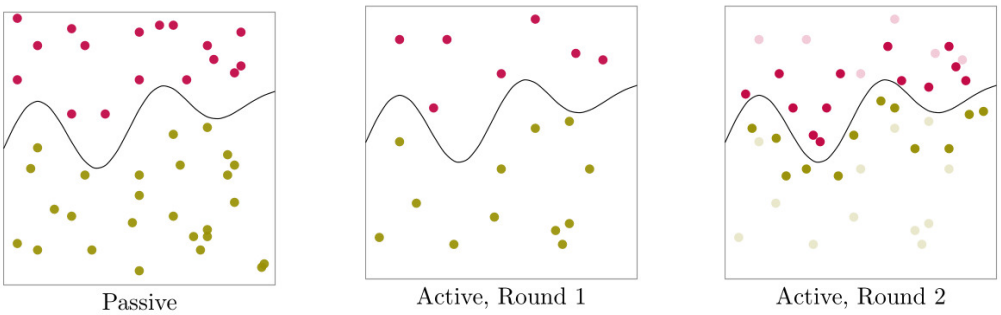

Figure 1. An illustration of the ideal outcome of the two-round active algorithm. In the "passive" setting, points are simply taken from the data-generating distribution. As is often the case in active learning, the goal is to eliminate parts of space where classification is likely easy. In the right-most figure, lower opacity designates first-round samples.

product of some Borel-regular measure μ,[1] the marginal of $\mathbb{P}$ over the instance space $\mathcal{X}$, and $\eta : \mathcal{X} \to [0, 1]$, which denotes the conditional probability function $\mathbb{P}(Y = 1|X = x)$. We assume that μ has full support on $\mathcal{X}$.

We assume that after n i.i.d. samples from $\mathbb{P}$, which are labeled by definition, we enter a second phase of the sampling with an additional label budget of m. In this phase, we may specify any $R \subseteq \mathcal{X}$ with $\mu(R) > 0$, and subsequently rejection sample from R, i.e. sample from the product of μ conditioned on R and η. The main sampling algorithm, formalized in Algorithm 1, makes use of this power to provide a specialized training set for a modified NN classifier, introduced in Section 2.3. Figure 1 gives an idea of the ideal outcome of the strategy.

3.2. k-Nearest Neighbors and Bayes Optimal Classifier

We use this section to briefly review the standard k-nearest neighbors algorithm. Given a set of points in a training set $\mathcal{S} := \{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} : i \in [n]\}$, the standard k-NN classifier can be stated as

$$g_{n,k}(x) := \begin{cases} 1, & \text{if } \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

where $Y_{(i)}(x)$ is the label of this i^{th} nearest training instance from x under ρ. In this way, k-NN provide a sort of local "vote" over what the prediction should be, with prediction at x going to the winning class.

We assume that nearest neighbor tie breaking is done by drawing independent uniform random variables from $[0, 1]$, with preference going to lower values of this uniform draw. The event that two sample points are equidistant from a new point x is a measure 0 event (over the draw of x) for many distributions, but we will make statements concerning all $x \in \mathcal{X}$ simultaneously, for which tie-breaking will be more impactful; see Appendix A for a primer on how this tie-breaking rule enters the analysis.

The Bayes optimal classifier is simply the best guess for classification given full knowledge of η, that is

$$g^*(x) := \begin{cases} 1, & \text{if } \eta(x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

There is well-established literature on the asymptotic relationship of k-NN and g^*. Perhaps the most intuitively useful is that when $k_n \to \infty$, $\frac{k_n}{n} \to 0$, and the classifier is in Euclidean space,

$$\mathbb{P}\{g_{n,k_n}(X) \neq Y\} \to \mathbb{P}\{g^*(X) \neq Y\}$$

almost surely over the sampling (Devroye et al., 1994). Weaker versions exist on general metric spaces (Cover, 1968; Stone, 1977).

3.3. A Modified k-Nearest Neighbors Classifier

We introduce and study a modified k-NN classifier to facilitate the investigation of our particular setting. The modification helps analytically by making it likely that the nearest neighbors of a new point x form an unbiased estimate of conditional probability in a neighborhood x in the second sampling round, as well as alleviating complications arising from dependencies carried between sampling rounds. For a given sample of size $n + m$, and a regions $R \subseteq R^+ \subseteq \mathcal{X}$ with $\mu(R^+) > 0$, we consider the modified k-NN classifier

$$g_{n,m,k}(x) := \begin{cases} \mathbb{1}[\frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x; R^+) \geq \frac{1}{2}], & \text{if } x \in R \\ \mathbb{1}[\frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x) \geq \frac{1}{2}], & \text{otherwise} \end{cases}$$

Here, we assume that R^+ has been used as the acceptance region in the second round of sampling, and its smaller cousin R is the set of points that use rejection samples for prediction. We use $Y_{(i)}(x; R^+)$ to denote the ith nearest neighbor to x out of all of samples rejection sampled from

[1] Thus, μ can measure all members of the σ-algebra generated by open sets under ρ.

R^+, and use $Y_{(i)}(x)$ to denote the ith nearest neighbor to x out of samples drawn from $\mathbb{P}$. Roughly speaking, the idea will be to treat prediction in R as a prediction problem on a new probability space with properties inherited from the original, larger space. We will see that if the region R^+ is well-chosen, and small enough under μ, this targeting procedure has the potential to be effective at improving label complexities.

3.4. Finite Sample Guarantees for k-NN in the Passive Setting

The main analytic tools used in this work are based around those introduced in (Chaudhuri & Dasgupta, 2014). We use this section to introduce some of the main analytic concepts of their work, as they are the basis for an intuitive understanding of our two-round sampling algorithm. In the passive setting, their results align with optimal rates of convergence for nonparametric classifiers under standard regularity conditions.

The core idea of their work is that the k-NN classifier may disagree with the Bayes optimal classifier in a region of space where the distribution is complex enough that n samples are not sufficient. The measure of this region, which is referred to as the "effective decision boundary", is a function of n and k, and is a controlling factor in the convergence of k-NN to the Bayes optimal classifier.

The most vital definition of (Chaudhuri & Dasgupta, 2014), is the smallest r such that the closed ball $B(x,r) := \{x^{'} \in \mathcal{X} : \rho(x,x^{'}) \leq r\}$ has mass at least p.

Definition 1. *For $p \in (0,1]$ the "probability radius" of a point x is the real number*

$$r(x;p) := \inf\{r | \mu\left(B(x,r)\right) \geq p\}.$$

The power of the probability radius is that it is adaptive to scale changes in μ, and so can uniformly be applied throughout instance space to anchor the definition of a region in which one can be confident that k-NN will predict correctly after seeing a certain number of samples. The "effective decision boundary" can then be stated as in terms of this quantity.

Definition 2. *For any $p \in (0,1]$ and $\Delta \in (0, \frac{1}{2}]$, the "effective interior" is the set*

$$\mathcal{X}^{+/-}_{p,\Delta} = \mathcal{X}^{+}_{p,\Delta} \cup \mathcal{X}^{-}_{p,\Delta},$$

where

$$\mathcal{X}^{+}_{p,\Delta} := \left\{ x \mid \eta(x) > \frac{1}{2} \ \wedge \ \eta(B(x,r)) - \frac{1}{2} \geq \Delta \right\},$$

$$\mathcal{X}^{-}_{p,\Delta} := \left\{ x \mid \eta(x) < \frac{1}{2} \ \wedge \ \frac{1}{2} - \eta(B(x,r)) \geq \Delta \right\}.$$

The "effective decision boundary" is the set

$$\partial_{p,\Delta} := \mathcal{X} \setminus \mathcal{X}^{+/-}_{p,\Delta}.$$

Here, $\eta(S) := \frac{1}{\mu(S)} \int_S \eta d\mu$, the probability of getting a 1 label conditional on an instance being in S with $\mu(S) > 0$. The main result is that the measure of the "effective decision boundary" controls the disagreement rate of the k-NN classifier and the Bayes optimal classifier with high probability.

Theorem 1 (**of** Chaudhuri and Dasgupta, 2014)**.** *Fix $\delta \in (0,1)$, and integers $k < n$. With probability $\geq 1 - \delta$ over the draw of n i.i.d. training samples from $\mathbb{P}$,*

$$\Pr_{X \sim \mu}\big(g^*(X) \neq g_{n,k}(X)\big) \leq \mu(\partial_{p,\Delta}) + \delta,$$

when for a constant c_δ we have

$$p = c_\delta \cdot \frac{k}{n+m}$$
$$\Delta = \min\left(\frac{1}{2}, \sqrt{\frac{1}{k} \cdot \log(2/\delta)}\right).$$

This work will largely focus on the role of the parameter p in determining guarantees, which intuitively corresponds to investigating the "locality" of the voting procedure deciding the prediction of the classifier. As such, we will fix Δ to its value in the above theorem statement throughout. The constant c_δ will also continue to pop up in our analysis, and so we make it explicit: $c_\delta := (1 - \sqrt{(4/k) \cdot \log(2/\delta)})^{-1}$. We describe this as a constant, as for the regimes we will be interested in, it will be $O(1)$; we reserve further discussion of this for later.

4. Active Algorithm

A classic algorithmic idea in both parametric and nonparametric active learning is to iteratively eliminate parts of the instance space where the learner can be sufficiently confident that no more samples are needed. Our algorithm operates under this idea as well. However, whereas the learner can directly exploit inductive bias to execute such a strategy in the in parametric setting, our algorithm implicitly relies on the smoothness of the distribution. We use the first round of sampling, in part, to detect a subset of instance space in which the k-NN agree strongly enough that the learner can be confident its predictions in this region are already correct. The learner then spends most of its labeling budget in the complement of this region in the second round of sampling.

More precisely, we first propose an algorithm that, based on the first round of samples, detects regions of space where the average labels of the k-nearest neighbors is close to

$1/2$; these are the regions of space where the classifier is not highly confident of the predicted label. We then show that under a natural smoothness assumption on η, which we describe below, that these are indeed the hard regions – in other words, the nearest neighbor classifier is expected to be accurate in the complement of these regions. The region from which rejection sample in the second round is a slightly augmented version of these hard regions, expanded so as to contain the nearest neighbors of points in the hard regions as well.

4.1. Definition of Smoothness

As discussed earlier, such a strategy inherently requires a smoothness condition for η, as if the data distribution can exhibit arbitrary behavior, agreement of the k-NN votes does not imply anything about the nature of η at finer resolutions. The notion of smoothness for η under which we operate was introduced in (Chaudhuri & Dasgupta, 2014), and is designed to be evocative of traditional notions of Hölder continuity in more general metric spaces.

Definition 3. *For $\alpha > 0, L > 0$, we say the conditional probability function η is "(α, L)-smooth" in $(\mathcal{X}, \rho, \mu)$ if for all $x, x^{'} \in \mathcal{X}$,*

$$\left|\eta(x) - \eta(x^{'})\right| \leq L\mu\left(B^o(x, \rho(x, x^{'}))\right)^{\alpha},$$

where $B^o(x, \rho(x, x^{'}))$ denotes the open ball centered at x with radius $\rho(x, x^{'})$.

When $\mathcal{X} = \mathbb{R}^d$ and μ has a density with respect to the Lebesgue measure that is non-trivially bounded below, this is a necessary condition for the Hölder continuity of η (Chaudhuri & Dasgupta, 2014).

4.2. What Regions are Hard?

When smoothness of η holds, it is possible to construct empirical estimates of the regions of space in which k-NN needs extra samples to classify accurately. We use the following region as the basis for the area of instance space in which the learner uses the majority of its label budget in the second round of sampling.

Definition 4. *Given a confidence parameter $\delta \in (0, 1)$, and a first-round sample $\mathcal{S}_n$, the "estimated hard region" after n samples for the k-NN classifier is the set*

$$\hat{\mathscr{H}}_{n,k} := \left\{x \in \mathcal{X} \mid \left|\frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x) - \frac{1}{2}\right| < 3\Delta_{\text{UC}}\right\},$$

where

$$\Delta_{\text{UC}} := c_0\sqrt{\frac{d\log(n) + \log(1/\delta)}{k}}.$$

Here, c_0 is a universal constant, and d denotes the VC-dimension of the set of all sets of the form

$$B'(x', r', z') := \Big\{(x, z) \in \mathcal{X} \times [0, 1] \mid \rho(x, x') < r \veebar \rho(x, x') = r, z < z'\Big\}.$$

.

The set of $B'(x', r', z')$ is a natural object of analysis for nearest neighbor tie-breaking. Some standard settings where d is on the order of the VC-dimension of the set of closed balls are discussed in the Appendix.

4.3. Utility of the Estimated Hard Region

The "estimated hard region" has some extremely useful properties. As mentioned in Section 2.4, the complement of the effective decision boundary for passive nearest neighbors can be thought of as the region of space where the learner can be confident that k-NN will predict correctly after n samples. It is a fundamental lemma of ours that the "estimated hard region" approximates the effective boundary from the outside when η is sufficiently smooth.

Lemma 1. *For all $k < n \in \mathbb{N}$, and all $\delta \in (0, 1)$, if η is sufficiently (α, L)-smooth in the sense that*

$$2L\left(\frac{k}{n} + c_2\sqrt{\frac{d + \log(1/\delta)}{n}}\right)^{\alpha} \leq \Delta_{\text{UC}},$$

for some absolutely constant c_2, then with probability $\geq 1 - \delta^2/8$ over the draw of n i.i.d. samples from $\mathbb{P}$,

$$\partial_{c_\delta \frac{k}{n}, \Delta} \subseteq \hat{\mathscr{H}}_{n,k}.$$

In other words, it's very likely that after the first round of sampling, the complement of the estimated hard region will be a subset of instance space in which the k-NN does not need more samples to classify well.

4.4. Specification of the Rejection Region

The the actual region used for rejection sampling is a modification of the estimated hard region, a larger version of this set which facilitates the analysis. The idea is that by augmenting the hard region, we can be confident that the nearest neighbors of any point in the estimated hard region are contained in this augmented region. This ensures that the expectation of the nearest neighbors vote is the conditional probability of 1 in nearest neighbors balls, an important property in our analysis.

Definition 5. *Given a sample $\mathcal{S}$ of size n and a parameter $\zeta \geq 0$, the "augmented estimated hard region" is the set*

$$\hat{\mathscr{H}}^{+}_{n,k} := \bigcup_{x \in \hat{\mathscr{H}}_{n,k}} B^o_{(\bar{k}_\zeta + 1)}(x),$$

where

$$\bar{k}_\zeta = \left\lceil c_\delta(1+\zeta)k + \Theta\left(\sqrt{n\left(\log(4/\delta)+d\right)}\right)\right\rceil,$$

and $B^o_{(\bar{k}_\zeta+1)}(x)$ denotes the largest open ball of centered at x with radius distance from x to $\bar{k}_\zeta + 1$-NN of x under ρ.

Thus, the augmented hard region contains all the points in the hard region and points close by to these points. We reserve a more precise discussion of the constants for the Appendix. We note that as a union of open sets, this is a measurable region for Borel regular μ.

The parameter ζ guards against the discontinuity of the distribution. It is intuitive and provably true that $\mu\left(B\left(x, r(x;p)\right)\right) \geq p$ always. However, to ensure some of our estimation is accurate, we need a two-sided condition on the measure of $B\left(x, r(x;p)\right)$, which might quite a bit larger than p if μ is not well-behaved. The following provides a reasonable way of quantifying such a condition.

Definition 6. *We say the measure μ is "ζ-regular" if there is a $\zeta \geq 0$ such that for all $x \in \mathcal{X}$ and $p \in (0,1)$,*

$$\mu\left(B\left(x, r(x;p)\right)\right) \leq (1+\zeta)p.$$

That is, the the measure of an $r(x;p)$ is never much greater than p. Note when $\mathcal{X} = \mathbb{R}^d$ and μ possesses a density with respect to the Lebesgue measure that is non-trivially bounded below, this condition is satisfied with $\zeta = 0$. We give our guarantees for the algorithm in terms of this parameter.

4.5. Algorithm

Algorithm 1 introduces the active learning scheme, which uses the first round of $i.i.d.$ samples from $\mathbb{P}$ to both estimate the hard region of space, and to do classification in its complement. We denote the product of measure $\mathbb{P}$ with itself n times in the Algorithm via $\mathbb{P}^{\bigotimes n}$. Ber$(\cdot)$ is used to denote a Bernoulli random variable with given parameter. All sample statements in the algorithm implicitly mean sample independently of the previous randomness in all cases.

We point out that although the learner can be confident that classification will succeed in the complement of the augmented hard region, the learner labels some points from this region in the second round to ensure consistency of the method. This is encoded in the algorithm via parameter π, which controls how many samples from the target distribution we take in the second round. We will see that $\pi > 0$ is roughly sufficient for the consistency of the method if n is fixed and $m \to \infty$. Without loss of generality, we implicitly assume that πm is an integer in our analysis.

Commencement of the second round of sampling is specified in terms of the distributional condition $\mu(\hat{\mathscr{H}}^+_{n,k}) > 0$.

Algorithm 1 Two-Round Active Algorithm

Input: $n, m \in \mathbb{N}$, $\pi \in (0,1]$, $\zeta \geq 0$, $\mathcal{S} = \emptyset$, $i = 0$

$\mathcal{S} \leftarrow \{(X_1,Y_1),\ldots,(X_n,Y_n)\} \sim \mathbb{P}^{\bigotimes n}$

if $\mu(\hat{\mathscr{H}}^+_{n,k,\zeta}) > 0$ **then**

  $\mathcal{S} \leftarrow \{(X_{n+1},Y_{n+1}),\ldots,(X_{n+\lfloor \pi m\rfloor},Y_{n+\lfloor \pi m\rfloor})\} \sim \mathbb{P}^{\bigotimes \lfloor \pi m\rfloor}$

  while $i < (1-\pi)m$ **do**

    $(X_i,Y_i) \leftarrow \emptyset$

    Sample $X \sim \mu$

    if $X \in \hat{\mathscr{H}}^+_{n,k,\zeta}$ **then**

      $X_i \leftarrow X$

      Sample $Y_i \sim$ Ber$(\eta(X))$

    end if

    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(X_i,Y_i)\}$

    $i \leftarrow |\mathcal{S}| - (n + \lfloor \pi m\rfloor)$

  end while

else

  $\mathcal{S} \leftarrow \mathcal{S} \cup \{(X_{n+1},Y_{n+1}),\ldots,(X_{n+m},Y_{n+m})\} \sim \mathbb{P}^{\bigotimes m}$

end if

Note that $\mu(\hat{\mathscr{H}}^+_{n,k})$ can be estimated to arbitrary accuracy using only unlabeled samples (which do not count towards label complexity). It is thus not uncommon to see such conditions in the active learning literature (Balcan et al., 2006; Minsker, 2012). More importantly, the measure need only be verified to be greater than 0 to proceed into the second round of sampling, which is strictly easier than estimating.

5. Guarantees

5.1. Bounding the Measure of the Acceptance Region

Looking at Algorithm 1, it is easy to see that the scheme will not yield any benefits if the labels disagree significantly everywhere in instance space; in this case, all unlabeled samples will be vacuously labeled in the second round of sampling. In other words, any kind of guarantee of speed up will have to involve a quantity which measures the fraction of the data distribution which is "easy" – and where η is bounded away from $1/2$. We begin with defining this region, and subsequently state our core result in terms of its measure.

Definition 7. *Fix $k < n$ and $\delta \in (0,1)$. We say the "easy region" of instance space $\mathscr{E}_{n,k}$ is the set defined via*

$$\mathscr{E}^+_{n,k} := \{x \mid \forall r \leq r(x;\bar{p}_{\text{UC}}),\ \eta(B(x,r)) - 1/2 \geq 5\Delta_{\text{UC}}\}$$
$$\mathscr{E}^-_{n,k} := \{x \mid \forall r \leq r(x;\bar{p}_{\text{UC}}),\ 1/2 - \eta(B(x,r)) \geq 5\Delta_{\text{UC}}\}$$
$$\mathscr{E}_{n,k} := \mathscr{E}^+_{n,k} \cup \mathscr{E}^-_{n,k}.$$

where $\bar{p}_{\text{UC}} = \frac{\bar{k}_\zeta}{n} + c_2\sqrt{\frac{d+\log(1/\delta)}{n}}$, for some constant c_2.

Given the condition bounding η away from $1/2$, we will eventually be able to simplify the measure of this set under the Tsybakov noise condition, and further elucidate conditions where speed up occurs.

5.2. Finite Sample Guarantee

The core result is a finite sample guarantee for the modified k-NN classifier $g_{n,m,k}$ trained on samples acquired in the method of Algorithm 1. The style of the result is meant to match Theorem 1 of (Chaudhuri & Dasgupta, 2014) above, facilitating comparison between the passive and active guarantees.

We note that in what follows, $g_{n,m,k}(\cdot)$ denotes the modified k-NN classifier using $R = \hat{\mathscr{H}}_{n,k}$ and $R^+ = \hat{\mathscr{H}}^+_{n,k,\zeta}$. In what follows, c_2 is the same constant as introduced in Lemma 1.

Theorem 2. *Fix $\delta \in (0,1)$, $\zeta \geq 0$, $\pi \in (0,1)$, and $k, n, m \in \mathbb{N}$, such that $\bar{k}_\zeta < n$ and $m(1-\pi)c_\delta \geq nc_{\delta/\sqrt{2}}$. Assume that μ is ζ-regular, and η is (α, L)-smooth such that*

$$3L(\bar{p}_{\text{UC}})^\alpha \leq \Delta_{\text{UC}}.$$

Then after a draw of $n+m$ samples via Algorithm 1 run with parameters π and ζ, with probability $\geq 1-\delta$ over the sampling,

$$\Pr_{X\sim\mu}\Big(g^*(X) \neq g_{n,m,k}(X)\Big) \leq \mu(\partial_{p',\Delta}) + \delta,$$

where

$$p' = c_{\delta/\sqrt{2}} \cdot \frac{k}{(1-\pi)m} \cdot \mu(\mathcal{X} \setminus \mathscr{E}_{n,k})\Big).$$

Recall that the main controlling factor in the passive setting, adjusted for the power of m extra samples from $\mathbb{P}$, is $c_\delta \frac{k}{n+m}$. It follows from the definition of the effective boundary that if $p' \leq p$, it holds that $\partial_{p',\Delta} \subseteq \partial_{p,\Delta}$. Thus, fixing the error tolerance δ, a sufficient condition for tighter guarantees than in the passive setting (in addition to the assumptions of Theorem 1) is that

$$c_{\delta/\sqrt{2}} \frac{k}{\frac{1-\pi}{\mu(\mathcal{X}\setminus\mathscr{E}_{n,k})}m} \leq c_\delta \frac{k}{n+m}.$$

This condition is a bit opaque, but in Section 4.3, we will see that it admits clean simplification under margin conditions on $\mathbb{P}$. For now, we note for $k \in \Omega(d\log(n))$, it holds that $c_\delta \approx c_{\delta/\sqrt{2}}$, and so if in addition $n \approx m$ and π is small (which is advantageous for tightening the finite sample bounds of Theorem 2), an approximate sufficient condition for speed up is $\mu(\mathscr{E}_{n,k}) > \frac{1}{2}$.

5.3. Consistency

A major upside of our method is that consistency is almost immediate. There are two intuitive reasons for this. Firstly, we do not permanently cut off any region of space from sampling as m increases and n stays fixed, which is roughly the condition presented by (Dasgupta, 2012) for making the standard k-NN classifier consistent under any active sampling strategy. The second stems from the form of our finite sample guarantees; by controlling the disagreement with the Bayes optimal classifier directly, we naturally converge to the Bayes-risk. To see this second fact, denote the risk of a classifier g trained with $n+m$ samples via

$$R_{n+m}(g) := \mathbb{P}\left(g(X) \neq Y\right),$$

and the risk of the Bayes optimal classifier be

$$R^* := \mathbb{P}\left(g^*(X) \neq Y\right).$$

Then, as noted in (Chaudhuri & Dasgupta, 2014), for every classifier g it holds that

$$R_{n+m}(g) - R^* \leq \Pr_{X\sim\mu}\big(g(X) \neq g^*\big).$$

Thus, Theorem 2 provides a bound on the excess risk that holds with high probability that under standard regularity conditions shrinks as the sample size increases. This is the main intuition as for why the following guarantee holds.

Theorem 3. *Fix $\zeta \geq 0$, $n \in \mathbb{N}$, and $\pi \in (0,1)$, such that $\bar{k}_\zeta < n$ when $k = k_n$. Suppose $(\mathcal{X}, \rho, \mu)$ satisfies the Lebesgue differentiation theorem, and the following conditions govern the growth of k_{n+m} as $m \to \infty$:*

$$k_{n+m} \to \infty$$
$$k_{n+m}/(n+m) \to 0.$$

Then for any first round sample $\mathcal{S}_n$ such that the boundary of the acceptance region $\partial\hat{\mathscr{H}}_{n,k}$ has $\mu(\partial\hat{\mathscr{H}}_{n,k}) = 0$, it holds that

$$\lim_{m\to\infty} \Pr_{\mathcal{S}_m}\left(R_{n+m}(g_{n,m,k}) - R^* > \epsilon\right) = 0,$$

that is, we convergence to the Bayes-risk in probability over the draw of the second m samples.

Here, the boundary $\partial\hat{\mathscr{H}}^+_{n,k}$ is the boundary of the set $\hat{\mathscr{H}}^+_{n,k}$ in the real-analytic sense, i.e. the set difference between the closure and interior of the set as given by open neighborhoods under ρ. Note that this consistency guarantee requires no assumptions on the smoothness of η, something that is notable for non-parametric active learning algorithms.

5.4. Speedup under Margin

Theorem 2 has the utility of presenting very general conditions under which our active scheme has tighter guarantees

than the passively trained k-NN classifier. However, with the generality comes a bit of opaqueness. Much of the complexity of the conditions under which our active scheme has tighter guarantees can be cleared up via investigation of Theorem 2 under the canonical Tsybakov noise conditions due to (Audibert & Tsybakov, 2005), which describe distributions with relatively low mass in regions where η is near $\frac{1}{2}$. We use the formulation of the noise condition given by (Chaudhuri & Dasgupta, 2014).

Definition 8. *We say $\mathbb{P}$ satisfies the "β-margin condition" if there is some $C > 0$ for which*

$$\mu\left(\left\{x \mid \left|\eta(x) - \frac{1}{2}\right| \leq t\right\}\right) \leq Ct^{\beta}.$$

The following corollary to Theorem 2 shows that under smoothness and the β-margin condition, our actively trained classifier admits tighter high probability bounds on the disagreement with the Bayes optimal classifier than the passive high probability guarantees of (Chaudhuri & Dasgupta, 2014).

Corollary 1. *Suppose the assumptions of Theorem 2 hold are satisfied by α, L, δ, π, ζ, n, $m = \Theta(n)$. If in addition $\mathbb{P}$ satisfies the β-margin condition, with probability $\geq 1 - \delta$ over the draw of $n + m$ training samples from Algorithm 1,*

$$\Pr_{X\sim\mu}\left(g^*(X) \neq g_{n,m,k}(X)\right) \leq \tilde{O}\left(\left(\sqrt{\frac{\log(1/\delta)}{k}} + \left(\frac{k}{n} \cdot \mu(\mathcal{X} \setminus \mathscr{E}_{n,k})\right)^{\alpha}\right)^{\beta}\right) + \delta,$$

where

$$\mu(\mathcal{X} \setminus \mathscr{E}_{n,k}) \leq \tilde{O}\left(\left(\sqrt{d\log(n)/k} + \left(k/n + \sqrt{d/n}\right)^{\alpha}\right)^{\beta}\right).$$

The corresponding guarantee for the (Chaudhuri & Dasgupta, 2014) analysis of passively trained nearest neighbors states that when β-margin condition holds, with probability $\geq 1 - \delta$ over the draw of $n + m$ samples from $\mathbb{P}$,

$$\Pr_{X\sim\mu}\left(g^*(X) \neq g_{n+m,k}(X)\right) \leq \tilde{O}\left(\left(\sqrt{\frac{\log(1/\delta)}{k}} + \left(\frac{k}{n}\right)^{\alpha}\right)^{\beta}\right) + \delta.$$

The $\tilde{O}$ notation here suppresses $\log(1/\delta)$ terms, as well as the "constant" c_δ; see the Appendix for the precise bound. This showcases that for fixed δ, if $k_{n+m} \in \omega(d\log(n))$, we have that $\mu(\mathcal{X} \setminus \mathscr{E}_{n,k}) \downarrow 0$, making the active rate tighter than the passive one, and increasingly so as $n + m$ grows.

We note that c_δ is $O(1)$ for virtually all regimes where we could expect our active algorithm to improve over passive. For example, if δ is fixed, then in order to have the estimated hard region guide any targeting over passive sampling, we need $k_{n+m} > \log(1/\delta) + \log(n)$. For such settings of k, c_δ is bounded above by a constant for sufficiently large n. In this case, the $\tilde{O}$ notation is only needed for the bound on $\mu(\mathcal{X} \setminus \mathscr{E}_{n,k})$.

5.5. Discussion of Optimality of the Scheme

The extent to which our scheme is optimal is not fully clear. As noted in our discussion of consistency, the disagreement of a classifier with the Bayes optimal is bounded below by the excess risk of the classifier, but bounds stemming from this observation are not tight, ruling out trivial comparisons to minimax lower bounds over active learners in the literature.

(Chaudhuri & Dasgupta, 2014) produces a passive lower bound for passive k-NN which states in part that when η is (α, L)-smooth, k-NN disagrees with the Bayes optimal on $x \in \mathcal{X}$ with

$$\left|\eta(x) - 1/2\right| < \frac{1}{\sqrt{k}} - L\left(\frac{k + \sqrt{k} + 1}{n}\right)^{\alpha},$$

which for small enough k reduces to

$$\left|\eta(x) - 1/2\right| < O\left(\frac{1}{\sqrt{k}}\right).$$

The purpose of the k-NN vote is a local estimate of η, with the prediction improving as the vote becomes more and more local. Roughly speaking, when η is smooth but not locally bounded away from $1/2$ by a margin of $\Omega(\frac{1}{\sqrt{k}})$, the vote over k neighbors will fall on the wrong side of $1/2$ (relative to the optimal prediction) with constant probability (Chaudhuri & Dasgupta, 2014). This strongly suggests that the term $1/\sqrt{k}$ in the bound of Corollary 1 is necessary for a voting-based classifier, whether passively or actively trained.

The second term in the guarantee of Corollary 1 corresponds to the locality of the vote; it shrinks as as the number of labels in the first round n increases, and the second round of labeling can be more and more targeted to a specific region of instance space. The factor of $\sqrt{d\log(n)/k}$ in upper bound on $\mu(\mathcal{X} \setminus \mathscr{E}_{n,k})$ is a relic of the use of uniform convergence over conditional probabilities. It is known that this factor cannot be reduced to $\sqrt{d/k}$, so this term is in some sense necessary for our algorithmic approach of estimating the effective boundary from the outside (Balsubramani et al., 2019). Whether different strategies can show further gains is an interesting future direction.

6. Conclusion

In this work, we have introduced a straightforwards and intuitive algorithm for active learning for k-NN, showed that it admits simply consistency guarantees, and exhibited conditions under which our guarantees are tighter than corresponding passive guarantees.

Our algorithm primarily works by improving the locality of the k-NN vote in regions where classification with k-NN can benefit from higher resolution sampling. An interesting extension to this idea might be to investigate adaptively modifying k throughout the instance space as the learner gains information on which regions of space admit strong classification guarantees, perhaps mirroring the passive strategy of (Balsubramani et al., 2019).

7. Acknowledgements

This work was supported by NSF under CNS 1804829 and ARO MURI W911NF2110317.

References

Angluin, D. Learning regular sets from queries and counterexamples. *Information and Computation*, 75(2):87–106, 1987.

Audibert, J.-Y. and Tsybakov, A. B. Fast learning rates for plug-in classifiers under the margin condition. *arXiv preprint math/0507180*, 2005.

Awasthi, P. and Kanade, V. Learning using local membership queries under smooth distributions. *CoRR*, abs/1211.0996, 2012.

Balcan, M.-F. and Urner, R. Active learning – modern learning theory. 2015. URL https://www.cs.cmu.edu/~ninamf/papers/al-survey-enc-algos.pdf.

Balcan, M.-F., Beygelzimer, A., and Langford, J. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 65–72, 2006.

Balsubramani, A., Dasgupta, S., Moran, S., et al. An adaptive nearest neighbor rule for classification. *Advances in Neural Information Processing Systems*, 32, 2019.

Beygelzimer, A., Dasgupta, S., and Langford, J. Importance weighted active learning. *CoRR*, abs/0812.4952, 2008.

Castro, R. M. and Nowak, R. D. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54 (5):2339–2353, 2008.

Chaudhuri, K. and Dasgupta, S. Rates of convergence for nearest neighbor classification. *CoRR*, 2014.

Cover, T. Rates of convergence for nearest neighbor procedures. In *Proceedings of The Hawaii International Conference on System Sciences*, 1968.

Dasgupta, S. Consistency of nearest neighbor classification under selective sampling. In *COLT*, 2012.

Dasgupta, S., Hsu, D., and Monteleoni, C. A general agnostic active learning algorithm. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, 2007.

Despres, C. J. J. The vapnik-chervonenkis dimension of cubes in $\mathbb{R}^d$, 2017.

Devroye, L., Gyorfi, L., Krzyzak, A., and Lugosi, G. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, 1994.

Hanneke, S. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 353–360, 2007.

Hanneke, S. and Yang, L. Minimax analysis of active learning. 2014.

Kääriäinen, M. Active learning in the non-realizable case. In *International Conference on Algorithmic Learning Theory*, pp. 63–77. Springer, 2006.

Kontorovich, A., Sabato, S., and Urner, R. Active nearest-neighbor learning in metric spaces, 2018.

Locatelli, A., Carpentier, A., and Samory, K. Adaptivity to noise parameters in nonparametric active learning. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

Minsker, S. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13(1), 2012.

Njike, B. N. and Siebert, X. Nonparametric adaptive active learning under local smoothness condition. *CoRR*, abs/2102.11077, 2021. URL https://arxiv.org/abs/2102.11077.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. 2014.

Stone, C. Consistent nonparametric regression. *Annals of Statistics*, 1977.

Zhang, C. and Chaudhuri, K. Beyond disagreement-based agnostic active learning. *CoRR*, abs/1407.2657, 2014.

A. Nearest Neighbor Tie-Breaking

We use the tie-breaking rule of (Chaudhuri & Dasgupta, 2014) to handle the case that two neighbors of x are the same distance from x under ρ. The idea is simply to augment the instance space to $\mathcal{X}' := \mathcal{X} \times [0,1]$, and assume training instances are drawn from the product measure $\mu' := \mu \times \lambda$, where λ is the Lebesgue measure on $[0,1]$. Preference then goes to lower values drawn from the uniform measure in tie-breaking. With probability 1 over the randomness in the tie breaking mechanism, ambiguity among nearest neighbors can be resolved in this way.

Given a training sample $\mathcal{S} = \mathcal{S}_n \cup \mathcal{S}_m$, and a new point $x \in \mathcal{X}$, let $X_{(i)}(x)$ denote the ith nearest neighbor to x under ρ out of samples drawn from $\mathbb{P}$ (where ties are broken according to the above), and let $Y_{(i)}(x)$ denote the label corresponding to $X_{(i)}(x)$. Let $X'_{(i)}(x) := (X_{(i)}(x), Z_{(i)}(x))$ be the ith nearest neighbor augmented with it's tie breaking draw, i.e. the ith nearest neighbor in the augmented instance space. Our analysis ideas center around balls in the original instance space, but to incorporate this tie breaking mechanism, we often consider balls in the augmented instances space. For $x' \in \mathcal{X}$, $r' \geq 0$, and $z' \in [0,1]$, we define

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

For a ball $B' = B'(x', r', z')$ let the conditional probability of a 1 given that an instance falls in the ball be

$$\eta(B') := \frac{1}{\mu'(B')} \int_{B'} \eta \, d\mu',$$

whenever $\mu'(B') > 0$, where we slightly abusively let $\eta(x, z) := \eta(x)$.

B. Complexity of Augmented Balls

Our results rely on the VC-dimension of the set of augmented balls in $\mathcal{X} \times [0,1]$, i.e. the VC-dimension of the set

$$\mathcal{B}' := \{B'(x', r', z') : x' \in \mathcal{X},\ r' \geq 0,\ z' \in [0,1]\},$$

being finite. As such, we find it important to show that this set can indeed have finite VC-dimension. It is intuitive that the VC-dimension of this set is closely related to the VC-dimension of the set of balls in $\mathcal{X}$. While we did not seriously pursue a general result to this end, we illustrate a simple and important example under a specific setting of ρ where the VC-dimension of the set of augmented balls grows linearly with that of balls in the original space $\mathcal{X}$.

Observation 1. *Suppose $\mathcal{X} = \mathbb{R}^d$, and $\rho(x, x') = \|x - x'\|_\infty$. Then $VC(\mathcal{B}') \leq 2d$.*

Proof. We give an argument based on (Despres, 2017). Let $P' = \{(x_i, z_i) \in \mathbb{R}^d \times [0,1]\}_{i=1}^K$ be a set of points that is shattered by augmented balls in $\mathcal{B}'$. For a given $(x_i, z_i) \in P'$ let $x_i = (x_i(1), \ldots x_i(d))$ denote the values of x_i in each of the dimensions. Note that for a given $(x_i, z_i) \in P'$, it holds that there is a dimension $l \leq d$ for which either $x_i(l) \geq \max_{j \neq i} x_j(l)$ or $x_i(l) \leq \min_{i \neq j} x_j(l)$, or WLOG if $x_i(l) = \max_{j \neq i} x_j(l)$, then $z_i \geq z_j$ for each $j \neq i$ with $x_j(l) = \max_{j \neq i} x_j(l)$. The reasoning is as follows: in l_∞, it holds that for some set of intervals of equal length $(l_1, u_1), \ldots, (l_d, u_d)$,

$$B'(x', r', z') = \Big\{(x, z) \in \mathbb{R}^d \times [0,1] : x(1) \in (x'(1) - r, x'(1) + r), \ldots, x(d) \in (x'(d) - r, x'(d) + r)$$
$$\veebar \max_i |x(i) - x'(i)| = r' \wedge z < z'\Big\}$$

so if (x_i, z_i) does not have this property, then there is no ball B' that has $(x_j, z_j) \in B'$ for all $j \neq i$ and $(x_i, z_i) \notin B'$. This is because any ball that contains maximal and minimal points in all dimensions and directions must be large in enough in each dimension to accommodate the maximal and minimal points. But notice that there can only be $2d$ maximal/minimal points in any set of points in augmented space, one for each dimension and direction, so $|P'| \leq 2d$. □

We note that this analysis essentially gives an l_1 result in $\mathbb{R}^2$ as well. Different geometries will likely require more complex arguments.

We finish this section with the observation that augmented balls have larger VC-dimension than their counterparts in the original instance space. This allows us to use $d = VC(\mathcal{B}')$ throughout in bounds regarding only the uniform convergence of balls in the original instance space $\mathcal{X}$.

Observation 2. *Suppose d is the VC-dimension of $\mathcal{B}'$ in $\mathcal{X} \times [0,1]$. Then the VC-dimension of the set of closed balls in $\mathcal{X}$, which we denote by $\mathcal{B}$, has $VC(\mathcal{B}) \leq d$.*

Proof. Suppose a set of points P of size K in $\mathcal{X}$ can be shattered by a set of closed balls $B_1, \ldots, B_{2^K}$. Then any augmented balls with $B_1, \ldots, B_{2^K}$ as their projection into $\mathcal{X}$ and have z parameters > 0 shatter the original points P each augmented with $z_i = 0$ in augmented space. □

C. Additional Notation for Proofs

We use this section to tie together some notation from the body with some further notation that will be necessary for our proofs.

C.1. Nearest Neighbors

Given a training sample $\mathcal{S} = \mathcal{S}_n \cup \mathcal{S}_m$, we implicitly assume it has been augmented with tie-breaking draws, such that nearest neighbors are well-defined. As introduced above, given a point $x \in \mathcal{X}$, we use $X_{(i)}(x)$ to specifically denote the ith nearest neighbor to x under the metric ρ of samples drawn from $\mathbb{P}$, and let $Y_{(i)}(x)$ denote the label corresponding to $X_{(i)}(x)$. Similarly, let $X_{(i)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta})$ denote the ith nearest neighbor to x drawn as a rejection sample from $\hat{\mathscr{H}}^{+}_{n,k,\zeta}$, and $Y_{(i)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta})$ it's corresponding label.

C.2. Balls: Open, Closed, etc.

We denote closed balls in the original instance space $\mathcal{X}$ via $B(x', r') := \{x \in \mathcal{X} : \rho(x', x) \leq r'\}$ and open balls via $B^o(x', r') := \{x \in \mathcal{X} : \rho(x', x) < r'\}$. Let

$$B_{(k+1)}(x) := B\left(x, \rho\left(x, X_{(k+1)}(x)\right)\right)$$

be shorthand for the closed ball of $k+1$ nearest neighbors to x drawn from $\mathbb{P}$ (where again, tie-breaking has already taken place in the augmented dimension). We use a similar notation for nearest neighbor balls of rejection samples:

$$B_{(k+1)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta}) := B\left(x, \rho\left(x, X_{(k+1)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta})\right)\right).$$

Both definitions may be modified to denote open balls via the include of the "o" in the superscript, e.g.

$$B^o_{(k+1)}(x) := B^o\left(x, \rho\left(x, X_{(k+1)}(x)\right)\right)$$

The associated augmented ball is a sort of "semi-open" ball that we denote via

$$B'_{(k+1)}(x) := B'\left(x, \rho\left(x, X_{(k+1)}(x)\right), Z_{k+1}(x)\right),$$

For a given $p \in (0,1]$ and some $x \in \mathcal{X}$, denote the closed ball around x with radius $r(x;p)$ via

$$B_p(x) := B\left(x, r(x;p)\right).$$

C.3. Constants and Parameter Settings

We use d to denote the VC-dimension of balls in the augmented space $\mathcal{X} \times [0,1]$. For given $k, n \in \mathbb{N}$, some particular parameter settings will also have special significance, and are fixed throughout:

$$c_\delta := \frac{1}{1 - \sqrt{(4/k)\ln(2/\delta)}}$$

$$\Delta := \min\left(\frac{1}{2}, \sqrt{\frac{\ln(2/\delta)}{k}}\right)$$

$$\Delta_{\text{UC}} := c_0 \sqrt{\frac{d\log(n) + \log(1/\delta)}{k}}.$$

where c_0 is a universal constant discussed in Lemma 7. Given an additional setting of $\zeta \geq 0$, the k-enlargement $\bar{k}_\zeta \in \mathbb{N}$ is given by, for some absolute constant c_1 (discussed in Lemma 2),

$$\bar{k}_\zeta := \left\lceil c_\delta(1+\zeta)k + c_1\sqrt{n\left(\log(4/\delta)+d\right)}\right\rceil.$$

Then we can also define:

$$p_{\text{UC}} := \frac{k}{n} + c_2\sqrt{\frac{d+\log(1/\delta)}{n}}$$
$$\bar{p}_{\text{UC}} := \frac{\bar{k}_\zeta}{n} + c_2\sqrt{\frac{d+\log(1/\delta)}{n}},$$

where c_2 a universal constant discussed in Lemma 3.

C.4. Empirical Measures

For a given measurable set $S \subseteq \mathcal{X}$, we use $\hat{\mu}(S)$ to denote the standard empirical of measure of mass of the set under μ, i.e. $\hat{\mu}(S) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}[X_i \in S]$. Similarly, for a given set $S' \subseteq \mathcal{X}'$ which contains at least one sample instance, we let $\hat{\eta}(S')$ be the standard empirical estimate of the conditional probability of a 1 in the set S', i.e.

$$\hat{\eta}(S') := \frac{1}{|\{X_i' \in S'\}|}\sum_{i=1}^n Y_i \mathbb{1}[X_i' \in S'],$$

where the samples X_i' are in the augmented space. In general, estimates of conditional probabilities take place in the augmented space, where there is no ambiguity about nearest neighbors.

D. Theorem Proofs

D.1. Estimation of Probability Balls, Radii

We begin with some basic results regarding the approximation of $r(x;p)$ uniformly across instance space. These results are the foundation for some results on set containment discussed later.

The first result to this end states that if $\bar{k}_\zeta$ is sufficiently large, with high probability, $\bar{k}_\zeta$-NN balls approximate $r(x; c_\delta k/n)$ balls from the outside uniformly across the instance space.

Lemma 2. *Fix $k, n \in \mathbb{N}$, and $\delta \in (0,1)$. Assume $d < \infty$, and μ is ζ-regular, and $\bar{k}_\zeta < n$. Then with probability $\geq 1 - \delta^2/16$ over the draw of n i.i.d. samples from $\mathbb{P}$, it holds that for all $x \in \mathcal{X}$,*

$$B_{c_\delta k/n}(x) \subseteq B^o_{(\bar{k}_\zeta+1)}(x).$$

Proof. For a given set $S \subseteq \mathcal{X}$, let $H(S) := \sum_{i=1}^n \mathbb{1}\{X_i \in S\}$ be the random variable counting the number of first round training samples with instances falling inside S. Let $\mathcal{B}$ denote the set of all closed balls in the space $\mathcal{X}$, and let $\mathcal{B}(c_\delta k/n) := \{B_{c_\delta k/n}(x) \mid x \in \mathcal{X}\}$. Note the relationship between events

$$\left\{\exists x \in \mathcal{X} \text{ s.t. } B_{c_\delta k/n}(x) \not\subseteq B^o_{(\bar{k}_\zeta+1)}(x)\right\} \subseteq \left\{\inf_{B \in \mathcal{B}(c_\delta k/n)} H(B^c) < n - \bar{k}_\zeta\right\}.$$

We upper bound the measure of this latter event. Recall that the VC dimension of a collection of sets is the same as that of the collection of complements of the sets in the original collection. It holds by the definition of $\bar{k}_\zeta$ and uniform convergence over the empirical probabilities of elements of $\mathcal{B}$ (as in (Shalev-Shwartz & Ben-David, 2014)) that there is some constant c_1

for which

$$\begin{aligned}
\Pr_{\mathcal{S}_n\sim\mathbb{P}^n}\left(\inf_{B\in\mathcal{B}(c_\delta k/n)} H(B^c) < n-\bar{k}_\zeta\right) &\leq \Pr_{\mathcal{S}_n\sim\mathbb{P}^n}\left(\sup_{B\in\mathcal{B}(c_\delta k/n)} -n\hat{\mu}(B^c) \geq \bar{k}_\zeta - n + 1\right) \\
&\leq \Pr_{\mathcal{S}_n\sim\mathbb{P}^n}\Bigg(\sup_{B\in\mathcal{B}(c_\delta k/n)} n\mu(B^c) - n\hat{\mu}(B^c) \\
&\qquad\qquad \geq \left(\bar{k}_\zeta - n + 1\right) + n\left(1-(1+\zeta)p\right)\Bigg) \\
&\leq \Pr_{\mathcal{S}_n\sim\mathbb{P}^n}\left(\sup_{B\in\mathcal{B}} |\mu(B^c)-\hat{\mu}(B^c)| \geq \frac{\bar{k}_\zeta+1}{n} + (1+\zeta)p)\right) \\
&\leq \exp(d)\cdot\exp\left(-n\left((\bar{k}_\zeta+1)/n - c_\delta k/n(1+\zeta)\right)^2/c_1^2\right) \\
&\leq \delta^2/16.
\end{aligned}$$

Again, we have made use of the uniform convergence of empirical measures of balls in space. □

Next, we show a sort of inverse of the above: we show that for a given k, if we pick the probability radius to be large enough, then with high probability, it holds that all k-NN balls are contained in balls with this probability radius uniformly across instance space $\mathcal{X}$. The analysis is independent of the choice of k, so an analogous statement holds for $\bar{k}_\zeta$.

Lemma 3. *Fix $k, n \in \mathbb{N}$ such that $k < n$, and $\delta \in (0,1)$. Assume $d < \infty$. Then with probability $\geq 1-\delta^2/16$ over the draw of n i.i.d. samples from $\mathbb{P}$, it holds that for all $x \in \mathcal{X}$.*

$$B_{(k+1)}(x) \subseteq B_{p_{\text{UC}}}(x).$$

Proof. The proof follows that of Lemma 2. As before, let $H(B)$ be the random variable counting the number of training samples with instances falling inside the ball B. In this case, we want to make sure this number is not too small. We again translate events via

$$\left\{\exists x \in \mathcal{X}\ s.t.\ B_{(k+1)}(x) \not\subseteq B_{p_{\text{UC}}}(x)\right\} \subseteq \left\{\inf_{B\in\mathcal{B}(p_{\text{UC}})} H(B) < k+1\right\},$$

and upper bound the measure of the latter event via

$$\begin{aligned}
\Pr_{\mathcal{S}_n\sim\mathbb{P}^n}\left(\inf_{B\in\mathcal{B}(p_{\text{UC}})} H(B) < k+1\right) &= \Pr_{\mathcal{S}_n\sim\mathbb{P}^n}\left(\sup_{B\in\mathcal{B}(p_{\text{UC}})} -H(B) \geq -k\right) \\
&\leq \Pr_{\mathcal{S}_n\sim\mathbb{P}^n}\left(\sup_{B\in\mathcal{B}} |\mu(B)-\hat{\mu}(B)| \geq p_{\text{UC}} - k/n\right) \\
&\leq \exp(d)\cdot\exp\left(-n(k/n-p_{\text{UC}})^2/c_2^2\right) \\
&\leq \delta^2/16.
\end{aligned}$$

□

Lemma 4. *Fix $n \in \mathbb{N}$ such that $\bar{k}_\zeta < n$, and $\delta \in (0,1)$. Assume $d < \infty$. Then with probability $\geq 1-\delta^2/16$ over the draw of n i.i.d. samples from $\mathbb{P}$,*

$$B_{(\bar{k}_\zeta+1)}(x) \subseteq B_{\bar{p}_{\text{UC}}}(x),$$

for all $x \in \mathcal{X}$.

Proof. The proof is identical to that of Lemma 3, just using $\bar{k}_\zeta$ and $\bar{p}_{\text{UC}}$. □

D.2. Estimation of the Effective Boundary

In this section, we show how to leverage the smoothness of η to show that the estimator $\hat{\mathscr{H}}_{n,k}$ is a useful approximation of the effective boundary $\partial_{c_\delta k/n,\Delta}$, the part of space in which we cannot expect k-NN to classify correctly after n samples from $\mathbb{P}$.

The first result to this end states that the conditional probability of a 1 inside an augmented ball is similar to the conditional probability of a 1 inside a ball in the original space $\mathcal{X}$ when radius of both of the balls is sufficiently small and smoothness on η is in effect.

Lemma 5. *Fix $x \in \mathcal{X}$, $z \in (0,1]$, $p \in (0,1)$ and $0 < r, \tilde{r} \leq r(x;p)$ (or $r = 0$ or $\tilde{r} = 0$ if $\mu(B(x,0)) > 0$). Then if $(\mathcal{X}, \rho, \mu)$ is (α, L)-smooth, it holds that*

$$\left|\eta\left(B'(x,r,z)\right) - \eta\left(B(x,\tilde{r})\right)\right| \leq 2Lp^\alpha.$$

Proof. When $supp(\mu) = \mathcal{X}$, we have $\mu(B,r) > 0$ for each $r > 0$. The result then follows from a loose analysis:

$$\begin{aligned}
|\eta\left(B'(x,r,z)\right) - \eta(B(x,r))| &\leq |\eta\left(B'(x,r,z)\right) - \eta(x)| + |\eta(x) - \eta(B(x,r'))| \\
&\leq Lp^\alpha + Lp^\alpha \\
&= 2Lp^\alpha,
\end{aligned}$$

where the the second inequality is implied by the definition of smoothness, as noted (in the case of non-augmented balls with $r > 0$) by (Chaudhuri & Dasgupta, 2014). We repeat the idea here for completeness, noting that

$$\begin{aligned}
|\eta\left(B'(x,r,z)\right) - \eta(x)| &= \left|\left(\frac{1}{\mu'(B'(x,r,z))}\int_{B'(x,r,z)} \eta(x',z')d\mu'\right) - \eta(x)\right| \\
&= \left|\frac{1}{\mu'(B'(x,r,z))}\int_{B'(x,r,z)} \left(\eta(x',z') - \eta(x)\right)d\mu'\right| \\
&= \left|\frac{1}{\mu'(B'(x,r,z))}\int_{B'(x,r,z)} \left(\eta(x') - \eta(x)\right)d\mu'\right| \\
&\leq \frac{1}{\mu'(B'(x,r,z))}\int_{B'(x,r,z)} |\eta(x') - \eta(x)|\, d\mu' \\
&\leq \frac{1}{\mu'(B'(x,r,z))}\int_{B'(x,r,z)} L\mu(B^o(x,r))^\alpha d\mu' \\
&= L\mu(B^o(x,r))^\alpha \\
&\leq Lp^\alpha.
\end{aligned}$$

The final line comes from (Chaudhuri & Dasgupta, 2014) Lemma 4. Note that such statements can be made for open balls as well with the same proof technique. □

Lemma 6. *Fix $p \in (0,1)$ and $\Delta \in (0,1/2]$. Then if μ has full support, and $(\mathcal{X}, \rho, \mu)$ is (α, L)-smooth, it holds that*

$$\mathcal{X}^+_{p,\Delta} = \{x \in \mathcal{X} : \forall r \leq r(x;p),\ \eta(B(x,r)) - 1/2 \geq \Delta\},$$

and

$$\mathcal{X}^-_{p,\Delta} = \{x \in \mathcal{X} : \forall r \leq r(x;p),\ 1/2 - \eta(B(x,r)) \geq \Delta\}.$$

Proof. By definition, when $supp(\mu) = \mathcal{X}$, we have $\mu(B,r) > 0$ for each $r > 0$, and we have

$$\mathcal{X}^+_{p,\Delta} = \{x \in \mathcal{X} : \eta(x) > 1/2 \wedge \forall r \leq r(x;p),\ \eta(B(x,r)) - 1/2 \geq \Delta\},$$

Thus, it suffices to show that $\eta(B(x,r)) - 1/2 \geq \Delta$ implies $\eta(x) > 0$ under smoothness of η. If $\mu(B(x,0)) > 0$, then this is trivial. So assume WLOG that $\mu(B(x,0)) = 0$. By the smoothness, we have for each r with $0 < r \leq r(x;p)$ that

$$|\eta(B(x,r)) - \eta(x)| \leq L\mu(B^o(x,r))^\alpha.$$

We can now write for such r that

$$\begin{aligned} 1/2 < \Delta/2 + 1/2 &< \Delta + 1/2 \\ &\leq \eta(B(x,r)) \\ &\leq |\eta(B(x,r)) - \eta(x)| + \eta(x) \\ &\leq L\mu(B^o(x,r))^\alpha + \eta(x) \\ &\leq L\mu(B(x,r))^\alpha + \eta(x). \end{aligned}$$

Thus we may conclude, for $0 < r \leq r(x;p)$, that

$$\eta(x) > \Delta/2 - L\mu(B^o(x,r))^\alpha + 1/2.$$

Thus, if there is some such r such that $\Delta/2 \geq L\mu(B^o(x,r))^\alpha$, we will show $\eta(x) > 1/2$. Let $B_n := B(x, r(x;p)/n)$ for $n = 1, 2, \ldots$. Then $\cap_{n\geq 1} B_n = \{x\}$. By continuity from above of the measure, we have that

$$\lim_{n\to\infty} \mu(B_n) = \mu\left(\cap_{n\geq 1} B_n\right) = \mu(\{x\}) = 0.$$

Thus, there is some N for which $n \geq N$ yields $\Delta/2 \geq L\mu(B^o(x,r))^\alpha$. □

The next result, due to (Balsubramani et al., 2019), describes the uniform convergence of k-NN votes to the true underlying conditional probability in augmented balls. The utility of this result for us is to ensure that for any new $x \in \mathcal{X}$ which is a candidate for rejection sampling, we can be sure that looking at it's k-NN votes, we get an accurate representation of the conditional probability nearby.

Lemma 7. *Let $\mathcal{B}'$ be the set of all augmented balls, i.e. $\mathcal{B}' = \{B(x,r,z) : x \in \mathcal{X},\ r > 0,\ z \in [0,1]\}$, and suppose this set has VC-dimension d in $\mathcal{X} \times [0,1]$. Fix $k < n \in \mathbb{N}$ and $\delta \in (0,1)$. With probability $\geq 1 - \delta^2/16$ over the draw of n i.i.d. samples from $\mathbb{P}$, we have that*

$$\sup_{B' \in \mathcal{B}'} |\eta(B') - \hat{\eta}(B')| < c_0 \sqrt{\frac{d\log(n) + \log(1/\delta)}{k}}$$

for a universal constant c_0.

Proof. This is the content of Theorem 8 of (Balsubramani et al., 2019). □

Finally, we show that the effective boundary after n samples is contained in our "estimated hard region" with high probability when η is smooth enough. We drop the condition that μ has full support in the following to simplify the result statements.

Lemma 1. *Fix $k < n \in \mathbb{N}$, and $\delta \in (0,1)$. Suppose we estimate*

$$\hat{\mathscr{H}}_{n,k} = \left\{ x \mid \left| \hat{\eta}(B'_{(k+1)}(x)) - 1/2 \right| < 3\Delta_{\text{UC}} \right\}.$$

Then if η is sufficiently (α, L)-smooth in the sense that

$$2L\left(p_{\text{UC}}\right)^\alpha \leq \Delta_{\text{UC}},$$

then with probability $\geq 1 - \delta^2/8$ over the draw of n i.i.d. samples from $\mathbb{P}$,

$$\partial_{c_\delta k/n, \Delta} \subseteq \hat{\mathscr{H}}_{n,k}.$$

Proof. We will show that when uniform convergence takes effect for the empirical estimates of the mass of balls in the space $\mathcal{X}$, as well as for conditional probabilities in augmented balls, that $\hat{\mathscr{H}}_{n,k}$ approximates the effective boundary from the outside. The necessary uniform convergence statements will hold with high probability over the sampling giving the result.

Fix an arbitrary sample $\mathcal{S}_n$. Suppose that for this sample, it holds for each $x \in \mathcal{X}$ that

$$\left|\hat{\eta}(B'_{(k+1)}(x)) - \eta(B'_{(k+1)}(x))\right| < \Delta_{\text{UC}}.$$

Suppose further that for each $x \in \mathcal{X}$, it holds that

$$B_{(k+1)}(x) \subseteq B_{p_{\text{UC}}}(x).$$

In this case, by Lemma 5 and the assumption on the smoothness of the distribution, we have for any $x \in \mathcal{X}$ and any $r \leq r(x; p_{\text{UC}})$ that

$$\left|\eta(B'_{(k+1)}(x)) - \eta(B(x,r))\right| \leq 2L\left(p_{\text{UC}}\right)^{\alpha} \leq \Delta_{\text{UC}}.$$

Suppose that $\hat{\eta}(B'_{(k+1)}(x)) \geq 1/2$. When η is sufficiently smooth, and the two notions of uniform convergence hold, it holds by the triangle inequality that for any $r \leq r(x; c_\delta k/n)$,

$$\begin{aligned}
\hat{\eta}(B'_{(k+1)}(x)) - 1/2 &\leq \left|\hat{\eta}(B'_{(k+1)}(x)) - \eta(B'_{(k+1)}(x))\right| + \eta(B'_{(k+1)}(x)) - 1/2 \\
&\leq \left|\hat{\eta}(B'_{(k+1)}(x)) - \eta(B'_{(k+1)}(x))\right| + \left|\eta(B'_{(k+1)}(x)) - \eta(B(x,r))\right| \\
&\quad + \eta(B(x,r)) - 1/2 \\
&\leq 2\Delta_{\text{UC}} + \eta(B(x,r)) - 1/2,
\end{aligned}$$

as such an r must respect $r \leq r(x; p_{\text{UC}})$ given that $r(x; c_\delta k/n) \leq r(x; p_{\text{UC}})$). Put differently,

$$\hat{\eta}(B'_{(k+1)}(x)) - 1/2 - 2\Delta_{\text{UC}} \leq \eta(B(x,r)) - 1/2.$$

Suppose $x \notin \hat{\mathscr{H}}_{n,k}$. Then by definition and our assumption on the estimate, $3\Delta_{\text{UC}} \leq \left|\hat{\eta}(B'_{(k+1)}(x)) - \frac{1}{2}\right| = \hat{\eta}(B'_{(k+1)}(x)) - 1/2$, which, given that $\Delta < \Delta_{\text{UC}}$, implies that

$$\Delta \leq \eta(B(x,r)) - 1/2.$$

Because this statement holds for each $r \leq r(x; c_\delta k/n)$, it holds by Lemma 6 and the definition of the effective boundary and Lemma 6 that $x \notin \partial_{c_\delta k/n, \Delta}$. The case where $\hat{\eta}(B'_{(k+1)}(x)) < 1/2$ can be handled analogously.

By Lemmas 7 and 3 respectively, the first two events take place each with probability $\geq 1 - \delta^2/16$ over the draw of the sample, and so by the union bound, the conclusion holds with probability $\geq 1 - \delta^2/8$. □

D.3. Formulation of Rejection Region, Controlling its Measure

As discussed in the main text, the actual region used for rejection sampling is a larger cousin of the estimated hard region. We will see that this allows us to more easily apply k-NN analysis techniques from the literature, as it will allow us to avoid analyzing cases where the k-NN form a biased estimate of the conditional probability of 1 inside any given nearest neighbors ball. A vital part of guaranteeing speedup under our scheme is an upper bound on the measure of this acceptance region. In this section, we show how to form a distributional quantity that provides such a bound.

Recall that given an estimated hard region $\hat{\mathscr{H}}_{n,k}$ and parameter ζ, we say its associated augmented estimate is the set

$$\hat{\mathscr{H}}^{+}_{n,k,\zeta} := \bigcup_{x \in \hat{\mathscr{H}}_{n,k}} B^{o}_{(\bar{k}_\zeta + 1)}(x),$$

and the "easy region" $\mathscr{E}_{n,k}$ is defined in the following manner:

$$\begin{aligned}
\mathscr{E}^{+}_{n,k} &:= \{x \mid \forall r \leq r(x; \bar{p}_{\text{UC}}),\ \eta(B(x,r)) - 1/2 \geq 5\Delta_{\text{UC}}\} \\
\mathscr{E}^{-}_{n,k} &:= \{x \mid \forall r \leq r(x; \bar{p}_{\text{UC}}),\ 1/2 - \eta(B(x,r)) \geq 5\Delta_{\text{UC}}\} \\
\mathscr{E}_{n,k} &:= \mathscr{E}^{+}_{n,k} \cup \mathscr{E}^{-}_{n,k}.
\end{aligned}$$

We first show that when the sample allows us to approximate probability radii well, and η is sufficiently smooth, membership in the acceptance region means that there is some point in the estimated hard region with a similar local conditional probability.

Lemma 8. *Fix $\zeta \geq 0$, $k \in \mathbb{N}$ such that $\bar{k}_\zeta < n \in \mathbb{N}$, and a sample $\mathcal{S}_n$. Suppose it holds for some $x \in \mathcal{X}$ that*

$$x \in \hat{\mathscr{H}}^+_{n,k,\zeta},$$

and the sample is such that for all $x' \in \mathcal{X}$,

$$B_{(k+1)}(x') \subseteq B_{p_{\text{UC}}}(x'),\ B_{(\bar{k}_\zeta+1)}(x') \subseteq B_{\bar{p}_{\text{UC}}}(x').$$

Then if η is (α, L)-smooth, there exists $\tilde{x} \in \hat{\mathscr{H}}_{n,k}$ for which

$$\left|\eta\left(B_{(\bar{k}_\zeta+1)}(x)\right) - \eta\left(B'_{(k+1)}(\tilde{x})\right)\right| \leq 3L(\bar{p}_{\text{UC}})^\alpha.$$

Proof. If $x \in \hat{\mathscr{H}}^+_{n,k,\zeta}$ and $B_{(\bar{k}_\zeta+1)}(x') \subseteq B_{\bar{p}_{\text{UC}}}(x')$, then $\exists \tilde{x} \in \hat{\mathscr{H}}_{n,\Delta}$ for which $\rho(\tilde{x}, x) \leq r_{\bar{p}_{\text{UC}}}(\tilde{x})$; by definition of the augmented estimate, there is some $\tilde{x} \in \hat{\mathscr{H}}_{n,k}$ for which $x \in B^o_{(\bar{k}_\zeta+1)}(\tilde{x})$, and so because $B^o_{(\bar{k}_\zeta+1)}(\tilde{x}) \subseteq B_{(\bar{k}_\zeta+1)}(\tilde{x}) \subseteq B_{\bar{p}_{\text{UC}}}(\tilde{x})$, the condition $\rho(\tilde{x}, x) \leq r_{\bar{p}_{\text{UC}}}(\tilde{x})$ follows by definition. Now we may write, using the definition of smoothness and the properties of smoothness discussed in Lemma 5,

$$\begin{aligned}
\left|\eta\left(B_{(\bar{k}_\zeta+1)}(x)\right) - \eta\left(B'_{(k+1)}(\tilde{x})\right)\right| &\leq |\eta\left(B_{(\bar{k}_\zeta+1)}(x)\right) - \eta(x)| + |\eta(x) - \eta(\tilde{x})| + |\eta(\tilde{x}) - \eta(B'_{(k+1)}(\tilde{x}))| \\
&\leq L(\bar{p}_{\text{UC}})^\alpha + L(\bar{p}_{\text{UC}})^\alpha + L(p_{\text{UC}})^\alpha \\
&\leq 3L(\bar{p}_{\text{UC}})^\alpha.
\end{aligned}$$

where the bounds on the first and second term of the first line come from the assumption that $\forall x \in \mathcal{X},\ B_{(\bar{k}_\zeta+1)}(x) \subseteq B_{\bar{p}_{\text{UC}}}(x)$, and the bound on the third term comes from the assumption that $\forall x \in \mathcal{X}$, $B_{(k+1)}(x) \subseteq B_{p_{\text{UC}}}(x)$. □

This result allows us to associate the easy region with the acceptance region of Algorithm 1. Lemma 9 gives us a natural bound on the measure of the acceptance region automatically.

Lemma 9. *Fix $\zeta \geq 0$ and $k \in \mathbb{N}$ such that $\bar{k}_\zeta < n \in \mathbb{N}$. Suppose η is sufficiently (α, L)-smooth in the sense that*

$$3L(\bar{p}_{\text{UC}})^\alpha \leq \Delta_{\text{UC}}.$$

Then with probability $\geq 1 - 3\delta^2/16$ over a draw of n i.i.d samples from $\mathbb{P}$,

$$\hat{\mathscr{H}}^+_{n,k,\zeta} \subseteq \mathcal{X} \setminus \mathscr{E}_{n,k}.$$

Proof. The proof technique is the same as that of Lemma 1. We suppose $x \in \hat{\mathscr{H}}^+_{n,k,\zeta}$, and show that $x \notin \mathscr{E}_{n,k}$ when certain conditions hold; finally, we argue by a union bound that the conditions occur simultaneously with probability $\geq 1 - \delta^2/8$.

Suppose it holds that, in addition to $3L(\bar{p}_{\text{UC}})^\alpha \leq \Delta_{\text{UC}}$, we have that for each $B' \in \mathcal{B}' := \{B(x, r, z) : x \in \mathcal{X}, r > 0, z \in [0, 1]\}$, that $|\eta(B') - \hat{\eta}(B')| < \Delta_{\text{UC}}$, and that for each $x \in \mathcal{X}$, both $B_{(\bar{k}_\zeta+1)}(x) \subseteq B_{\bar{p}_{\text{UC}}}(x)$ and $B_{(k+1)}(x) \subseteq B_{p_{\text{UC}}}(x)$. In this case, the following string of inequalities holds, where $\tilde{x} \in \hat{\mathscr{H}}_{n,k}$.

$$\begin{aligned}
\left|\eta\left(B_{(\bar{k}_\zeta+1)}(x)\right) - \frac{1}{2}\right| &\leq \left|\eta\left(B'_{(k+1)}(\tilde{x})\right) - \frac{1}{2}\right| + 3L(\bar{p}_{\text{UC}})^\alpha \\
&\leq \left|\eta\left(B'_{(k+1)}(\tilde{x})\right) - \frac{1}{2}\right| + \Delta_{\text{UC}} \\
&\leq \left|\hat{\eta}\left(B'_{(k+1)}(\tilde{x})\right) - \frac{1}{2}\right| + 2\Delta_{\text{UC}} \\
&< 3\Delta_{\text{UC}} + 2\Delta_{\text{UC}} \\
&= 5\Delta_{\text{UC}},
\end{aligned}$$

where the first line is a consequence of Lemma 8 and the triangle inequality, the second is by $3L(\bar{p}_{\text{UC}})^\alpha \leq \Delta$, the third is is by the uniform convergence of conditional probabilities in the augmented balls, and the second to last is from the definition of what it means for $\tilde{x} \in \hat{\mathscr{H}}_{n,k}$. This means that under these conditions, $\rho(x, X_{\bar{k}_\zeta}(x))$ is a radius that is sufficiently contentious to make $x \notin \mathscr{E}_{n,k}$, given that in this case $\rho(x, X_{\bar{k}_\zeta}(x)) \leq r(x; \bar{p}_{\text{UC}})$. By Lemmas 1, 2 and 4, and a union bound, the conditions hold with probability $\geq 1 - 3\delta^2/16$, and so the result holds as claimed. □

There are a lot of moving parts in our formulation of the acceptance region after the first n samples. We introduce a couple of definitions to help summarize the events we need to take place over the first n samples for our rejection sampling/ classification scheme to have desirable performance properties. We then show that these events take place with high probability.

Definition 9. *Given $\delta \in (0,1)$, a parameter setting of $\zeta \geq 0$ in Algorithm 1 and $k \in \mathbb{N}$ such that $\bar{k}_\zeta < n \in \mathbb{N}$, we define the event that the first n samples are "good for estimation" to be*

$$\text{GE}(\mathcal{S}_n) := \left\{ \partial_{c_\delta k/n,\Delta} \subseteq \hat{\mathscr{H}}_{n,k} \subset \hat{\mathscr{H}}^+_{n,k,\zeta} \subseteq \mathcal{X} \setminus \mathscr{E}_{n,k} \wedge \ \forall x \in \mathcal{X},\ B_{c_\delta k/n}(x) \subset B^o_{(\bar{k}_\zeta+1)}(x) \right\}.$$

Analogously we say that the sample S_n is "bad for estimation" when this event doesn't take place, and write

$$\text{BE}(\mathcal{S}_n) := \text{GE}(\mathcal{S}_n)^c.$$

Lemma 10. *Fix $k < n \in \mathbb{N}$. Suppose μ is ζ-regular, and η is sufficiently (α, L)-smooth in the sense that*

$$3L(\bar{p}_{\text{UC}})^\alpha \leq \Delta_{\text{UC}}.$$

Then it holds that

$$\mathbb{E}_{\mathcal{S}_n \sim \mathbb{P}^{\otimes n}} \left[\mathbb{1}\ \text{BE}(\mathcal{S}_n) \right] \leq \delta^2/4.$$

Proof. By definition, if it holds that

$$\partial_{c_\delta k/n,\Delta} \subseteq \hat{\mathscr{H}}_{n,k},$$

and

$$\hat{\mathscr{H}}^+_{n,k,\zeta} \subseteq \mathcal{X} \setminus \mathscr{E}_{n,k},$$

and

$$B_{c_\delta k/n}(x) \subseteq B^o_{(\bar{k}_\zeta+1)}(x),$$

then the sample is 'good for estimation." These events occur simultaneously with probability $\geq 1 - \delta^2/4$ by Lemmas 1, 9 and 2 respectively, and the union bound. This is because there are only 4 fundamental uniform convergence results on which these results are built, so the fact each of these holds with probability $\geq 1 - \delta^2/16$ is sufficient for the result. □

D.4. Further Notation for Analyzing Second-Round Samples

Roughly speaking, we have so far gathered the results required to ensure that the first round of sampling is likely to go to plan in terms of our estimate of what a good rejection region will be. What we are left to show is that if the first round goes well, then it is likely that the second round samples lead to an accurate classifier.

To do this, we essentially treat prediction using the rejection samples as a new prediction problem in a smaller space. In this section, we introduce some definitions that help us to that end, as well as some complex bad events that describe how the modified k-NN classifier introduce in the Preliminaries can predict incorrectly. As we did with the estimation events, we will look to control the measure of these events over the two rounds of the sampling to obtain a bound on the probability that the modified classifier disagrees with the Bayes optimal.

Recall the definition of the modified k-NN classifier, applied with $R = \hat{\mathscr{H}}_{n,k}$ and $R^+ = \hat{\mathscr{H}}^+_{n,k,\zeta}$:

$$g_{n,m,k}(x) := \begin{cases} \mathbb{1}[\frac{1}{k}\sum_{i=1}^k Y_{(i)}(x; \hat{\mathscr{H}}^+_{n,k,\zeta}) \geq \frac{1}{2}], & \text{if} x \in \hat{\mathscr{H}}_{n,k} \\ \mathbb{1}[\frac{1}{k}\sum_{i=1}^k Y_{(i)}(x) \geq \frac{1}{2}], & \text{otherwise.} \end{cases}$$

A related useful notion will be that of the effective boundary in a subspace. To introduce this, we must introduce some related ideas.

Definition 10. *Given $x \in \mathcal{X}$, the "closed ball of radius r inside the set S" is the set*

$$B^{\cap S}(x, r) := \{x' \in S : \rho(x, x') \leq r\} = B(x, r) \cap S.$$

Definition 11. *For a given $p \in (0, 1]$ and measurable $S \subseteq \mathcal{X}$, the "probability radius in S" of some $x \in S$ is*

$$r^{\cap S}(x; p) := \inf \left\{ r \mid \mu \left(B^{\cap S}(x, r) \right) \geq p \right\},$$

whenever $\mu(S) \geq p$, and 0 whenever $\mu(S) < p$.

Definition 12. *For a given $p \in (0, 1)$, $S \subseteq \mathcal{X}$ with $\mu(S) \geq p$, the "effective interior in S" is the set*

$$\begin{aligned}\mathcal{X}^S_{p,\Delta} := &\left\{ x \in S : \eta(x) > 1/2 \ \wedge \ \forall r \leq r^{\cap S}(x; p), \ \ \eta(B^{\cap S}(x, r)) - 1/2 \geq \Delta \right\} \\ &\cup \left\{ x \in S : \eta(x) < 1/2 \ \wedge \ \forall r \leq r^{\cap S}(x; p), \ \ 1/2 - \eta(B^{\cap S}(x, r)) \geq \Delta \right\}.\end{aligned}$$

Lemma 11. *Fix $p' \in (0, 1)$ and $\Delta \in (0, 1/2]$ and a measurable $S \subseteq \mathcal{X}$. If it further holds holds that $B(x, r(x; p')) \subseteq S$, then*

$$r^{\cap S}(x; p') = r(x; p').$$

Proof. Let $R := \{r : \mu(B(x, r) \geq p'\}$ and $R^{\cap S} := \{r : \mu(B^{\cap S}(x, r) \geq p'\}$. Note that if $r \geq r(x; p')$, then $r \in R$ and $r \in R^{\cap S}$, as $B^{\cap S}(x, r(x; p')) = B(x, r(x; p'))$. If $r < r(x; p')$, then it holds that $\mu(B(x, r)) = \mu(B^{\cap S}(x, r)) < p'$, or by definition of the infimum, $r(x; p')$ is not a lower bound for each element in R, and so cannot be the infimum. Thus $R = R^{\cap S}$, so they have the same infimum.

□

Lemma 12. *Fix $p' \in (0, 1]$ and $\Delta \in (0, 1/2]$ and a measurable $S \subseteq \mathcal{X}$. If it further holds holds that $B(x, r(x; p')) \subseteq S$, μ has full support, and (η, ρ, μ) is (α, L)-smooth, then the condition*

$$\forall r \leq r^{\cap S}(x; p), \ \ \eta(B^{\cap S}(x, r)) - 1/2 \geq \Delta$$

implies that $\eta(x) > 1/2$.

Proof. When $B(x, r(x; p')) \subseteq S$, we have that $r(x; p') = r^{\cap S}(x; p')$ by Lemma 11, and for each $r \leq r^{\cap S}(x; p'$, we have $B^{\cap S}(x, r) = B(x, r)$. Thus, we may repeat the argument of Lemma 6. □

Definition 13. *For a given $p \in (0, 1]$, $S \subseteq \mathcal{X}$ with $\mu(S) \geq p$, the "effective boundary in S" is the set*

$$\partial^S_{p,\Delta} := \mathcal{X} \setminus \mathcal{X}^S_{p,\Delta}.$$

D.4.1. Bad Prediction Events for the Modified k-NN Classifier

The following introduce the ways in which the modified k-NN classifier can err. Essentially, they are more complex versions of the original ideas of (Chaudhuri & Dasgupta, 2014). The idea of their analysis is that for a new point in instance space that is not in the effective boundary, the k-NN classifier can only make a mistake if one of the two events take place: the k-NN of x are further away than one would expect with n samples, or the empirical estimate of the conditional probability in the k-NN ball has a large deviation. The simplicity of this analysis follows directly from the definition of the effective boundary. For us, our modified k-NN classifier is essentially two k-NN classifiers predicting on different probability spaces. Thus, our idea is to repeat these analytical ideas on two separate spaces at once. The event of non-local neighbors or non-representative k-NN votes is a compound event that takes place when non-locality or non-representativeness takes place in either of the two spaces.

Fix $x \in \mathcal{X}$, $\zeta \geq 0$, $k \in \mathbb{N}$ such that $\bar{k}_\zeta < n \in \mathbb{N}$, $m \in \mathbb{N}$, and a sample $\mathcal{S} = \mathcal{S}_n \cup \mathcal{S}_m$. Denote the event that the sample is "non-local" relative to x laying in the estimated hard region to be

$$\mathrm{NL}_H(x, \mathcal{S}) := \left\{ \rho \left(x, X_{(k+1)}(x; \hat{\mathscr{H}}^+_{n,k,\zeta}) \right) > r^{\cap \hat{\mathscr{H}}^+_{n,k,\zeta}}(x; p'), \ x \in \hat{\mathscr{H}}_{n,k} \right\},$$

where $X_{(k+1)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta})$ is the $k+1^{st}$ nearest instance to x drawn as a rejection sample from $\hat{\mathscr{H}}^{+}_{n,k,\zeta}$. Similarly, we consider the event that the sample $\mathcal{S}$ is "non-local" relative to x laying outside the estimated hard region to be

$$\text{NL}_{\neg H}(x, \mathcal{S}) := \left\{ \rho\left(x, X_{(k+1)}(x)\right) > r(x; c_\delta k/n),\ x \notin \hat{\mathscr{H}}_{n,k} \right\}.$$

We summarize the event that the nearest neighbors of x are "not local" via the indicator function

$$\mathbb{1}\ \text{NL}(x, \mathcal{S}) := \mathbb{1}\ \text{NL}_H(x, \mathcal{S}) \cdot \mathbb{1}\text{GE}(\mathcal{S}_n) + \mathbb{1}\ \text{NL}_{\neg H}(x, \mathcal{S})$$

We also define some more complex events corresponding to non-representative estimation of conditional probabilities. Let the event that the conditional probability of an augmented ball defined by the $k+1$-NN to $x \in \mathcal{X}$ is underestimated by a margin $\geq \Delta$ based on the samples from $\mathbb{P}$ be

$$\text{BV}_{\neg H,1}(x, \mathcal{S}) := \left\{ \eta\left(B'_{(k+1)}(x)\right) - \hat{\eta}\left(B'_{(k+1)}(x)\right) \geq \Delta \right\},$$

and analogously when the estimation is done with samples from the second sampling round

$$\text{BV}_{H,1}(x, \mathcal{S}) := \left\{ \eta\left(B'_{(k+1)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta})\right) - \hat{\eta}\left(B'_{(k+1)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta})\right) \geq \Delta \right\}.$$

Here, $B'_{(k+1)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta})$ denotes the "augmented ball in $\hat{\mathscr{H}}^{+}_{n,k,\zeta}$" of $k+1$-NN to x drawn as a rejection sample from $\hat{\mathscr{H}}^{+}_{n,k,\zeta}$. Formally speaking,

$$\begin{aligned} B'_{(k+1)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta}) := \Big\{ & x' \in \hat{\mathscr{H}}^{+}_{n,k,\zeta}, z' \in [0,1] : \rho(x, x') < \rho\left(x, X_{(k+1)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta})\right) \veebar \\ & \rho(x, x') = \rho\left(x, X_{(k+1)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta})\right) \wedge z' < Z_{(k+1)}(x; \hat{\mathscr{H}}^{+}_{n,k,\zeta}) \Big\}. \end{aligned}$$

Similarly, we consider the event that the conditional probability is locally overestimated by a margin $\geq \Delta$ by samples from $\mathbb{P}$,

$$\text{BV}_{\neg H,0}(x, \mathcal{S}) := \left\{ \hat{\eta}(B'_{(k+1)}(x)) - \eta(B'_{(k+1)}(x)) \geq \Delta \right\},$$

and by samples from $\hat{\mathscr{H}}^{+}_{n,k,\zeta}$,

$$\text{BV}_{H,0}(x, \mathcal{S}) := \left\{ \hat{\eta}(B'_{(k+1)}(x, \hat{\mathscr{H}}^{+}_{n,k,\zeta})) - \eta(B'_{(k+1)}(x, \hat{\mathscr{H}}^{+}_{n,k,\zeta})) \geq \Delta \right\}.$$

The indicator on the event that the voting "goes badly" around x is then defined as

$$\begin{aligned} \mathbb{1}\ \text{BV}(x, \mathcal{S}) := & \mathbb{1}\{x \in \mathcal{X}_0\}\Big(\mathbb{1}\text{BV}_{H,0}(x, \mathcal{S}) \cdot \mathbb{1}\{x \in \hat{\mathscr{H}}_{n,k}\} \cdot \mathbb{1}\text{GE}(\mathcal{S}_n) + \mathbb{1}\ \text{BV}_{\neg H,0}(x, \mathcal{S}) \Big) + \\ & \mathbb{1}\{x \in \mathcal{X}_1\}\Big(\mathbb{1}\text{BV}_{H,1}(x, \mathcal{S}) \cdot \mathbb{1}\{x \in \hat{\mathscr{H}}_{n,k}\} \cdot \mathbb{1}\text{GE}(\mathcal{S}_n) + \mathbb{1}\ \text{BV}_{\neg H,1}(x, \mathcal{S}) \Big), \end{aligned}$$

where $\mathcal{X}_0 := \left\{ x \in \mathcal{X} \mid \eta(x) < \frac{1}{2} \right\}$, and $\mathcal{X}_1 := \left\{ x \in \mathcal{X} \mid \eta(x) \geq \frac{1}{2} \right\}$.

D.5. Towards a Finite Sample Bound

Given these event definitions, we will show that their measures can be bounded over the sampling, leading to a finite sample bound for the k-NN classifier. First, we show how to bound the measure of a "bad vote" for our modified k-NN classifier. In subsequent lemma, we show that the non-locality event is also a rare event. We assume WLOG that $(1-\pi)m$ is an integer to facilitate the analysis.

Lemma 13. *Fix $\delta \in (0,1)$, $k, n, m \in \mathbb{N}$, and $x \in \mathcal{X}$. Suppose $\mathcal{S} = \mathcal{S}_n \cup \mathcal{S}_m$ is drawn according to Algorithm 1 run with parameters $\pi > 0$ and $\zeta \geq 0$, with $k < (1-\pi)m$ and $\bar{k}_\zeta < n$. Then*

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{1}\ \text{BV}(x, \mathcal{S})\right] \leq 2\exp(-2k\Delta^2).$$

Proof. Suppose that $x \in \mathcal{X}_0$. Then it holds that

$$\mathbb{1}\ \mathrm{BV}(x, \mathcal{S}) := \mathbb{1}\mathrm{BV}_{H,0}(x, \mathcal{S}) \cdot \mathbb{1}\{x \in \hat{\mathscr{H}}_{n,k}\} \cdot \mathbb{1}\mathrm{GE}(\mathcal{S}_n) + \mathbb{1}\ \mathrm{BV}_{\neg H,0}(x, \mathcal{S}).$$

We have via the argument in (Chaudhuri & Dasgupta, 2014) Lemma 10 (tightened via a one-sided Hoeffding bound) that

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{1}\ \mathrm{BV}_{\neg H,0}(x, \mathcal{S})\right] \leq \exp(-2k\Delta^2).$$

The idea here is that averaging over the labels of the k-NN of x forms an unbiased estimate of $\eta\left(B'(x, X_{(k+1)}(x), Z_{(k+1)}(x))\right)$, and so the probability that this estimate deviates by more than Δ from it's true mean can be controlled via a Hoeffding bound.

The first term can be handled similarly if we consider the sampling procedure in the two stages. Suppose after the first n samples, we have

$$\mathbb{1}[x \in \hat{\mathscr{H}}_{n,k}] \cdot \mathbb{1}\ \mathrm{GE}(\mathcal{S}_n) = 1.$$

In this case, the classifier $g_{n,m,k}$ will only use the second m samples to make predictions on x. It it treats prediction on x as prediction in a new space defined by conditioning on $\hat{\mathscr{H}}^{+}_{n,k,\zeta}$. Note that when the initial sample $\mathcal{S}_n$ is "good for estimation", it holds that $B_{c_\delta k/n}(x) \subseteq B^{o}_{(\bar{k}_\zeta+1)}(x)$, and so

$$\mu(\hat{\mathscr{H}}^{+}_{n,k}) \geq \mu(B^{o}_{(\bar{k}_\zeta+1)}(x)) \geq c_\delta k/n > 0,$$

so this space is non-degenerate and can be sampled from; further, as $k < (1-\pi)m$, we are guaranteed that $X_{(k+1)}(x; \hat{\mathscr{H}}^{+}_{n,k})$ exists. Thus, given $x \in \hat{\mathscr{H}}_{n,k}$, we can reapply the argument given to bound the second term, just in this new space from which we get m samples. Then, in more formal language, we have argued that

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\left[\mathbb{1}\ \mathrm{BV}_{H,0}(x, \mathcal{S}) \cdot \mathbb{1}\{x \in \hat{\mathscr{H}}_{n,k}\} \cdot \mathbb{1}\mathrm{GE}(\mathcal{S}_n)\right] &\leq \Pr_{\mathcal{S}}\left(\mathrm{BV}_{H,0}(x, \mathcal{S}) \ \mid\ x \in \hat{\mathscr{H}}_{n,k},\ \mathrm{GE}(\mathcal{S}_n)\right) \\ &\leq \exp(-2k\Delta^2).\end{aligned}$$

The result follows from repeating the argument for the case that $x \in \mathcal{X}_1$, and noting that

$$2\exp\left(-2k\Delta^2\right)\left(\mathbb{1}\ \{x \in \mathcal{X}_0\} + \mathbb{1}\ \{x \in \mathcal{X}_1\}\right) \leq 2\exp(-2k\Delta^2).$$

□

Lemma 14. *Fix $x \in \mathcal{X}$, $\zeta \geq 0$, $k, n, m \in \mathbb{N}$, $p' \in (0, c_\delta k/n]$, and $\gamma \in (0, 1]$. Suppose $\mathcal{S} = \mathcal{S}_n \cup \mathcal{S}_m$ is drawn in accordance with Algorithm 1 run with parameters $\pi > 0$ and ζ, such that $k \leq (1-\gamma)\frac{(1-\pi)}{\mu(\mathcal{X}\setminus\mathscr{E}_{n,k})}mp'$ and $\bar{k}_\zeta < n$. Then we have*

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{1}\ \mathrm{NL}_H(x, \mathcal{S}) \cdot \mathbb{1}\mathrm{GE}(\mathcal{S}_n)\right] \leq \exp(-k\gamma^2/2).$$

Proof. The idea is similar to the one employed in the previous lemma. We condition on events depending only on the first n samples, and then note that given those events, non-locality can be controlled by Binomial deviation bounds.

In particular, if after the first n samples it holds that $\mathbb{1}\{x \in \hat{\mathscr{H}}_{n,k}\} \cdot \mathbb{1}\ \mathrm{GE}(\mathcal{S}_n) = 1$, then prediction takes place in the smaller space. Given that the initial sample $\mathcal{S}_n$ is "good for estimation", it holds that $B_{c_\delta k/n}(x) \subseteq B^{o}_{(\bar{k}_\zeta+1)}(x)$, and so, as above,

$$\mu(\hat{\mathscr{H}}^{+}_{n,k}) \geq \mu(B^{o}_{(\bar{k}_\zeta+1)}(x)) \geq c_\delta k/n > 0.$$

This implies that we can rejection sample from $\hat{\mathscr{H}}^{+}_{n,k,\zeta}$, and the probability of any one of the $(1-\pi)m$ accepted samples from the second round falling in the ball

$$B^{\cap\hat{\mathscr{H}}^{+}_{n,k,\zeta}}\left(x, r^{\cap\hat{\mathscr{H}}^{+}_{n,k,\zeta}}(x; p')\right)$$

is $\geq \frac{p'}{\mu(\mathcal{X}\setminus\mathscr{E}_{n,k})}$; when $x \in \hat{\mathscr{H}}_{n,k}$ and $B_{c_\delta k/n}(x) \subseteq B^o_{(\bar{k}_\zeta+1)}(x)$ we have that

$$B^{\cap\hat{\mathscr{H}}^+_{n,k,\zeta}}(x, r^{\cap\hat{\mathscr{H}}^+_{n,k,\zeta}}(x;p')) = B(x, r(x;p')),$$

and we are conditioning on $\hat{\mathscr{H}}^+_{n,k,\zeta}$. Then we can use the argument from (Chaudhuri & Dasgupta, 2014) Lemma 9 to get the result: given that $\mu(B(x, r(x;p')) \geq p'$, with probability $\geq 1 - \exp(-k\gamma^2/2)$ more than k of the $(1-\pi)m$ rejection samples will fall inside the set (by the multiplicative Chernoff bound). □

We need one final result before giving the main, finite sample bound. This is a technical result that will allow us to relate the effective boundary inside the acceptance region to the effective boundary in the instance space $\mathcal{X}$, which gives a simplification of our finite sample bound that facilitates comparison with passive guarantees.

Lemma 15. *Fix $x \in \mathcal{X}$, $\zeta \geq 0$, $k, n, m \in \mathbb{N}$, such that $\bar{k}_\zeta < n$ and $p' \leq c_\delta k/n$. For all first round samples $\mathcal{S}_n$ from Algorithm 1 run with parameter ζ, it holds that*

$$\left\{x \in \partial^{\hat{\mathscr{H}}^+_{n,k,\zeta}}_{p',\Delta},\ x \in \hat{\mathscr{H}}_{n,k}\right\} \cap \mathrm{GE}(\mathcal{S}_n) \subseteq \left\{x \in \partial_{p',\Delta}\right\}.$$

Proof. If the sample is such that $\mathbb{1}\mathrm{GE}(\mathcal{S}_n) = 1$ and it further holds that $x \in \hat{\mathscr{H}}_{n,k}$, then it holds from the definition of the event that the sample is "good for estimation" that

$$B\left(x, r(x; c_\delta k/n)\right) \subseteq \hat{\mathscr{H}}^+_{n,k,\zeta}.$$

By $p' \leq c_\delta k/n$, we have $r(x;p') \leq r(x; c_\delta k/n)$, and so it further holds that

$$B(x, r(x;p')) \subseteq \hat{\mathscr{H}}^+_{n,k,\zeta},$$

and that $r^{\hat{\mathscr{H}}^+_{n,k,\zeta}}(x;p') = r(x;p')$. Suppose by contradiction that $x \notin \partial_{p',\Delta}$. Then WLOG, $x \in \mathcal{X}^+_{p',\Delta} = \{x : \forall r \leq r(x;p'),\ \eta(B(x,r)) - 1/2 \geq \Delta\}$. Because $B(x, r(x;p')) \subseteq \hat{\mathscr{H}}^+_{n,k,\zeta}$, we must have for each $r \leq r(x;p')$,

$$B(x,r) = B^{\cap\hat{\mathscr{H}}^+_{n,k,\zeta}}(x,r),$$

and so, this would imply via Lemma 12 that $x \notin \partial^{\hat{\mathscr{H}}^+_{n,k,\zeta}}_{p',\Delta}$, which is a contradiction. □

The main finite sample bound can now be stated. The analysis techniques come straight from (Chaudhuri & Dasgupta, 2014), where we have more complex "bad events" to ensure don't take place.

Theorem 2. *Fix $\delta \in (0,1)$, $\zeta \geq 0$, $\pi \in (0,1)$, and $k, n, m \in \mathbb{N}$, such that $\bar{k}_\zeta < n$ and $m(1-\pi)c_\delta \geq nc_{\delta/\sqrt{2}}$. Assume that μ is ζ-regular, and η is (α, L)-smooth such that*

$$3L(\bar{p}_{\mathrm{UC}})^\alpha \leq \Delta_{\mathrm{UC}}.$$

Then, with probability $\geq 1-\delta$ over the draw of $n+m$ samples from Algorithm 1 run with parameters π and ζ, we have

$$\Pr_{X\sim\mu}\left(g^*(X) \neq g_{n,m,k}(X)\right) \leq \mu(\partial_{p',\Delta}) + \delta,$$

where

$$p' = c_{\delta/\sqrt{2}} \cdot \frac{k}{\frac{1-\pi}{\mu(\mathcal{X}\setminus\mathscr{E}_{n,k})}m}.$$

Proof. Fix $x_0 \in \mathcal{X}$ arbitrarily. We assert that for every sample $\mathcal{S} = \mathcal{S}_n \cup \mathcal{S}_m$, the following bound holds:

$$\begin{aligned}
\mathbb{1}\left\{g^*(x_0) \neq g_{n,m,k}(x_0)\right\} \leq & \mathbb{1}\left\{x_0 \in \partial^{\hat{\mathscr{H}}^+_{n,\Delta}}_{p',\Delta},\ x_0 \in \hat{\mathscr{H}}_{n,k}\right\} \cdot \mathbb{1}\mathrm{GE}(\mathcal{S}_n) + \\
& \mathbb{1}\left\{x_0 \in \partial_{c_\delta k/n,\Delta},\ x_0 \notin \hat{\mathscr{H}}_{n,\Delta}\right\} \cdot \mathbb{1}\mathrm{GE}(\mathcal{S}_n) + \\
& \mathbb{1}\,\mathrm{BE}(\mathcal{S}_n) + \\
& \mathbb{1}\,\mathrm{NL}(x_0, \mathcal{S}) + \\
& \mathbb{1}\,\mathrm{BV}(x_0, \mathcal{S}).
\end{aligned}$$

To see this, fix a sample $\mathcal{S}$. Assume $\mathcal{S}$ is such that none of the first three terms in the upper bound evaluate to 1, or the bound holds automatically for this sample. Suppose WLOG that we further have $x_0 \in \hat{\mathscr{H}}_{n,\Delta}$. If, in addition, the sample is such that $\mathbb{1}\text{NL}_H(x_0, \mathcal{S}) = 1$, by assumption it must hold that $\mathbb{1}\text{NL}(x_0, \mathcal{S}) = 1$ (because $\mathbb{1}\text{GE}(\mathcal{S}_n) = 1$ by assumption), validating the bound for this $\mathcal{S}$. So further assume $\mathbb{1}\text{NL}_H(x_0, \mathcal{S}) = 0$. WLOG assume $x_0 \in \mathcal{X}_0$. In this case, because we have that

$$x_0 \notin \partial_{p',\Delta}^{\hat{\mathscr{H}}^+_{n,\Delta}},$$

by definition, we have

$$\forall r \leq r^{\cap \hat{\mathscr{H}}^+_{n,k,\zeta}}(x_0; p'),\ \frac{1}{2} - \eta\left(B^{\cap \hat{\mathscr{H}}^+_{n,k,\zeta}}(x_0, r)\right) \geq \Delta,$$

and thus under smoothness,

$$\frac{1}{2} - \eta(x_0) > \Delta.$$

Thus, the only way our classifier disagrees with the Bayes optimal classifier is if the voting overestimates the conditional probability of the k-NN ball, i.e. $\mathbb{1}\ \text{BV}_{H,0}(x_0, \mathcal{S}) = 1$, validating the bound for this sample $\mathcal{S}$; see (Chaudhuri & Dasgupta, 2014) for some formalism regarding the handling of augmented balls here. The full claim can be proved with identical arguments, by working through variations of the cases $x_0 \in \hat{\mathscr{H}}_{n,k}$ and $x_0 \in \mathcal{X}_0$. Note that by the fact that $m(1-\pi)c_\delta \geq nc_{\delta/\sqrt{2}}$ and $k < \bar{k}_\zeta < n$, the $k+1$-nearest neighbors are well-defined for both rejection samples, and for samples from $\mathbb{P}$.

We note further that this bound can be further simplified:

$$\begin{aligned}\mathbb{1}\left\{g^*(x_0) \neq g_{n,m,k}(x_0)\right\} \leq &\mathbb{1}\left\{x_0 \in \partial_{p',\Delta}\right\} + \\ &\mathbb{1}\ \text{BE}(\mathcal{S}_n) + \\ &\mathbb{1}\ \text{NL}(x_0, \mathcal{S}) + \\ &\mathbb{1}\ \text{BV}(x_0, \mathcal{S}).\end{aligned}$$

To see this why this holds, two observations need to be made. Firstly, the first term of the original bound admits transformation into the first term of this simplified bound by Lemma 15 when the condition $m(1-\pi)c_\delta \geq nc_{\delta/\sqrt{2}}$ holds (as this condition implies that $p' \leq c_\delta k/n$). Further, the second term of the original bound can be eliminated entirely. This elimination comes from the fact that

$$\left\{x_0 \in \partial_{c_\delta k/n,\Delta},\ x_0 \notin \hat{\mathscr{H}}_{n,k}\right\} \cap \text{GE}(\mathcal{S}_n) = \emptyset,$$

as when $\mathbb{1}\text{GE}(\mathcal{S}_n) = 1$, we have by definition $\partial_{c_\delta k/n,\Delta} \subseteq \hat{\mathscr{H}}_{n,k}$. Now, by Lemma 10, we have that

$$\mathbb{E}_{\mathcal{S}_n}\left[\mathbb{1}\ \text{BE}(\mathcal{S}_n)\right] \leq \delta^2/4.$$

By Lemma 14, and by the setting of the parameter p' and $c_\delta k/n$, we have

$$\mathbb{E}_{\mathcal{S}_n,\mathcal{S}_m}\left[\mathbb{1}\ \text{NL}(x_0, \mathcal{S})\right] \leq \delta^2/4,$$

where by $\mathbb{E}_{\mathcal{S}_n,\mathcal{S}_m}$ we mean the expectation over the sampling scheme defined by Algorithm 1. By Lemma 13 and the choice of Δ,

$$\mathbb{E}_{\mathcal{S}_n,\mathcal{S}_m}\left[\mathbb{1}\text{BV}(x_0, \mathcal{S})\right] \leq \delta^2/2.$$

Thus, for a new $X_0 \sim \mu$,

$$\mathbb{E}_{X_0}\mathbb{E}_{\mathcal{S}}\left[\mathbb{1}\ \text{BE}(\mathcal{S}) + \text{NL}(X_0, \mathcal{S}) + \text{BV}(X_0, \mathcal{S})\right] \leq \delta^2$$

Thus, by Markov's inequality, their sum is $\leq \delta$ with probability $\geq 1-\delta$. This gives the result given that $\mathbb{E}_{X_0\sim\mu}\left[\mathbb{1}\left\{X_0 \in \partial_{p',\Delta}\right\}\right] = \mu(\partial_{p',\Delta})$. □

Corollary 1. *Suppose the assumptions of Theorem 2 hold are satisfied by α, L, δ, π, ζ, n, $m = \Theta(n)$. If in addition $\mathbb{P}$ satisfies the β-margin, then with probability $\geq 1-\delta$ over the draw of $n+m$ training samples from Algorithm 1,*

$$\text{Pr}_{X\sim\mu}\left(g^*(X) \neq g_{n,m,k}(X)\right) \leq O\left(\left(\sqrt{\frac{\log(1/\delta)}{k}} + \left(c_\delta \cdot \frac{k}{n} \cdot \mu(\mathscr{E}_{n,k})\right)^\alpha\right)^\beta\right) + \delta,$$

where

$$\mu(\mathcal{X} \setminus \mathscr{E}_{n,k}) \leq O\left(\left(\sqrt{\frac{d\log(n)+\log(1/\delta)}{k}} + \left(c_\delta \cdot \frac{k}{n} + \sqrt{\frac{d+\log(1/\delta)}{n}}\right)^\alpha\right)^\beta\right).$$

The corresponding guarantee for the (Chaudhuri & Dasgupta, 2014) analysis of passively trained nearest neighbors states, with probability $\geq 1-\delta$,

$$\Pr_{X\sim\mu}\left(g^*(X) \neq g_{n,m,k}(X)\right) \leq O\left(\left(\sqrt{\frac{\log(1/\delta)}{k}} + \left(c_\delta \cdot \frac{k}{n}\right)^\alpha\right)^\beta\right) + \delta.$$

Proof. Under (α, L)-smoothness and μ having full support, as noted by (Chaudhuri & Dasgupta, 2014) Lemma 4,

$$\partial_{p,\Delta} \subset \left\{x \in \mathcal{X} : \left|\eta(x) - \frac{1}{2}\right| \leq \Delta + Lp^\alpha\right\}.$$

The result, up to the decay of the mass of $\mathcal{X} \setminus \mathscr{E}_{n,k}$, follows from the direct application of the β-margin condition with margin $\Delta + Lp^\alpha$, where p is adjusted for the active case as in Theorem 2.

The decay of the of the mass of $\mathcal{X} \setminus \mathscr{E}_{n,k}$ follows the same lines. Note that from the definitions,

$$\partial_{\bar{p}_{\text{UC}},\Delta_{\text{UC}}} = \mathcal{X} \setminus \mathscr{E}_{n,k},$$

and so the idea from above can be directly reapplied. □

D.6. Consistency

For consistency results, we first start by showing an alternative finite sample guarantee that is more conducive for our later arguments. To this end, we introduce a few further distribution quantities relating to a slightly different notion of effective boundary. The intuition as to why such a consistency result can be shown is as given above; our slightly different notion of effective boundary still useful for controlling the probability of disagreement with the Bayes optimal classifier, and can be tightened with increasing m.

Definition 14. *For $p \in (0,1]$, $\Delta \in (0, \frac{1}{2}]$, the "effective interior" of (Chaudhuri & Dasgupta, 2014) is the set*

$$\mathcal{X}_{p,\Delta} := \mathcal{X} \setminus \partial_{p,\Delta}.$$

Definition 15. *For a given set $S \subseteq \mathcal{X}$, and for $p \in (0,1]$, let the "probability p-interior" of S be*

$$\mathrm{I}_p(S) := \{x \mid B(x, r(x;p)) \subseteq S\}.$$

These two definitions enable the specification of the alternative notion of effective boundary we use in our consistency guarantees.

Definition 16. *For $\vec{p} = (p, p') \in (0,1) \times (0,1)$, and a given region $S \subseteq \mathcal{X}$, let the "effective boundary in $\mathcal{X}$ relative to S' and $\vec{p}$" be*

$$\tilde{\partial}^S_{\vec{p},\Delta} := \mathcal{X} \setminus \left(\left(\mathcal{X}_{p',\Delta} \cap \mathrm{I}_{p'}(S)\right) \sqcup \left(\mathcal{X}_{p,\Delta} \cap \mathrm{I}_p(S^c)\right)\right).$$

This alternative effective boundary definition turns out to have controlling properties for the disagreement of the modified k-NN classifier with the Bayes optimal classifier. This is the content of the next lemma.

Lemma 16. *Pick $\delta \in (0,1)$, $\zeta \geq 0$, $\pi \in (0,1)$, and $k, n, m \in \mathbb{N}$ such that $\bar{k}_\zeta < n$. Suppose $\mathcal{S} = \mathcal{S}_n \cup \mathcal{S}_m$ is drawn according to Algorithm 1 with parameters π and ζ. With probability $\geq 1-\delta$ over the sampling,*

$$\Pr_{X\sim\mu}\left(g^*(X) \neq g_{n,m,k}(X)\right) \leq \mu\left(\tilde{\partial}^{\mathscr{H}_{n,k}}_{\vec{p}_m,\Delta}\right) + \delta,$$

where $\vec{p}_m = (p_{\mathrm{I}}, p'_{\mathrm{I}})$, and $p_{\mathrm{I}} = c_\delta \frac{k}{n+\pi m}$, $p'_{\mathrm{I}} = c_\delta \frac{k}{(1-\pi)m}$.

Proof. The proof is very similar to that of Theorem 2. Fix $x_0 \in \mathcal{X}$ and a sample $\mathcal{S}$. Define the event

$$\mathrm{NL}_{H,\mathrm{I}}(x_0, \mathcal{S}) := \left\{ \rho(x_0, X_{(k+1)}(x_0; \hat{\mathscr{H}}^{+}_{n,k,\zeta})) > r(x_0; p'_{\mathrm{I}}),\ x_0 \in \mathrm{I}_{p'_{\mathrm{I}}}(\hat{\mathscr{H}}_{n,k}) \right\},$$

the event

$$\mathrm{NL}_{\neg H,\mathrm{I}}(x_0, \mathcal{S}) := \left\{ \rho(x_0, X_{(k+1)}(x_0)) > r(x_0; p_{\mathrm{I}}),\ x_0 \in \mathrm{I}_{p_{\mathrm{I}}}\left((\hat{\mathscr{H}}_{n,k})^c \right) \right\},$$

and summarize via

$$\mathrm{NL}_{\mathrm{I}}(x_0, \mathcal{S}) := \mathrm{NL}_{H,I}(x_0, \mathcal{S}) \cup \mathrm{NL}_{\neg H,I}(x_0, \mathcal{S}).$$

Further, define

$$\begin{aligned}\mathbb{1}\mathrm{BV}_{\mathrm{I}}(x_0, S) :=& \mathbb{1}\{x_0 \in \mathcal{X}_0\}\Big(\mathbb{1}\mathrm{BV}_{H,0}(x_0, \mathcal{S}) \cdot \mathbb{1}\{x_0 \in \mathrm{I}_{p'_{\mathrm{I}}}(\hat{\mathscr{H}}_{n,k})\} + \mathbb{1}\ \mathrm{BV}_{\neg H,0}(x_0, \mathcal{S}) \Big) + \\ & \mathbb{1}\{x_0 \in \mathcal{X}_1\}\Big(\mathbb{1}\mathrm{BV}_{H,1}(x_0, \mathcal{S}) \cdot \mathbb{1}\{x_0 \in \mathrm{I}_{p'_{\mathrm{I}}}(\hat{\mathscr{H}}_{n,k})\} + \mathbb{1}\ \mathrm{BV}_{\neg H,1}(x_0, \mathcal{S}) \Big),\end{aligned}$$

Similarly to Theorem 2, we show that for any sample $\mathcal{S}$ of size $n + m$,

$$\begin{aligned}\mathbb{1}\left\{ g^*(x_0) \neq g_{n,m,k}(x_0) \right\} \leq & \mathbb{1}\left\{ x_0 \in \tilde{\partial}^{\hat{\mathscr{H}}_{n,k}}_{\vec{p},\Delta} \right\} + \\ & \mathbb{1}\mathrm{NL}_{\mathrm{I}}(x_0, \mathcal{S}) + \\ & \mathbb{1}\mathrm{BV}_{\mathrm{I}}(x_0, \mathcal{S}).\end{aligned}$$

The reasoning is analogous. Suppose for a given sample that $x_0 \notin \tilde{\partial}^{\hat{\mathscr{H}}_{n,k}}_{\vec{p}_m,\Delta}$. Then WLOG $x_0 \in \mathrm{I}_{p'_{\mathrm{I}}}(\hat{\mathscr{H}}_{n,k}) \cap \mathcal{X}_{p'_{\mathrm{I}},\Delta}$. Then if more than k of the second round rejection samples fall in $B(x_0, r(x_0; p'_{\mathrm{I}}))$, then the only way $g_{n,m,k}$ classifies different from the Bayes optimal is for the voting to be non-representative by a margin $\geq \Delta$ under the k nearest rejection samples.

We then bound the probability of these events under a draw from Algorithm 1. First, consider the event $\mathrm{NL}_{H,\mathrm{I}}(x_0, \mathcal{S})$. As before, we bound the conditional probability that $\rho(x_0, X_{(k+1)}(x_0; \hat{\mathscr{H}}^{+}_{n,k,\zeta})) > r(x_0; p'_{\mathrm{I}})$ given that $x_0 \in \mathrm{I}_{p'_{\mathrm{I}}}(\hat{\mathscr{H}}_{n,k})$. Conditioned on the fact that $B(x_0, r(x_0; p'_{\mathrm{I}}))$ is contained in the acceptance region, we can borrow the idea of Lemma 14 to say that with probability $\geq 1 - \delta^2/4$, we have that $\geq k + 1$ of the points drawn as rejection samples fall in $B(x_0, r(x_0; p'_{\mathrm{I}}))$, where the adjustment to p'_{I} comes from the fact that we don't have an upper bound on the measure of acceptance region.

Note that $\mathrm{NL}_{\neg H,\mathrm{I}}(x_0, \mathcal{S}) \subseteq \left\{ \rho(x_0, X_{(k+1)}(x_0)) > r(x_0; p_{\mathrm{I}}) \right\}$. Thus, as before, we can directly apply the multiplicative Chernoff bound to control the probability that $< k+1$ of the $n + \pi m$ samples from $\mathbb{P}$ do no fall within the ball $B(x, r(x; p_{\mathrm{I}}))$ (which has mass $\geq p_{\mathrm{I}}$).

Whether prediction is taking place in $\hat{\mathscr{H}}_{n,k}$ or its complement, the probability that the voting is non-representative of the conditional probability of 1 in an augmented ball by a margin $\geq \Delta$ is bounded by $\delta^2/2$ by the argument of Lemma 13. Note that the $k + 1$-NN from either rejection samples or from samples from $\mathbb{P}$ are always well-defined. One can finish the result the same way as in Theorem 2 by using Markov's inequality. □

We will use this result in combination with a couple of results of (Chaudhuri & Dasgupta, 2014) to obtain our final consistency guarantee. In what follows, we use $\partial\hat{\mathscr{H}}_{n,k}$ to denote the boundary of the set $\hat{\mathscr{H}}_{n,k}$ in $\mathcal{X}$. Formally, we let

$$\partial\hat{\mathscr{H}}_{n,k} := \hat{\mathscr{H}}_{n,k} \setminus \mathrm{Int}(\hat{\mathscr{H}}_{n,k}),$$

where we define $\mathrm{Int}(\hat{\mathscr{H}}_{n,k})$ as the set of points in $\hat{\mathscr{H}}_{n,k}$ for which there is some $r > 0$ for which $B^o(x, r) \subset \hat{\mathscr{H}}_{n,k}$. In what follows, we set $\Delta_m := \sqrt{(4/k_{n+m}) \log(2/\delta)}$, and $\vec{p}_m = (c_\delta \frac{k}{n+\pi m}, c_\delta \frac{k}{(1-\pi)m})$, and let $\delta_0 := \{x \in \mathcal{X} : \eta(x) = 1/2\}$.

Lemma 17. *Suppose $(\mathcal{X}, \mu, \rho)$ satisfies the Lebesgue Differentiation Theorem. Fix $\mathcal{S}_n$ such that $\mu(\partial\hat{\mathscr{H}}_{n,k}) = 0$, and further suppose that we have a rule for the growth of k with that of m that obeys $k_{n+m} \to \infty$ as $m \to \infty$. Then there exists $\mathcal{X}_{\text{null}} \subset \mathcal{X}$ with $\mu(\mathcal{X}_{\text{null}}) = 0$ such that*

$$x \in \mathcal{X} \setminus \mathcal{X}_{\text{null}},\ \eta(x) \neq \frac{1}{2} \implies x \in \mathcal{X} \setminus \tilde{\partial}^{\hat{\mathscr{H}}_{n,k}}_{\vec{p},\Delta}$$

for some $\vec{p}$ and some $\Delta > 0$.

Proof. Very similarly to (Chaudhuri & Dasgupta, 2014) Lemma 12, set

$$\mathcal{X}_{\text{null, ne}} := \left\{ x \in \mathcal{X} : \lim_{r \downarrow 0} \eta\left(B(x,r)\right) \neq \eta(x) \right\}.$$

By the Lebesgue Differentiation Theorem, $\mu(\mathcal{X}_{\text{null, ne}}) = 0$. Let

$$\mathcal{X}_{\text{null}} := \partial \hat{\mathscr{H}}_{n,k} \cup \mathcal{X}_{\text{null, ne}}.$$

Note that $\mu(\mathcal{X}_{\text{null}}) = 0$ as well.

Now fix $x \notin \mathcal{X}_{\text{null}}$ with $\eta(x) \neq \frac{1}{2}$. Suppose WLOG that $\eta(x) > 1/2$. Set $\Delta = (\eta(x) - 1/2)/2 > 0$ by assumption. By the fact that $\lim_{r \downarrow 0} \eta\left(B(x,r)\right) = \eta(x)$, it holds that there is some r_0 for which $r \leq r_0$ implies that $\eta(B(x,r)) - \frac{1}{2} \geq \Delta$. Because $x \notin \mathcal{X}_{\text{null}}$ means that $x \notin \partial \hat{\mathscr{H}}_{n,k}$, it holds WLOG that $x \in \text{int}(\hat{\mathscr{H}}_{n,k})$, given that $\mathcal{X} = \text{int}(\hat{\mathscr{H}}_{n,k}) \cup \text{int}(\hat{\mathscr{H}}_{n,k}^c) \cup \partial \hat{\mathscr{H}}_{n,k}$. Thus, there is some ball $B^o(x, r_0')$ such that $B^o(x, r_0') \subset \hat{\mathscr{H}}_{n,k}$. Then note that $B(x, r_0'/2) \subseteq B^o(x, r_0')$, so if we choose

$$p = \min\left(\mu(B(x, r_0)), \mu(B(x, r_0'/2)) \right),$$

it holds that $x \in \mathrm{I}_p(\hat{\mathscr{H}}_{n,k}) \cap \mathcal{X}_{p,\Delta}$, and thus that $x \in \mathcal{X} \setminus \tilde{\partial}_{\vec{p},\Delta}^{\hat{\mathscr{H}}_{n,k}}$. □

Lemma 18. *For the settings of $\vec{p}_m$ and Δ_m above, we have*

$$\lim_{m \to \infty} \mu\left(\tilde{\partial}_{\vec{p}_m, \Delta_m}^{\hat{\mathscr{H}}_{n,k}} \setminus \partial_0 \right) = 0.$$

Proof. The proof is identical to that of (Chaudhuri & Dasgupta, 2014) Lemma 13. By definition, $\vec{p}_m \downarrow (0,0)$ and $\Delta_m \downarrow 0$. Let

$$A_m := \tilde{\partial}_{\vec{p}_m, \Delta_m}^{\hat{\mathscr{H}}_{n,k}} \setminus \partial_0.$$

Then $A_1 \supset A_2 \supset \ldots$. By Lemma 17, for $x \in \mathcal{X} \setminus (\mathcal{X}_{\text{null}} \cup \partial_0)$, there is some m such that $x \notin \tilde{\partial}_{\vec{p}_m, \Delta_m}^{\hat{\mathscr{H}}_{n,k}}$. Thus

$$\cap_{m \geq 1} A_m \subset \mathcal{X}_{\text{null}},$$

and so the result follows by the fact that $\mu(\mathcal{X}_{\text{null}}) = 0$ and using continuity from above of the measure μ. □

Theorem 3. *Fix $\zeta \geq 0$, $n \in \mathbb{N}$, and $\pi \in (0,1)$, such that $\bar{k}_\zeta < n$ when $k = k_n$. Suppose $(\mathcal{X}, \rho, \mu)$ satisfies the Lebesgue differentiation theorem, and the following conditions govern the growth of k_{n+m} as $m \to \infty$:*

$$k_{n+m} \to \infty$$
$$k_{n+m}/(n+m) \to 0.$$

Then for any first round sample $\mathcal{S}_n$ such that the boundary of the acceptance region $\partial \hat{\mathscr{H}}_{n,k}$ has $\mu(\partial \hat{\mathscr{H}}_{n,k}) = 0$, it holds that

$$\lim_{m \to \infty} \Pr_{\mathcal{S}_m}\left(R_{n+m}(g_{n,m,k}) - R^* > \epsilon \right) = 0,$$

that is, we convergence to the Bayes-risk in probability over the draw of the second m samples.

Proof. The proof follows the argument of Theorem 2 of (Chaudhuri & Dasgupta, 2014). Firstly, as they note, it holds that

$$R_{n+m}(g_{n,m,k}) - R^* \leq \Pr_{X \sim \mu}\left(\eta(X) \neq \frac{1}{2},\ g_{n,m,k}(X) \neq g^*(X) \right).$$

By $k_{n+m}/(n+m) \to 0$, it must hold that for large enough M_1 that $k_{n+m} < \min\left((1-\pi)m, n + \pi m\right)$ if $m \geq M_1$. Thus, for any choice of δ, if $m \geq M_1$ it holds with probability $\geq 1 - \delta$ that

$$R_{n+m}(g_{n,m,k}) - R^* \leq \mu\left(\tilde{\partial}_{\vec{p}_m, \Delta_m}^{\hat{\mathscr{H}}_{n,k}} \setminus \partial_0 \right) + \delta.$$

Choose $\delta_m = \exp(-k_{n+m}^{1/2})$. Fix $\epsilon > 0$. By Lemma 18, we can choose some M_2 such that $\delta_m \leq \epsilon/2$ and $\mu\left(\tilde{\partial}_{\vec{p}_m,\Delta_m}^{\hat{\mathscr{H}}_{n,k}} \setminus \partial_0\right) \leq \epsilon/2$ for all $m \geq M_2$. Then if $m \geq \max(M_1, M_2)$,

$$\Pr_{\mathcal{S}_m}\left(R_{n+m}(g_{n,m,k}) - R^* > \epsilon\right) \leq \delta_m.$$

The result follows by taking the limit as $m \to \infty$.

□