# Nonparametric Active Learning, Part 1: Smooth Regression Functions

**Steve Hanneke**

*e-mail:* steve.hanneke@gmail.com

**Abstract:** This article presents a general approach to noise-robust active learning for classification problems, based on performing sequential hypothesis tests, modified with an early cut-off stopping criterion. It also proposes a specific instantiation of this approach suitable for learning under a smooth regression function, and proves that this method is minimax optimal (up to logarithmic factors) in this setting, under Tsybakov's noise assumption and a regularity assumption on the marginal density function. Furthermore, the achieved rates are strictly faster than the corresponding minimax rates for learning from random samples (passive learning).

**AMS 2000 subject classifications:** Primary 62L05, 68Q32, 62H30, 68T05; secondary 68T10, 62L10, 68Q25, 68W40, 62G08.
**Keywords and phrases:** active learning, sequential design, selective sampling, nonparametric statistics, Hölder class, Tsybakov noise, statistical learning theory, classification.

## 1. Introduction

Active learning is a sequential design protocol for supervised learning problems, in which a learning algorithm initially has access to a pool of unlabeled data (i.e., just the covariates $X_i$ are observed), and may then sequentially select instances $X_i$ from that pool and request the values of the corresponding labels (response variables $Y_i$), one at a time. The objective is to learn a low-risk predictor $\hat{f}$ mapping any $X$ to an estimate of the corresponding $Y$. We are interested in bounding the achievable guarantees on the risk as a function of the number $n$ of label requests. Such a bound is particularly interesting when it is significantly smaller than the analogous results obtainable for $n$ *random* $(X, Y)$ samples (a setting we refer to as *passive* learning). In the present work, we focus on the case of binary classification (i.e., pattern recognition), where $Y \in \{-1, 1\}$, and the results established here are proven under the assumptions that the regression function is smooth and satisfies Tsybakov's noise condition, and the density of $X$ satisfies the strong density assumption (bounded away from 0 on its support, which is a regular set).

The practical motivation for using active learning is that, in many applications of supervised learning, the bottleneck in time and effort is the process of labeling the collection of unlabeled samples. For instance, if we desire a classifier to automatically label webpages by whether they are about politics or not, we can very easily obtain a large collection of webpages ($X_i$ samples), but to annotate them with the corresponding labels $Y_i$ (whether they are about politics or not) requires a human labeler to read the pages individually and provide the corresponding label. Obtaining the required large number of such labeled samples necessary to train a modern high-dimensional classifier can be an extremely time-consuming process. The hope is that, by *sequentially* selecting the samples $X_i$ to be labeled, we can focus the labeler's efforts on only the most informative and non-redundant samples given the labels he or she has already provided, and thereby reduce the total number of labels required to train a classifier of a desired accuracy.

The objective in this work is to propose an active learning strategy that is *practical*, *general*, and can provide *near-optimal* rates of convergence under standard conditions. Toward these ends, we present a general abstract method for active learning, which can be instantiated in a variety of natural ways by specifying various subroutines. The method is based on the idea of using label requests for samples in local neighborhoods, and performing sequential hypothesis tests to identify the sign of the regression function in those neighborhoods. We find that having these kinds of well-placed tightly-clustered pockets of labeled samples can be more valuable to a learning method than as many random samples. However, if the labels in some regions are very noisy, then we must be careful not to exhaust too many label requests in those regions. For this reason, we employ a *cut-off*, so that if the sequential test does not halt within some $\kappa$ number of label requests, we simply give up on identifying the optimal classification in that region. We also allow this cut-off $\kappa$ to vary over time as the algorithm runs. We refer to the resulting method as using a *Tiered Cut-off in a Test for the Optimal Classification*, which admits the convenient acronym TICTOC.

The present work instantiates this general approach for the specific scenario of learning under a Hölder smooth regression function. We select the locations of the local neighborhoods by using estimates of the regression function at previous locations, together with the smoothness assumption, to identify a random location at which the optimal classification cannot already be inferred from the data collected so far. We then use the above TICTOC strategy to attempt to identify the optimal classification at the selected location, if possible, with a number of label requests not exceeding a well-chosen cut-off value. This repeats for a number of rounds until we reach a set budget on the number of label requests. At that point, we construct a classifier based on the inferred optimal classifications at the chosen locations. Specifically, we use these as a training set in a *nearest neighbor* classifier.

We prove that, under the assumption of a Hölder smooth regression function, together with Tsybakov's noise condition and an assumption that the density of $X$ is bounded away from zero and has regular support, this method obtains the minimax optimal rate of convergence in expected excess classification risk (up to log factors), in terms of the allowed number of label requests $n$. Furthermore, this minimax rate can be significantly faster than the corresponding optimal rate of convergence achievable with $n$ random labeled samples (i.e., passive learning), established by Audibert & Tsybakov [1].

## 2. Main Results

To state the results formally, we first introduce a few basic definitions. Let $(\mathcal{X}, \rho)$ be a metric space, where $\mathcal{X}$ is referred to as the *instance space*, and is equipped with the Borel $\sigma$-algebra generated by $\rho$ to define the measurable sets. Let $\mathcal{Y} = \{-1, 1\}$

denote the *label space*. For any probability measure $P$ over $\mathcal{X} \times \mathcal{Y}$, and $(X, Y) \sim P$, denote by $\eta(x; P) = \mathbb{E}[Y|X = x]$, the *regression function* at the point $x \in \mathcal{X}$, and $f_P^\star(x) = \mathrm{sign}(\eta(x; P)) \in \mathcal{Y}$, the *optimal classification* at $x$. Also, for any measurable $f : \mathcal{X} \to \mathcal{Y}$ (called a *classifier*), denote $\mathcal{R}(f; P) = P((x, y) : f(x) \neq y)$, the *error rate* (or *classification risk*) of $f$. Also, we generally denote by $P_X$ the marginal distribution of $P$ over $\mathcal{X}$.

In the learning problem, for some $u \in \mathbb{N}$, there is an implicit $(\mathcal{X} \times \mathcal{Y})^u$-valued random variable $\mathcal{Z}_u = \{(X_1, Y_1), \ldots, (X_u, Y_u)\}$: the data. The active learning setting, described informally in the previous section, may then be formalized as follows. An *active learning* algorithm is an estimator, taking as input a *label budget* $n \in \mathbb{N}$, and which is permitted access to the data $\mathcal{Z}_u$ via the following sequential protocol. Initially, only the *unlabeled* data $X_1, \ldots, X_u$ are accessible to the algorithm. The algorithm may then select an index $i_1 \in \{1, \ldots, u\}$ and "request" access to the associated label $Y_{i_1}$, the value of which it is then permitted access to. It may then select another index $i_2 \in \{1, \ldots, u\}$ and "request" access to $Y_{i_2}$, and so on. This continues for up to $n$ label requests. To be clear, if the algorithm requests the same label $Y_{i_t}$ twice, the second request is redundant (i.e., there is only one copy of each label). Finally, the algorithm produces a classifier $\hat{f}_{n,u}$. We denote by $\mathbb{A}_a$ the set of all active learning algorithms.

For the sake of comparison, we will also discuss *passive* learning algorithms, which produce a classifier $\hat{f}$ based on $n$ labeled samples: $(X_1, Y_1), \ldots, (X_n, Y_n)$. To unify the notation, for our purposes we can equivalently define a passive learning algorithm as a special type of active learning algorithm, which for any $n \leq u$, always chooses $i_t = t$ for $t \in \{1, \ldots, n\}$, and has no dependence on $\{X_i : i > n\}$ (though this last part can be removed without affecting the claims below). We then denote by $\mathbb{A}_p$ the set of all passive learning algorithms.

For a given distribution $P$, we will be interested in guarantees on the error rate $\mathcal{R}(\hat{f}_{n,u}; P)$ of the classifier produced by a given active (or passive) learning algorithm, under the condition that $\mathcal{Z}_u \sim P^u$. To express such guarantees, for any $\mathcal{A}_a \in \mathbb{A}_a$ and $n, u \in \mathbb{N}$, define the random variable $\mathcal{R}(n, u; \mathcal{A}_a, P)$ as the value $\mathcal{R}(\hat{f}_{n,u}; P)$, for $\hat{f}_{n,u}$ the classifier returned by $\mathcal{A}_a(n)$ when we define $\mathcal{Z}_u$ to have distribution $P^u$.

As mentioned, in this article, we are interested in the achievable risk guarantees under certain assumptions on the distribution $P$. These assumptions are taken directly from the work of Audibert and Tsybakov [1] (who determine the optimal rates for passive learning under these assumptions). Specifically, we work under the assumption that the regression function is Hölder smooth, and that $P$ satisfies Tsybakov's noise condition, along with an additional assumption on the density function of the marginal $P_X$. Although the method presented below should often be reasonable under some suitable generalization of these conditions to general metric spaces, for simplicity we state the theoretical results for the specific case where $\mathcal{X} = \mathbb{R}^d$, for any $d \in \mathbb{N}$, with $\rho$ the Euclidean metric. The assumptions are formalized as follows.

**Hölder smoothness assumption** For any finite constants $\beta \in (0, 1]$ and $L \geq 1$, define the *Hölder space* $\Sigma(L, \beta)$ as the set of functions $f : \mathcal{X} \to [-1, 1]$ such that, $\forall x, x' \in \mathcal{X}$,

$$|f(x) - f(x')| \leq L\rho(x, x')^\beta.$$

Hölder spaces commonly arise in the literatures on nonparametric regression and density estimation, as they provide a natural and well-quantified notion of *smoothness* for a function. See, for instance, [1, 7, 21, 24], for further discussion of the properties of Hölder spaces, and their relevance to nonparametric statistics.

**Strong density assumption** Let $\lambda$ denote the Lebesgue measure on $\mathbb{R}^d$, and for any $x \in \mathcal{X}$ and $r \geq 0$, denote $\mathrm{B}(x,r) = \{x' \in \mathcal{X} : \rho(x,x') \leq r\}$. For $c_0, r_0 \in (0,1]$, we say a Lebesgue measurable set $A \subseteq \mathbb{R}^d$ is $(c_0, r_0)$*-regular* if $\forall x \in A$, $\forall r \in (0, r_0]$, $\lambda(A \cap \mathrm{B}(x,r)) \geq c_0 \lambda(\mathrm{B}(x,r))$. For constants $\mu_{\min}, c_0, r_0 \in (0,1]$, let $\mathrm{SD}(\mu_{\min}, c_0, r_0)$ denote the set of probability measures $P$ over $\mathbb{R}^d$ having a density $p$ with respect to $\lambda$ with $(c_0, r_0)$-regular support $\mathrm{supp}(p) = \{x \in \mathbb{R}^d : p(x) > 0\}$, which satisfies $p(x) \geq \mu_{\min}$ for all $x \in \mathrm{supp}(p)$. The distributions in $\mathrm{SD}(\mu_{\min}, c_0, r_0)$ are said to satisfy the *strong density assumption*. We note that this is actually a slightly weaker version of this assumption than originally studied by Audibert and Tsybakov [1]; however, one can easily verify that the result attributed to that work below (namely (1)) remains valid under this weaker version.

**Tsybakov's noise assumption** Finally, we will quantify the noisiness of the distribution $P$ using Tsybakov's noise condition. Specifically, for constants $a \in [1, \infty)$ and $\alpha \in (0,1)$, let $\mathrm{TN}(a, \alpha)$ denote the set of probability measures $P$ over $\mathcal{X} \times \mathcal{Y}$ such that $\forall \varepsilon > 0$,

$$P_X(x : |\eta(x;P)| \leq \varepsilon) \leq a^{\frac{1}{1-\alpha}} \varepsilon^{\frac{\alpha}{1-\alpha}}.$$

This convenient condition (in various forms) has been studied a great deal in recent years, in both the passive and active learning literatures (e.g., [4, 5, 9–13, 15, 17–20, 22, 25]). Various interpretations and sufficient conditions for this assumption have appeared in the literature. In combination with the other assumptions above, it can often (very loosely) be interpreted as restricting how "flat" the regression function $\eta(\cdot;P)$ can typically be, near the decision boundary of $f_P^\star$: larger $\alpha$ values indicate that $\eta(x;P)$ typically makes a steep transition as it changes sign (i.e., as $x$ approaches the decision boundary), while smaller $\alpha$ values allow for $\eta(x;P)$ to hug 0 more closely as $x$ approaches the decision boundary. See the above references for various alternative forms of the assumption (some stronger, some weaker), and further discussions of sufficient conditions for it to holds.

We will be interested in distributions $P$ satisfying the above conditions. Denote $\Xi = (0,1] \times [1,\infty) \times (0,1]^3 \times [1,\infty) \times (0,1)$. For any $\xi = (\beta, L, \mu_{\min}, c_0, r_0, a, \alpha) \in \Xi$, let us denote by $\mathcal{P}(\xi)$ the set of all probability measures $P$ over $\mathcal{X} \times \mathcal{Y}$ contained in $\mathrm{TN}(a,\alpha)$, with marginal $P_X$ contained in the set $\mathrm{SD}(\mu_{\min}, c_0, r_0)$, and with regression function $\eta(\cdot;P)$ contained in $\Sigma(L,\beta)$. Audibert and Tsybakov [1] have established that, $\forall \xi = (\beta, L, \mu_{\min}, c_0, r_0, a, \alpha) \in \Xi$,

$$\inf_{\mathcal{A}_p \in \mathbb{A}_p} \sup_{P \in \mathcal{P}(\xi)} \mathbb{E}[\mathcal{R}(n,u;\mathcal{A}_p,P)] - \mathcal{R}(f_P^\star;P) = \Theta\left(n^{-\frac{\beta}{(2\beta+d)(1-\alpha)}}\right). \tag{1}$$

In contrast, the following is the main result of the present work, establishing an improved optimal rate for active learning methods. In particular, we will show that this rate is achieved by a simple method presented in Section 4, based on the TicToc strategy sketched above.

**Theorem 1.** *For any $\xi = (\beta, L, \mu_{\min}, c_0, r_0, a, \alpha) \in \Xi$ satisfying $\frac{\alpha}{1-\alpha} \leq \frac{d}{\beta}$,*

$$\inf_{\mathcal{A}_a \in \mathbb{A}_a} \inf_{u \in \mathbb{N}} \sup_{P \in \mathcal{P}(\xi)} \mathbb{E}[\mathcal{R}(n, u; \mathcal{A}_a, P)] - \mathcal{R}(f_P^\star; P) = \tilde{\Theta}\left(n^{-\frac{\beta}{(2\beta+d)(1-\alpha)-\alpha\beta}}\right).$$

*Furthermore, this rate remains valid with $u$ bounded by an $\tilde{O}\left(n^{\frac{(2\beta+d)(1-\alpha)}{(2\beta+d)(1-\alpha)-\alpha\beta}}\right)$ sequence.*

Note that this represents an improvement over the rate (1) established by Audibert and Tsybakov [1] for passive learning. The lower bound in Theorem 1 was previously established by Minsker [19, 20]. An upper bound, matching up to logarithmic factors, was also established by Minsker, but under stronger assumptions on $P$, and via a somewhat more-specialized method (though, interestingly, also allowing an extension to Hölder classes with higher-order smoothness, beyond what is studied in the present work). The main contribution of the present work is therefore to develop a *simple* and *general* method, and a corresponding analysis establishing near-optimality under the above conditions (without additional restrictions). The hope is that this method is simple enough that it (or a suitable variant of it) may actually be useful in practice.

To be clear, the asymptotic notations $\Theta(f(n))$ and $O(f(n))$ treat all values except the number of labels $n$ as constant. The same will be true of asymptotic claims below involving a variable $\varepsilon$ taken to approach 0. Also, the modifications $\tilde{O}$ and $\tilde{\Theta}$ indicate that there may be additional logarithmic factors. The actual logarithmic factors obtained in the upper bound will be made explicit in Theorem 2 below.

We can equivalently express Theorem 1 as a result on the *sample complexity* (and indeed, this is the form in which we prove the result below). Specifically, for the active learning method $\mathcal{A}_a = $ ActiveAlg introduced below, there are values $n$ and $u$ sufficient to achieve $\mathbb{E}[\mathcal{R}(n, u; \mathcal{A}_a, P)] - \mathcal{R}(f_P^\star; P) \leq \varepsilon$, which satisfy

$$n = \tilde{O}\left(\left(\frac{1}{\varepsilon}\right)^{\frac{d}{\beta}(1-\alpha)+2-3\alpha}\right) = \tilde{O}\left(\left(\frac{1}{\varepsilon}\right)^{-\alpha+\frac{(2\beta+d)(1-\alpha)}{\beta}}\right)$$

and

$$u = \tilde{O}\left(\left(\frac{1}{\varepsilon}\right)^{\frac{(2\beta+d)(1-\alpha)}{\beta}}\right).$$

Theorem 1 further implies that a value of $n$ of the above form is *minimal* such that the minimax expected excess risk of active learning is bounded by $\varepsilon$. Furthermore, from (1), we see that a value of $u$ of the above form is also minimal such that the minimax expected excess risk of *passive* learning is bounded by $\varepsilon$ when $n = u$ [1]. This indicates that the active learning method below effectively achieves the same excess error guarantee $\varepsilon$ as an optimal passive learning method that requires *all* $u$ samples to be labeled, but does so while requesting only an $\tilde{O}(\varepsilon^{\alpha})$ fraction of these $u$ labels. Interestingly, this is precisely the *same* type of improvement of active over passive achievable (at best) in the case of learning with general VC classes [13].

## 3. Relation to Prior Work

The subject of optimal rates of convergence in classification under smoothness conditions on the regression function was studied by Audibert and Tsybakov [1], with the motivation of providing an analysis of *plug-in* learning rules: that is, classifiers that predict according to the sign of an estimate of the regression function. They established a general result for such plug-in rules, whereby one can convert rates of convergence for a tail bound on the point-wise risk of a regression estimator into results for the classification error of the corresponding plug-in classifier, under Tsybakov's noise condition. Plugging in such a bound for an appropriate nonparametric regression estimator, holding under the above assumptions on the regression function $\eta(\cdot; P)$ and density of $P_X$, they immediately arrive at the upper bound in (1) above. In addition to (1), they also establish an extension of this result for Hölder classes with *higher-order* smoothness: that is, with bounded derivatives up to a given order, and the Hölder smoothness condition above holding for the derivatives of that order. They additionally study weaker forms of the density assumption $\mathrm{SD}(\mu_{\min}, c_0, r_0)$, though the resulting optimal rates are slower in that case. The present work leaves open the interesting questions of extension of Theorem 1 to include higher-order smoothness assumptions or the "mild" density assumption of Audibert and Tsybakov [1].

Györfi [8] has proposed a generalization of the Hölder smoothness assumption, in the context of analyzing $k$-nearest neighbors regression and classification estimators. Specifically, he defines a notion of smoothness of $\eta(\cdot; P)$ *relative* to the marginal distribution $P_X$, based on the difference between $\eta(x; P)$ and the *average* value of $\eta(\cdot; P)$ in a ball $\mathrm{B}(x, r)$ if a given probability. This generalization admits a much greater variety of distributions $P$, while retaining the essential features of the Hölder smoothness assumption needed for the analysis of many nonparametric regression estimators such as the $k$-nearest neighbors estimator. Chaudhuri and Dasgupta [6] have recently carried out a very general (and fairly tight) analysis of the $k$-nearest neighbors estimator in general metric spaces. They also propose a generalization of Györfi's smoothness condition to general metric spaces, and discuss the implications of their general analysis under this condition. In particular, their bounds imply that the $k$-nearest neighbors classifier achieves a rate of convergence on the order of (1) in the special case of the conditions stated in the previous section. While the method proposed in the present work is well-defined for general metric spaces, and appears to be quite reasonable in that setting when the regression function is smooth in that metric, our analysis is restricted to the specific setting of the Euclidean metric on $\mathbb{R}^d$ for simplicity. Furthermore, we leave for future work the problem of instantiating the general method in a way that admits analysis under the more-general $P_X$-relative smoothness assumptions studied by [6, 8].

In the context of active learning, Minsker [19, 20] studies a setting close to that described above, and establishes an optimal rate of the same form as that in Theorem 1. Indeed, as mentioned above, the lower bound in Theorem 1 is directly supplied by Minsker's result. However, due to his reliance on more-restrictive assumptions, the upper bound above does not follow from that work. Specifically, Minsker assumes $P \in \mathrm{TN}(a, \alpha)$ and $\eta(\cdot; P) \in \Sigma(L, \beta)$ as above, plus an assumption on $P_X$ that can be viewed as analogous to the strong density assumption above. However, he also makes an assumption on $\eta(\cdot; P)$, relating the $L_2$ and $L_\infty$ approximation losses of certain piecewise constant or polynomial approximations of $\eta(\cdot; P)$ in the vicinity of the optimal decision boundary. This assumption is quite specific to the requirements of the analysis of the active learning method proposed in that work. As such, one of the contributions of the present work is establishing these upper bounds under the original assumptions of Audibert and Tsybakov [1], without additional restrictions. In addition to this, the method proposed below seems to enjoy some practical advantages, in its simplicity, and its milder reliance on the specific assumptions of the analysis. Interestingly, unlike the present work, Minsker [20] also establishes a result under the higher-order smoothness assumptions studied by Audibert and Tsybakov [1]. Whether or not these rates for active learning with higher-order smoothness remain valid, without the additional restrictions of [20], remains an interesting open question.

In a recent article, Kontorovich, Sabato and Urner [16] propose an active learning method, which bears a number of similarities in its form to the method presented below. Like the method below, it is based on using a number of label requests in local regions to attempt to identify the optimal classification. Additionally, the final classifier produced by both methods is based on a nearest neighbor rule, constructed by using seed points and inferred region classifications. However, the details of the setting, analysis, and results are entirely different from the present work, and the method differs from that discussed here in a number of important ways. In particular, the method of Kontorovich, Sabato and Urner [16] uses the same number of label requests in each local region, and the locations of the seed points are determined a priori (so as to form a cover) given an appropriate resolution for the cover. In contrast, the method proposed in the present work uses sequential tests (modified to involve a cut-off) to adaptively determine the number of label requests needed for each local region, which is important for adapting to the variability in the noise rate across regions, characteristic of Tsybakov's noise condition. Furthermore, the locations of the seed points are also chosen adaptively in the method below, effectively allowing the resolution of the cover to vary per region as needed. One can show that both of these types of adaptivity are necessary to achieve the optimal rate in Theorem 1 (though see Section 5.2 for a related discussion). However, it is worth noting that the method of Kontorovich, Sabato and Urner [16] does enjoy the favorable property that it couples well with a model selection method they propose, and thereby can be made adaptive to the optimal value of a certain parameter appearing in their results (namely, the resolution of the cover used to select their seed points for the local regions). It would be interesting to determine whether a related technique might enable the method of the present work to adapt to some of the parameters of the assumptions above.

Perhaps the most directly relevant work to the approach proposed here is the analysis of optimal rates for active learning with general VC classes under various noise conditions by Hanneke and Yang [13]. The upper bounds established in that work, for several of the noise models they study, are proven by analyzing a general active learning method, which can be viewed as a variant of the TicToc strategy studied in the present work. There are a number of additional details in that work, necessary for the method to apply to general VC classes, but the essence of the strategy remains the same. Indeed, since this strategy appears to yield near-optimal rates for active learning in a variety of settings, one of the objectives of the present work is to distill this approach into a simple and general form, which can then be instantiated in each specific context by specifying certain subroutines.

## 4. Active Learning with TicToc

We now present the abstract form of the new active learning algorithm, for a given $n$, $u$, and $\mathcal{Z}_u$. The algorithm is described in terms of several subroutines (namely, GetSeed, TicToc, and Learn), the specifications of which will affect the behavior of the algorithm; we explain the naming and roles of these below. The specific method achieving the rate in Theorem 1 will be a version of this abstract algorithm, characterized by a particular choice of these subroutines, as described below. To be clear, all of these subroutines are allowed access to $X_1, \ldots, X_u$, and the TicToc subroutine makes label requests as well. For notational simplicity, we make this dependence on $X_1, \ldots, X_u$ implicit, so that these values are not explicitly stated as arguments to the subroutines.

```
Algorithm: ActiveAlg
Input: Label budget n
Output: Classifier f̂_{n,u}.
-------------------------------------------
1. t ← 0, 𝕃 ← {}
2. For m = 1, 2, ...
3.    s_m ← GetSeed(𝕃, m, t, n)
4.    (ℒ_m, t) ← TicToc(s_m, m, t, n)
5.    𝕃 ← 𝕃 ∪ {(s_m, ℒ_m)}
6.    If t = n or m = u
7.       Return f̂_{n,u} ← Learn(𝕃)
```

The general idea behind this approach is that GetSeed chooses indices $s_m$ of points $X_{s_m}$ from the unlabeled pool that seem in some sense *informative*. These are referred to as the *seeds*. These seeds are then given to the TicToc subroutine, which is perhaps the most important part of this algorithm. This subroutine is charged with the task of identifying the optimal classification $f_P^\star(X_{s_m})$ of these seed points, if it can do so within a reasonable number of label requests. In particular, this step is the source of the claimed advantages over passive learning, and is the subject of much discussion and analysis below. The final step then uses the accumulated labeled data set to run a standard passive learning method, thereby producing the returned classifier.

### 4.1. *The* TicToc *Subroutine*

Since it is central to the algorithm, being the source of labeled samples used by the other subroutines, we begin by describing the general form of the TicToc subroutine. As mentioned, the TicToc subroutine is intended to attempt to identify the optimal classification $f_P^\star(X_{s_m})$ of the seed point $X_{s_m}$. Generally, the *unachievable ideal* for this would be to somehow obtain a sequence of conditionally independent *copies* of $Y_{s_m}$ given $(s_m, X_{s_m})$, and then perform a *sequential hypothesis test* to determine whether $\mathbb{E}[Y_{s_m}|(s_m, X_{s_m})]$ is positive or negative. Indeed, if one could somehow obtain such copies (for instance, in the case of $Y_{s_m}$ being the result of a randomized computer simulation), the instantiation of this subroutine presented below can be significantly simplified. However, since such independent copies would not be available in most applications of active learning, we instead propose using the *nearest neighbors* of $X_{s_m}$ in the unlabeled pool as *surrogate points*. This is reasonable in the present context, given that we are interested in the case of *smooth* regression functions $\eta(\cdot; P)$. Specifically, for any $x \in \mathcal{X}$ and $k \in \{1, \ldots, u\}$, define

$$N_k(x) = \underset{i \in \{1,\ldots,u\} \setminus \{N_{k'}(x) : k' < k\}}{\operatorname{argmin}} \rho(X_i, x),$$

where we may break ties by any (consistent, deterministic) means that depends on $x$ and the $X_i$ sequence (including their indices), but is independent of the $Y_i$ sequence given $x$ and the $X_i$ sequence. For completeness, also define, for integers $k > u$, $N_k(x) = N_1(x)$ (or any other arbitrary index in $\{1, \ldots, u\}$). Then we propose to perform the same kind of sequential hypothesis test as mentioned above, except instead of independent copies of $Y_{s_m}$, we request the values of $Y_{N_k(X_{s_m})}$ for $k = 1, 2, \ldots$ until we can determine the sign of $\eta(X_{s_m}; P)$. We refer to this idea as a *Test for the Optimal Classification*, abbreviated TOC, using *nearest neighbors surrogate points*.

However, we must be careful with the above strategy, since seed points $X_{s_m}$ with $\eta(X_{s_m}; P)$ very close to 0 can potentially require a very large number of label requests to determine their optimal classifications: roughly proportional to $|\eta(X_{s_m}; P)|^{-2}$. In a sense, this fact is doubly important because points $x$ with $\eta(x; P)$ close to 0 are also less influential to the error rate of a classifier: that is, for a classifier $f$ with a given value of $P_X(x : f(x) \neq f_P^\star(x))$, the excess error rate $\mathcal{R}(f; P) - \mathcal{R}(f_P^\star; P)$ will be smaller in the case where the points in $\{x : f(x) \neq f_P^\star(x)\}$ have $\eta(x; P)$ close to 0, compared to the case where these points have $\eta(x; P)$ far from 0. For these reasons, we need to modify the above sequential hypothesis test so that it does not waste too many label requests on such unimportant highly-noisy points. Specifically, we will enforce a *cut-off*, such that if the sequential test has not identified the optimal classification within a given number of label requests (called the *cut-off threshold*), the subroutine terminates anyway and returns whatever labeled data it has accumulated.

In the present setting, it turns out it is in fact possible to achieve the near-optimal rates from Theorem 1 using this combination of a sequential test plus cut-off. However, in many other settings, such as the analysis of general VC classes studied by Hanneke and Yang [13], an additional modification of this strategy is required. Specifically, if the label budget allows for it, we want *some* of the very-noisy points to have accurate estimates of their optimal classifications, so rather than fixing this cut-off threshold to be constant, we can instead monotonically vary the cut-off in tiers as the algorithm proceeds. The effect of this in the overall algorithm is that the set of seed points $X_{s_m}$ whose test for their optimal classification runs to completion contains a disproportionate number of less-noisy points compared to the $P$ distribution. Though we do not make use of this latter capability in the present work (instead using only a single tier, aside from minor adjustments in logarithmic terms), it is worth mentioning, as it can be a crucial aspect of this strategy when applied in other settings.

Taking the above ideas together, we may concisely refer to this general strategy as using a *TIered Cut-off in a Test for the Optimal Classification* (with nearest-neighbor surrogate points), which admits the convenient acronym TICTOC, from which the subroutine takes its name. Formally, the above motivation leads to the following specification of the TICTOC subroutine; we discuss the various quantities referenced in the subroutine below.

Subroutine: TICTOC
Input: Seed point index $s_m$, integers $m, t, n$
Output: Labeled data set $\mathcal{L}_m$, total query counter $t$

1. $k \leftarrow 0$, $\mathcal{L}_m \leftarrow \{\}$
2. While $k < \min\{n-t, \kappa(m,s_m,t,n)\}$ and $\left|\sum_{(x,y)\in\mathcal{L}_m} y\right| < c_T \zeta(\mathcal{L}_m,s_m,m,n)$
3. $\quad k \leftarrow k+1$
4. $\quad$ Request $Y_{N_k(X_{s_m})}$ and set $\mathcal{L}_m \leftarrow \mathcal{L}_m \cup \{(X_{N_k(X_{s_m})}, Y_{N_k(X_{s_m})})\}$
5. Return $(\mathcal{L}_m, t+k)$

Appropriate values of the scalars $\kappa(m,s_m,t,n)$ and $\zeta(\mathcal{L}_m,s_m,m,n)$ referenced in the subroutine are generally based on an analysis of the concentration properties of the sum of requested labels. Generally, we will have $\zeta(\mathcal{L}_m,s_m,m,n)$ on the order of $\sqrt{|\mathcal{L}_m|\log\log(|\mathcal{L}_m|)}$, based on the law of the iterated logarithm. For the specific method below that achieves the rate in Theorem 1, we will take $\kappa(m,s_m,t,n)$ on the order of $n^{\frac{2(1-\alpha)\beta}{(2\beta+d)(1-\alpha)-\alpha\beta}}$, chosen so that points $X_{s_m}$ having reasonably large $|\eta(X_{s_m};P)|$ values will result in termination due to the *second* condition in Step 2. The detailed definitions of these quantities, as used in the analysis, are given below. The constant $c_T$ is generally a numerical constant. For our purposes, it suffices to take $c_T = 2$, though the general analysis below requires only that $c_T > 1$. The general idea is that, with $c_T = 1$, the second condition in Step 2 is strictly providing a sequential test for the optimal classification of $X_{s_m}$, whereas with $c_T > 1$, it provides a slightly stronger guarantee: a nontrivial lower bound on $|\eta(X_{s_m};P)|$ when this quantity is sufficiently far from 0.

For the analysis below, we specifically take the following definitions of $\zeta$ and $\kappa$. For any $x \in [0,\infty)$, denote $\mathrm{Log}(x) = \ln(\max\{x,e\})$. Fix a numerical constant $c_e \in [1,\infty)$; for our purposes, a value $c_e = 9$ will suffice. Then for any $k,s \in \mathbb{N}$, $a \in [1,\infty)$, $\alpha \in (0,1)$, and any $\varepsilon,\delta \in (0,1)$, denoting $\gamma_\varepsilon = \max\{\varepsilon, a^{-1}\varepsilon^{1-\alpha}\}$, define

$$\tilde{\zeta}_\delta(k,s) = \begin{cases} \infty, & \text{if } k < \tilde{c}_1 \mathrm{Log}\left(\frac{2c_e s^2}{\delta}\right) \\ \sqrt{\tilde{c}_2 k \left(2\mathrm{LogLog}(6k) + \mathrm{Log}\left(\frac{c_e s^2}{\delta}\right)\right)}, & \text{otherwise} \end{cases}$$

and

$$\tilde{\kappa}_{\varepsilon,\delta}(s;a,\alpha) = \frac{\tilde{c}_3}{\gamma_\varepsilon^2}\left(\mathrm{LogLog}\left(\frac{\tilde{c}_4}{\gamma_\varepsilon}\right) + \mathrm{Log}\left(\frac{3c_e s^2}{\delta}\right)\right),$$

where for our purposes it suffices to take $\tilde{c}_1 = \frac{173}{4}$, $\tilde{c}_2 = 12$, and $\tilde{c}_3$ and $\tilde{c}_4$ are numerical constants whose sufficient values we discuss below. The specific choice of these last two constants is only important to the constant factors (some depending on $\xi$) in the bound below. The analysis below is carried out with abstract specifications of $\tilde{c}_3$ and $\tilde{c}_4$ subject to constraints (see the discussion at the start of Section 5.1). The definition of $\tilde{\zeta}$ is inspired by the very-recent work of Balsubramani and Ramdas [3] on sequential hypothesis testing based on the recently-established finite-sample version of the law of the iterated logarithm by Balsubramani [2].

For the results established below, we take $\zeta(\mathcal{L}_m, s_m, m, n) = \tilde{\zeta}_\delta(|\mathcal{L}_m|, s_m)$, and $\kappa(m, s_m, t, n) = \tilde{\kappa}_{\varepsilon,\delta}(s_m; a, \alpha)$, for choices of $\varepsilon$ and $\delta$ specified in the theorem below.

### 4.2. Specification of GETSEED

Next, we turn to the specification of an appropriate GETSEED subroutine. There are many reasonable choices for the GETSEED subroutine, and generally it perhaps makes the most sense to use *another active learning method* in this subroutine. In this way, the above algorithm can be viewed as a technique for helping other active learning methods *handle label noise* more effectively. Indeed, a (more sophisticated) variant of this kind of noise-robustification approach underlies the proof by Hanneke and Yang [13] that established the minimax rates for active learning with general VC classes. We now define a specification of GETSEED that is reasonable for our present purposes of learning under the assumption of a smooth regression function. However, as we also discuss below, for certain ranges of the parameters (namely, for $\alpha < 2/3$), we can in fact achieve the optimal asymptotic rate with even the simplistic GETSEED that merely returns a uniform *random* sample (without replacement) from the unlabeled pool.

The general idea is to base GETSEED on a kind of active learning algorithm, but one that typically expects the responses to its queries to contain information sufficient to identify the *noise-free* $f_P^\star(X_{s_m})$ labels, rather than the usual (noisy) $Y_{s_m}$ labels. However, it should also be tolerant to the possibility that if it chooses a point $X_{s_m}$ that is too noisy (i.e., $|\eta(X_{s_m}; P)|$ close to 0), then the response might fall short of identifying this $f_P^\star(X_{s_m})$ value. In the specification below, GETSEED in fact uses slightly more information than merely the inferred $f_P^\star(X_{s_m})$ values, using also a coarse estimate of the magnitude of $|\eta(X_{s_m}; P)|$. Specifically, for our present purposes, consider the following definition. For simplicity, define $s_0 = 0$, which only comes up in the case $m = 1$ (which, in fact, also implies we always have $s_1 = 1$ with this subroutine).

Subroutine: GETSEED
Input: Sequence of pairs $\mathbb{L} = \{(s_{m'}, \mathcal{L}_{m'})\}_{m'<m}$, integers $m, t, n$
Output: Index $s_m$

1. For each $m' < m$, define $\hat{\gamma}_{m'} = \frac{1}{|\mathcal{L}_{m'}|}\left(\left|\sum_{(x,y)\in\mathcal{L}_{m'}} y\right| - \zeta(\mathcal{L}_{m'}, s_{m'}, m', n)\right)$
2. For $s = s_{m-1}+1, s_{m-1}+2, \ldots, u$
3. If $\nexists m' < m$ s.t. $\hat{\gamma}_{m'} \geq 0$ and $\rho(X_{s_{m'}}, X_s) \leq (c_g \hat{\gamma}_{m'}/L)^{1/\beta}$
4. Return $s_m = s$
5. Return $s_m = u$

The constant $c_g$ will be discussed in the analysis; for our purposes, it suffices to choose $c_g = 1/8$. The essential feature of this subroutine is that it uses the labeled data sets constructed so far (i.e., the output of previous calls to TICTOC) and calculates a confidence lower-bound $\hat{\gamma}_{m'} \leq |\eta(X_{s_{m'}}; P)|$ for each of these previous seed points (this requires that the critical value $c_T\zeta$ in TICTOC be defined to give a slightly stronger guarantee on $\sum_{(x,y)\in\mathcal{L}_{m'}} y$ than merely that it have the same sign as $f_P^\star(X_{s_{m'}})$, as we discuss in the analysis below). It then identifies the next point $X_s$ in the unlabeled data sequence such that the value of $f_P^\star(X_s)$ (and indeed, also a lower-bound on $|\eta(X_s;P)|$) is not logically entailed from these confidence lower-bounds. This represents the next point for which we have a certain degree of uncertainty about $f_P^\star(X_s)$ (or, more strictly, about $|\eta(X_s;P)|$).

In the event that the algorithm runs out of unlabeled samples, GETSEED will return in Step 5. This case is completely inconsequential to the result below, and the return value in Step 5 can be set arbitrarily; we have chosen $u$ as a default return value purely to simplify the statement of certain results below. In practice, in this event, one might consider using the remaining label requests in other ways, such as altering the parameters (e.g., reducing $\varepsilon$ or $\delta$) and re-cycling through the unlabeled samples to make additional label requests.

### *4.3. Learning a 1-Nearest Neighbor Classifier*

Finally, we turn to the LEARN subroutine. As with GETSEED, the specification of the LEARN subroutine generally depends on the learning problem. In the analysis of general VC classes, Hanneke and Yang [13] found it appropriate to define LEARN as an empirical risk minimization algorithm. However, for our present context of learning under the assumption of a smooth regression function, we find it most appropriate to use a LEARN subroutine that constructs a flexible nonparametric classifier. In particular, one particularly simple instantiation of the LEARN subroutine is to construct a 1-*nearest neighbor* classifier, using as data set the points $(X_{s_m}, \hat{Y}_{s_m})$, for $\hat{Y}_{s_m} = \mathrm{sign}\left(\sum_{(x,y)\in\mathcal{L}_m} y\right)$, for only those $s_m$ indices for which a sequential test for their optimal classification is able to form a definite conclusion about their $f_P^\star$ value. In our context, this corresponds to those $(X_{s_m}, \hat{Y}_{s_m})$ with $\left|\sum_{(x,y)\in\mathcal{L}_m} y\right| \geq \zeta(\mathcal{L}_m, s_m, m, n)$. Altogether, this would correspond to a kind of locally-adaptive modification of an active learning method of [16], enabling it to adapt to local noise conditions in its choice of queries and placement of centroids. This method is simple in both its form and its analysis.

Formally, for any $t \in \{1,\ldots,u\}$, any distinct $j_1,\ldots,j_t \in \{1,\ldots,u\}$, and any $y_{j_1},\ldots,y_{j_t} \in \mathcal{Y}$, for $\mathcal{S} = \{(X_{j_1},y_{j_1}),\ldots,(X_{j_t},y_{j_t})\}$, and for any $x \in \mathcal{X}$, denote

$$N_1(x;\mathcal{S}) = \underset{j\in\{j_1,\ldots,j_t\}}{\operatorname{argmin}} \rho(X_j, x),$$

where we may break ties arbitrarily (but consistently). We then define the *1-nearest neighbor classifier*

$$\hat{f}_{\mathrm{NN}}(x;\mathcal{S}) = y_{N_1(x;\mathcal{S})}.$$

For completeness, also define $\hat{f}_{\mathrm{NN}}(x;\{\}) = 1$ for all $x$. Then consider the following subroutine.

Subroutine: $\textsc{Learn}_{1\text{NN}}$
Input: Sequence of (index, data set) pairs $\mathbb{L} = \{(s_m, \mathcal{L}_m)\}_m$
Output: Classifier $\hat{f}$

1. $\mathcal{S} \leftarrow \{\}$
2. For $m = 1, \ldots, |\mathbb{L}|$
3. If $\left|\sum_{(x,y)\in\mathcal{L}_m} y\right| \geq \zeta(\mathcal{L}_m, s_m, m, n)$
4. $\hat{Y}_{s_m} \leftarrow \mathrm{sign}\left(\sum_{(x,y)\in\mathcal{L}_m} y\right)$
5. $\mathcal{S} \leftarrow \mathcal{S} \cup \{(X_{s_m}, \hat{Y}_{s_m})\}$
6. Return the 1-nearest neighbor classifier $\hat{f}_{\mathrm{NN}}(\cdot; \mathcal{S})$

In the analysis below, we take $\textsc{Learn} = \textsc{Learn}_{1\text{NN}}$.

## 5. Analysis

We now show that the rate in Theorem 1 is achieved by ActiveAlg with the above specifications of subroutines. Specifically, we have the following result; see the discussion following the theorem for descriptions of how the numerical constants ($c_e$, $\tilde{c}_3$, etc.) should be set for this theorem to hold.

**Theorem 2.** *For any $\xi = (\beta, L, \mu_{\min}, c_0, r_0, a, \alpha) \in \Xi$ satisfying $\frac{\alpha}{1-\alpha} \leq \frac{d}{\beta}$, there exist finite constants $C_1, C_2 \geq 1$ such that, for any $\varepsilon, \delta \in (0, 1/2)$, for $\mathcal{A}_a =$ ActiveAlg (with the above specifications of subroutines), and with $\zeta(\mathcal{L}_m, s_m, m, n) = \tilde{\zeta}_\delta(|\mathcal{L}_m|, s_m)$ and $\kappa(m, s_m, t, n) = \tilde{\kappa}_{\varepsilon,\delta}(s_m; a, \alpha)$, for any $P \in \mathcal{P}(\xi)$ and any $n, u \in \mathbb{N}$ satisfying*

$$u \geq C_1 \left(\frac{1}{\varepsilon}\right)^{2-2\alpha+\frac{d}{\beta}(1-\alpha)} \log\left(\frac{1}{\varepsilon\delta}\right)$$

*and*

$$n \geq C_2 \left(\frac{1}{\varepsilon}\right)^{2-3\alpha+\frac{d}{\beta}(1-\alpha)} \log^2\left(\frac{1}{\varepsilon\delta}\right),$$

*with probability at least $1-\delta$, it holds that $\mathcal{R}(n, u; \mathcal{A}_a, P) - \mathcal{R}(f_P^\star; P) \leq \varepsilon$.*

In particular, this implies that for any $\xi$ as in the theorem, there exist finite constants $C, C', C'' > 0$ such that, for any sequence $u_n \geq C'' n^{\frac{(2\beta+d)(1-\alpha)}{(2\beta+d)(1-\alpha)-\alpha\beta}}$, letting $\varepsilon_n = C'\left(\frac{\log^2(n)}{n}\right)^{\frac{\beta}{(2\beta+d)(1-\alpha)-\alpha\beta}}$, using $\mathcal{A}_a =$ ActiveAlg (with the above specifications of the subroutines), and with $\zeta(\mathcal{L}_m, s_m, m, n) = \tilde{\zeta}_{\varepsilon_n}(|\mathcal{L}_m|, s_m)$ and $\kappa(m, s_m, t, n) = \tilde{\kappa}_{\varepsilon_n,\varepsilon_n}(s_m; a, \alpha)$, for any $P \in \mathcal{P}(\xi)$,

$$\mathbb{E}[\mathcal{R}(n, u_n; \mathcal{A}_a, P)] - \mathcal{R}(f_P^\star; P) \leq C\left(\frac{\log^2(n)}{n}\right)^{\frac{\beta}{(2\beta+d)(1-\alpha)-\alpha\beta}}.$$

The upper bound in Theorem 1 then follows immediately from this, so that (when combined with the lower bound of Minsker [19] discussed above) establishing Theorem 2 will also complete the proof of Theorem 1.

As discussed by Audibert and Tsybakov [1] and Minsker [19], the restriction in the theorem to $\frac{\alpha}{1-\alpha} \leq \frac{d}{\beta}$ is merely a convenience. One can in fact show that the case $\frac{\alpha}{1-\alpha} > \frac{d}{\beta}$ is a fairly *trivial* case, given the other assumptions above. Specifically, note that by the strong density assumption and Hölder smoothness assumption, if $\inf_{x_0 \in \mathrm{supp}(p)} |\eta(x_0; P)| = 0$, then for all $\varepsilon \in (0, 2Lr_0^{\beta}]$, taking any $x_0 \in \mathrm{supp}(p)$ with $|\eta(x_0; P)| \leq \varepsilon/2$,

$$\begin{aligned}
&P_X(x : |\eta(x; P)| \leq \varepsilon) \geq P_X\Big(\mathrm{B}\Big(x_0, (\varepsilon/(2L))^{1/\beta}\Big)\Big) \\
&\geq \mu_{\min}\lambda\Big(\mathrm{B}\Big(x_0, (\varepsilon/(2L))^{1/\beta}\Big) \cap \mathrm{supp}(p)\Big) \\
&\geq \mu_{\min}c_0\lambda\Big(\mathrm{B}\Big(x_0, (\varepsilon/(2L))^{1/\beta}\Big)\Big) = \mu_{\min}c_0\frac{\pi^{d/2}}{\Gamma((d/2)+1)}(\varepsilon/(2L))^{d/\beta},
\end{aligned}$$

and therefore $P$ cannot satisfy Tsybakov's noise condition with a value $\alpha$ satisfying $\frac{\alpha}{1-\alpha} > \frac{d}{\beta}$. Thus, if $\frac{\alpha}{1-\alpha} > \frac{d}{\beta}$, then it must be that $\inf_{x_0 \in \mathrm{supp}(p)} |\eta(x_0; P)| > 0$. In particular, this would imply that $P$ satisfies Tsybakov's noise assumption with values of $\alpha$ *arbitrarily close* to 1 (while maintaining that $a$ is bounded, and in particular $a \leq 1/\inf_{x_0 \in \mathrm{supp}(p)} |\eta(x_0; P)|$). Nearly all of the analysis below (except only Lemma 11, due to some minor simplifying calculations in the proof) holds without the restriction to $\frac{\alpha}{1-\alpha} \leq \frac{d}{\beta}$. Based on careful examination of the (unsimplified) bound on the number of queries in the proof of Lemma 11 below (namely, (26)), we may conclude that when $\alpha$ may be taken arbitrarily close to 1 (while $a$ remains bounded), it is possible to achieve $\mathcal{R}(n, u; \mathcal{A}_a, P) - \mathcal{R}(f_P^{\star}; P) \leq \varepsilon$ with probability at least $1 - \delta$, using any budget $n \geq C\mathrm{Log}^2(\frac{1}{\varepsilon\delta})$, for a constant $C$ depending on $d, \beta, L, \mu_{\min}, c_0, r_0$, and the bound on $a$.

### 5.1. *Proof of Theorem 2*

We will prove Theorem 2 via a sequence of lemmas. For the remainder of this section, fix any $\xi = (\beta, L, \mu_{\min}, c_0, r_0, a, \alpha) \in \Xi$ and $P \in \mathcal{P}(\xi)$, and let $p$ denote the density function of $P_X$ from the strong density assumption. To simplify the notation, we omit the $P$ argument in certain notation below; specifically, $\eta(\cdot)$ abbreviates $\eta(\cdot; P)$, and $f^{\star}$ abbreviates $f_P^{\star}$. We will in fact establish a slightly more general result, allowing certain constant factors to be abstractly specified. Specifically, let $c_e = 9$, and fix any finite constants $c_T > 1$, $\tilde{c}_3, \tilde{c}_4 > 0$, and $\check{C}_0, \check{c}_0, c_g, c_b, \bar{c} \in (0,1)$ such that $c_g(2-\check{c}_0)\check{C}_0 < 1-\check{C}_0$, $c_g(2-\check{c}_0) < 1$, $\bar{c} \leq 1-\check{C}_0\left(1+c_g(2-\check{c}_0)\right)$, $\check{C}_0 \leq c_b \leq 1-\bar{c}-\frac{c_g(2-\check{c}_0)(1+\bar{c})}{1-c_g(2-\check{c}_0)}$, $\tilde{c}_2 c_T^2 \geq 4$, $\tilde{c}_3 \geq \max\Big\{\frac{4\tilde{c}_2 c_T^2}{\check{c}_0^2 c_b^2}, \tilde{c}_1\Big\}$, and $\tilde{c}_4 \geq \frac{\sqrt{24\tilde{c}_2}c_T}{\check{c}_0 c_b}$. For instance, to satisfy these constraints it would suffice to take $\check{C}_0 = 1/8$, $\check{c}_0 = 1/16$, $c_b = 1/4$, $c_g = 1/8$, $\bar{c} = 153/512$, $c_T = 2$, $\tilde{c}_3 = e^{14}$, and $\tilde{c}_4 = e^8$, though it should be possible to further reduce the constant factors in the bound by a more careful choice of these constants. Fix $\varepsilon, \delta \in (0, 1/2)$, and let $\kappa(m, s_m, t, n)$ and $\zeta(\mathcal{L}_m, s_m, m, n)$ be as in theorem statement. Also, for any $s \in \{1, \ldots, u\}$ and $k \in \mathbb{N}$, introduce the abbreviations $\zeta_{k,s} = \tilde{\zeta}_{\delta}(k, s)$ and $\kappa_s = \lceil \tilde{\kappa}_{\varepsilon,\delta}(s; a, \alpha) \rceil$. Also recall from above the definition $\gamma_{\varepsilon} = \max\big\{\varepsilon, a^{-1}\varepsilon^{1-\alpha}\big\}$.

**Lemma 3.** *There exist finite constants $\check{C}_1, \check{C}_2, \check{C}_3 \geq 1$ such that, if*

$$u \geq \check{C}_1 + \check{C}_2 \left(\frac{1}{\varepsilon}\right)^{\frac{(2\beta+d)(1-\alpha)}{\beta}} \mathrm{Log}\left(\frac{\check{C}_3}{\varepsilon\delta}\right), \tag{2}$$

*then on an event $E_1$ of probability at least $1-\delta/c_e$, $\{X_1,\ldots,X_u\} \subseteq \mathrm{supp}(p)$, and for every $s \in \{1,\ldots,u\}$ with $|\eta(X_s)| \geq \check{C}_0\gamma_\varepsilon$, for every $i \in \{1,\ldots,\kappa_s\}$, it holds that $f^\star(X_{N_i(X_s)}) = f^\star(X_s)$ and*

$$\check{c}_0|\eta(X_s)| \leq f^\star(X_s)\eta(X_{N_i(X_s)}) \leq (2-\check{c}_0)\,|\eta(X_s)|,$$

*whereas for every $s \in \{1,\ldots,u\}$ with $|\eta(X_s)| < \check{C}_0\gamma_\varepsilon$, for every $i \in \{1,\ldots,\kappa_s\}$, it holds that*

$$|\eta(X_{N_i(X_s)})| < \check{C}_0(2-\check{c}_0)\gamma_\varepsilon.$$

*Furthermore, if (2) holds, then $u \geq \kappa_s$ for every $s \in \{1,\ldots,u\}$.*

*Proof.* Let $\check{\rho}_\varepsilon = (\check{C}_0(1-\check{c}_0)\gamma_\varepsilon/L)^{1/\beta}$ and fix any $x_0 \in \mathrm{supp}(p)$. Note that the strong density assumption implies that

$$\begin{aligned} P_X(\mathrm{B}(x_0,\check{\rho}_\varepsilon)) &\geq P_X(\mathrm{B}(x_0,\min\{r_0,\check{\rho}_\varepsilon\})) \\ &\geq \mu_{\min}\lambda(\mathrm{supp}(p) \cap \mathrm{B}(x_0,\min\{r_0,\check{\rho}_\varepsilon\})) \\ &\geq \mu_{\min}c_0\lambda(\mathrm{B}(x_0,\min\{r_0,\check{\rho}_\varepsilon\})) = \frac{\mu_{\min}c_0\pi^{d/2}}{\Gamma((d/2)+1)}\min\{r_0,\check{\rho}_\varepsilon\}^d. \end{aligned} \tag{3}$$

Also note that, for any $s \in \mathbb{N}$, a bit of algebra reveals that

$$\kappa_s - 1 \geq 4\mathrm{Log}\left(\frac{2c_e s^2}{\delta}\right).$$

Therefore, by the Chernoff bound, if

$$u - 1 \geq \frac{\Gamma((d/2)+1)}{\pi^{d/2}} \frac{2}{\mu_{\min}c_0\min\{r_0,\check{\rho}_\varepsilon\}^d}\left(\kappa_s - 1\right), \tag{4}$$

then with probability at least $1-\delta/(2c_e s^2)$,

$$|\{X_{s'} : s' \in \{1,\ldots,u\}\setminus\{s\}\} \cap \mathrm{B}(x_0,\check{\rho}_\varepsilon)| \geq (1/2)(u-1)P_X(\mathrm{B}(x_0,\check{\rho}_\varepsilon)) \geq \kappa_s - 1,$$

where the last inequality is by (3). Furthermore, note that if (4) is satisfied for every $s \in \{1,\ldots,u\}$, it follows immediately that $u \geq \kappa_s$ for every $s \in \{1,\ldots,u\}$.

For any $s \in \{1,\ldots,u\}$, since the points $\{X_{s'} : s' \in \{1,\ldots,u\}\setminus\{s\}\}$ are independent of $X_s$, we may apply the above argument to $x_0 = X_s$ under the conditional distribution given $X_s$, on the event that $X_s \in \mathrm{supp}(p)$. Together with the law of total probability, the fact that $X_s \in \mathrm{supp}(p)$ with probability one, and the fact that we always have $X_s \in \mathrm{B}(X_s,\check{\rho}_\varepsilon)$, this implies that with probability at least $1-\delta/(2c_e s^2)$, $X_s \in \mathrm{supp}(p)$ and

$$|\{X_{s'} : s' \in \{1,\ldots,u\}\} \cap \mathrm{B}(X_s,\check{\rho}_\varepsilon)| \geq \kappa_s. \tag{5}$$

By the union bound, this holds simultaneously for all $s \in \{1, \ldots, u\}$ with probability at least $1 - \sum_{s \le u} \delta/(2c_e s^2) \ge 1 - \delta/c_e$. In particular, for any $s \in \{1, \ldots, u\}$, when (5) holds, it must be that

$$\max_{1 \le i \le \kappa_s} \rho\big(X_s, X_{N_i(X_s)}\big) \le \check{\rho}_\varepsilon.$$

Together with the Hölder smoothness assumption and our definition of $\check{\rho}_\varepsilon$ above, this implies

$$\max_{1 \le i \le \kappa_s} \big|\eta(X_{N_i(X_s)}) - \eta(X_s)\big| \le \check{C}_0(1 - \check{c}_0)\gamma_\varepsilon. \tag{6}$$

Therefore, if $|\eta(X_s)| \ge \check{C}_0\gamma_\varepsilon$, then every $i \in \{1, \ldots, \kappa_s\}$ has

$$\begin{aligned} f^\star(X_s)\eta(X_{N_i(X_s)}) &\le f^\star(X_s)\eta(X_s) + \check{C}_0(1 - \check{c}_0)\gamma_\varepsilon \\ &\le f^\star(X_s)\eta(X_s) + (1 - \check{c}_0)|\eta(X_s)| = (2 - \check{c}_0)|\eta(X_s)| \end{aligned}$$

and

$$\begin{aligned} f^\star(X_s)\eta(X_{N_i(X_s)}) &\ge f^\star(X_s)\eta(X_s) - \check{C}_0(1 - \check{c}_0)\gamma_\varepsilon \\ &\ge f^\star(X_s)\eta(X_s) - (1 - \check{c}_0)|\eta(X_s)| = \check{c}_0|\eta(X_s)|. \end{aligned}$$

In particular, since this last quantity is at least $\check{c}_0\check{C}_0\gamma_\varepsilon$, which is strictly positive, and since $f^\star(X_{N_i(X_s)}) = \mathrm{sign}(\eta(X_{N_i(X_s)}))$ and $f^\star(X_s) \in \{-1, 1\}$, we have $f^\star(X_{N_i(X_s)}) = f^\star(X_s)$.

On the other hand, if $|\eta(X_s)| < \check{C}_0\gamma_\varepsilon$, then (6) implies that $\forall i \in \{1, \ldots, \kappa_s\}$,

$$|\eta(X_{N_i(X_s)})| \le |\eta(X_s)| + \check{C}_0(1 - \check{c}_0)\gamma_\varepsilon < \check{C}_0(2 - \check{c}_0)\gamma_\varepsilon.$$

To complete the proof, it remains only to argue that there exists a choice of the constants $\check{C}_1, \check{C}_2, \check{C}_3$ so that (2) suffices to guarantee (4) holds for every $s \in \{1, \ldots, u\}$. The rest of the proof is devoted to establishing this fact. For any $s \in \{1, \ldots, u\}$, we note that

$$1 + \frac{\Gamma((d/2)+1)}{\pi^{d/2}} \frac{2\,(\kappa_s - 1)}{\mu_{\min} c_0 \min\{r_0, \check{\rho}_\varepsilon\}^d} \le \frac{\Gamma((d/2)+1)}{\pi^{d/2}} \frac{2\kappa_s}{\mu_{\min} c_0 \min\{r_0, \check{\rho}_\varepsilon\}^d},$$

and denoting $C_1' = \frac{\Gamma((d/2)+1)}{\pi^{d/2}} \frac{4\tilde{c}_3}{\mu_{\min} c_0}$, this is at most

$$\frac{C_1'}{\gamma_\varepsilon^2}\left(\mathrm{LogLog}\left(\frac{\tilde{c}_4}{\gamma_\varepsilon}\right) + \mathrm{Log}\left(\frac{3c_e s^2}{\delta}\right)\right) \max\left\{\frac{1}{r_0^d}, \left(\frac{L}{\check{C}_0(1 - \check{c}_0)\gamma_\varepsilon}\right)^{\frac{d}{\beta}}\right\}.$$

Letting $C_2' = \frac{C_1'\check{C}_0^2(1-\check{c}_0)^2}{L^2 r_0^{2\beta+d}}$, $C_3' = \mathrm{LogLog}\Big(\frac{\tilde{c}_4(\check{C}_0(1-\check{c}_0))}{L r_0^\beta}\Big) + \mathrm{Log}(3c_e)$, and $C_4' = C_1' a^{2+\frac{d}{\beta}} \left(\frac{L}{\check{C}_0(1-\check{c}_0)}\right)^{\frac{d}{\beta}}$, the above expression is at most

$$\max\begin{cases} C_2'C_3' + 2C_2'\mathrm{Log}\big(\frac{u}{\delta}\big) \\ C_4'\left(\frac{1}{\varepsilon}\right)^{\frac{(2\beta+d)(1-\alpha)}{\beta}}\left(\mathrm{LogLog}\big(\frac{\tilde{c}_4}{\varepsilon}\big) + \mathrm{Log}\Big(\frac{3c_e u^2}{\delta}\Big)\right) \end{cases}.$$

Furthermore (see e.g., Corollary 4.1 of [23]), $u$ is guaranteed to be at least this large as long as it satisfies

$$u \geq \max \begin{cases} 2C_2'C_3' + 4C_2'\mathrm{Log}(C_2') + 4C_2' \ln\left(\frac{1}{\delta}\right) \\ 2C_4' \left(\frac{1}{\varepsilon}\right)^{\frac{(2\beta+d)(1-\alpha)}{\beta}} \left(\ln\left(\frac{C_5'}{\delta}\right) + C_6' \ln\left(\frac{1}{\varepsilon}\right)\right) \end{cases},$$

where $C_5' = 3c_e\tilde{c}_4\tilde{c}_4(C_4')^2$ and $C_6' = \left(1 + \frac{2(2\beta+d)(1-\alpha)}{\beta}\right)$. Since the terms in this maximum are both nonnegative, we can upper bound its value by the sum of the two terms. Thus, the lemma follows (with some loss in the constant factors compared to the above sufficient size of $u$) by taking $\check{C}_1 = 2C_2'C_3' + 4C_2'\mathrm{Log}(C_2')$, $\check{C}_2 = 4C_2' + 2C_4'C_6'$, and $\check{C}_3 = (C_5')^{1/C_6'}$. $\square$

The next lemma follows immediately from Theorem 4 of Balsubramani [2], a finite-sample version of the law of the iterated logarithm for martingales. Its statement is included here for the purpose of self-containment of this article; the interested reader is referred to the original article of Balsubramani [2] for the proof.

**Lemma 4.** *Let $b_1, b_2, \ldots$ be finite positive constants, and let $\{M_i\}_{i=0}^{\infty}$ be a martingale such that $M_0 = 0$ and $\forall t \in \mathbb{N}$, $|M_t - M_{t-1}| \leq b_t$. For any $\delta' \in (0,1)$, with probability at least $1-\delta'$, $\forall t \in \mathbb{N}$ with $\sum_{i=1}^{t} b_i^2 \geq 173\mathrm{Log}\left(\frac{4}{\delta'}\right)$,*

$$|M_t| \leq \sqrt{3\left(\sum_{i=1}^{t} b_i^2\right)\left(2\mathrm{LogLog}\left(\frac{3}{2\max\{|M_t|, b_t\}}\sum_{i=1}^{t} b_i^2\right) + \mathrm{Log}\left(\frac{2}{\delta'}\right)\right)}.$$

We use this lemma to establish the following result.

**Lemma 5.** *For each $s, k \in \{1, \ldots, u\}$, define*

$$\gamma_{s,k}^{\star} = \left(f^{\star}(X_s)\frac{1}{k}\sum_{i=1}^{k} Y_{N_i(X_s)}\right) - \frac{1}{k}\zeta_{k,s}.$$

*There is an event $E_2$ of probability at least $1 - 4\delta/c_e$ such that, if $u$ satisfies (2), then on $E_1 \cap E_2$, for every $s \in \{1, \ldots, u\}$ and every $k \in \{1, \ldots, \kappa_s\}$, each of the following claims holds:*

- *If $|\eta(X_s)| \geq \check{C}_0\gamma_\varepsilon$, then*

$$(2 - \check{c}_0)\,|\eta(X_s)| > \gamma_{s,k}^{\star}. \tag{7}$$

- *For every finite $c \geq 1$, if $|\eta(X_s)| \geq \check{C}_0\gamma_\varepsilon$ and $\left|\sum_{i=1}^{k} Y_{N_i(X_s)}\right| \geq c\zeta_{k,s}$, then*

$$\gamma_{s,k}^{\star} \geq \frac{c-1}{c+1}\check{c}_0|\eta(X_s)|. \tag{8}$$

- *If $|\eta(X_s)| < \check{C}_0\gamma_\varepsilon$, then*

$$\left|\frac{1}{k}\sum_{i=1}^{k} Y_{N_i(X_s)}\right| - \frac{1}{k}\zeta_{k,s} < \check{C}_0(2-\check{c}_0)\gamma_\varepsilon. \tag{9}$$

*Proof.* Suppose $u$ satisfies (2) and fix any $s \in \{1, \ldots, u\}$. First note that, under the conditional distribution given $X_1, \ldots, X_u$, the sequence

$$M_k = \sum_{i=1}^{k} (Y_{N_i(X_s)} - \eta(X_{N_i(X_s)})),$$

$k \in \{1, \ldots, \kappa_s\}$, forms a martingale (with the convention $M_0 = 0$), satisfying $|M_k - M_{k-1}| \leq 2$. Therefore, applying Lemma 4, together with the law of total probability, we have that, on an event $E_{2,s}$ of probability at least $1 - 2\delta/(c_e s^2)$, every $k \in \{1, \ldots, \kappa_s\}$ with $k \geq \frac{173}{4}\mathrm{Log}\left(\frac{2c_e s^2}{\delta}\right)$ satisfies

$$\left| \sum_{i=1}^{k} (Y_{N_i(X_s)} - \eta(X_{N_i(X_s)})) \right| < \sqrt{12k \left( 2\mathrm{LogLog}(6k) + \mathrm{Log}\left( \frac{c_e s^2}{\delta} \right) \right)}.$$

By our definition of $\zeta_{k,s}$, this implies that on $E_{2,s}$, every $k \in \{1, \ldots, \kappa_s\}$ satisfies

$$\left| \sum_{i=1}^{k} (Y_{N_i(X_s)} - \eta(X_{N_i(X_s)})) \right| < \zeta_{k,s}. \tag{10}$$

Note that the left hand side of this inequality equals

$$\left| \left( f^\star(X_s) \sum_{i=1}^{k} Y_{N_i(X_s)} \right) - \left( \sum_{i=1}^{k} f^\star(X_s) \eta(X_{N_i(X_s)}) \right) \right|.$$

Therefore, (10) implies that, on $E_{2,s}$, every $k \in \{1, \ldots, \kappa_s\}$ satisfies

$$\frac{1}{k} \sum_{i=1}^{k} f^\star(X_s) \eta(X_{N_i(X_s)}) > \left( f^\star(X_s) \frac{1}{k} \sum_{i=1}^{k} Y_{N_i(X_s)} \right) - \frac{1}{k} \zeta_{k,s} = \gamma^\star_{s,k}.$$

Furthermore, Lemma 3 implies that, on $E_1$, if $|\eta(X_s)| \geq \check{C}_0 \gamma_\varepsilon$, then the leftmost expression above is at most $(2 - \check{c}_0)\, |\eta(X_s)|$, so that (7) holds.

In the other direction, (10) also implies that, on $E_{2,s}$, every $k \in \{1, \ldots, \kappa_s\}$ satisfies

$$f^\star(X_s) \sum_{i=1}^{k} Y_{N_i(X_s)} > \left( \sum_{i=1}^{k} f^\star(X_s) \eta(X_{N_i(X_s)}) \right) - \zeta_{k,s}.$$

Lemma 3 then implies that, on $E_1$, if $|\eta(X_s)| \geq \check{C}_0 \gamma_\varepsilon$, then the right hand side of this inequality is at least $k\check{c}_0 |\eta(X_s)| - \zeta_{k,s}$, so that

$$f^\star(X_s) \sum_{i=1}^{k} Y_{N_i(X_s)} > k\check{c}_0 |\eta(X_s)| - \zeta_{k,s}. \tag{11}$$

In particular, since $k\check{c}_0 |\eta(X_s)| \geq 0$, (11) implies $f^\star(X_s) \sum_{i=1}^{k} Y_{N_i(X_s)} > -\zeta_{k,s}$, or equivalently $-f^\star(X_s) \sum_{i=1}^{k} Y_{N_i(X_s)} < \zeta_{k,s}$. Thus, since

$$\left| \sum_{i=1}^{k} Y_{N_i(X_s)} \right| = \max\left\{ f^\star(X_s) \sum_{i=1}^{k} Y_{N_i(X_s)}, -f^\star(X_s) \sum_{i=1}^{k} Y_{N_i(X_s)} \right\},$$

if $\left|\sum_{i=1}^{k} Y_{N_i(X_s)}\right| \geq \zeta_{k,s}$ and (11) holds, then it must be that $\left|\sum_{i=1}^{k} Y_{N_i(X_s)}\right| = f^{\star}(X_s)\sum_{i=1}^{k} Y_{N_i(X_s)}$. We therefore have that, if $\left|\sum_{i=1}^{k} Y_{N_i(X_s)}\right| \geq c\zeta_{k,s}$ (for any $c \geq 1$) and (11) holds, then

$$\begin{aligned}
\gamma_{s,k}^{\star} &= \left(f^{\star}(X_s)\frac{1}{k}\sum_{i=1}^{k} Y_{N_i(X_s)}\right) - \frac{2c}{c+1}\frac{1}{k}\zeta_{k,s} + \frac{c-1}{c+1}\frac{1}{k}\zeta_{k,s} \\
&\geq \left(f^{\star}(X_s)\frac{1}{k}\sum_{i=1}^{k} Y_{N_i(X_s)}\right) - \frac{2c}{c+1}\frac{1}{k}\frac{1}{c}\left(f^{\star}(X_s)\sum_{i=1}^{k} Y_{N_i(X_s)}\right) + \frac{c-1}{c+1}\frac{1}{k}\zeta_{k,s} \\
&= \left(1 - \frac{2}{c+1}\right)\frac{1}{k}\left(f^{\star}(X_s)\sum_{i=1}^{k} Y_{N_i(X_s)}\right) + \frac{c-1}{c+1}\frac{1}{k}\zeta_{k,s},
\end{aligned}$$

and (11) implies this last expression is at least as large as

$$\left(1 - \frac{2}{c+1}\right)\left(\check{c}_0|\eta(X_s)| - \frac{1}{k}\zeta_{k,s}\right) + \frac{c-1}{c+1}\frac{1}{k}\zeta_{k,s} = \frac{c-1}{c+1}\check{c}_0|\eta(X_s)|.$$

Altogether, we have that on the event $E_1 \cap E_{2,s}$, if $|\eta(X_s)| \geq \check{C}_0\gamma_{\varepsilon}$, then every $k \in \{1,\dots,\kappa_s\}$ with $\left|\sum_{i=1}^{k} Y_{N_i(X_s)}\right| \geq c\zeta_{k,s}$ satisfies (8).

Finally we turn to establishing (9). For any $k \in \{1,\dots,\kappa_s\}$, (10) also implies

$$\left|\frac{1}{k}\sum_{i=1}^{k} Y_{N_i(X_s)}\right| - \frac{1}{k}\zeta_{k,s} < \left|\frac{1}{k}\sum_{i=1}^{k} \eta(X_{N_i(X_s)})\right| \leq \frac{1}{k}\sum_{i=1}^{k} |\eta(X_{N_i(X_s)})|.$$

Lemma 3 further implies that, on $E_1$, if $|\eta(X_s)| < \check{C}_0\gamma_{\varepsilon}$, then every $i \in \{1,\dots,k\}$ has $|\eta(X_{N_i(X_s)})| < \check{C}_0(2-\check{c}_0)\gamma_{\varepsilon}$. Altogether, we have that on $E_1 \cap E_{2,s}$, if $|\eta(X_s)| < \check{C}_0\gamma_{\varepsilon}$, then every $k \in \{1,\dots,\kappa_s\}$ satisfies (9).

The result now follows by defining $E_2 = \bigcap_{s=1}^{u} E_{2,s}$, which has probability at least $1 - \sum_{s=1}^{u} 2\delta/(c_e s^2) \geq 1 - 4\delta/c_e$ by the union bound. □

**Lemma 6.** *Let $\hat{\mathbb{L}}$ denote the final set $\mathbb{L}$ in* ActiveAlg$(n)$ *(with subroutines as in Theorem 2), and let $\hat{\mathcal{S}}$ denote the final set $\mathcal{S}$ in* $\text{Learn}_{\text{1NN}}(\hat{\mathbb{L}})$. *If $u$ satisfies (2), then on $E_1 \cap E_2$, every $(X_s, \hat{Y}_s) \in \hat{\mathcal{S}}$ with $|\eta(X_s)| \geq \check{C}_0\gamma_{\varepsilon}$ has $\hat{Y}_s = f^{\star}(X_s)$.*

*Proof.* Note that every $(X_s, \hat{Y}_s) \in \hat{\mathcal{S}}$ has $s$ equal to some $s_m$ for some $m$ encountered in the execution of ActiveAlg$(n)$, and furthermore (by definition of $\text{Learn}_{\text{1NN}}$), these values $s_m$ satisfy

$$\left|\sum_{(x,y)\in\mathcal{L}_m} y\right| \geq \zeta_{|\mathcal{L}_m|,s_m}.$$

Also recall (from the definition of TicToc) that $\mathcal{L}_m = \{(X_{N_i(X_{s_m})}, Y_{N_i(X_{s_m})}) : i \leq |\mathcal{L}_m|\}$ and $|\mathcal{L}_m| \leq \kappa_{s_m}$ for each such $m$. Thus, by taking $c = 1$ in (8), Lemma 5 implies that if $u$ satisfies (2), then on $E_1 \cap E_2$, for each of these values $s_m$ with $(X_{s_m}, \hat{Y}_{s_m}) \in \hat{\mathcal{S}}$, if $|\eta(X_{s_m})| \geq \check{C}_0\gamma_\varepsilon$, then

$$\left( f^\star(X_{s_m}) \frac{1}{|\mathcal{L}_m|} \sum_{(x,y)\in\mathcal{L}_m} y \right) - \frac{1}{|\mathcal{L}_m|}\zeta_{|\mathcal{L}_m|,s_m} \geq 0.$$

In particular, since $\frac{1}{|\mathcal{L}_m|}\zeta_{|\mathcal{L}_m|,s_m} > 0$, this implies $\left( f^\star(X_{s_m}) \frac{1}{|\mathcal{L}_m|} \sum_{(x,y)\in\mathcal{L}_m} y \right) > 0$, which means $\operatorname{sign}\left( \sum_{(x,y)\in\mathcal{L}_m} y \right) = f^\star(X_{s_m})$, and therefore $\hat{Y}_{s_m} = f^\star(X_{s_m})$, as claimed. □

**Lemma 7.** *For each $s \in \{1, \ldots, u\}$, define*

$$Q_s = \left\lceil \max\left\{ \frac{4\tilde{c}_2 c_T^2}{\check{c}_0^2 |\eta(X_s)|^2}, \tilde{c}_1 \right\} \left( \operatorname{LogLog}\left( \frac{\sqrt{24\tilde{c}_2} c_T}{\check{c}_0 |\eta(X_s)|} \right) + \operatorname{Log}\left( \frac{3c_e s^2}{\delta} \right) \right) \right\rceil,$$

*with the convention that $Q_s = \infty$ when $\eta(X_s) = 0$. There is an event $E_3$ of probability at least $1 - 2\delta/c_e$ such that, if $u$ satisfies (2), then on $E_1 \cap E_3$, for every $s \in \{1, \ldots, u\}$ with $|\eta(X_s)| \geq c_b\gamma_\varepsilon$ (for $c_b$ defined above, at the top of Section 5.1), for every $t, m \in \mathbb{N}$, the pair $(\mathcal{L}, t')$ that would be returned from* $\textsc{TicToc}(s, m, t, \infty)$ *satisfies $t' - t \leq Q_s$ and*

$$\left| \sum_{(x,y)\in\mathcal{L}} y \right| \geq c_T \zeta_{|\mathcal{L}|,s}.$$

*Proof.* Suppose $u$ satisfies (2) and fix any $s \in \{1, \ldots, u\}$. Recalling that $c_b \geq \check{C}_0$, Lemma 3 implies that on $E_1$, if $|\eta(X_s)| \geq c_b\gamma_\varepsilon$, then every $i \in \{1, \ldots, \kappa_s\}$ has $f^\star(X_s)\eta(X_{N_i(X_s)}) \geq \check{c}_0|\eta(X_s)|$. Also note that, by our choice of the constants $\tilde{c}_1$ and $\tilde{c}_3$, and the fact that $\gamma_\varepsilon \leq 1$, if $|\eta(X_s)| \geq c_b\gamma_\varepsilon$, then $Q_s \leq \kappa_s$ and $Q_s \geq \tilde{c}_1 \operatorname{Log}\left(\frac{2c_e s^2}{\delta}\right)$.

Therefore, by applying Hoeffding's inequality under the conditional distribution given $X_1, \ldots, X_u$, together with the law of total probability, there is an event $E_{3,s}$ of probability at least $1 - \delta/(c_e s^2)$ such that, on $E_1 \cap E_{3,s}$, if $|\eta(X_s)| \geq c_b\gamma_\varepsilon$, then

$$f^\star(X_s) \sum_{i=1}^{Q_s} Y_{N_i(X_s)} \geq \check{c}_0 |\eta(X_s)| Q_s - \sqrt{2Q_s \ln\left(\frac{c_e s^2}{\delta}\right)}. \tag{12}$$

Note that if $Q_s$ satisfies

$$Q_s \geq \frac{1}{\check{c}_0^2 |\eta(X_s)|^2} \left( \sqrt{2\ln\left(\frac{c_e s^2}{\delta}\right)} + c_T \sqrt{\tilde{c}_2 \left( 2\operatorname{LogLog}(6Q_s) + \operatorname{Log}\left(\frac{c_e s^2}{\delta}\right) \right)} \right)^2, \tag{13}$$

then the right hand side of (12) is at least

$$c_T \sqrt{\tilde{c}_2 Q_s \left( 2\operatorname{LogLog}(6Q_s) + \operatorname{Log}\left(\frac{c_e s^2}{\delta}\right) \right)} = c_T \zeta_{Q_s,s}.$$

By a bit of calculus to handle the $\mathrm{LogLog}$ term, one can verify that $Q_s$ indeed satisfies the inequality (13); in fact, its definition is primarily motivated by this inequality, aside from the $\tilde{c}_1$ term which guarantees $Q_s \geq \tilde{c}_1 \mathrm{Log}\left(\frac{2c_e s^2}{\delta}\right)$. In particular, this means that (12) implies $\left|\sum_{i=1}^{Q_s} Y_{N_i(X_s)}\right| \geq c_T \zeta_{Q_s,s}$.

Thus, on $E_1 \cap E_{3,s}$, if $|\eta(X_s)| \geq c_b \gamma_\varepsilon$, then denoting by $k_s$ the *smallest* $k \in \{1,\ldots,\kappa_s\}$ with $k \geq \tilde{c}_1 \mathrm{Log}\left(\frac{2c_e s^2}{\delta}\right)$ and $\left|\sum_{i=1}^{k} Y_{N_i(X_s)}\right| \geq c_T \zeta_{k,s}$, we have that such a $k_s$ exists and satisfies $k_s \leq Q_s \leq \kappa_s$. It follows that, for any $t, m \in \mathbb{N}$, the execution of $\textsc{TicToc}(s,m,t,\infty)$ terminates with $k$ equal this $k_s$ value, and therefore the returned pair $(\mathcal{L}, t')$ has $t' - t = |\mathcal{L}| = k_s \leq Q_s$ and

$$\left|\sum_{(x,y)\in\mathcal{L}} y\right| = \left|\sum_{i=1}^{k_s} Y_{N_i(X_s)}\right| \geq c_T \zeta_{k_s,s} = c_T \zeta_{|\mathcal{L}|,s}.$$

The proof is completed by defining $E_3 = \bigcap_{s=1}^{u} E_{3,s}$, and noting that this has probability at least $1 - \sum_{s=1}^{u} \delta/(c_e s^2) \geq 1 - 2\delta/c_e$ by the union bound. $\square$

**Lemma 8.** *Denote $\rho_\varepsilon = (\bar{c}\gamma_\varepsilon/L)^{1/\beta}$ (for $\bar{c}$ as defined at the beginning of Section 5.1), and define $q_\varepsilon = \frac{\pi^{d/2}}{\Gamma((d/2)+1)} \mu_{\min} c_0 \min\{r_0, \rho_\varepsilon/2\}^d$ and*

$$s_{\varepsilon,\delta} = \left\lceil \frac{1}{q_\varepsilon} \ln\left(\frac{2^d c_e}{q_\varepsilon \delta}\right)\right\rceil.$$

*If $u \geq s_{\varepsilon,\delta}$, then there is an event $E_4$ of probability at least $1 - \delta/c_e$, on which*

$$\sup_{x_0 \in \mathrm{supp}(p)} \min_{s \in \{1,\ldots,s_{\varepsilon,\delta}\}} \rho(x_0, X_s) < \rho_\varepsilon. \tag{14}$$

*Proof.* Let $x_1,\ldots,x_M$ denote a maximal $(\rho_\varepsilon/2)$-packing in $\mathrm{supp}(p)$ under the metric $\rho$: that is, a set of points in $\mathrm{supp}(p)$ of maximal cardinality such that $\min_{i\neq j} \rho(x_i,x_j) \geq \rho_\varepsilon/2$. Then (as is well known, e.g. [14]) it also supplies a $(\rho_\varepsilon/2)$-cover of $\mathrm{supp}(p)$: that is, $\mathrm{supp}(p) \subseteq \bigcup_{i\leq M} \mathrm{B}(x_i, \rho_\varepsilon/2)$. Therefore, by the triangle inequality, to satisfy (14) it suffices to have

$$\{X_1,\ldots,X_{s_{\varepsilon,\delta}}\} \cap \{x \in \mathcal{X} : \rho(x,x_i) < \rho_\varepsilon/2\} \neq \emptyset$$

for every $i \in \{1,\ldots,M\}$.

Now, for any constant $c \in (0,1/2]$, by the strong density assumption (which also implies $P_X$ is absolutely continuous with respect to $\lambda$), $\forall i \in \{1,\ldots,M\}$,

$$\begin{aligned} P_X(x : \rho(x,x_i) < c\rho_\varepsilon) &= P_X(\mathrm{B}(x_i, c\rho_\varepsilon)) \geq \mu_{\min} \lambda(\mathrm{B}(x_i, c\rho_\varepsilon) \cap \mathrm{supp}(p)) \\ &\geq \mu_{\min} c_0 \lambda(\mathrm{B}(x_i, \min\{r_0, c\rho_\varepsilon\})) \geq (2c)^d q_\varepsilon. \end{aligned} \tag{15}$$

In particular, with $c = 1/4$, (15) implies that every $i \in \{1,\ldots,M\}$ satisfies $P_X(x : \rho(x,x_i) < \rho_\varepsilon/4) \geq 2^{-d} q_\varepsilon$. Furthermore, since $x_1,\ldots,x_M$ is a $(\rho_\varepsilon/2)$-packing, the triangle inequality implies the sets $\{x : \rho(x,x_i) < \rho_\varepsilon/4\}$ are disjoint over $i \in \{1, \ldots, M\}$. Thus,

$$1 \geq P_X\left(\bigcup_{i \leq M} \{x : \rho(x, x_i) < \rho_\varepsilon/4\}\right) = \sum_{i \leq M} P_X(x : \rho(x, x_i) < \rho_\varepsilon/4) \geq M 2^{-d} q_\varepsilon,$$

from which it immediately follows that $M \leq \frac{2^d}{q_\varepsilon}$.

Additionally, with $c = 1/2$, (15) implies that each $i \in \{1, \ldots, M\}$ satisfies $P_X(x : \rho(x, x_i) < \rho_\varepsilon/2) \geq q_\varepsilon$. Therefore, $\min_{s \in \{1, \ldots, s_{\varepsilon,\delta}\}} \rho(x_i, X_s) < \rho_\varepsilon/2$ holds with probability at least

$$1 - (1 - q_\varepsilon)^{s_{\varepsilon,\delta}} \geq 1 - \exp\{-q_\varepsilon s_{\varepsilon,\delta}\} \geq 1 - \frac{\delta q_\varepsilon}{2^d c_e}.$$

Finally, by the union bound, $\min_{s \in \{1, \ldots, s_{\varepsilon,\delta}\}} \rho(x_i, X_s) < \rho_\varepsilon/2$ holds simultaneously for all $i \in \{1, \ldots, M\}$ with probability at least $1 - \frac{\delta q_\varepsilon}{2^d c_e} M \geq 1 - \frac{\delta}{c_e}$. □

**Lemma 9.** *Let $\hat{\mathcal{S}}$ be as in Lemma 6. Let $\hat{m}$ denote the random variable representing the final value of $m$ upon termination of* ActiveAlg$(n)$ *(due to satisfying the condition in Step 6). If $u$ satisfies (2), on the event $E_1 \cap E_2 \cap E_3 \cap E_4$, if $s_{\hat{m}} > s_{\varepsilon,\delta}$, then for every $x_0 \in \operatorname{supp}(p)$ with $|\eta(x_0)| \geq \gamma_\varepsilon$,*

$$\hat{f}_{\mathrm{NN}}(x_0; \hat{\mathcal{S}}) = f^\star(x_0).$$

*Proof.* The claim is vacuous if $u \leq s_{\varepsilon,\delta}$ (since the definition of GETSEED implies $s_{\hat{m}} \leq u$), so suppose $u > s_{\varepsilon,\delta}$. Also suppose $u$ satisfies (2), that the event $E_1 \cap E_2 \cap E_3 \cap E_4$ holds, and that $s_{\hat{m}} > s_{\varepsilon,\delta}$. Fix any $x_0 \in \operatorname{supp}(p)$ with $|\eta(x_0)| \geq \gamma_\varepsilon$ and denote by $s(x_0) = \operatorname{argmin}_{s \leq s_{\varepsilon,\delta}} \rho(X_s, x_0)$. By Lemma 8, we have $\rho(x_0, X_{s(x_0)}) < \rho_\varepsilon$. The Hölder smoothness assumption then implies $|\eta(X_{s(x_0)}) - \eta(x_0)| < \bar{c}\gamma_\varepsilon$, which entails

$$(1 - \bar{c})\gamma_\varepsilon \leq |\eta(x_0)| - \bar{c}\gamma_\varepsilon < |\eta(X_{s(x_0)})| < |\eta(x_0)| + \bar{c}\gamma_\varepsilon \leq (1 + \bar{c})|\eta(x_0)|. \quad (16)$$

Since $s(x_0) \leq s_{\varepsilon,\delta} < s_{\hat{m}}$, there is a (unique) index $m$ for which $s$ takes the value $s(x_0)$ in the execution of GETSEED$(\mathbb{L}, m, t, n)$ during the execution of ActiveAlg$(n)$; denote this unique index as $m(x_0)$. Now there are two cases to consider. First (case 1), if $s_{m(x_0)} \neq s(x_0)$, then $\exists m' < m(x_0)$ with $\hat{\gamma}_{m'} \geq 0$ and

$$\rho(X_{s_{m'}}, X_{s(x_0)}) \leq (c_g \hat{\gamma}_{m'}/L)^{1/\beta}. \quad (17)$$

Note that we always have $|\mathcal{L}_{m'}| \leq \kappa_{s_{m'}}$.

Now, if it were true that $|\eta(X_{s_{m'}})| < \check{C}_0\gamma_\varepsilon$, then (9) of Lemma 5 would imply $\hat{\gamma}_{m'} < \check{C}_0(2 - \check{c}_0)\gamma_\varepsilon$. Together with the Hölder smoothness assumption and (17), this would imply

$$|\eta(X_{s(x_0)})| < |\eta(X_{s_{m'}})| + c_g\check{C}_0(2 - \check{c}_0)\gamma_\varepsilon < (1 + c_g(2 - \check{c}_0))\,\check{C}_0\gamma_\varepsilon \leq (1 - \bar{c})\gamma_\varepsilon,$$

where this last inequality is based on the restriction on $\bar{c}$ discussed above, at the start of Section 5.1. But since (16) implies $|\eta(X_{s(x_0)})| > (1 - \bar{c})\gamma_\varepsilon$, this would obtain a contradiction.

We may therefore conclude that $|\eta(X_{s_{m'}})| \geq \check{C}_0\gamma_\varepsilon$. Combining this with the fact that $\hat{\gamma}_{m'} \geq 0$ (and the fact that $\mathcal{L}_{m'} = \{(X_{N_i(X_{s_{m'}})}, Y_{N_i(X_{s_{m'}})}) : i \leq |\mathcal{L}_{m'}|\}$, from the definition of TicToc), (8) of Lemma 5 (with $c = 1$) implies $\gamma^\star_{s_{m'},|\mathcal{L}_{m'}|} \geq 0$. Since $\hat{\gamma}_{m'}$ is equal either $\gamma^\star_{s_{m'},|\mathcal{L}_{m'}|}$ or $-\gamma^\star_{s_{m'},|\mathcal{L}_{m'}|} - \frac{2}{|\mathcal{L}_{m'}|}\zeta_{|\mathcal{L}_{m'}|,s_{m'}}$, and at most one of these can be nonnegative, the facts that $\hat{\gamma}_{m'} \geq 0$ and $\gamma^\star_{s_{m'},|\mathcal{L}_{m'}|} \geq 0$ together imply that $\hat{\gamma}_{m'} = \gamma^\star_{s_{m'},|\mathcal{L}_{m'}|}$. Combining this with the fact that $|\eta(X_{s_{m'}})| \geq \check{C}_0\gamma_\varepsilon$, (7) of Lemma 5 implies $\hat{\gamma}_{m'} < (2 - \check{c}_0)|\eta(X_{s_{m'}})|$. Plugging this into (17) yields

$$\rho(X_{s_{m'}}, X_{s(x_0)}) < (c_g(2 - \check{c}_0)|\eta(X_{s_{m'}})|/L)^{1/\beta}. \tag{18}$$

By the Hölder smoothness assumption, this implies

$$|\eta(X_{s(x_0)})| > |\eta(X_{s_{m'}})| - c_g(2 - \check{c}_0)|\eta(X_{s_{m'}})| = (1 - c_g(2 - \check{c}_0))\,|\eta(X_{s_{m'}})|.$$

Recalling that $c_g(2 - \check{c}_0) < 1$ (from the definitions of these constants at the beginning of Section 5.1), and plugging this back into (18), we obtain

$$\rho(X_{s_{m'}}, X_{s(x_0)}) < \left(\frac{c_g(2 - \check{c}_0)}{1 - c_g(2 - \check{c}_0)}\frac{|\eta(X_{s(x_0)})|}{L}\right)^{1/\beta}.$$

Together with (16), this implies

$$\rho(X_{s_{m'}}, X_{s(x_0)}) < \left(\frac{c_g(2 - \check{c}_0)(1 + \bar{c})}{1 - c_g(2 - \check{c}_0)}\frac{|\eta(x_0)|}{L}\right)^{1/\beta}.$$

By the triangle inequality, we therefore have that

$$\begin{aligned}\rho(x_0, X_{s_{m'}}) &\leq \rho(x_0, X_{s(x_0)}) + \rho(X_{s_{m'}}, X_{s(x_0)}) \\ &< \rho_\varepsilon + \left(\frac{c_g(2 - \check{c}_0)(1 + \bar{c})}{1 - c_g(2 - \check{c}_0)}\frac{|\eta(x_0)|}{L}\right)^{1/\beta}.\end{aligned} \tag{19}$$

On the other hand (case 2), if $s_{m(x_0)} = s(x_0)$, then since $\rho(x_0, X_{s(x_0)}) < \rho_\varepsilon$, it trivially follows that $\rho(x_0, X_{s_{m(x_0)}})$ is less than (19). Thus, in either case, $\exists m < \hat{m}$ such that $\rho(x_0, X_{s_m})$ is less than (19). Plugging in the definition of $\rho_\varepsilon$, together with the fact that $|\eta(x_0)| \geq \gamma_\varepsilon$, and using monotonicity of the $\ell_p$ norm in $p$ (more specifically, $(|x|^{1/\beta} + |y|^{1/\beta})^\beta \leq |x| + |y|$ for any $x, y \in \mathbb{R}$), we have that

$$\begin{aligned}\rho(x_0, X_{s_m}) &< \left(\bar{c}\frac{|\eta(x_0)|}{L}\right)^{1/\beta} + \left(\frac{c_g(2 - \check{c}_0)(1 + \bar{c})}{1 - c_g(2 - \check{c}_0)}\frac{|\eta(x_0)|}{L}\right)^{1/\beta} \\ &\leq \left(\left(\bar{c} + \frac{c_g(2 - \check{c}_0)(1 + \bar{c})}{1 - c_g(2 - \check{c}_0)}\right)\frac{|\eta(x_0)|}{L}\right)^{1/\beta}.\end{aligned} \tag{20}$$

Together with the Hölder smoothness assumption and the fact that $|\eta(x_0)| \geq \gamma_\varepsilon$, this implies

$$\begin{aligned}|\eta(X_{s_m})| &> |\eta(x_0)| - \left(\bar{c} + \frac{c_g(2 - \check{c}_0)(1 + \bar{c})}{1 - c_g(2 - \check{c}_0)}\right)|\eta(x_0)| \\ &\geq \left(1 - \bar{c} - \frac{c_g(2 - \check{c}_0)(1 + \bar{c})}{1 - c_g(2 - \check{c}_0)}\right)\gamma_\varepsilon \geq c_b\gamma_\varepsilon,\end{aligned} \tag{21}$$

where this last inequality follows from the restrictions on these constant values discussed above (at the top of Section 5.1).

Now let $t_m$ denote the value of $t$ in the execution of ActiveAlg$(n)$ upon reaching Step 4, at which point the algorithm executes $\textsc{TicToc}(s_m, m, t_m, n)$, which returns a pair $(\mathcal{L}_m, t')$. Note that, since $m < \hat{m}$, it must be that the condition in Step 6 of ActiveAlg$(n)$ is not satisfied for this index $m$, which in particular implies $t' < n$. This, in turn, implies that the constraint "$k < n - t$" in Step 2 of TicToc is satisfied throughout the execution of $\textsc{TicToc}(s_m, m, t_m, n)$. Since, with the above definitions of $\kappa$ and $\zeta$, this constraint is the only place the value of $n$ appears in TicToc, we conclude that $\textsc{TicToc}(s_m, m, t_m, n) = \textsc{TicToc}(s_m, m, t_m, \infty)$ for this particular $m$. Combining this fact with (21), Lemma 7 implies

$$\left| \sum_{(x,y)\in\mathcal{L}_m} y \right| \geq c_T \zeta_{|\mathcal{L}_m|, s_m} \geq \zeta_{|\mathcal{L}_m|, s_m}.$$

Therefore, by the definition of $\textsc{Learn}_{\mathrm{1NN}}$, for $\hat{Y}_{s_m} = \mathrm{sign}\left(\sum_{(x,y)\in\mathcal{L}_m} y\right)$, we have $(X_{s_m}, \hat{Y}_{s_m}) \in \hat{\mathcal{S}}$.

Combining this with (20), and denoting by $(X_{\hat{s}}, \hat{Y}_{\hat{s}})$ the element of $\hat{\mathcal{S}}$ with $\hat{s} = N_1(x_0; \hat{\mathcal{S}})$, we have

$$\begin{aligned} \rho(x_0, X_{\hat{s}}) &= \min_{(x,y)\in\hat{\mathcal{S}}} \rho(x_0, x) \leq \rho(x_0, X_{s_m}) \\ &< \left( \left( \bar{c} + \frac{c_g(2-\check{c}_0)(1+\bar{c})}{1-c_g(2-\check{c}_0)} \right) \frac{|\eta(x_0)|}{L} \right)^{1/\beta}. \end{aligned}$$

Together with the Hölder smoothness assumption, this implies

$$\begin{aligned} f^\star(x_0)\eta(X_{\hat{s}}) &> f^\star(x_0)\eta(x_0) - \left( \bar{c} + \frac{c_g(2-\check{c}_0)(1+\bar{c})}{1-c_g(2-\check{c}_0)} \right) |\eta(x_0)| \\ &= \left( 1 - \bar{c} - \frac{c_g(2-\check{c}_0)(1+\bar{c})}{1-c_g(2-\check{c}_0)} \right) |\eta(x_0)| \geq c_b \gamma_\varepsilon, \end{aligned}$$

where this last inequality follows as in (21) above. In particular, since $c_b\gamma_\varepsilon > 0$, this implies $\mathrm{sign}(\eta(X_{\hat{s}})) = \mathrm{sign}(f^\star(x_0))$, so that $f^\star(X_{\hat{s}}) = f^\star(x_0)$, and also implies that $|\eta(X_{\hat{s}})| \geq c_b\gamma_\varepsilon$. Recalling that $c_b \geq \check{C}_0$, Lemma 6 then implies $\hat{Y}_{\hat{s}} = f^\star(X_{\hat{s}})$. Altogether, we have that $\hat{Y}_{\hat{s}} = f^\star(x_0)$, as claimed. □

**Lemma 10.** *Let* $c_v = 2^{-\beta}\frac{c_T-1}{c_T+1}\check{c}_0 c_g$. *For every* $\gamma > 0$, *define*

$$S_\gamma = \frac{\Gamma((d/2)+1)}{\pi^{d/2}} \frac{a^{\frac{1}{1-\alpha}}(2+c_v)^{\frac{\alpha}{1-\alpha}}}{\mu_{\min} c_0} \max\left\{ \frac{1}{r_0^d}, \left(\frac{L}{c_v\gamma}\right)^{\frac{d}{\beta}} \right\} \gamma^{\frac{\alpha}{1-\alpha}}.$$

*If* $u$ *satisfies* (2), *then on the event* $E_1 \cap E_2 \cap E_3$, *for every* $\gamma \geq c_b\gamma_\varepsilon$, *there are at most* $\max\{S_\gamma, 1\}$ *indices* $m$ *in the execution of* ActiveAlg$(n)$ *such that* $s_m < u$ *and* $\gamma < |\eta(X_{s_m})| \leq 2\gamma$.

*Proof.* Suppose $u$ satisfies (2) and that the event $E_1 \cap E_2 \cap E_3$ holds, and fix any $\gamma \geq c_b \gamma_\varepsilon$. Let $r_\gamma = 2(c_v \gamma / L)^{1/\beta}$. The proof proceeds in two parts: first arguing that any $r_\gamma$-packing in $\{x \in \mathrm{supp}(p) : \gamma < |\eta(X_{s_m})| \leq 2\gamma\}$ has size at most $S_\gamma$, and second proving that the points $X_{s_m}$ with $s_m < u$ and $\gamma < |\eta(X_{s_m})| \leq 2\gamma$ comprise an $r_\gamma$-packing.

The first part proceeds similarly to part of the proof of Lemma 8, with a few important modifications to incorporate Tsybakov's noise assumption. Let $x_1, \ldots, x_M$ be any $r_\gamma$-packing of $\{x \in \mathrm{supp}(p) : \gamma < |\eta(x)| \leq 2\gamma\}$: that is, each $x_i$ is contained in $\mathrm{supp}(p)$ and has $\gamma < |\eta(x_i)| \leq 2\gamma$, and if $M > 1$ then $\min_{j \neq i} \rho(x_i, x_j) \geq r_\gamma$. The strong density assumption implies that, $\forall i \in \{1, \ldots, M\}$,

$$
\begin{aligned}
&P_X(x : \rho(x_i, x) < r_\gamma/2) = P_X(\mathrm{B}(x_i, r_\gamma/2)) \geq \mu_{\min} \lambda(\mathrm{B}(x_i, r_\gamma/2) \cap \mathrm{supp}(p)) \\
&\geq \mu_{\min} c_0 \lambda(\mathrm{B}(x_i, \min\{r_0, r_\gamma/2\})) = \frac{\pi^{d/2}}{\Gamma((d/2)+1)} \mu_{\min} c_0 \min\{r_0, r_\gamma/2\}^d.
\end{aligned}
$$

Since $x_1, \ldots, x_M$ is an $r_\gamma$-packing, it follows that the sets $\{x : \rho(x_i, x) < r_\gamma/2\}$ are disjoint over $i \in \{1, \ldots, M\}$. Therefore,

$$
\begin{aligned}
P_X\left(\bigcup_{i \leq M} \{x : \rho(x_i, x) < r_\gamma/2\}\right) &= \sum_{i \leq M} P_X(x : \rho(x_i, x) < r_\gamma/2) \\
&\geq \frac{\pi^{d/2}}{\Gamma((d/2)+1)} \mu_{\min} c_0 \min\{r_0, r_\gamma/2\}^d M. \quad (22)
\end{aligned}
$$

Furthermore, by the Hölder smoothness assumption, for every $i \in \{1, \ldots, M\}$ and every $x \in \mathcal{X}$ with $\rho(x_i, x) < r_\gamma/2$, we have

$$
|\eta(x)| < |\eta(x_i)| + c_v \gamma \leq (2 + c_v)\gamma.
$$

Combining this with Tsybakov's noise assumption, we have that

$$
\begin{aligned}
P_X\left(\bigcup_{i \leq M} \{x : \rho(x_i, x) < r_\gamma/2\}\right) &\leq P_X(x : |\eta(x)| < (2 + c_v)\gamma) \\
&\leq a^{\frac{1}{1-\alpha}} (2 + c_v)^{\frac{\alpha}{1-\alpha}} \gamma^{\frac{\alpha}{1-\alpha}}. \quad (23)
\end{aligned}
$$

Combining (22) with (23), it immediately follows that

$$
M \leq \frac{\Gamma((d/2)+1)}{\pi^{d/2}} \frac{a^{\frac{1}{1-\alpha}} (2 + c_v)^{\frac{\alpha}{1-\alpha}}}{\mu_{\min} c_0} \frac{\gamma^{\frac{\alpha}{1-\alpha}}}{\min\{r_0, r_\gamma/2\}^d} = S_\gamma.
$$

Since $x_1, \ldots, x_M$ was an arbitrary $r_\gamma$-packing in $\{x \in \mathrm{supp}(p) : \gamma < |\eta(x)| \leq 2\gamma\}$, we conclude that this is an upper bound on the size of any such packing.

Lemma 3 implies that every $X_s \in \mathrm{supp}(p)$. Therefore, to complete the proof, it suffices to establish that the points $X_{s_m}$ with $s_m < u$ and $\gamma < |\eta(X_{s_m})| \leq 2\gamma$ are $r_\gamma$-separated (if any such points exist). We will in fact establish that this is true of the (potentially larger) set of points $X_{s_m}$ with $s_m < u$ and $|\eta(X_{s_m})| > \gamma$.

Fix any $m_0 < \hat{m}$ with $s_{m_0} < u$ and $|\eta(X_{s_{m_0}})| > \gamma$; if there is no such $m_0$, then the result trivially follows, so for the remainder we suppose such an $m_0$ exists. Consider the round of ActiveAlg$(n)$ with $m = m_0$, and let $t_m$ denote the value of $t$ upon reaching Step 4 with this index $m = m_0$, at which point the algorithm executes TICTOC$(s_m, m, t_m, n)$. Since $m < \hat{m}$, as in the proof of Lemma 9 it holds that TICTOC$(s_m, m, t_m, n) =$ TICTOC$(s_m, m, t_m, \infty)$. Therefore, since $\gamma \geq c_b \gamma_\varepsilon$, Lemma 7 implies that the pair $(\mathcal{L}_m, t')$ returned by TICTOC$(s_m, m, t_m, n)$ satisfies

$$\left| \sum_{(x,y)\in\mathcal{L}_m} y \right| \geq c_T \zeta_{|\mathcal{L}_m|, s_m}.$$

Combining this with (8) of Lemma 5 and the definition of $\mathcal{L}_m$ from TICTOC (and recalling that $c_b \geq \check{C}_0$, $c_T > 1$, and $|\mathcal{L}_m| \leq \kappa_{s_m}$), we obtain that

$$\gamma^\star_{s_m, |\mathcal{L}_m|} \geq \frac{c_T - 1}{c_T + 1} \check{c}_0 |\eta(X_{s_m})| > \frac{c_T - 1}{c_T + 1} \check{c}_0 \gamma.$$

Furthermore, since the rightmost quantity is strictly positive and $\zeta_{|\mathcal{L}_m|, s_m} > 0$, it follows from the definitions of $\gamma^\star_{s_m, |\mathcal{L}_m|}$ and $\mathcal{L}_m$ that $f^\star(X_{s_m}) \sum_{(x,y)\in\mathcal{L}_m} y > 0$, so that $\hat{\gamma}_m = \gamma^\star_{s_m, |\mathcal{L}_m|} > 0$, where $\hat{\gamma}_m = \frac{1}{|\mathcal{L}_m|}\left( \left| \sum_{(x,y)\in\mathcal{L}_m} y \right| - \zeta_{|\mathcal{L}_m|, s_m} \right)$, as defined in the GETSEED subroutine.

In particular, this implies that any $s \in \{s_{m_0} + 1, \ldots, u\}$ with $\rho(X_{s_{m_0}}, X_s) \leq r_\gamma$ necessarily has

$$\rho(X_{s_{m_0}}, X_s) \leq (c_g \hat{\gamma}_{m_0} / L)^{1/\beta}.$$

Noting that (by a simple inductive proof based on the definition of GETSEED and the fact that $s_{m_0} < u$) every $s_m$ with $m \in \{m_0 + 1, \ldots, \hat{m}\}$ has $s_m > s_{m_0}$, and that every such $m$ for which $s_m < u$ has (by the condition in Step 3 of GETSEED) $\rho(X_{s_{m_0}}, X_{s_m}) > (c_g \hat{\gamma}_{m_0} / L)^{1/\beta}$, we conclude that every $m \in \{m_0 + 1, \ldots, \hat{m}\}$ with $s_m < u$ satisfies $\rho(X_{s_{m_0}}, X_{s_m}) > r_\gamma$.

Since this holds for any choice of $m_0 < \hat{m}$ with $s_{m_0} < u$ and $|\eta(X_{s_{m_0}})| > \gamma$, it follows that, defining $\mathcal{M}_\gamma = \{m \in \{1, \ldots, \hat{m}\} : s_m < u, |\eta(X_{s_m})| > \gamma\}$, we have that either $|\mathcal{M}_\gamma| \leq 1$ (in which case the corresponding set of points $X_{s_m}$ is trivially an $r_\gamma$-packing) or else

$$\min_{\substack{m, m' \in \mathcal{M}_\gamma : \\ m \neq m'}} \rho(X_{s_m}, X_{s_{m'}}) = \min_{\substack{m_0, m \in \mathcal{M}_\gamma : \\ m_0 < m}} \rho(X_{s_{m_0}}, X_{s_m}) > r_\gamma,$$

which completes the proof. □

**Lemma 11.** *Suppose $\frac{\alpha}{1-\alpha} \leq \frac{d}{\beta}$. There exist finite constants $\hat{C}_1, \hat{C}_2 \geq 1$ such that, if $u$ satisfies (2) and $u > s_{\varepsilon,\delta}$, and $n$ satisfies*

$$n \geq \hat{C}_1 \left( \frac{1}{\varepsilon} \right)^{2 - 3\alpha + \frac{d}{\beta}(1-\alpha)} \mathrm{Log}^2\left( \frac{\hat{C}_2}{\varepsilon\delta} \right), \tag{24}$$

*then there is an event $E_5$ of probability at least $1 - \delta/c_e$ such that, on the event $E_1 \cap E_2 \cap E_3 \cap E_5$, the random variable $\hat{m}$ (from Lemma 9) satisfies $s_{\hat{m}} > s_{\varepsilon,\delta}$.*

*Proof.* Suppose $u$ satisfies (2) and $u > s_{\varepsilon,\delta}$. For each $s \in \{1,\ldots,u\}$, note that the pair $(\mathcal{L}, t')$ that would be returned from $\textsc{TicToc}(s,m,t,\infty)$ has the property that the difference $t'-t$ is invariant to the values of $m$ and $t$ (since our definitions of $\kappa$ and $\zeta$ are independent of these arguments); denote by $\hat{Q}_s$ this value of $t'-t$. Since $s_m$ is strictly increasing in $m$ when $s_m < u$, in the event that ActiveAlg$(n)$ terminates with $m = u$ satisfied in Step 6, we may immediately conclude that $s_{\hat{m}} = u > s_{\varepsilon,\delta}$, so that the result trivially holds in this case. Otherwise, if ActiveAlg$(n)$ terminates with $m < u$, then it must be that it terminates with $t = n$ in Step 6. In this case, the conclusion that $s_{\hat{m}} > s_{\varepsilon,\delta}$ will immediately follow if we can establish that

$$n > \sum_{\substack{m \leq \hat{m}: \\ s_m \leq s_{\varepsilon,\delta}}} \hat{Q}_{s_m}. \tag{25}$$

Denote $j_\varepsilon = \lfloor \max\{\log_2(1/(c_b\gamma_\varepsilon)), 0\} \rfloor$ and fix any $j \in \{1,\ldots,j_\varepsilon\}$ (if $j_\varepsilon \neq 0$). By Lemma 7, on $E_1 \cap E_3$, any $m \in \{1,\ldots,\hat{m}\}$ with $s_m \leq s_{\varepsilon,\delta}$ and $2^{-j} < |\eta(X_{s_m})| \leq 2^{1-j}$ satisfies

$$\hat{Q}_{s_m} \leq Q_{s_m} < 2^{2j}\hat{C}_1' \left( \mathrm{LogLog}\left( \frac{\hat{C}_2'}{\varepsilon} \right) + \mathrm{Log}\left( \frac{3c_e s_{\varepsilon,\delta}}{\delta} \right) \right),$$

where $\hat{C}_1' = 2\max\left\{ \frac{4\tilde{c}_2 c_T^2}{\check{c}_0^2}, \tilde{c}_1 \right\}$ and $\hat{C}_2' = \frac{\sqrt{24\tilde{c}_2}c_T}{\check{c}_0 c_b}$. Also, by Lemma 10, on the event $E_1 \cap E_2 \cap E_3$,

$$\left| \left\{ m \in \{1,\ldots,\hat{m}\} : s_m \leq s_{\varepsilon,\delta}, 2^{-j} < |\eta(X_{s_m})| \leq 2^{1-j} \right\} \right| \leq \max\{S_{2^{-j}}, 1\}.$$

Furthermore, for any $m \in \{1,\ldots,\hat{m}\}$ with $s_m \leq s_{\varepsilon,\delta}$ and $|\eta(X_{s_m})| \leq 2^{-j_\varepsilon}$, the definition of $\textsc{TicToc}$ always guarantees $\hat{Q}_{s_m} \leq \kappa_{s_m} \leq \kappa_{s_{\varepsilon,\delta}}$. Additionally, by the Chernoff bound and Tsybakov's noise assumption, there is an event $E_5$ of probability at least $1 - \delta/c_e$, on which

$$\begin{aligned} \left| \left\{ s \in \{1,\ldots,s_{\varepsilon,\delta}\} : |\eta(X_s)| \leq 2^{-j_\varepsilon} \right\} \right| &\leq \log_2\left( \frac{c_e}{\delta} \right) + 2eP_X(x : |\eta(x)| \leq 2^{-j_\varepsilon}) s_{\varepsilon,\delta} \\ &\leq \log_2\left( \frac{c_e}{\delta} \right) + 2ea^{\frac{1}{1-\alpha}} 2^{-j_\varepsilon\left(\frac{\alpha}{1-\alpha}\right)} s_{\varepsilon,\delta}. \end{aligned}$$

Altogether, we have that on the event $E_1 \cap E_2 \cap E_3 \cap E_5$,

$$\begin{aligned} \sum_{\substack{m \leq \hat{m}: \\ s_m \leq s_{\varepsilon,\delta}}} \hat{Q}_{s_m} \leq{} & \left( \log_2\left( \frac{c_e}{\delta} \right) + 2ea^{\frac{1}{1-\alpha}} 2^{-j_\varepsilon\left(\frac{\alpha}{1-\alpha}\right)} s_{\varepsilon,\delta} \right) \kappa_{s_{\varepsilon,\delta}} \\ & + \sum_{j=1}^{j_\varepsilon} \max\{S_{2^{-j}}, 1\} 2^{2j} \hat{C}_1' \left( \mathrm{LogLog}\left( \frac{\hat{C}_2'}{\varepsilon} \right) + \mathrm{Log}\left( \frac{3c_e s_{\varepsilon,\delta}}{\delta} \right) \right). \end{aligned} \tag{26}$$

Thus, to satisfy (25) (and hence have $s_{\hat{m}} > s_{\varepsilon,\delta}$ on the event $E_1 \cap E_2 \cap E_3 \cap E_5$), it suffices to take $n$ greater than the right hand side of (26). All that remains is to show that the expression on the right hand side of (26) can be relaxed into the form (24).

Specifically, by plugging in the definitions of these various quantities, simplifying the resulting expression via basic inequalities, noting that the summation in the second term is bounded by a geometric series (using the fact that $\frac{\alpha}{1-\alpha} \leq \frac{d}{\beta}$), and coalescing the constant factors in the final expression, one can straightforwardly verify that the right hand side of (26) is less than the expression on the right hand side of (24) if we choose

$$\hat{C}_1 = \left(\frac{2d}{\beta} + \hat{C}_3' \frac{2^{d+3}ed}{\beta \bar{c}^{d/\beta}} c_b^{\frac{\alpha}{1-\alpha}} a^{1+\frac{d}{\beta}}\right) 4\tilde{c}_3 \left(1 + \frac{d}{\beta}(1-\alpha)\right) a^2 + \frac{1}{3}\hat{C}_1' \hat{C}_3' a^{\frac{1}{1-\alpha}} (2+c_v)^{\frac{\alpha}{1-\alpha}} c_v^{-\frac{d}{\beta}} \left(\frac{2a}{c_b}\right)^{2+\frac{d}{\beta}-\frac{\alpha}{1-\alpha}} \left(2 + \frac{d}{\beta}(1-\alpha)\right)$$

and

$$\hat{C}_2 = \max \begin{cases} \bar{c}^{-1}\left(4^d c_e \hat{C}_3'\right)^{\frac{\beta}{d}} \\ \left(\frac{3c_e\tilde{c}_4}{\bar{c}}\left(\hat{C}_3' \frac{2^{d+1}d}{\beta\bar{c}^{d/\beta}} a^{\frac{d}{\beta}}\right)^2 \left(4^d c_e \hat{C}_3'\right)^{\frac{\beta}{d}}\right)^{1/\left(2+2\frac{d}{\beta}(1-\alpha)\right)} \\ \bar{c}^{-1}\left(3c_e \hat{C}_2' \hat{C}_3' \frac{2^{d+1}d}{\beta\bar{c}^{d/\beta}} a^{\frac{d}{\beta}} \left(4^d c_e \hat{C}_3'\right)^{\frac{\beta}{d}}\right)^{1/\left(2+\frac{d}{\beta}(1-\alpha)\right)} \end{cases}.$$

where $\hat{C}_3' = \frac{\Gamma((d/2)+1)L^{d/\beta}}{\pi^{d/2}\mu_{\min} c_0 r_0^d}$ We should note, however, that this simplification of form comes at the expense of some loss of precision in the dependence on certain constant factors compared to (26). □

We are now ready to piece these lemmas together into a proof of Theorem 2.

*Proof of Theorem 2.* Suppose $\frac{\alpha}{1-\alpha} \leq \frac{d}{\beta}$, that $u$ satisfies (2) and has $u > s_{\varepsilon,\delta}$, that $n$ satisfies (24), and that the event $\bigcap_{i=1}^5 E_i$ holds. Let $\hat{f}_{n,u}$ denote the classifier returned by ActiveAlg$(n)$, and note that $\hat{f}_{n,u}(\cdot) = \hat{f}_{\mathrm{NN}}(\cdot;\hat{\mathcal{S}})$. By Lemma 11, we have $s_{\hat{m}} > s_{\varepsilon,\delta}$. Therefore, by Lemma 9, every $x_0 \in \mathrm{supp}(p)$ with $|\eta(x_0)| \geq \gamma_\varepsilon$ has $\hat{f}_{n,u}(x_0) = f^\star(x_0)$. This implies

$$\mathcal{R}(\hat{f}_{n,u};P) - \mathcal{R}(f^\star;P) = \int \mathbb{1}\left[\hat{f}_{n,u}(x) \neq f^\star(x)\right] |\eta(x)| P_X(\mathrm{d}x)$$
$$\leq \int \mathbb{1}[|\eta(x)| < \gamma_\varepsilon] \, |\eta(x)| P_X(\mathrm{d}x) \leq P_X(x : |\eta(x)| < \gamma_\varepsilon)\gamma_\varepsilon.$$

If $\gamma_\varepsilon = \varepsilon$, then this last expression is trivially at most $\varepsilon$. Otherwise, if $\gamma_\varepsilon = a^{-1}\varepsilon^{1-\alpha}$, then Tsybakov's noise assumption implies

$$P_X(x : |\eta(x)| < \gamma_\varepsilon)\gamma_\varepsilon \leq (a\gamma_\varepsilon)^{\frac{1}{1-\alpha}} = \varepsilon.$$

To complete the proof, we note that the event $\bigcap_{i=1}^5 E_i$ has probability at least $1 - 9\delta/c_e = 1 - \delta$ by the union bound, and that (by basic inequalities) any $n$ and

$u$ satisfying the size constraints stated in Theorem 2 will necessarily satisfy (2), (24), and $u > s_{\varepsilon,\delta}$, if we choose $C_2 = 4\hat{C}_1 \mathrm{Log}^2(\hat{C}_2)$ and

$$C_1 = \max\begin{cases} \check{C}_1 + 2\check{C}_2 \mathrm{Log}(\check{C}_3) \\ \frac{2^{d+2}\Gamma((d/2)+1) d L^{d/\beta}}{\pi^{d/2}\mu_{\min} c_0 r_0^d \beta \bar{c}^{d/\beta}} a^{\frac{d}{\beta}} \mathrm{Log}\left(\left(\frac{4^d c_e \Gamma((d/2)+1)}{\pi^{d/2}\mu_{\min} c_0 r_0^d}\right)^{\frac{\beta}{d}} \frac{L}{\bar{c}}\right) \end{cases}. \qquad \square$$

### 5.2. *Using Random Seeds*

For $\alpha < 2/3$, one can verify that the result in Theorem 2 in fact remains valid even when we replace GETSEED with the *trivial* subroutine GETSEED$(\mathbb{L}, m, t, n) = m$. Only the constant factors are affected. However, this variant of GETSEED gives a suboptimal rate for $\alpha > 2/3$, yielding a label complexity bound $\tilde{\Theta}\left(\left(\frac{1}{\varepsilon}\right)^{\frac{d}{\beta}(1-\alpha)}\right)$, which is worse than optimal by a factor $\left(\frac{1}{\varepsilon}\right)^{3\alpha-2}$. For brevity, we leave the details of the analysis of this alternative method as an exercise for the interested reader.

## References


[1] J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. 1, 2, 2, 2, 2, 2, 3, 5

[2] A. Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. arXiv*:1405.2639*, 2015. 4.1, 5.1

[3] A. Balsubramani and A. Ramdas. Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, 2016. 4.1

[4] R. M. Castro and R. D. Nowak. Upper and lower error bounds for active learning. In *The 44th Annual Allerton Conference on Communication, Control and Computing*, 2006. 2

[5] R. M. Castro and R. D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, July 2008. 2

[6] K. Chaudhuri and S. Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems* 27, 2014. 3

[7] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999. 2

[8] L. Györfi. The rate of convergence of $k_n$-NN regression estimates and classification rules. *IEEE Transactions on Information Theory*, 27(3):362–364, 1981. 3

[9] S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39 (1):333–361, 2011. 2

[10] S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13(5):1469–1587, 2012. 2

[11] S. Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2–3):131–309, 2014. 2

[12] S. Hanneke. Refined error bounds for several learning algorithms. *Journal of Machine Learning Research*, 17(135):1–55, 2016. 2

[13] S. Hanneke and L. Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 16(12):3487–3602, 2015. 2, 2, 3, 4.1, 4.2, 4.3

[14] A. N. Kolmogorov and V. M. Tikhomirov. $\varepsilon$-entropy and $\varepsilon$-capacity of sets in function spaces. *American Mathematical Society Translations, Series 2*, 17:277–364, 1961. 5.1

[15] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. 2

[16] A. Kontorovich, S. Sabato, and R. Urner. Active nearest-neighbor learning in metric spaces. In *Advances in Neural Information Processing Systems* 29, 2016. 3, 4.3

[17] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999. 2

[18] P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006. 2

[19] S. Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13(1):67–90, 2012. 2, 2, 3, 5

[20] S. Minsker. *Non-asymptotic Bounds for Prediction Problems and Density Estimation*. PhD thesis, School of Mathematics, Georgia Institute of Technology, 2012. 2, 2, 3

[21] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. 2

[22] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. 2

[23] M. Vidyasagar. *Learning and Generalization with Applications to Neural Networks*. Springer-Verlag, second edition, 2003. 5.1

[24] L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006. 2

[25] C. Zhang and K. Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems* 27, 2014. 2