Nuances in Margin Conditions Determine Gains in Active Learning

Samory Kpotufe
Columbia University

Gan Yuan
Columbia University

Yunfan Zhao
Columbia University

Abstract

We consider nonparametric classification with smooth regression functions, where it is well known that notions of margin in $\mathbb{E}[Y|X]$ determine fast or slow rates in both active and passive learning. Here we elucidate a striking distinction between the two settings. Namely, we show that some seemingly benign nuances in notions of margin—somehow involving the uniqueness of the Bayes classifier, and which have no apparent effect on rates in passive learning—determine whether or not *any* active learner can outperform passive learning rates. In particular, for *Audibert-Tsybakov's margin condition* (allowing general situations with non-unique Bayes classifiers), no active learner can gain over passive learning in commonly studied settings where the marginal on X is near uniform. Our results thus negate the usual intuition from past literature that active rates should generally improve over passive rates in nonparametric settings.

1 INTRODUCTION

Margin conditions, i.e., conditions quantifying the gap between class probabilities, have been known to determine the hardness of classification both in passive learning, i.e., where the learner only has access to i.i.d. data (Mammen and Tsybakov, 1999; Tsybakov, 2004; Massart and Nédélec, 2006; Audibert and Tsybakov, 2007), and in active learning where the learner can adaptively query labels (Castro and Nowak, 2008; Hanneke, 2011; Koltchinskii, 2010; Minsker, 2012; Hanneke and Yang, 2015; Wang and Singh, 2016; Yan et al., 2016; Locatelli et al., 2017, 2018). Naturally, a main concern in active learning is in guaranteeing savings over passive learning, and here we show that some basic distinctions between margin conditions—seemingly having to do with the uniqueness of the Bayes classifier, and which appear to have gone unnoticed—determine whether savings are possible at all over passive rates in nonparametric settings.

Here we consider the setting of nonparametric classification with smooth regression functions, i.e., one where $\eta_y(x) \doteq \mathbb{P}[Y = y|X = x]$ is α-Hölder continuous for every label $y \in [L]$. Two main notions of margin have appeared interchangeably in passive learning in this setting; assume $y = 1$ or 2:

$$\text{(i) } \mathbb{P}(|\eta_1 - \eta_2| \leq \tau) \lesssim \tau^\beta, \text{(ii) } \mathbb{P}(0 < |\eta_1 - \eta_2| \leq \tau) \lesssim \tau^\beta,$$

for some *margin parameter* $\beta > 0$. Both definitions are termed *Tsybakov's low noise or margin condition* without distinction in the literature. However, excluding 0 as in (ii) is more natural since any classifier $\hat{h}$ has the same error as Bayes in those regions where $\eta_1 = \eta_2$, i.e., where the Bayes is not unique. On the other hand, (i) implies uniqueness (up to measure 0) of the Bayes classifier, as seen by letting $\tau \to 0$. As such, (ii) admits more general settings with non-unique Bayes, and is thus preferred in the seminal result of Audibert and Tsybakov (2007) on margins in nonparametrics.

Interestingly, using (i) or (ii), the minimax risk is the same in passive learning, e.g., $O(n^{-\alpha(\beta+1)/(2\alpha+d)})$ when P_X is uniform, see Audibert and Tsybakov (2007). However, as we show, a sharp distinction emerges in active learning, where condition (ii) leads to two regimes in terms of savings:

- Under the common *strong density* assumption, relaxing uniform P_X, no active learner can achieve a better rate—beyond constants—than the minimax passive rate (Theorem 1). In contrast, as first shown in Minsker (2012), condition (i) always leads to strictly faster rates than passive.
- For general P_X, active learners can strictly gain over the worst case passive rate (Theorem 3). Our rates for (ii) are then similar to those under (i) shown in Locatelli et al. (2017).

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

Previous work in nonparametric active learning invariably adopted condition (i) which makes sense in light of our results since savings cannot be shown otherwise. Our results in fact further highlight two sources of savings in active learning, owing to the distinction between the above two bulletted regimes: a), an active learner can evenly sample the decision boundary while i.i.d. samples might miss it under general P_X, and b), an active learner can quickly stop sampling in those regions where there is little to gain in excess error over the Bayes, having discovered a label or labels with sufficiently low excess error. Under near uniform P_X, the source of saving a) is gone since even i.i.d. data has good coverage of the decision boundary, while b) remains, although in a limited form: an active learner can only significantly benefit from regions of high margin, while it cannot effectively identify regions where multiple labels are nearly equivalent (e.g., non-unique Bayes) which it should in fact also give up on.

Here we emphasize that our results do not preclude limited gains in practice under uniform P_X, since minimax rates fail to identify constants. In particular, we can refine the margin conditions to distinguish between regions of high margin and those with equivalent labels, and derive a refined upper-bound, under uniform P_X, that highlight such limited gains over passive learning (Theorem 2).

Finally, our upper-bounds are for general multi-class active learning, requiring minor modification over past algorithms (e.g., those of Locatelli et al., 2017), namely additional book-keeping (Section 3.2), and refined correctness arguments. On the other hand, our main Theorem 1 requires considerable new technicality over usual lower-bound arguments for active learning, involving careful *randomization* of hard regions of space (see discussion in Section 3.1).

Our results leave open whether similar nuances in regimes of gain exist in parametric settings, e.g., under bounded VC classes, where many active learners have been shown to gain under *sharp* margin conditions such as (i) (Hanneke, 2011; Koltchinskii, 2010; Wang and Singh, 2016).

Paper Outline. We start in Section 2 with technical setup, followed by an overview of main results in Section 3, and analysis in Section 4. Due to space constraints, some proofs are relegated to the appendix.

2 PROBLEM SETTING

We consider a joint distribution $P_{X,Y}$ on $[0,1]^d \times [L]$, where we use the short notation $[L] \doteq \{1,\ldots,L\}$ for $L \in \mathbb{N}$. Define the regression function $\eta(x) \doteq (\eta_1(x),\ldots,\eta_L(x))$ where $\eta_y(x) \doteq \mathbb{P}(Y = y|X = x)$ for $y \in [L]$.

Definition 1. *The regression function η is (λ,α)-***Hölder** *continuous for some $\alpha \in (0,1], \lambda > 0$, if.:*

$$\forall x, x' \in [0,1]^d, \quad \|\eta(x) - \eta(x')\|_\infty \leq \lambda \|x - x'\|_\infty^\alpha .$$

Remark 1. For simplicity of presentation, we assume $\alpha \leq 1$ in Theorem 2. The case of $\alpha > 1$, can be handled simply by replacing the averaging in each cell with higher order polynomial regression (as done e.g. in Locatelli et al. 2017), but does not add much to the main message despite the added technicality.

Definition 2. *For $r = 2^{-k}$ for $k \in \mathbb{N}$, define the partition $\mathscr{C}_r$ of $[0,1]^d$ as the collection of hypercubes $\mathcal{C}$ of the form $\prod_{i\in d}[(l_i-1)r, l_i r)$, $l_i \in [1/r]$. We call $\mathscr{C}_r$ a* **dyadic partition** *at* level r.

Definition 3. *P_X is said to satisfy a* **strong density condition** *if there exists some $c_d > 0$ such that $\forall r \in \{2^{-k} : k \in \mathbb{N}\}$ and $\mathcal{C} \in \mathscr{C}_r$ with $P_X(\mathcal{C}) > 0$, we have*

$$P_X(\mathcal{C}) \geq c_d \cdot r^d .$$

The condition clearly holds for $P_X = \mathcal{U}[0,1]^d$, or simply has lower-bounded density, and is adapted from other works on active learning (Minsker, 2012; Locatelli et al., 2017).

2.1 Active Learning

We consider active learning under a fixed budget n of queries. At each sampling step, the learner may query the label of any point $x \in \text{support}(P_X)$ and a label Y is returned according to the conditional $P_{Y|X=x}$. We let $S \equiv \{(X_i, Y_i)\}_{i=1}^n$ denote the resulting sample. A classifier $\hat{h}_n = \hat{h}_n(S) : [0,1]^d \mapsto [L]$ is then returned.

We evaluate the performance of an active learner by the excess risk of the final classifier $\hat{h}_n$ it outputs. Throughout the paper, we use the notation $\hat{h}$ for the active learning algorithm, and $\hat{h}_n$ for the final classifier the algorithm $\hat{h}$ returns.

Definition 4. *We consider the 0-1 risk of a classifier $h : [0,1]^d \mapsto [L]$, namely $R(h) \doteq \mathbb{P}(h(X) \neq Y)$, which is minimized by the so-called Bayes classifier $h^*(x) \in$* $\text{argmax}_y\, \mathbb{P}(Y = y|X = x)$. *The* **excess risk** *$\mathcal{E}(h) \doteq R(h) - R(h^*)$ is then given by:*

$$\mathcal{E}(h) = \mathbb{E}\ [\max_{y\in[L]} \eta_y(X) - \eta_{h(X)}(X)].$$

2.2 Margin Assumption

We start with a notion of *soft* margin.

Definition 5. *Let $\eta_{(1)} \geq \cdots \geq \eta_{(L)}$ denote order statistics on $\eta_y, y \in [L]$. The* **margin** *at x is defined as*

$\mathcal{M}(x) \doteq \eta_{(1)}(x) - \max_{y:\eta_y(x)\neq\eta_{(1)}(x)} \eta_y(x)$. *In the case where* $\forall y \in [L], \eta_y(x) = 1/L$, *we use the convention that* max *of empty set is* $-\infty$ *so that* $\mathcal{M}(x) = \infty$.

Definition 6. $P_{X,Y}$ *satisfies the* **Tsybakov's margin condition** *(TMC) with* $C_\beta > 0$, $\beta \geq 0$, *if :*

$$\forall \tau > 0, \quad P_X\left(\{x : \mathcal{M}(x) \leq \tau\}\right) \leq C_\beta \tau^\beta. \qquad (1)$$

The above extends TMC for $L = 2$ to general L: when $L = 2$, the margin $\mathcal{M}(x) = |\eta_1(x) - \eta_2(x)|$ when $\eta_1(x) \neq \eta_2(x)$ and $\mathcal{M}(x) = \infty$ when $\eta_1(x) = \eta_2(x) = 1/2$. The above thus coincides condition (ii) of Section 1, i.e., admits non-unique Bayes as in Audibert and Tsybakov (2007), but here we allows general $L \geq 2$.

3 OVERVIEW OF RESULTS

3.1 No Gain under Strong Density Condition

Surprisingly, under the Audibert-Tsybakov's margin condition, no active learner can gain in excess risk rate over their passive counterparts when we assume the strong density condition for P_X. For simplicity, we consider the binary case.

Theorem 1. *Consider a binary classification problem, i.e,* $L = 2$. *Let* $c_d, \alpha \in (0, 1], \lambda, \beta > 0, C_\beta > 1$ *with* $\alpha\beta \leq d$ *and* $\Xi = (c_d, \lambda, \alpha, C_\beta, \beta)$. *Let* $\mathcal{P}(\Xi)$ *denote the class of distributions on* $[0, 1]^d \times \{0, 1\}$ *such that:*

- P_X *satisfies a strong density condition with* c_d*;*
- *the regression function* $\eta(x)$ *is* (λ, α)*-Hölder;*
- $P_{X,Y}$ *satisfies TMC with parameter* (β, C_β)*.*

Then, $\exists C_1 > 0$, *independent on* n, *such that:*

$$\inf_{\hat{h}} \sup_{P_{X,Y}\in\mathcal{P}(\Xi)} \mathbb{E}\ \mathcal{E}(\hat{h}_n) \geq C_1 n^{-\frac{\alpha(\beta+1)}{2\alpha+d}},$$

where the infimum is taken over all active learners, and the expectation is taken over the sample distribution, determined by P *and* $\hat{h}$ *jointly.*

Following the seminal results of Audibert and Tsybakov (2007), it is easy to show that a simple plug-in passive learner (e.g., a tree-based classifier) achieves the rate of $n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}$ for any $P_{X,Y} \in \mathcal{P}(\Xi)$.

Our main arguments depart from usual lower-bounds arguments in active learning Castro and Nowak (2008); Minsker (2012); Locatelli et al. (2017) in that we do not work directly on constructing a suitable subset of $\mathcal{P}(\Xi)$, but rather move to a larger class Σ with non empty intersection Σ_β with $\mathcal{P}(\Xi)$. We then put a suitable measure on Σ that concentrates on Σ_β; importantly, this measure also encodes regions of $[0, 1]^d$ where the Bayes is unique. We then show that for any fixed sampling mechanism $\hat{h}$, the excess error of the classifier $\hat{h}_n$ is lower-bounded as in Theorem 1, in expectation under our measure on Σ, implying the statement of Theorem 1 by concentration on Σ_β. A main difficulty remains in removing dependencies inherent in the observed sample S: this is done by decoupling the sampling $\hat{h}$ from the eventual classifier $\hat{h}_n$ by a reduction to simpler Neyman-Pearson type classifier h_n^*—with the same sampling mechanism as $\hat{h}$—whose error can be localized to regions of $[0, 1]^d$ and depends just on local Y values, thanks to our choice of distributions in Σ where little information is leaked across regions of space. This is all presented in Section 4.1.

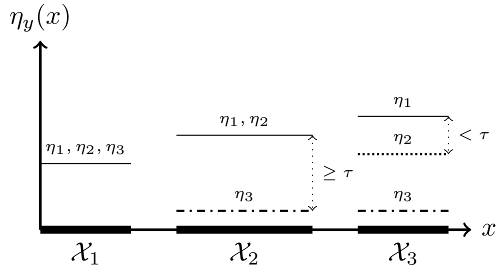

Figure 1: Different types of margin over space.

3.2 Upper-Bounds

Theorem 1 indicates that the classical TMC is not enough to guarantee gains over passive learning, under strong density. Nonetheless, some gain can be shown under a refined margin condition that better isolates regions of space with unique Bayes label (Theorem 2). Furthermore, under more general P_X, we show in Theorem 3 that a better rate than passive can always be attained even under classical TMC. Both results are established using the same procedure, which we present first. We assume smooth η in all that follows.

Assumption 1. $\eta(x)$ *is* (λ, α)*-Hölder for some known* $\lambda > 0$, *and some unknown* $\alpha \in (0, 1]$.

As in prior work Minsker (2012); Locatelli et al. (2017), we assume access to λ or any upper-bound thereof.

3.2.1 An Adaptive Procedure

The detailed approach is presented in Algorithm 1, and follows an adaptation strategy of Locatelli et al. (2017, 2018) for unknown smoothness α. This procedure repeatedly calls a non-adaptive subroutine, Algorithm 2, for a sequence of increasing values of α, i.e. $\{\alpha_i\}_{i=1}^{\lfloor\log(n)\rfloor^3}$ with $\alpha_i = i/\lfloor\log(n)\rfloor^3$.

In a departure from the binary case ($L = 2$) studied in prior work, both procedures operate by maintaining a set of candidate labels via local elimination (requir-

Algorithm 1 Meta Algorithm

```
Input: $n, \delta, \lambda$
Initialization:
• Set $\alpha_0 = 0$, $n_0 = \frac{n}{\lfloor \log(n) \rfloor^3}$, $\delta_0 = \frac{\delta}{\lfloor \log(n) \rfloor^3}$
• Set minimum level $r_0 = 2^{\lfloor \log_2(n_0^{-1/d}) \rfloor}$
• Set final candidate labels $\mathcal{L}_{\mathcal{C}} = [L], \forall \mathcal{C} \in \mathscr{C}_{r_0}$

for $i = 1, ..., \lfloor \log(n) \rfloor^3$ do
    // Run the non-adaptive subroutine
    Set $\alpha_i = \frac{i}{\lfloor \log(n) \rfloor^3}$
    Run Algorithm 2 with $(n_0, \delta_0, \alpha_i, \lambda, r_0)$
        to obtain candidate labels $\{\mathcal{L}_{\mathcal{C}}^{\alpha_i}\}_{\mathcal{C} \in \mathscr{C}_{r_0}}$
    // Aggregate candidate labels
    if $\forall \mathcal{C} \in \mathscr{C}_{r_0}, \mathcal{L}_{\mathcal{C}} \cap \mathcal{L}_{\mathcal{C}}^{\alpha_i} \neq \emptyset$ then
        $\forall\ \mathcal{C} \in \mathscr{C}_{r_0}$, set $\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}} \cap \mathcal{L}_{\mathcal{C}}^{\alpha_i}$
    end if
end for
Output: $\hat{h}_n(x) = \min \mathcal{L}_{\mathcal{C}}$ for $x \in \mathcal{C} \in \mathscr{C}_{r_0}$
```

ing new book-keeping), and remaining labels are then aggregated at the end to return a final classifier.

Next, we discuss the non-adaptive subroutine, Algorithm 2, that assumes a known α. It operates top down on dyadic partitions $\mathscr{C}_r$, $r = 1/2 \to 0$, and aims to quickly detect cells $\mathcal{C} \in \mathscr{C}_r$ with large sharp margin and stops sampling there; all cells with at least two remaining candidate labels are deemed *active*, and form a set $\mathcal{A}_r \subset \mathscr{C}_r$ of cells which are then refined.

The budget is tracked throughout, by sampling as little as $n_{r,\alpha}$ points in each $\mathcal{C} \in \mathscr{C}_r$, for

$$n_{r,\alpha} \doteq 2 \log \left(\frac{2L}{\delta_0 r^{d+1}} \right) \Big/ (\lambda r^{\alpha})^2. \tag{2}$$

This sample is used to estimate η in each cell $\mathcal{C}$ as

$$\hat{\eta}_y(\mathcal{C}) = n_{r,\alpha}^{-1} \sum_{i=1}^{n_{r,\alpha}} \mathbb{I}(Y_i^{\mathcal{C}} = y), \tag{3}$$

and eliminate labels y whenever $\hat{\eta}_{(1)}(\mathcal{C}) - \hat{\eta}_y(\mathcal{C}) \geq \tau_{r,\alpha}$, where we define

$$\hat{\eta}_{(1)}(\mathcal{C}) \doteq \max_y \hat{\eta}_y(\mathcal{C}), \quad \text{and } \tau_{r,\alpha} \doteq 6\lambda r^{\alpha}. \tag{4}$$

3.2.2 Rates Under Strong Density Condition.

We start with the following definition.

Definition 7. *The* **sharp margin** *on η is defined as $\mathcal{M}'(x) \doteq \eta_{(1)}(x) - \eta_{(2)}(x)$, where we have $\eta_{(1)} = \eta_{(2)}$ when the Bayes label is not unique at x.*

Assumption 2. *$P_{X,Y}$ satisfies a* **refined margin condition** *(RMC) with $\varepsilon_0, C_\beta, \beta, \beta' > 0$ with $\beta' \geq \beta$:*

$$\forall \tau > 0, \quad P_X(\{x : \mathcal{M}(x) \leq \tau\}) \leq C_\beta \tau^{\beta}; \text{ and}$$

$$\forall \tau > 0, \quad P_X(\{x : \mathcal{M}'(x) \leq \tau\}) \leq \varepsilon_0 + C_\beta \tau^{\beta'}.$$

Algorithm 2 Non-adaptive Algorithm

```
Input: $n_0, \delta_0, \alpha, \lambda, r_0$
Initialization:
• Initial level: $r = 1/2$
• Active cells: $\mathcal{A}_r = \mathscr{C}_r$
• Budget up to level $r$: $m_r = |\mathcal{A}_r| n_{r,\alpha}$ (see (2))
• Candidate labels: $\mathcal{L}_{\mathcal{C}}^{\alpha} = [L], \forall\ \mathcal{C} \in \mathscr{C}_r$
while $(m_r \leq n_0)$ and $(|\mathcal{A}_r| > 0)$ do
    // Eliminate bad labels
    for each $\mathcal{C} \in \mathcal{A}_r$ do
        Samples $(X_i^{\mathcal{C}}, Y_i^{\mathcal{C}})_{j \leq n_{r,\alpha}}$ in cell $\mathcal{C}$
        Compute $\{\hat{\eta}_y(\mathcal{C})\}_{y \in [L]}$ by (3)
        Set $\mathcal{L}_{\mathcal{C}}^{\alpha} = \mathcal{L}_{\mathcal{C}}^{\alpha} \backslash \{y : \hat{\eta}_{(1)}(\mathcal{C}) - \hat{\eta}_y(\mathcal{C}) \geq \tau_{r,\alpha}\}$(4)
    end for
    // Pass information to the next level
    $\forall \mathcal{C}' \in \mathscr{C}_{r/2}$ with $\mathcal{C}' \subset \mathcal{C}$, set $\mathcal{L}_{\mathcal{C}'}^{\alpha} = \mathcal{L}_{\mathcal{C}}^{\alpha}$
    Set $\mathcal{A}_{r/2} = \cup\{\mathcal{C}' \in \mathscr{C}_{r/2} : \mathcal{C}' \subset \mathcal{C}$ for some
                    $\mathcal{C} \in \mathcal{A}_r$ with $|\mathcal{L}_{\mathcal{C}}^{\alpha}| \geq 2\}$
    Set $r = r/2$ // Go to next level
    Set $m_{r/2} = m_r + |\mathcal{A}_r| n_{r,\alpha}$ // Update the budget used
end while
Set $r_{\min} = 2r$ // The minimum level reached
Set $\mathcal{L}_{\mathcal{C}}^{\alpha} = \mathcal{L}_{\mathcal{C}'}^{\alpha}, \forall\ \mathcal{C} \in \mathscr{C}_{r_0}$ with $\mathcal{C} \subset \mathcal{C}' \in \mathscr{C}_{r_{\min}}$
Output: $\{\mathcal{L}_{\mathcal{C}}^{\alpha}\}_{\mathcal{C} \in \mathscr{C}_{r_0}}$
```

Remark 2. The two conditions in Assumption 2 differ when the Bayes is not unique, i.e., when $\mathbb{P}(\mathcal{M}' = 0) \doteq \varepsilon_0 > 0$, otherwise $\mathcal{M} = \mathcal{M}'$ a.e., and we may choose $\beta = \beta'$. For illustration, consider the example of Figure 1 with $L = 3$. We have $\{x : \mathcal{M}(x) \leq \tau\} = \mathcal{X}_3$, while $\{x : \mathcal{M}'(x) \leq \tau\} = \cup_{i=1}^{3} \mathcal{X}_i$. In particular, $\varepsilon_0 = P_X(\mathcal{X}_1 \cup \mathcal{X}_2)$, as $\mathcal{M}' = 0$ on $\mathcal{X}_1 \cup \mathcal{X}_2$.

The upper-bound shown in Theorem 2 below depends on ε_0, and recovers existing bounds (for the binary case) when $\varepsilon_0 = 0$, namely $\widetilde{O}\left(n^{-\alpha(\beta'+1)/(2\alpha+d-\alpha\beta')}\right)$ as shown e.g. in Minsker (2012); Locatelli et al. (2017) under sharp margin. This is an improvement over the passive learners, and matches the active lower-bound in Minsker (2012) under strong density condition with $\alpha\beta \leq d$. For large $\varepsilon_0 > 0$, the first term $\widetilde{O}\left(n^{-\alpha(\beta+1)/(2\alpha+d)}\right)$ dominates, matching our lower-bound of Theorem 1.

Theorem 2. *Let $n \in \mathbb{N}$ and $\alpha \in (0, 1]$ and $\alpha\beta' \leq d$. Let $\hat{h}_n$ denote the classifier returned by Algorithm 1 with input n, λ and $0 < \delta < 1$. Under Assumption 1 and 2, and assume further that strong density condition holds for some $c_d > 0$, then with probability* *at least* $1-\delta$,

$$\mathcal{E}\left(\hat{h}_n\right) \leq C_2 \left(\varepsilon_0^{\frac{\alpha(\beta+1)}{2\alpha+d}} \left(\frac{\lambda^{\frac{d}{\alpha}} \log^3(n) \log\left(\frac{4L\lambda^2 n}{\delta}\right)}{n} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}} + \left(\frac{\lambda^{\frac{d}{\alpha}\vee\beta'} \log^3(n) \log\left(\frac{4L\lambda^2 n}{\delta}\right)}{n} \right)^{\frac{\alpha(\beta'+1)}{2\alpha+d-\alpha\beta'}} \right)$$

for some constant $C_2 > 0$ *independent of* $n, \delta, \lambda, L, \varepsilon_0$.

Remark 3. The bound is trivial for $\alpha < \frac{1}{\log(n)}$, since $n^{-\alpha} \geq n^{-1/\log(n)} = \frac{1}{e}$. Thus, we only need to show for $\alpha \geq \frac{1}{\log(n)}$.

A main novelty in the analysis is to separately consider parts of space with unique Bayes, determined by ε_0 and β', and those parts of space where the Bayes might not be unique, but which still have margin, determined by β. Furthermore, our consideration of general multiclass, together with non-unique Bayes, brings in a bit of added technicality due largely to additional book-keeping. In particular, while in Minsker (2012); Locatelli et al. (2017), the main correctness argument involved showing that all labeled parts of space (i.e. cells with a single label left) have 0 excess error w.h.p., we additionally have to show that in fact, remaining labels in most active cells are close in error to Bayes.

3.2.3 Rates for General Densities

For general P_X, on the other hand, Algorithm 2 has an excess risk rate of order $\widetilde{O}(n^{-(\alpha(\beta+1))/(2\alpha+d)})$, which is always faster than the lower minimax rate $O(n^{-(\alpha(\beta+1))/(2\alpha+d+\alpha\beta)})$ for passive learning of Audibert and Tsybakov (2007) under the same conditions.

In other words, under TMC, which allows non-unique Bayes classifiers, active learning guarantees savings over the worst-case rate of passive learning, given the ability to evenly sample the decision boundary.

Theorem 3. *Let* $n \in \mathbb{N}$ *and* $\alpha \in (0,1]$ *and* $\alpha\beta' \leq d$*. Let* $\hat{h}_n$ *denote the classifier returned by Algorithm 1 with input* n*,* λ *and* $0 < \delta < 1$*. Under Assumption 1 and 2, with probability at least* $1-\delta$*,*

$$\mathcal{E}\left(\hat{h}_n\right) \leq C_3 \left(\frac{\log^3(n) \lambda^{\frac{d}{\alpha}} \log\left(\frac{4L\lambda^2 n}{\delta}\right)}{n} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}}$$

for some constant $C_3 > 0$ *that does not depend on* $n, \delta, \lambda, L, \varepsilon_0$.

The proof ideas follow similar outlines as for Theorem 2, though more direct.

4 ANALYSIS

4.1 Proof of Theorem 1

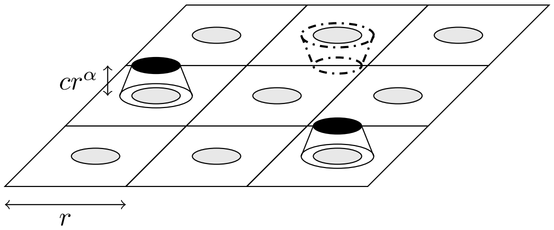

Figure 2: Construction for Theorem 1 onto a partition $\mathscr{C}_r$ of $[0,1]^d$, for a critical $r = r(n, \alpha, \beta, \lambda)$. Two *coins* are thrown in each cell $\mathcal{C}$, one $z_\mathcal{C}$ with some bias determining whether the Bayes is unique, the other $\sigma_\mathcal{C}$ determining the Bayes label. The regression function is constructed as $\eta_C \approx 1/2 \pm r^\alpha$, and together with P_X forces any $\hat{h}$ to mostly rely on local information.

4.1.1 Construction of Joint Distributions

We again operate over a dyadic partition $\mathscr{C}_r$ of the unit cube $[0,1]^d$. Let $r = c_1 n^{-\frac{1}{2\alpha+d}}$, where $c_1 = \frac{64}{\lambda^2}$. Without loss of generality, we assume that $-\log_2 r \in \mathbb{N}$. Furthermore, we denote the barycenter of any $\mathcal{C} \in \mathscr{C}_r$ as $x_\mathcal{C}$. The marginal distribution P_X has the density with respect to the Lebesgues measure:

$$f(x) \doteq \begin{cases} 4^d & \text{if } \|x - x_\mathcal{C}\| < r/8 \text{ for some } \mathcal{C} \in \mathscr{C}_r; \\ 0 & \text{otherwise.} \end{cases}$$

where $\|\cdot\|$ is the supnorm. Let $\boldsymbol{z} = (z_\mathcal{C})_{\mathcal{C}\in\mathscr{C}_r} \in \{0,1\}^{|\mathscr{C}_r|}$ and $\boldsymbol{\sigma} = (\sigma_\mathcal{C})_{\mathcal{C}\in\mathscr{C}_r} \in \{\pm 1\}^{|\mathscr{C}_r|}$. Define:

$$\eta_{\boldsymbol{z},\boldsymbol{\sigma}}(x) \doteq 1/2 + c_\eta \sum_{\mathcal{C}\in\mathscr{C}_r} z_\mathcal{C} \cdot \sigma_\mathcal{C} \cdot \phi_\mathcal{C}(x),$$

where $c_\eta = \lambda/8$, and

$$\phi_\mathcal{C}(x) = \min\left\{ (2r^\alpha - 8r^{\alpha-1}\|x - x_\mathcal{C}\|)_+, r^\alpha \right\}.$$

For each pair $(\boldsymbol{z},\boldsymbol{\sigma})$, one can define a joint probability distribution $P_{\boldsymbol{z},\boldsymbol{\sigma}}$ characterized by P_X and $\mathbb{E}[Y|X = x] = \eta_{\boldsymbol{z},\boldsymbol{\sigma}}(x)$. See Figure 2 for an example of $P_{\boldsymbol{z},\boldsymbol{\sigma}}$ for $d = 2$ and $r = 3$. In particular, P_X is uniformly distributed within its support, which is the area shaded in gray. In a cell $\mathcal{C} \in \mathscr{C}_r$ where $z_\mathcal{C} = 1$, we have a small bump in regression function, of which the direction is determined by $\sigma_\mathcal{C}$. By construction, $\eta_{\boldsymbol{z},\boldsymbol{\sigma}}$ is always a constant in the intersection of $\mathcal{C}$ and the support of P_X, with the only possible values being $1/2$ and $1/2 \pm c_\eta r^\alpha$.

Remark 4. Our construction in fact satisfies the strong density assumption of Audibert and Tsybakov (2007): their assumption requires lower-bounded densities only on the distribution support which is allowed to be disconnected, as constructed here.

4.1.2 Establishing the Lower-bound

The proof of the Theorem 1 is divided and conquered by Proposition 1 to 4. Let $\Sigma \doteq \{P_{\boldsymbol{z},\boldsymbol{\sigma}} : (\boldsymbol{z},\boldsymbol{\sigma}) \in \{0,1\}^{|\mathscr{C}_r|} \times \{\pm 1\}^{|\mathscr{C}_r|}\}$ and $\Sigma_\beta \doteq \{P_{\boldsymbol{z},\boldsymbol{\sigma}} : (\boldsymbol{z},\boldsymbol{\sigma}) \in \Theta_\beta\}$ where $\Theta_\beta \doteq \{(\boldsymbol{z},\boldsymbol{\sigma}) : \forall \tau > 0, P_X(\{x : 0 < |2\eta_{\boldsymbol{z},\boldsymbol{\sigma}}(x) - 1| \leq \tau\}) \leq C_\beta \tau^\beta\}$.

Proposition 1. $\Sigma_\beta \subset \mathcal{P}(\Xi)$. *Consequently,*

$$\inf_{\hat{h}} \sup_{P_{X,Y} \in \mathcal{P}(\Xi)} \mathbb{E}\ \mathcal{E}(\hat{h}_n) \geq \inf_{\hat{h}} \sup_{P_{X,Y} \in \Sigma_\beta} \mathbb{E}\ \mathcal{E}(\hat{h}_n).$$

where the infimum is taken over all active learners.

Proof. Let $P_{\boldsymbol{z},\boldsymbol{\sigma}} \in \Sigma_\beta$. The TMC is satisfied by construction, and it is trivial to show that strong density condition holds for $c_d = 1$. It is left to show that $\eta_{\boldsymbol{z},\boldsymbol{\sigma}}$ is (λ, α)-Hölder. In fact, this hold for all $P_{\boldsymbol{z},\boldsymbol{\sigma}} \in \Sigma$.

Let $x, x' \in [0,1]^d$. If they are in a common cell $\mathcal{C}$, then

$$\begin{aligned} |\eta_{\boldsymbol{z},\boldsymbol{\sigma}}(x) - \eta_{\boldsymbol{z},\boldsymbol{\sigma}}(x')| &\leq z_{\mathcal{C}} c_\eta (8r^{\alpha-1}\|x - x'\|) \\ &\leq \lambda \|x - x'\|^\alpha, \end{aligned}$$

where the last inequality is due to the fact $r/\|x - x'\| \geq 1$ and $\alpha - 1 < 0$. If they are in different cells, $|\eta_{\boldsymbol{z},\boldsymbol{\sigma}}(x) - \eta_{\boldsymbol{z},\boldsymbol{\sigma}}(x')| = 0$ if $\|x - x'\| < r/4$. Therefore,

$$|\eta_{\boldsymbol{z},\boldsymbol{\sigma}}(x) - \eta_{\boldsymbol{z},\boldsymbol{\sigma}}(x')| \leq 2c_\eta r^\alpha \leq \lambda \|x - x'\|^\alpha.$$

Therefore, $\eta_{\boldsymbol{z},\boldsymbol{\sigma}}$ is $(\lambda, \alpha)-$Hölder. □

Let $\boldsymbol{z} \in \{0,1\}^{|\mathscr{C}_r|} \overset{\text{i.i.d}}{\sim} \text{Ber}(r^{\alpha\beta})$, and $\boldsymbol{\sigma} \in \{\pm 1\}^{|\mathscr{C}_r|} \overset{\text{i.i.d}}{\sim}$ Radamacher(1/2), $\boldsymbol{z} \perp\!\!\!\perp \boldsymbol{\sigma}$.

Proposition 2. *Let $\hat{h}$ be any active learner. Then,*

$$\begin{aligned} \sup_{P_{X,Y} \in \Sigma_\beta} \mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}} \mathcal{E}(\hat{h}_n) \geq \mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}} & \mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}} \mathcal{E}(\hat{h}_n) \\ & - \exp(-c_2 r^{-(d-\alpha\beta)}), \end{aligned}$$

for some $c_2 > 0$, where $\mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}}(\cdot)$ is expectation taken over sample S, under the sampling distribution $P_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}}$ determined by $P_{\boldsymbol{z},\boldsymbol{\sigma}}$ and $\hat{h}$ jointly, and $\mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}}(\cdot)$ is the expectation taken over $\boldsymbol{z},\boldsymbol{\sigma}$.

Proof. By construction, $|2\eta_{\boldsymbol{z},\boldsymbol{\sigma}}(x) - 1|$ is either 0 or bounded from below by $2c_\eta r^\alpha$ almost surely. Thus, we only need to consider $\tau = tc_\eta r^\alpha$ for $t \geq 2$. For given $\boldsymbol{z}$, $P_X(\{x : 0 < |\eta_{\boldsymbol{z},\boldsymbol{\sigma}}(X) - 1/2| \leq tc_\eta r^\alpha\}) \leq r^d \mathbf{1}^\top \boldsymbol{z}$. By Chernoff bound (Lemma B.1),

$$\mathop{\mathbb{P}}_{\boldsymbol{z}}\left(r^d \mathbf{1}^\top \boldsymbol{z} \leq C_\beta r^{\alpha\beta}\right) \geq 1 - \exp\left(c_2 r^{-(d-\alpha\beta)}\right),$$

where $c_2 = (C_\beta - 1)^2/3$. Therefore, $\mathop{\mathbb{P}}_{\boldsymbol{z},\boldsymbol{\sigma}}((\boldsymbol{z},\boldsymbol{\sigma}) \in \Theta_\beta) \geq 1 - \exp(c_2 r^{-(d-\alpha\beta)})$ and

$$\begin{aligned} \sup_{P_{\boldsymbol{z},\boldsymbol{\sigma}} \in \Sigma_\beta} \mathbb{E}\, \mathcal{E}(\hat{h}_n) &\geq \mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}} \left[\mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}} \mathcal{E}(\hat{h}_n) \middle| (\boldsymbol{z},\boldsymbol{\sigma}) \in \Theta_\beta \right] \\ &\geq \mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}} \mathcal{E}(\hat{h}_n) - \mathop{\mathbb{P}}_{\boldsymbol{z},\boldsymbol{\sigma}}((\boldsymbol{z},\boldsymbol{\sigma}) \notin \Theta_\beta) \\ &\geq \mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}} \mathcal{E}(\hat{h}_n) - \exp(c_2 r^{-(d-\alpha\beta)}). \end{aligned}$$

□

Definition 8. *The* ***conditional Neyman-Pearson learner*** *$\hat{h}^*$ is the active learner that makes the same sampling decision $\pi_{\hat{h}}$ as $\hat{h}$, and labels according to the following rules for each $\mathcal{C} \in \mathscr{C}_r$. Conditional on the sample $S_{\mathcal{C}} = (X_i^{\mathcal{C}}, Y_i^{\mathcal{C}})_{i=1}^{n_{\mathcal{C}}}$ in $\mathcal{C}$,*

$$\hat{h}_n^*(x) = \left(1 + \operatorname*{argmax}_{\sigma \in \{\pm 1\}} \prod_{i=1}^{n_{\mathcal{C}}} P_{z_{\mathcal{C}}=1,\sigma_{\mathcal{C}}=\sigma}(Y_i^{\mathcal{C}} | X_i^{\mathcal{C}})\right) \Big/ 2,$$

for all $x \in \mathcal{C}$, where $P_{z_{\mathcal{C}},\sigma_{\mathcal{C}}}(Y_i^{\mathcal{C}} | X_i^{\mathcal{C}})$ is the probability of $Y_i^{\mathcal{C}}$ given $X_i^{\mathcal{C}}$, $z_{\mathcal{C}}$ and $\sigma_{\mathcal{C}}$.

Proposition 3. *Let $\hat{h}$ be any active learner, and $\hat{h}^*$ be the corresponding conditional Neyman-Pearson learner, then*

$$\mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}} \mathcal{E}(\hat{h}_n) \geq \mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}} \mathcal{E}(\hat{h}_n^*).$$

Proof. We can decompose the excess risk as:

$$\mathcal{E}(\hat{h}_n) = \sum_{\mathcal{C} \in \mathscr{C}_r} \mathcal{E}_{\mathcal{C}}(\hat{h}_n); \tag{5}$$

with $\mathcal{E}_{\mathcal{C}}(\hat{h}_n) \doteq \int_{\mathcal{C} \cap \{\hat{h}_n \neq (1+\sigma_{\mathcal{C}})/2\}} |2\eta_{\boldsymbol{z},\boldsymbol{\sigma}}(x) - 1| dP_X(x)$. Thus, we only need to show that for $\mathcal{C} \in \mathscr{C}_r$,

$$\mathop{\mathbb{E}}_{S|\hat{h}} \mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}|S,\hat{h}} \mathcal{E}_{\mathcal{C}}(\hat{h}_n^*) \leq \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}|S,\hat{h}} \mathcal{E}_{\mathcal{C}}(\hat{h}_n),$$

where $\mathbb{E}_{S|\hat{h}}$ is the expectation taken over the distribution of S given $\hat{h}$ and $\mathbb{E}_{\boldsymbol{z},\boldsymbol{\sigma}|S,\hat{h}}$ is the taken over the conditional distribution of $(\boldsymbol{z},\boldsymbol{\sigma})$ given S and $\hat{h}$. In the following proof, we suppress the dependency on $\hat{h}$ in notation for simplicity. Note that

$$\mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}|S} \left[\mathcal{E}_{\mathcal{C}}(\hat{h}_n) | z_{\mathcal{C}} = 0\right] = 0; \text{ and}$$

$$\mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}|S} \left[\mathcal{E}_{\mathcal{C}}(\hat{h}_n) | z_{\mathcal{C}} = 1, \sigma_{\mathcal{C}}\right] = 2c_\eta r^{\alpha+d} \mathbb{I}\left(\hat{h}_n \neq (1 + \sigma_{\mathcal{C}})/2\right).$$

Therefore,

$$\begin{aligned} \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}|S} \mathcal{E}_{\mathcal{C}}(\hat{h}_n) = c_\eta r^{d+\alpha(\beta+1)} - c_\eta r^\alpha \mathop{\mathbb{E}}_{S} \Big[& \mathbb{I}\left(\hat{h}_n = 1\right) \\ & (\mathbb{P}(z_{\mathcal{C}} = 1, \sigma_{\mathcal{C}} = 1 | S) - \mathbb{P}(z_{\mathcal{C}} = 1, \sigma_{\mathcal{C}} = -1 | S)) \Big] \end{aligned}$$

is minimized if $\hat{h}_n(x) = 1$ when

$$\frac{\mathbb{P}(z_\mathcal{C} = 1, \sigma_\mathcal{C} = 1|S)}{\mathbb{P}(z_\mathcal{C} = 1, \sigma_\mathcal{C} = -1|S)} \geq 1,$$

and $\hat{h}_n(x) = 0$ otherwise. Finally, notice that

$$\begin{aligned}\frac{\mathbb{P}(z_\mathcal{C} = 1, \sigma_\mathcal{C} = 1|S)}{\mathbb{P}(z_\mathcal{C} = 1, \sigma_\mathcal{C} = -1|S)} &= \frac{dP_{S|z_\mathcal{C}=1,\sigma_\mathcal{C}=1}(S)}{dP_{S|z_\mathcal{C}=1,\sigma_\mathcal{C}=-1}(S)} \\ &= \frac{\prod_{i=1}^{n_\mathcal{C}} P_{z_\mathcal{C}=1,\sigma_\mathcal{C}=1}(Y_i^\mathcal{C}|X_i^\mathcal{C})}{\prod_{i=1}^{n_\mathcal{C}} P_{z_\mathcal{C}=1,\sigma_\mathcal{C}=-1}(Y_i^\mathcal{C}|X_i^\mathcal{C})}.\end{aligned}$$

where the last step is clear from the definition

$$\begin{aligned}dP_{S|z_\mathcal{C},\sigma_\mathcal{C}}(S) =& \prod_{i=1}^{n} \pi_{\hat{h}}(X_i|\{X_j, Y_j\}_{j<i}) \\ &\cdot \underset{\boldsymbol{z}_{(\mathcal{C})},\boldsymbol{\sigma}_{(\mathcal{C})}}{\mathbb{E}} \prod_{\mathcal{C}'\in\mathscr{C}_r} \prod_{i=1}^{n_{\mathcal{C}'}} P_{z_{\mathcal{C}'},\sigma_{\mathcal{C}'}}(Y_i^{\mathcal{C}'}|X_i^{\mathcal{C}'})dS\end{aligned}$$

Hence, the labeling decision of $\hat{h}^*$ minimize $\underset{\boldsymbol{z},\boldsymbol{\sigma}}{\mathbb{E}}\ \underset{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}}{\mathbb{E}} \mathcal{E}_\mathcal{C}(\hat{h})$ for each $\mathcal{C}$, hence $\underset{\boldsymbol{z},\boldsymbol{\sigma}}{\mathbb{E}}\ \underset{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}}{\mathbb{E}} \mathcal{E}(\hat{h})$. □

Notation: For any distribution P on S, we use $dP(S)/dS$ to denote the joint density of continuous $\{X_i\}_{i=1}^n$ and discrete $\{Y_i\}_{i=1}^n$.

Remark 5. Proposition 3 shows that we only need to lower-bound the excess risk rate for the collection of Neyman-Pearson classifiers. Further, since $\mathcal{E}_\mathcal{C}(\hat{h})$ is a function of $z_\mathcal{C}, \sigma_\mathcal{C}$ and $S_\mathcal{C}$, we have $\mathbb{E}_{\boldsymbol{z},\boldsymbol{\sigma}} \mathbb{E}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}^*} \mathcal{E}_\mathcal{C}(\hat{h}_n^*) = \mathbb{E}_{z_\mathcal{C},\sigma_\mathcal{C}} \mathbb{E}_{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}} \mathcal{E}_\mathcal{C}(\hat{h}_n^*)$ where $\mathbb{E}_{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}$ is the expectation over the distribution $P_{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}$ of $S_\mathcal{C}$ given $z_\mathcal{C}, \sigma_\mathcal{C}$ (where we have marginalized out the randomness in other cells). Furthermore, one can decompose $P_{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}$ into the sampling location decision $P^\mathcal{C}_{X|\hat{h}}$ and the labeling distribution $P_{Y|X,z_\mathcal{C},\sigma_\mathcal{C}}$:

$$\begin{aligned}&dP_{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}(S_c) \\ =& \prod_{j=1}^{n_\mathcal{C}} dP^\mathcal{C}_{X|\hat{h}}(X_j^\mathcal{C}|\{X_i^\mathcal{C}, Y_i^\mathcal{C}\}_{i\leq j}) P_{Y|X,z_\mathcal{C},\sigma_\mathcal{C}}(Y_j^\mathcal{C}|X_j^\mathcal{C}).\end{aligned}$$

Proposition 4. *Let $\hat{h}^*$ be any conditional Neyman-Pearson learner. Then,*

$$\underset{\boldsymbol{z},\boldsymbol{\sigma}}{\mathbb{E}}\ \underset{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}^*}{\mathbb{E}} \mathcal{E}(\hat{h}_n^*) \geq C_1 n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}.$$

for some $C_1 > 0$.

Proof. By (5) and the Remark 5, we have

$$\underset{\boldsymbol{z},\boldsymbol{\sigma}}{\mathbb{E}}\ \underset{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}^*}{\mathbb{E}} \mathcal{E}(\hat{h}_n^*) = \sum_{\mathcal{C}\in\mathscr{C}_r} \underset{z_\mathcal{C},\sigma_\mathcal{C}}{\mathbb{E}}\ \underset{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}{\mathbb{E}} \mathcal{E}_\mathcal{C}(\hat{h}_n^*).$$

Let $m \doteq r^d n/2 \equiv (c_\eta r^\alpha)^{-2}/2$,

$$\begin{aligned}&\underset{z_\mathcal{C},\sigma_\mathcal{C}}{\mathbb{E}}\ \underset{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}{\mathbb{E}} \mathcal{E}_\mathcal{C}(\hat{h}_n^*) \\ \geq& \underset{z_\mathcal{C},\sigma_\mathcal{C}}{\mathbb{E}} \sum_{n_\mathcal{C}=1}^{m} \underset{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}{\mathbb{E}} [\mathcal{E}_\mathcal{C}(\hat{h}_n^*) \mid |S_\mathcal{C}|= n_\mathcal{C}] \\ &\qquad\qquad \cdot \underset{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}{\mathbb{P}} (|S_\mathcal{C}|= n_\mathcal{C}) \\ \geq& c_3 r^{d+\alpha} \underset{z_\mathcal{C},\sigma_\mathcal{C}}{\mathbb{E}}\ \underset{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}{\mathbb{P}} \left(z_\mathcal{C} = 1; |S_\mathcal{C}|\leq m\right),\end{aligned}$$

where the last inequality by Lemma 2. Furthermore,

$$\begin{aligned}&\sum_{\mathcal{C}\in\mathscr{C}_r} \underset{z_\mathcal{C},\sigma_\mathcal{C}}{\mathbb{E}}\ \underset{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}{\mathbb{P}} (z_\mathcal{C} = 1; |S_\mathcal{C}|\leq m) \\ =& \sum_{\mathcal{C}\in\mathscr{C}_r} \mathbb{P}(|S_\mathcal{C}|\leq m)\, \mathbb{P}\left(z_\mathcal{C} = 1||S_\mathcal{C}|\leq m)\right) \\ \geq& \frac{r^{\alpha\beta}}{1+c_4} \sum_{\mathcal{C}\in\mathscr{C}_r} \mathbb{P}(z_\mathcal{C} = 1||S_\mathcal{C}|\leq m) \geq \frac{r^{\alpha\beta-d}}{2(1+c_4)},\end{aligned}$$

where the second last inequality is due to Lemma 3, and the last inequality is from the choice of m. Finally,

$$\begin{aligned}\underset{\boldsymbol{z},\boldsymbol{\sigma}}{\mathbb{E}}\ \underset{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}}{\mathbb{E}} \mathcal{E}(\hat{h}_n^*) &= \sum_{\mathcal{C}\in\mathscr{C}_r} \underset{z_\mathcal{C},\sigma_\mathcal{C}}{\mathbb{E}}\ \underset{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}{\mathbb{E}} \mathcal{E}_\mathcal{C}(\hat{h}_n^*) \\ &= (c_3 r^{d+\alpha}) \left(\frac{r^{\alpha\beta-d}}{2(1+c_4)}\right) \geq C_1 n^{-\frac{\alpha(\beta+1)}{2\alpha+d}},\end{aligned}$$

where $C_1 = \frac{c_3(\lambda^2/64)^{\frac{\alpha(\beta+1)}{2\alpha+d}}}{2(1+c_4)} > 0$. □

4.1.3 Supporting lemmas

Lemma 1. *Condition on $z_\mathcal{C}$, $\sigma_\mathcal{C}$ and $|S_\mathcal{C}|= n_\mathcal{C}$, $\mathbf{Y}_\mathcal{C} = \{Y_j^\mathcal{C}\}_{j=1}^{n_\mathcal{C}} \overset{i.i.d}{\sim}$ Ber$(1/2 + z_\mathcal{C}\sigma_\mathcal{C} c_\eta r^\alpha)$.*

Proof. The conditional probability mass of $\mathbf{Y}_\mathcal{C}$ is

$$\begin{aligned}&P_{\mathbf{Y}_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}(\mathbf{Y}_\mathcal{C}) \\ =& \frac{\prod_{j=1}^{n_\mathcal{C}} dP^\mathcal{C}_{X|\hat{h}}(X_j^\mathcal{C}|\{X_i^\mathcal{C}, Y_i^\mathcal{C}\}_{i\leq j}) P_{Y|X,z_\mathcal{C},\sigma_\mathcal{C}}(Y_j^\mathcal{C}|X_j^\mathcal{C})}{\prod_{j=1}^{n_\mathcal{C}} dP^\mathcal{C}_{X|\hat{h}}(X_j^\mathcal{C}|\{X_i^\mathcal{C}, Y_i^\mathcal{C}\}_{i\leq j})} \\ =& \prod_{j=1}^{n_\mathcal{C}} P_{Y|X,z_\mathcal{C},\sigma_\mathcal{C}}(Y_j^\mathcal{C}|X_j^\mathcal{C}) = \prod_{j=1}^{n_\mathcal{C}} (1/2 + z_\mathcal{C}\sigma_\mathcal{C} Y_j^\mathcal{C} c_\eta r^\alpha),\end{aligned}$$

which concludes the proof. □

Lemma 2. *Let $n_\mathcal{C} \leq m = (c_\eta r^\alpha)^{-2}/2$ and $\hat{h}^*$ be a conditional Neyman-Pearson learner. Then, in cell $\mathcal{C}$, for any combination of $(z_\mathcal{C}, \sigma_\mathcal{C})$,*

$$\underset{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}}{\mathbb{E}} [\mathcal{E}_\mathcal{C}(\hat{h}_n^*) \mid |S_\mathcal{C}|= n_\mathcal{C}] \geq c_3 r^{d+\alpha} \mathbb{I}(z_\mathcal{C} = 1).$$

for some $c_3 > 0$.

Proof. When $z_{\mathcal{C}} = 0$, the inequality holds trivially. When $z_{\mathcal{C}} = 1$,

$$\mathcal{E}_{\mathcal{C}}(\hat{h}_n^*) = r^{d+\alpha}\mathbb{I}\left(\sigma_{\mathcal{C}}\left[\frac{1}{n_{\mathcal{C}}}\sum_{j=1}^{n_{\mathcal{C}}} Y_j^{\mathcal{C}} - \frac{1}{2}\right] < 0\right),$$

the inequality holds by Lemma 1 and the anti-concentration inequality (Lemma B.2). □

Lemma 3. *Let $S_{\mathcal{C}} = (X_j^{\mathcal{C}}, Y_j^{\mathcal{C}})_{j=1}^{n_{\mathcal{C}}}$ be such that $n_{\mathcal{C}} = |S_{\mathcal{C}}| \le m$. Then,*

$$\frac{\underset{\sigma_{\mathcal{C}}}{\mathbb{E}}\, dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=0,\sigma_{\mathcal{C}},\hat{h}}(S_{\mathcal{C}})}{\underset{\sigma_{\mathcal{C}}}{\mathbb{E}}\, dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=1,\sigma_{\mathcal{C}},\hat{h}}(S_{\mathcal{C}})} \le c_4,$$

for some absolute constant $c_4 > 0$. Consequently,

$$\mathbb{P}\left(z_{\mathcal{C}} = 1 || S_{\mathcal{C}}| \le m\right) \ge \frac{r^{\alpha\beta}}{1+c_4}.$$

Proof. By definition,

$$\begin{aligned}
&\underset{\sigma_{\mathcal{C}}}{\mathbb{E}}\, dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=0,\sigma_{\mathcal{C}},\hat{h}}(S_{\mathcal{C}}) \\
=&\left(\frac{1}{2}\right)^{n_{\mathcal{C}}} \prod_{j=1}^{n_{\mathcal{C}}} dP^{\mathcal{C}}_{X|\hat{h}}\left(X_j^{\mathcal{C}} | (X_i^{\mathcal{C}}, Y_i^{\mathcal{C}})_{i \le j}\right), \\
&\underset{\sigma_{\mathcal{C}}}{\mathbb{E}}\, dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=1,\sigma_{\mathcal{C}},\hat{h}}(S_{\mathcal{C}}) \\
\ge&\frac{1}{2}\left(\frac{1}{2} + c_\eta r^\alpha\right)^{n_{\mathcal{C}}/2}\left(\frac{1}{2} - c_\eta r^\alpha\right)^{n_{\mathcal{C}}/2} \\
&\cdot \prod_{j=1}^{n_{\mathcal{C}}} dP^{\mathcal{C}}_{X|\hat{h}}\left(X_j^{\mathcal{C}} | (X_i^{\mathcal{C}}, Y_i^{\mathcal{C}})_{i \le j}\right).
\end{aligned}$$

Thus,

$$\frac{\underset{\sigma_{\mathcal{C}}}{\mathbb{E}}\, dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=0,\sigma_{\mathcal{C}},\hat{h}}(S_{\mathcal{C}})}{\underset{\sigma_{\mathcal{C}}}{\mathbb{E}}\, dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=1,\sigma_{\mathcal{C}},\hat{h}}(S_{\mathcal{C}})} \le 2(1-4/m)^{-m/2} \le c_4,$$

for $c_4 = 16e^2$. Consequently,

$$\begin{aligned}
&\mathbb{P}\left(z_{\mathcal{C}} = 1 || S_{\mathcal{C}}| \le m\right) \\
=&\frac{\mathbb{P}(z_{\mathcal{C}} = 1, |S_{\mathcal{C}}| \le m)}{\mathbb{P}(|S_{\mathcal{C}}| \le m)} \\
=&\frac{\mathbb{P}(z_{\mathcal{C}} = 1, |S_{\mathcal{C}}| \le m)}{\mathbb{P}(z_{\mathcal{C}} = 1, |S_{\mathcal{C}}| \le m) + \mathbb{P}(z_{\mathcal{C}} = 0, |S_{\mathcal{C}}| \le m)} \\
\ge&\frac{r^{\alpha\beta}}{1+c_4}.
\end{aligned}$$

□

4.2 Proof of Upper-bounds

In this section, we establish the upper bounds on excess risk rates for Algorithm 1. Due to space limit, we only outline the proof of the results under strong density condition and relegate the more direct proof under general density and other technical details in the supplementary materials. We start with a guarantee on the subroutine.

Proposition 5 (Guarantees for Algorithm 2)**.** *Let $n_0 \in \mathbb{N}$ and $\alpha\beta' \le d$. Let $\{S_{\mathcal{C}}\}_{\mathcal{C} \in r_0}$ be the outputs of Algorithm 2 with input n_0, λ, α and $\delta_0 \in (0,1)$, and $\hat{h}_{n_0,\alpha}$ be any classifier that satisfies $\hat{h}_{n_0,\alpha}(x) \in S_{\mathcal{C}}, \forall x \in \mathcal{C} \in \mathscr{C}_{r_0}$. Under Assumption 1 and 2 and strong density condition, with probability at least $1-\delta_0$,*

$$\begin{aligned}
\mathcal{E}\left(\hat{h}_{n_0,\alpha}\right) \le C_5 &\left(\varepsilon_0^{\frac{\alpha(\beta+1)}{2\alpha+d}}\left(\frac{\lambda^{\frac{d}{\alpha}}\log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}}\right. \\
&\left. + \left(\frac{\lambda^{\frac{d}{\alpha}\vee\beta'}\log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0}\right)^{\frac{\alpha(\beta'+1)}{2\alpha+d-\alpha\beta'}}\right)
\end{aligned}$$

for some constant $C_5 > 0$, which are independent of $n_0, \lambda, L, \varepsilon_0$ and δ_0.

Proof. Under some favorable event ξ_α with probability at least $1-\delta_0$ (Lemma A.1), the following holds:

- Algorithm 2 will reach level

$$\begin{aligned}
r_{\min} \le \max &\left\{\left(\frac{c_7\lambda^{-2}\varepsilon_0\log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0}\right)^{\frac{1}{2\alpha+d}},\right. \\
&\left.\left(\frac{c_7\lambda^{\beta'-2}\log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0}\right)^{\frac{1}{2\alpha+d-\alpha\beta'}}\right\} \\
&\doteq \max\{Q_1, Q_2\};
\end{aligned}$$

for some $c_7 > 0$ (Lemma A.4);

- Algorithm 2 never eliminates Bayes labels (Lemma A.2);

- $\forall \mathcal{C} \in \mathscr{C}_{r_0}, \forall x \in \mathcal{C}, \forall y \in S_{\mathcal{C}}$, $\eta_{(1)}(x) - \eta_y(x) \le 10\lambda r_{\min}^\alpha$, and $S_{\mathcal{C}}$ contains only Bayes labels in regions where $\mathcal{M} > 10\lambda r_{\min}^\alpha$ (Lemma A.3);

When $Q_1 \le Q_2$, $\varepsilon_0 + C_\beta r_{\min}^{\alpha\beta'} \le c_8 r_{\min}^{\alpha\beta'}$ for some $c_8 > 0$,

$$\begin{aligned}
\mathcal{E}(\hat{h}_{n_0,\alpha}) &\le P_X(\{x : \mathcal{M}'(x) \le 10\lambda r_{\min}^\alpha\})(10\lambda r_{\min}^\alpha) \\
&\le C_5'\left(\frac{\lambda^{\frac{d}{\alpha}\vee\beta'}\log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0}\right)^{\frac{\alpha(\beta'+1)}{2\alpha+d-\alpha\beta'}},
\end{aligned}$$

for some $C_5' > 0$. When $Q_1 > Q_2$,

$$\begin{aligned}\mathcal{E}(\hat{h}_{n_0,\alpha}) &\le P_X(\{x : \mathcal{M}(x) \le 10\lambda r_{\min}^{\alpha}\})(10\lambda r_{\min}^{\alpha}) \\ &\le C_5'' \varepsilon_0^{\frac{\alpha(\beta+1)}{2\alpha+d}} \left(\frac{\lambda^{\frac{d}{\alpha}} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}}.\end{aligned}$$

for some $C_5'' > 0$. We then conclude the proof by choosing $C_5 = \max\{C_5', C_5''\}$. □

Outline of Proof for Theorem 2 and 3. The *Correctness of aggregation* relies on the fact that Algorithm 1 a) never adds back removed labels, and b) stops aggregating labels when all labels are about to be removed from a cell – this ensures the final candidate set $\mathcal{L}_{\mathcal{C}}$ contains no bad labels and is non-empty. By Proposition 6, we have excess risk bounds for all $\alpha_i \le \alpha$, among which the largest one satisfies $\alpha - \alpha_i \le 1/\lfloor \log(n) \rfloor^3$. Direct calculation shows that the excess risk bound with α_{i^*} is only a constant factor away from the one with α.

5 CONCLUSION

In this paper, we have shown that simple nuances in notions of margin—seemingly having to do with uniqueness of the Bayes classifier—affect whether any active learner can gain over passive learning. Our main result is the lower bound (Theorem 1), which requires proof techniques quite different from the usual lower bounds arguments in active learning, e.g. Minsker (2012), Locatelli et al. (2017). We also show that savings remain possible in the worst case over P_X, and also under a refined margin condition in regimes with small sampling budget.

Our main Theorem 1 is shown here for the binary case, which does not distinguish between uniqueness of the Bayes and all labels being equivalent; as such it leaves open the possibility of a more refined picture in the case of multiple labels, i.e., whether allowing multiple labels (but not all) to be equivalent is enough to preclude savings over passive learning.

Finally, while our results concern the nonparametric setting of active learning, it remains open whether similar nuances in achievable rates occur in parametric settings with bounded VC classes.

Acknowledgements

The three authors, Samory Kpotufe, Gan Yuan, Yunfan Zhao, are listed in alphabetical order. Samory Kpotufe acknowledges support under a Sloan fellowship, and NSF Grant Id 1739809. He is also a visiting faculty at Google AI Princeton.

References

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

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

Hanneke, S. (2011). Rates of convergence in active learning. *The Annals of Statistics*, pages 333–361.

Hanneke, S. and Yang, L. (2015). Minimax analysis of active learning. *J. Mach. Learn. Res.*, 16(12):3487–3602.

Koltchinskii, V. (2010). Rademacher complexities and bounding the excess risk in active learning. *The Journal of Machine Learning Research*, 11:2457–2485.

Locatelli, A., Carpentier, A., and Kpotufe, S. (2017). Adaptivity to noise parameters in nonparametric active learning. *Proceedings of Machine Learning Research*, 65:1–34.

Locatelli, A., Carpentier, A., and Kpotufe, S. (2018). An adaptive strategy for active learning with smooth decision boundary. In *Algorithmic Learning Theory*, pages 547–571. PMLR.

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

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

Minsker, S. (2012). Plug-in approach to active learning. *Journal of Machine Learning Research*, 13:67–90.

Mousavi, N. (2010). How tight is chernoff bound? `https://ece.uwaterloo.ca/~nmousavi/Papers/Chernoff-Tightness.pdf`.

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

Wang, Y. and Singh, A. (2016). Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Yan, S., Chaudhuri, K., and Javidi, T. (2016). Active learning from imperfect labelers. *Advances in Neural Information Processing Systems*, 29:2128–2136.

Supplementary Material: Nuances in Margin Conditions Determine Gains in Active Learning

A Proof of the Theorem 2 and 3

To begin with, we define some quantities and notions that will be used in the lemmas.

Definition 9. *Let A be any measurable subset of $[0,1]^d$ and $y \in [L]$. We define the regression function in A for label y as $\eta_y(A) \doteq \left[\int_A \eta_y(x)dx\right] / \left[\int_A dx\right]$.*

Given n_A independent samples $\{(X_j^A, Y_j)\}_{j=1}^{n_A}$ in A, an unbiased estimator of $\eta_y(A)$ is

$$\hat{\eta}_y(A) \doteq \frac{1}{n_A}\sum_{i=1}^{n_A} \mathbb{I}(Y_i = y).$$

To get the high probability bound, we focus the discussion on a subset under which the estimation error of $\hat{\eta}$ at each cell is small throughout the proof. We consider a favorable event $\xi_\alpha \doteq \bigcap_{r\in\mathcal{I}_r, \mathcal{C}\in\mathscr{C}_r} \xi_{\mathcal{C},r,\alpha}$, where

$$\mathcal{I}_r \doteq \{1/2, 1/4, \ldots, r_{\min}, r_{\min}/2\},$$
$$\xi_{\mathcal{C},r,\alpha} \doteq \left\{\|\hat{\eta}(\mathcal{C}) - \eta(\mathcal{C})\|_\infty \leq \lambda r^\alpha\right\}.$$

The following lemma shows that ξ_α is indeed a high probability event.

Lemma 4. $\mathbb{P}(\xi_\alpha) \geq 1 - \delta_0$.

Proof. By Hoeffding's inequality, for each $y \in [L]$,

$$\mathbb{P}\left(|\hat{\eta}_y(\mathcal{C}) - \eta_y(\mathcal{C})| \geq \lambda r^\alpha\right) \leq \frac{\delta_0 r^{d+1}}{L}.$$

By union bound, $\mathbb{P}(\xi_{\mathscr{C},r,\alpha}) \geq 1 - \sum_{y=1}^L (\delta_0 r^{d+1})/L = 1 - \delta_0 r^{d+1}$. Another application of union bound yields $\mathbb{P}(\xi_\alpha) \geq 1 - \sum_{r\in\mathcal{I}_r} r^{-d}\delta_0 r^{d+1} \geq 1 - \delta_0$. □

Next, we show some desired properties of Algorithm 2 on the favorable event ξ_α. In particular, Lemma 5 shows that, Algorithm 2 never eliminate Bayes labels; Lemma 6 shows that Algorithm 2 predicts only Bayes labels in the area where soft margin is large enough; Lemma 7 shows that the algorithm will at least reach some certain level $r_{\min}$ of partition.

Lemma 5. *On the event ξ_α, suppose that Algorithm 2 is in the depth that the partition is of sidelength r. For any $x \in [0,1]^d$, we have $\eta_y(x) < \eta_{(1)}(x)$ for any $y \notin S_{\mathcal{C}}$, where $x \in \mathcal{C} \in \mathscr{C}_r$. That is, the algorithm never eliminate Bayes labels.*

Proof. For any $y \in [L]$, by definition of ξ_α and smoothness assumption, we have

$$|\hat{\eta}_y(\mathcal{C}) - \eta_y(\mathcal{C})| \leq ||\hat{\eta}(\mathcal{C}) - \eta(\mathcal{C})||_\infty \leq \lambda r^\alpha;$$

$$|\eta_y(x) - \eta_y(\mathcal{C})| \leq ||\eta(x) - \eta(\mathcal{C})||_\infty \leq \lambda r^\alpha.$$

By the algorithm design, $\hat{\eta}_{(1)}(x) - \hat{\eta}_y(x) \geq 6\lambda r^\alpha$. Therefore, $\eta_{(1)}(x) - \eta_y(x) \geq \hat{\eta}_{(1)}(\mathcal{C}) - \hat{\eta}_y(\mathcal{C}) - 4\lambda r^\alpha > 0$. □

Lemma 6. *On the event ξ_α, suppose that Algorithm 2 is in the depth that the partition is of side length r. If $\eta_{(1)}(x) - \eta_y(x) \geq \Delta_r = 10\lambda r^\alpha$ for some $x \in [0,1]^d$ and $y \in [L]$, then for the cell $\mathcal{C} \in \mathscr{C}_r$ that contains x, the label y will be eliminated. Consequently, for any $x \in [0,1]^d$ with $\mathcal{M}(x) > \Delta_r$, $S_\mathcal{C}$ contains only Bayes labels.*

Proof. For any $y \in [L]$, by Assumption 1, $\eta_{(1)}(x) - \eta_y(x) \geq \eta_{(1)}(\mathcal{C}) - \eta_y(\mathcal{C}) - 2\lambda r^\alpha$. By the definition of ξ_α, we have $|\eta_y(\mathcal{C}) - \hat{\eta}_y(\mathcal{C})| \leq \lambda r^\alpha$, and hence

$$\begin{aligned}\hat{\eta}_{(1)}(\mathcal{C}) - \hat{\eta}_y(\mathcal{C}) \geq & |\eta_{(1)}(\mathcal{C}) - \eta_y(\mathcal{C})| - |\eta_y(\mathcal{C}) - \hat{\eta}_y(\mathcal{C})| - |\eta_{(1)}(\mathcal{C}) - \hat{\eta}_{(1)}(\mathcal{C})| \\ \geq & |\eta_{(1)}(\mathcal{C}) - \eta_y(\mathcal{C})| - 2\lambda r^\alpha \\ \geq & 6\lambda r^\alpha\,.\end{aligned}$$

□

Lemma 7. *On the event ξ_α,*

i) Under Assumption 1 and 2, then the finest partition Algorithm 2 can reach satisfies

$$r_{\min} \leq \left(\frac{c_6 \lambda^{-2} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{1/(2\alpha+d)} ;$$

for some $c_6 > 0$;

ii) Under Assumption 1 and 2, and assume further that strong density condition holds for $c_d > 0$, then

$$r_{\min} \leq \max\left\{ \left(\frac{c_7 \lambda^{-2} \varepsilon_0 \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{1}{2\alpha+d}}, \left(\frac{c_7 \lambda^{\beta'-2} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{1}{2\alpha+d-\alpha\beta'}} \right\},$$

for some $c_7 > 0$.

Proof. i) The total budget is not sufficient for a finer partition than length $r_{\min}$, hence

$$\begin{aligned} n_0 \leq \sum_{r \in \mathcal{I}_r} |\mathcal{A}_r| n_r &\leq \sum_{r \in \mathcal{I}_r} r^{-d} \cdot \frac{2\log(2L/\delta_0 r^{d+1})}{\lambda^2 r^{2\alpha}} \\ &\leq \frac{2(d+1)\log 2}{\lambda^2} \log(2L/(\delta_0 r_{min}^{d+1})) \sum_{r \in \mathcal{I}_r} r^{-(2\alpha+d)} \\ &\leq \frac{2(d+1)\log 2}{\lambda^2} \log(2L/(\delta_0 r_{\min}^{d+1})) \left(\frac{r_{\min}^{-(2\alpha+d)} 4^{2\alpha+d}}{2^{2\alpha+d} - 1} \right) \\ &\leq \frac{4(d+1)\log 2}{\lambda^2} \log(2L/(\delta_0 r_{\min}^{d+1})) \left(\frac{r_{\min}^{-(2\alpha+d)} 4^{2\alpha+d}}{2\alpha+d} \right), \end{aligned}$$

where the last equality is from the inequality $2^u - 1 \geq \frac{u}{2}$ for $u \in \mathbb{R}^+$. We now prove an upper bound on $\log(2L/(\delta_0 r_{\min}^{d+1}))$. Use the trivial bound

$$\frac{\log\left(2L/(\delta_0 r_{\min}^{d+1})\right)}{2\lambda^2 r_{\min}^{2\alpha}} = n_{r_{\min}} \leq n_0$$

and $\delta_0 r_{\min}^{d+1} < \delta_0/2 \leq 2e^{-1}$, we have

$$\frac{\log(L)}{2\lambda^2 r_{\min}^{2\alpha}} \leq n_0$$

which implies

$$r_{\min} \geq \left(\frac{\log(L)}{2\lambda^2 n_0} \right)^{1/2\alpha}$$

and therefore

$$\log(2L/(\delta_0 r_{\min}^{d+1})) \leq \log\left(\frac{2L}{\delta_0} \left(\frac{2\lambda^2 n_0}{\log(L)} \right)^{(d+1)/2\alpha} \right) \leq \frac{d+1}{2\alpha} \log\left(\frac{4L\lambda^2 n_0}{\delta_0} \right) \tag{6}$$

With this upper bound on $\log(L/(\delta_0 r_{\min}^{d+1}))$, we now proceed to upper bound $r_{\min}$. Clearly,

$$n_0 \leq \frac{c_6}{\lambda^2} \log\left(\frac{4L\lambda^2 n_0}{\delta_0} \right) r_{\min}^{-(2\alpha+d)}$$

where $c_6 = \frac{4(d+1)^2 4^{2\alpha+d} \log 2}{2\alpha(2\alpha+d)}$. Therefore,

$$r_{\min} \leq \left(\frac{c_6}{\lambda^2 n_0} \log\left(\frac{4L\lambda^2 n_0}{\delta_0} \right) \right)^{1/(2\alpha+d)}.$$

ii) From the strong density condition and Lemma 6, we have a tighter bound on the number of active cells:

$$|\mathcal{A}_r| \leq \frac{\varepsilon_0 + C_\beta (6\lambda r^\alpha)^{\beta'}}{c_d r^d}.$$

Using similar argument as in i), we have

$$\begin{aligned}
n_0 &\leq \sum_{r \in \mathcal{I}_r} |\mathcal{A}_r| n_r \\
&\leq \sum_{r \in \mathcal{I}_r} \frac{\varepsilon_0 + C_\beta (6\lambda r^\alpha)^{\beta'}}{c_d r^d} \cdot \frac{2 \log(2L/\delta_0 r^{d+1})}{\lambda^2 r^{2\alpha}} \\
&\leq \frac{4(d+1)\log 2}{c_d \lambda^2} \log(2L/(\delta_0 r_{\min}^{d+1})) \left(\varepsilon_0 \frac{r_{\min}^{-(2\alpha+d)} 4^{2\alpha+d}}{2\alpha+d} + C_\beta (6\lambda)^{\beta'} \frac{r_{\min}^{-(2\alpha+d-\alpha\beta')} 4^{2\alpha+d-\alpha\beta'}}{2\alpha+d-\alpha\beta'} \right) \\
&\leq c_7 \lambda^{-2} \log\left(\frac{4L\lambda^2 n_0}{\delta_0} \right) \max\left\{ \varepsilon_0 r_{\min}^{-(2\alpha+d)}, \lambda^{\beta'} r_{\min}^{-(2\alpha+d-\alpha\beta')} \right\}.
\end{aligned}$$

where $c_7 = \frac{4(d+1)^2 4^{2\alpha+d} \log 2}{c_d \alpha(2\alpha+d-\alpha\beta')} \max\{1, C_\beta 6^{\beta'}\}$, and the last step is from (6). Therefore,

$$r_{\min} \leq \max\left\{ \left(\frac{c_7 \lambda^{-2} \varepsilon_0 \log\left(\frac{4L\lambda^2 n_0}{\delta_0} \right)}{n_0} \right)^{\frac{1}{2\alpha+d}}, \left(\frac{c_7 \lambda^{\beta'-2} \log\left(\frac{4L\lambda^2 n_0}{\delta_0} \right)}{n_0} \right)^{\frac{1}{2\alpha+d-\alpha\beta'}} \right\}.$$

□

Now we prove rates for Algorithm 2. The proposition below is a generalized version of Proposition 5, and it includes rates under strong density condition.

Proposition 6 (Guarantees for Algorithm 2)**.** *Let $n_0 \in \mathbb{N}$ and $\alpha\beta' \leq d$. Let $\{S_\mathcal{C}\}_{\mathcal{C} \in r_0}$ be the outputs of Algorithm 2 with input n_0, λ, α and $\delta_0 \in (0,1)$, and $\hat{h}_{n_0,\alpha}$ be any classifier that satisfies $\hat{h}_{n_0,\alpha}(x) \in S_\mathcal{C}, \forall x \in \mathcal{C} \in \mathscr{C}_{r_0}$. Under Assumption 1 and 2,*

i) With probability at least $1 - \delta_0$,

$$\mathcal{E}\left(\hat{h}_{n_0,\alpha} \right) \leq C_4 \left(\frac{\lambda^{\frac{d}{\alpha}} \log\left(\frac{4L\lambda^2 n_0}{\delta_0} \right)}{n_0} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}}$$

ii) Suppose further that strong density condition holds with some $c_d > 0$, then with probability at least $1 - \delta_0$,

$$\mathcal{E}\left(\hat{h}_{n_0,\alpha}\right) \le C_5 \left(\varepsilon_0^{\frac{\alpha(\beta+1)}{2\alpha+d}} \left(\frac{\lambda^{\frac{d}{\alpha}} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}} + \left(\frac{\lambda^{\frac{d}{\alpha}\vee\beta'} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{\alpha(\beta'+1)}{2\alpha+d-\alpha\beta'}} \right)$$

for some constant $C_4, C_5 > 0$, which are independent of $n_0, \lambda, L, \varepsilon_0$ and δ_0.

Proof of Proposition 6.

i) On ξ_α with probability at least $1 - \delta_0$, we have by Part i) of Lemma 7,

$$\Delta_{r_{\min}} = 10\lambda r_{\min}^{\alpha} \le 10\lambda \left(\frac{c_6 \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{\lambda^2 n_0} \right)^{\frac{\alpha}{2\alpha+d}} \le 10 \left(\frac{c_6 \lambda^{\frac{d}{\alpha}} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{\alpha}{2\alpha+d}}.$$

By Lemma 6, the classifier $\hat{h}_{n_0,\alpha}$ makes no error at $\{x : \mathcal{M}(x) > \Delta_{r_{\min}}\}$, and thus

$$\mathcal{E}(\hat{h}_{n_0,\alpha}) \le \underset{X}{\mathbb{P}}(\mathcal{M}(x) \le \Delta_{r_{\min}}) \cdot \Delta_{r_{\min}} \le C_\beta \Delta_{r_{\min}}^{\beta+1} \le C_4 \left(\frac{\lambda^{\frac{d}{\alpha}} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}},$$

where $C_4 = C_\beta 10^{\beta+1} c_6^{\frac{\alpha(\beta+1)}{2\alpha+d}}$.

ii) On ξ_α with probability at least $1 - \delta_0$, we have by Part ii) of Lemma 7,

$$\begin{aligned}\Delta_{r_{\min}} &\le 10 \max\left\{ \varepsilon_0^{\frac{\alpha}{2\alpha+d}} \left(\frac{c_7 \lambda^{\frac{d}{\alpha}} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{\alpha}{2\alpha+d}}, \quad \left(\frac{c_7 \lambda^{\frac{d}{\alpha}\vee\beta'} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{\alpha}{2\alpha+d-\alpha\beta'}} \right\} \\ &\doteq 10\max\{Q_1, Q_2\}.\end{aligned}$$

Case 1: $Q_1 \le Q_2$

Under this case, it is clear that $\varepsilon_0 \le c_8 \Delta_{r_{\min}}^{\beta'}$ for some $c_8 > 0$.

Therefore,

$$\begin{aligned}\mathcal{E}(\hat{h}_{n,\alpha}) &\le \underset{X}{\mathbb{P}}(\mathcal{M}(x) \le \Delta_{r_{\min}})\Delta_{r_{\min}} \\ &\le \underset{X}{\mathbb{P}}(\mathcal{M}'(x) \le \Delta_{r_{\min}})\Delta_{r_{\min}} \\ &\le C_\beta(\varepsilon_0 + \Delta_{r_{\min}}^{\beta'})\Delta_{r_{\min}} \\ &\le C_\beta(c_8+1)\Delta_{r_{\min}}^{\beta'+1} \\ &\le C_5' \left(\frac{\lambda^{\frac{d}{\alpha}\vee\beta'} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{\alpha(\beta'+1)}{2\alpha+d-\alpha\beta'}},\end{aligned}$$

where $C_5' = C_\beta(c_8+1)10^{\beta'+1} c_7^{\frac{\alpha(\beta'+1)}{2\alpha+d-\alpha\beta'}}$.

Case 2: $Q_1 > Q_2$

Under this case,

$$\mathcal{E}(\hat{h}_{n_0,\alpha}) \le \underset{X}{\mathbb{P}}(\mathcal{M}(x) \le \Delta_{r_{\min}})\Delta_{r_{\min}} \le C_\beta \Delta_{r_{\min}}^{\beta+1} \le C_5'' \varepsilon_0^{\frac{\alpha(\beta+1)}{2\alpha+d}} \left(\frac{\lambda^{\frac{d}{\alpha}} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}},$$

where $C_5'' = C_\beta 10^{\beta+1} c_7^{\frac{\alpha(\beta+1)}{2\alpha+d}}$. Finally, set $C_5 = \max\{C_5', C_5''\}$ and the desired result follows.

□

Proof of Theorem 2 and 3.
Due to their similarity, we only prove Theorem 2, and omit the proof of Theorem 3. The bound is trivial for $\alpha < \frac{1}{\log(n)}$, since $n^{-\alpha} \geq n^{-1/\log(n)} \geq \frac{1}{e}$. Thus, we will consider $\alpha \geq \frac{1}{\log(n)}$. Let $\delta_0 = \delta/\big(\lfloor \log(n) \rfloor^3\big)$ and $\alpha_i = i/\lfloor \log(n) \rfloor^3$ for $i \in [\lfloor \log(n) \rfloor^3]$, as defined in Algorithm 2. Let i^* be the largest integer $i \in [\lfloor \log(n) \rfloor^3]$ such that $\alpha_i \leq \alpha$. By Lemma 4 and 5, on ξ_{α_i} with probability at least $1 - \delta_0$, we have

$$\forall \mathcal{C} \in \mathscr{C}_{r_0}, \forall x \in \mathcal{C}, \operatorname*{argmax}_y \eta_y(x) \in \mathcal{L}_{\mathcal{C}}^{\alpha_i}$$

By a union bound, with probability at least $1 - \lfloor \log(n) \rfloor^3 \delta_0 = 1 - \delta$, above holds jointly for all $i \leq i^*$. Thus, with probability at least $1 - \delta$,

$$\forall \mathcal{C} \in \mathscr{C}_{r_0}, \forall x \in \mathcal{C}, \operatorname*{argmax}_y \eta_y(x) \subseteq \cap_{i \leq i^*} \mathcal{L}_{\mathcal{C}}^{\alpha_i},$$

and hence $\cap_{i \leq i^*} \mathcal{L}_{\mathcal{C}}^{\alpha_i} \neq \emptyset$. Therefore, $\mathcal{L}_{\mathcal{C}} \subset \mathcal{L}_{\mathcal{C}}^{\alpha_{i^*}}$ for any $\mathcal{C} \in \mathscr{C}_{r_0}$. By proposition 6 and the fact that budget for each α_i is $n_0 = \frac{n}{\lfloor \log(n) \rfloor^3}$, we have

$$\mathcal{E}\left(\hat{h}_n\right) \leq C_5 \left(\varepsilon_0^{\frac{\alpha_{i^*}(\beta+1)}{2\alpha_{i^*}+d}} \left(\frac{\lambda^{\frac{d}{\alpha_{i^*}}} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{\alpha_{i^*}(\beta+1)}{2\alpha_{i^*}+d}} + \left(\frac{\lambda^{\frac{d}{\alpha_{i^*}} \vee \beta'} \log\left(\frac{4L\lambda^2 n_0}{\delta_0}\right)}{n_0} \right)^{\frac{\alpha_{i^*}(\beta'+1)}{2\alpha_{i^*}+d-\alpha_{i^*}\beta'}} \right)$$

It remains to argue that going from α_{i^*} to α, we add at most a constant multiplicative factor to the excess risk bound. Notice that

$$\frac{\alpha(1+\beta)}{2\alpha+d} - \frac{\alpha_{i^*}(1+\beta)}{2\alpha_{i^*}+d} \leq \frac{1+\beta}{2\alpha \lfloor \log(n) \rfloor^3} \leq \frac{1+\beta}{2\log^2(n)} \cdot \frac{\log^3(n)}{\lfloor \log(n) \rfloor^3}$$

where the last step is due to $\alpha \geq \frac{1}{\log(n)}$. Similarly,

$$\begin{aligned}
\frac{\alpha(1+\beta')}{2\alpha+d-\alpha\beta'} - \frac{\alpha_{i^*}(1+\beta')}{2\alpha_{i^*}+d-\alpha_{i^*}\beta'} \leq & \frac{(1+\beta')(\alpha-\alpha_{i^*})(2\alpha+d)}{(2\alpha+d-\alpha\beta')^2} \\
\leq & \frac{(1+\beta')(2\alpha+d)}{\log^3(n)(2\alpha+d-\alpha\beta')^2} \cdot \frac{\log^3(n)}{\lfloor \log(n) \rfloor^3} \\
\leq & \frac{(1+\beta')(2\alpha+d)}{\log^3(n)(2\alpha)^2} \cdot \frac{\log^3(n)}{\lfloor \log(n) \rfloor^3} \\
\leq & \frac{(1+\beta')(2+d)}{4\log^3(n)\alpha^2} \cdot \frac{\log^3(n)}{\lfloor \log(n) \rfloor^3} \\
\leq & \frac{(1+\beta')(2+d)}{4\log(n)} \cdot \frac{\log^3(n)}{\lfloor \log(n) \rfloor^3}
\end{aligned}$$

where the last step is due to $\alpha \geq \frac{1}{\log(n)}$. Therefore, for n sufficiently large,

$$\begin{aligned}
\left(\frac{\log^3(n) \lambda^{\frac{d}{\alpha_{i^*}}} \log\left(\frac{4L\lambda^2 n}{\delta}\right)}{n} \right)^{-\frac{\alpha(1+\beta)}{2\alpha+d} + \frac{\alpha_{i^*}(1+\beta)}{2\alpha_{i^*}+d}} & \leq 2e^{\frac{1+\beta}{2\log(n)}}, \\
\left(\frac{\log^3(n) \lambda^{\frac{d}{\alpha_{i^*}} \vee \beta'} \log\left(\frac{4L\lambda^2 n}{\delta}\right)}{n} \right)^{-\frac{\alpha(1+\beta')}{2\alpha+d-\alpha\beta'} + \frac{\alpha_{i^*}(1+\beta')}{2\alpha_{i^*}+d-\alpha_{i^*}\beta'}} & \leq 2e^{(1+\beta')(2+d)/4}
\end{aligned}$$

and hence Theorem 2 holds with for $C_2 = 2e^{(1+\beta')(2+d)} C_5$. □

B Technical Lemmas for the Lower-bound

Lemma 8 (Chernoff bound)**.** *Suppose* $Y_1, \ldots, Y_m$ *be independent random variables taking values in* $\{0, 1\}$ *and* $\bar{Y} = (\sum_{i=1}^m Y_i)/\, m$*. Then, for* $\varepsilon > 0$*,*

$$\mathbb{P}\left(\bar{Y} \geq (1+\varepsilon)\,\mathbb{E}\,\bar{Y}\right) \leq \exp\left(-m\varepsilon^2\,\mathbb{E}\,\bar{Y}/3\right).$$

Lemma 9 (Anti-concentration inequality)**.** *Let* $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} Ber(1/2+\delta)$ *for some* $0 < \delta < 1/2$*. If* $m \leq \delta^{-2}/2$*, then*

$$\mathbb{P}\left(\frac{1}{m}\sum_{j=1}^{m} Y_j < \frac{1}{2}\right) \geq c_3,$$

for some absolute constant $c_3 > 0$*.*

Proof. It follows directly from Theorem 2 (ii) of Mousavi (2010). □