Journal of Machine Learning Research 25 (2024) 1-31 Submitted 2/23; Revised 10/23; Published 3/24

Regimes of No Gain in Multi-class Active Learning

Gan Yuan GAN.YUAN@COLUMBIA.EDU
Department of Statistics
Columbia University in the City of New York
New York, NY 10027, USA

Yunfan Zhao YZ3685@COLUMBIA.EDU
Department of Industrial Engineering and Operations Research
Columbia University in the City of New York
New York, NY 10027, USA

Samory Kpotufe SKK2175@COLUMBIA.EDU
Department of Statistics
Columbia University in the City of New York
New York, NY 10027, USA

Editor: Aarti Singh

Abstract

We consider nonparametric classification with smooth regression functions, where it is well known that notions of margin in $\mathbb{P}(Y = y|X = x)$ determine fast or slow rates in both active and passive learning. Here we elucidate a striking distinction—most relevant in multi-class settings—between active and passive learning. Namely, we show that some seemingly benign nuances in notions of margin—involving the uniqueness of the Bayes classes, which have no apparent effect on rates in passive learning—determine whether or not *any* active learner can outperform passive learning rates. While a shorter conference version of this work already alluded to these nuances, it focused on the binary case and thus failed to be conclusive as to the source of difficulty in the multi-class setting: we show here that it suffices that the Bayes classifier fails to be unique, as opposed to needing *all classes to be Bayes optimal*, for active learning to yield no gain over passive learning.

More precisely, we show that for *Tsybakov's margin condition* (allowing general situations with non-unique Bayes classifiers), no active learner can gain over passive learning in terms of worst-case rate 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 improve over passive rates in nonparametric classification; as such these nuances allow to better characterize the actual sources of gain in active over passive learning.

Keywords: active learning, margin conditions, minimax lower bound, multi-class classification, non-parameteric classification

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 (Audibert and Tsybakov (2007); Mammen and Tsybakov (1999); Massart and Nédélec (2006); Tsybakov (2004)), and in active learning

where the learner can adaptively query labels Castro and Nowak (2008); Hanneke (2011); Hanneke and Yang (2015); Koltchinskii (2010); Locatelli et al. (2017, 2018); Minsker (2012); Wang and Singh (2016); Yan et al. (2016). 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—having to do with the uniqueness of the Bayes classes, which seemingly have gone un-noticed—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 $y \in [L]$, where $[L] \doteq \{1, \ldots, L\}$. Two main notions of margin have appeared interchangeably in passive learning in this setting; assume for now, for simplicity that $y = 1$ or 2:

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

for some *margin parameter* $\beta > 0$. Both definitions are termed *Tsybakov's low noise or margin condition* without distinction in the literature (e.g., Castro and Nowak (2008); Minsker (2012) for (i), and Audibert and Tsybakov (2007) for (ii)). 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 classifier is not unique. On the other hand, (i) implies uniqueness (up to measure 0) of the Bayes classes, as seen by letting $\tau \to 0$. As such, (ii) admits more general settings with non-unique Bayes classes, and is thus preferred in the seminal result of Audibert and Tsybakov (2007) on margins in nonparametrics. More generally, in multi-class, the distinction is whether we view the margin as 0 if the Bayes classifier is not unique (corresponding to (i)), or consider the margin between unique values of $\{\eta_y\}_{y\in[L]}$ (corresponding to (ii), and which seems more natural).

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 11). In contrast, as first shown in Minsker (2012), condition (i) always leads to strictly faster rates than passive. While this was shown for binary classification, this is also true in multi-class.

- For general P_X, active learners always gain over the worst case passive rate under either conditions (Theorem 23). As it turns out, the active learning rates are the same under both conditions, matching known rates under (i) in Locatelli et al. (2017) (when $L = 2$).

The comparison of worst-case performance for passive and active learners can be summarized as in Table 1.

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 bulleted regimes: a), an active learner can evenly sample the decision

	Margin Condition (i)	Margin Condition (ii)
Nearly Uniform P_X	$\mathcal{E}_A \ll \mathcal{E}_P$ (Theorem 21)	$\mathcal{E}_A \asymp \mathcal{E}_P$ (Theorem 11)
General P_X	$\mathcal{E}_A \ll \mathcal{E}_P$ (Theorem 23)	

Table 1: Comparison of minimax rates for active learning and passive learning under different regimes. Here, we define $\mathcal{E}_A \doteq \inf_{\hat{h}\text{: active learner}} \sup_{P_{X,Y}} \mathbb{E}_{P_{X,Y}} \mathcal{E}(\hat{h})$ and $\mathcal{E}_P \doteq \inf_{\hat{h}\text{: passive learner}} \sup_{P_{X,Y}} \mathbb{E}_{P_{X,Y}} \mathcal{E}(\hat{h})$ as the active and passive minimax excess risk rates.

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 classifier, having discovered some label with sufficiently low excess error. Under near uniform P_X, the first source of saving a) does not apply 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 where a single label is clearly better, i.e., has margin as in (i), while it cannot effectively identify regions where multiple labels are nearly equivalent (e.g., non-unique Bayes), where it may end up wasting queries while it should preferably give up; in fact we show in our main Theorem 11 that no active learning procedure can automatically decide when to give up on such a priori unknown regions, which forces a label complexity of the same order as in passive learning.

We emphasize that our results do not preclude limited gains in practice under uniform P_X, since minimax rates fail to fully capture constants. In fact, we can refine the margin conditions to account for regions with non-unique Bayes classes—where an active learner may still save over passive, and derive a refined upper-bound, under uniform P_X, that highlight such limited gains over passive learning (Theorem 21). Our upper-bounds require minor modification over past algorithms (e.g., those in Locatelli et al. (2017)), namely additional book-keeping (Section 3.2), and refined correctness arguments required in the multi-class setting.

Main Differences from Binary Case. In a preliminary conference version (Kpotufe et al., 2022), we considered the binary setting ($L = 2$), and showed that active learning has no gain (in terms of minimax rate) when *both* labels are Bayes in parts of the space. However, it leaves open for the multi-class setting whether it suffices that *some but not all* of the classes are Bayes optimal in parts of spaces. In the present work, we will show that any non-uniqueness in parts of the space can already prevent active learning from gaining through a refined lower-bound for the general multi-class setting. Here, we emphasize that a more delicate construction of difficult distributions is required, as the immediate extension for the binary case as in Kpotufe et al. (2022) would require *all* of the L labels to be Bayes equivalent and remains silent about the case more likely in practice where *some but not all* of the classes are Bayes optimal in parts of space. As a contrast, our new construction allows an arbitrary number of Bayes classes in the region where the Bayes classifier is non-unique and is still able to get the same minimax rate as passive learning.

Furthermore, the more flexible new construction allows us to accurately capture rates dependence a notion of *effective number of classes* (c.f., Definition 5 and Remark 22). Such dependence which appears to be missing in the literature (including our preliminary work

Kpotufe et al. (2022)). We are thus able to match upper and lower-bounds, in terms of both sample size n and effective number of classes L^*.

Paper Outline. We start in Section 2 with technical setup, followed by an overview of the main results in Section 3. The proofs of the main theorems are in Section 4, and a simulation study is presented in Section 5. Section 6 concludes the paper with some open questions.

2. Problem Setting

We consider a joint distribution $P_{X,Y}$ on $[0,1]^d \times [L]$, where the short notation $[L] \doteq \{1, \dots, L\}$ for $L \in \mathbb{N}$. Let P_X be the marginal for X with $\text{Support}(P_X) \subset [0,1]^d$. Define the regression function $\eta(x) \doteq (\eta_y(x))_{y \in [L]}$, where $\eta_y(x) \doteq \mathbb{P}(Y = y | X = x)$ for $y \in [L]$.

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

$$\forall x, x' \in \mathit{Support}(P_X) \quad \|\eta(x) - \eta(x')\|_\infty \leq \lambda \|x - x'\|_\infty^\alpha \,,$$

where $\|\cdot\|_\infty$ is the maximum element of a vector.

Remark 2. *For simplicity of presentation, we assume $\alpha \leq 1$. 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. As in prior work Locatelli et al. (2017); Minsker (2012), we assume access to λ or any upper-bound thereof.*

Our adaptive active learning algorithm is built on a dyadic partition of the unit cube.

Definition 3. *For $r = 2^{-k}$, $k \in \mathbb{N}$, define the partition $\mathscr{C}_r$ of $[0,1]^d$ as the collection of hypercubes 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.

Now, we are ready to define the strong density condition. The following definition is adapted from other works on active learning (Locatelli et al., 2017; Minsker, 2012).

Definition 4. *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 strong density condition clearly holds for $P_X = \mathcal{U}[0,1]^d$, or simply has lower-bounded density. Note that it allows a disconnected support $\mathcal{X}$, such as in our lower-bound construction in Section 4.1.1.

Finally, we note that the labels with low probabilities are less relevant to the difficulty of the classification problem. Thus, we introduce the notion of effective classes that filter out these low-probability labels.

Definition 5. *A class $y \in [L]$ is an effective class at x if $\eta_y(x) \geq \max_{l \in [L]} \eta_l(x)/2$. We let $L^*(x)$ denote the number of effective classes at x.*

The number of effective classes differentiates real multi-class situations and the degenerate one where all but two classes have positive probabilities.

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 $x \in [0,1]^d$, 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 6. *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 $h^*(x) \in \operatorname{argmax}_y \mathbb{P}(Y = y|X = x)$. 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 7. *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 8. *$P_{X,Y}$ satisfies the* **Tsybakov's margin condition** *(TMC) with $C_\beta > 0$, $\beta \geq 0$, if:*

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

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 with condition (ii) of the introduction, 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 TMC, no active learner can gain in excess risk rate over their passive counterparts when we assume the strong density condition for P_X. We start our discussion by defining the family of distributions for which active learning has no gains.

Definition 9. *Let $c_d, \lambda, C_\beta > 0$, $\alpha \in [0,1), \beta \geq 0, \gamma \geq 1$, $2 \leq L^* \leq L$, and $\mathcal{P}(c_d, \lambda, \alpha, C_\beta, \beta, L^*, \gamma)$ denote the family of joint distributions $P_{X,Y}$ on $[0,1]^d \times [L]$ where:*

(i) P_X satisfies a strong density condition with c_d;

(ii) the regression function $\eta(x)$ is (λ, α)-Hölder;

(iii) $P_{X,Y}$ *satisfies TMC with parameter* (β, C_β)*;*

(iv) $\forall x \in$ *Support*(P_X)*,* $L^* \le L^*(x) \le \min\{L, (L^*)^\gamma\}$*;*

(v) $\forall x \in$ *Support*(P_X)*, every class* $y \in [L]$ *has a positive probability.*

Remark 10. *Condition (iv) of Definition 9 implies that:* $\forall x \in$ *Support*(P_X)*, the log number of effective classes at* x *satisfies:* $\log L^* \le \log L^*(x) \le \gamma \log L^*$*. As we will see later in Theorem 11, 21, and 23, such condition for changing* $L^*(x)$ *over the space ensures the same rate, despite a difference in the leading constant.*

Theorem 11. *Let* $2 \le L^* \le L$*,* $c_d \in (0,1], \alpha \in (0,1], \lambda, \beta, C_\beta > 0$ *with* $\alpha\beta \le d$*, and* $\mathcal{P} = \mathcal{P}(c_d, \lambda, \alpha, C_\beta, \beta, \gamma)$*. Suppose that* $n \ge \lambda^{(1/\alpha - 2)(2\alpha+d)} (L^*)^{(\alpha+d)/\alpha} \log L^*$*, then* $\exists C_1 > 0$*, independent of* n, L *and* L^**, such that*

$$\inf_{\hat{h}} \sup_{P_{X,Y} \in \mathcal{P}} \mathbb{E}\ \mathcal{E}(\hat{h}_n) \ge C_1 \left(\frac{\log L^*}{L^*}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}} \left(\frac{1}{n}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}}; \tag{2}$$

where the infimum is taken over all (potentially active) learners, and the expectation is taken over the sample distribution, determined by $P_{X,Y}$ *and* $\hat{h}$ *jointly.*

Remark 12. *The proof of Theorem 11 requires a more involved construction of difficult distributions than the binary construction in Kpotufe et al. (2022). Here, we incorporate our novel concept of effective class in the lower-bound, and try to capture the full dependence on* L^* *that may arise. This means that we cannot use constructions with degenerate multi-class situations where only two classes have non-zero probabilities for some* $x \in [0,1]^d$ *as previous work by Reeve and Brown (2017) does. Furthermore, when* $(L^*)^\gamma < L$*, our result shows that active learning has no gain if some but not all classes are Bayes classes in parts of the space with a positive mass.*

Remark 13. *As a corollary to the Theorem 11, since* $\hat{h}$ *is any learner, including the passive ones, the rate in* (2) *is also a lower-bound on passive learning for multi-class classification. As such we know of no other work that so clearly integrates the number of (effective) classes into the learning rate.*

Remark 14. *We will show later in Theorem 21 an upper-bound that matches the rate in* (2) *up to logarithmic terms in* n *and* L^**. With other factors (e.g., margin conditions and smoothness) fixed, the rate in* (2) *gets faster when* L^* *increases (the dependence is of form* $\log L^*/L^*$*). This might be counter-intuitive at first, but can be simply explained by the fact that a larger* L^* *will lead to a smaller variance in the estimation of the regression function* $\eta_y(x)$ *(see more in Remark 19 and Lemma 20).*

The details of the proof of Theorem 11 are presented in Section 4. Our main arguments depart from usual lower-bounds arguments in active learning Castro and Nowak (2008); Minsker (2012); Locatelli et al. (2017). If we followed such information-theoretic constructions where the regions with non-unique Bayes classes are fixed in parts of the space, the learner would know the location of these regions and can achieve a fast rate by simply giving up on sampling in such parts of the space. Instead of working directly on constructing a

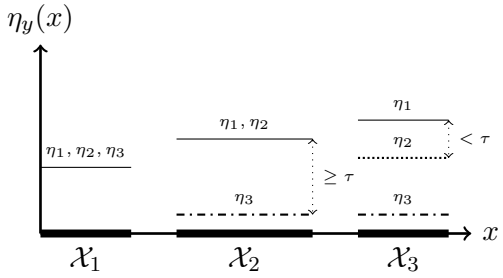

Figure 1: Different types of margin over space. Here, we consider an example with $d = 1$, and $L = 3$. The marginal P_X is supported on $\mathcal{X} = \cup_{i=1}^{3} \mathcal{X}_i$ (the thick lines on the x-axis). We have $\{x : \mathcal{M}(x) \leq \tau\} = \mathcal{X}_3$, while $\{x : \mathcal{M}'(x) \leq \tau\} = \mathcal{X}$. In particular, RMC is satisfied with $\varepsilon_0 = P_X(\mathcal{X}_1 \cup \mathcal{X}_2)$, as $\{\mathcal{M}' = 0\} = \mathcal{X}_1 \cup \mathcal{X}_2$.

suitable subset of $\mathcal{P}(c_d, \lambda, \alpha, C_\beta, \beta, L^*, \gamma)$, we move to a larger class Σ with non-empty intersection Σ_β with $\mathcal{P}(c_d, \lambda, \alpha, C_\beta, \beta, L^*, \gamma)$. We then randomize the construction by putting a suitable measure on Σ that concentrates on Σ_β. Importantly, this measure also encodes regions of $[0,1]^d$ where the Bayes classifier is unique. We show that for any fixed sampling mechanism of $\hat{h}$, the excess error of the classifier $\hat{h}_n$ is lower-bounded as in Theorem 11, in expectation under our measure on Σ, implying the statement of Theorem 11 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.

3.2 Upper-Bounds

Theorem 11 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 class (Theorem 21). Furthermore, under more general P_X, we show in Theorem 23 that a better rate than passive can always be attained even under classical TMC. Both results are established using the same procedure proposed in Section 3.2.1. We start with the following definition.

Definition 15. *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 class is not unique at x.*

Definition 16. *$P_{X,Y}$ is said to satisfy a* **refined margin condition** *(RMC) with $\varepsilon_0 \geq 0, C_\beta > 0$, $\beta' \geq \beta \geq 0$ if:*

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

Remark 17. *The two conditions in Definition 16 differ only when there are non-unique Bayes classes in parts of the space with positive mass, i.e., when* $\mathbb{P}(\mathcal{M}' = 0) \doteq \varepsilon_0 > 0$, *otherwise* $\mathcal{M} = \mathcal{M}'$, P_X*-a.s., and we may choose* $\beta = \beta'$ *and both conditions are equivalent. See the example of Figure 1 for a detailed illustration.*

Next, we define the classes of distributions for upper-bounds.

Definition 18. *Let* $1 \le L^*_{\min} \le L^*_{\max} \le L$, $c_d, \lambda, C_\beta > 0$, $\alpha \in [0,1), \beta' \ge \beta \ge 0, \varepsilon_0 \ge 0$. *We use* $\mathcal{P}_1(\lambda, \alpha, C_\beta, \beta, L^*_{\min}, L^*_{\max})$ *to denote the family of distributions on* $[0,1]^d \times [L]$ *where:*

(i) the regression function $\eta(x)$ *is* (λ, α)*-Hölder;*

(ii) $P_{X,Y}$ *satisfies TMC with parameter* (β, C_β)*;*

(iii) $L^*_{\min} \le L^*(X) \le L^*_{\max}$ *holds* P_X*-a.s.;*

We use $\mathcal{P}_2(c_d, \lambda, \alpha, \varepsilon_0, C_\beta, \beta, \beta', L^*_{\min}, L^*_{\max})$ *to denote the subclass of* $\mathcal{P}_1(\lambda, \alpha, C_\beta, \beta, L^*_{\min}, L^*_{\max})$ *where the following additional conditions are satisfied:*

(iv) P_X *satisfies a strong density condition with* c_d*;*

(v) $P_{X,Y}$ *satisfies RMC with parameter* $(\varepsilon_0, C_\beta, \beta, \beta')$.

Remark 19. *The parameters* $L^*_{\min}$ *and* $L^*_{\max}$ *control the hardness of the problem. First,* $L^*_{\max}$ *is an upper bound for the number of effective classes* $L^*(x)$ *at any* $x \in \mathcal{X}$*, which is also the number of class probabilities* $\eta_y(x)$ *one needs to estimate simultaneously at each* x*. Second, as it turns out the magnitude of each* $\eta_y(x)$ *also affects how well we may estimate it. In particular, note that estimating the regression function* $\eta_y(x)$ *is essentially estimating a Bernoulli distribution with variance* $\eta_y(x)(1-\eta_y(x)) \le O(1/L^*_{\min})$*. Therefore, larger* $L^*_{\min}$ *yields a smaller variance bound, and hence a faster concentration rate for estimating each* $\eta_y(x)$*. More specifically, Lemma 20 below shows that we need a sample size of at least* $3\log(2/\delta)/(\varepsilon^2 L^*_{\min})$ *to guarantee that one can estimate each* $\eta_y(x)$ *within an error of* ε *with* $1-\delta$ *for any* $\varepsilon, \delta > 0$.

Lemma 20. *(Mousavi (2010), Chernoff bound with small deviation) Let* $Z_1, Z_2, \ldots, Z_m$ *be i.i.d random variables taking values 0 or 1, and* $P(Z_i = 1) = p$*. Then, for any* $0 \le \varepsilon \le mp$,

$$P\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - p\right| > \varepsilon\right) \le 2\exp\left(-\frac{m\varepsilon^2}{3p}\right).$$

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, we require an additional booking-keeping procedure. In the binary case, once a label is eliminated in a cell, it is guaranteed that the other label is the Bayes label with high probability. Therefore,

Algorithm 1 Meta Algorithm

Input: n, δ, λ
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 $x \in \mathcal{C} \in \mathscr{C}_{r_0}$

we can quickly stop sampling there by marking this cell "non-active". However, in the multi-class case, even after some labels are eliminated, the remaining labels may still contain some non-Bayes ones. To that end, one needs to keep tracking not only which cells are active, but also a set of candidate labels for each active cell until all but one labels are eliminated.

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

$$n_{r,\alpha} \doteq \min\left\{1, \frac{1}{|\mathcal{L}_{\mathcal{C}}^{\alpha}|} + \tau_{2r,\alpha}\right\} \log\left(\frac{8|\mathcal{L}_{\mathcal{C}}^{\alpha}|}{\delta_0 r^{d+1}}\right) \Big/ 2(\lambda r^{\alpha})^2 \tag{3}$$

where $\tau_{r,\alpha} \doteq 6\lambda r^{\alpha}$. This sample is used to estimate η in each cell $\mathcal{C}$ as

$$\hat{\eta}_y(\mathcal{C}) = n_{r,\alpha}^{-1} \sum_{i=1}^{n_{r,\alpha}} \mathbb{1}(Y_i^{\mathcal{C}} = y), \tag{4}$$

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

$$\hat{\eta}_{(1)}(\mathcal{C}) \doteq \max_{y \in [L]} \hat{\eta}_y(\mathcal{C}) \tag{5}$$

3.2.2 Rates Under Strong Density Condition.

We first consider the excess risk rate for the adaptive algorithm under the strong density condition (Defintion 4).

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 (3))
• Candidate labels: $\mathcal{L}_{\mathcal{C}}^{\alpha} = [L], \forall\, \mathcal{C} \in \mathscr{C}_r$
while $(m_r \le n_0)$ and $(|\mathcal{A}_r| > 0)$ **do**
  // Eliminate bad labels
  for each $\mathcal{C} \in \mathcal{A}_r$ **do**
    Samples $(X_i^{\mathcal{C}}, Y_i^{\mathcal{C}})_{j \le n_{r,\alpha}}$ in cell $\mathcal{C}$ and compute $\{\hat{\eta}_y(\mathcal{C})\}_{y\in[L]}$ by (4)
    Set $\mathcal{L}_{\mathcal{C}}^{\alpha} = \mathcal{L}_{\mathcal{C}}^{\alpha} \setminus \{y : \hat{\eta}_{(1)}(\mathcal{C}) - \hat{\eta}_y(\mathcal{C}) \ge \tau_{r,\alpha}\}$(5)
  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 $|\mathcal{L}_{\mathcal{C}}^{\alpha}| \ge 2\}$
  Set $r = r/2$, $m_{r/2} = m_r + |\mathcal{A}_r| n_{r,\alpha}$ *// Go to next level and 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: $\{\mathcal{L}_{\mathcal{C}}^{\alpha}\}_{\mathcal{C}\in\mathscr{C}_{r_0}}$

Theorem 21. *Let $1 \le L^*_{\min} \le L^*_{\max} \le L$, $c_d, \lambda, C_\beta > 0$, $\alpha \in [0,1), \beta' \ge \beta \ge 0, \varepsilon_0 \ge 0$, and $\alpha\beta' \le d$. Let $\hat{h}_n$ denote the classifier returned by Algorithm 1 with input $n > 0$, $\lambda > 0$ and $0 < \delta < 1$. Suppose $P_{X,Y} \in \mathcal{P}_2(c_d, \lambda, \alpha, \varepsilon_0, C_\beta, \beta, \beta', L^*_{\min}, L^*_{\max})$, then with probability at least $1 - \delta$,*

$$\mathcal{E}\left(\hat{h}_n\right) \le C_2 \left(\varepsilon_0^{\frac{\alpha(\beta+1)}{2\alpha+d}} \left(\frac{\log L^*_{\max}}{L^*_{\min}}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}} \left(\frac{\lambda^{\frac{d}{\alpha}} \log^3(n) \log\left(\frac{8\lambda^2 n}{\delta}\right)}{n}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}} + \left(\frac{\log L^*_{\max}}{L^*_{\min}}\right)^{\frac{\alpha(\beta'+1)}{2\alpha+d-\alpha\beta'}} \left(\frac{\lambda^{\frac{d}{\alpha}} \log^3(n) \log\left(\frac{8\lambda^2 n}{\delta}\right)}{n}\right)^{\frac{\alpha(\beta'+1)}{2\alpha+d-\alpha\beta'}} \right)$$

for some constant $C_2 > 0$ independent of $n, \delta, \lambda, \varepsilon_0, L, L^*_{\min}, L^*_{\max}$.

Remark 22. *The upper-bound shown in Theorem 21 depends on ε_0, and recovers existing bounds (for the binary case) when $\varepsilon_0 = 0$, namely $\tilde{O}(n^{-\alpha(\beta'+1)/(2\alpha+d-\alpha\beta')})$ as shown e.g. in Locatelli et al. (2017); Minsker (2012). This is an improvement over the passive learners and matches the active lower-bound in Minsker (2012) under the strong density condition with $\alpha\beta \le d$. For $\varepsilon_0 \gtrsim \tilde{O}((nL^*_{\min})^{-\alpha\beta'/(2\alpha+d-\alpha\beta')})$, the first term dominates and the upper-bound matches the passive minimax rate. In particular, note that the distribution class $\mathcal{P} \equiv \mathcal{P}(c_d, \lambda, \alpha, C_\beta, \beta, L^*, \gamma)$ (c.f., Definition 9) is a subset of $\mathcal{P}_2 \equiv$*

$\mathcal{P}_2(c_d, \lambda, \alpha, 1, C_\beta, \beta, \infty, L^*, (L^*)^\gamma)$. *The upper-bound here for the class* $\mathcal{P}_2$ *almost matches the one in Theorem 11 for* $\mathcal{P}$*, ignoring the polylogarithmic terms.*

A main novelty in the analysis is to separately consider parts of space with unique Bayes classes, determined by ε_0 and β', and those parts of space where the Bayes classes might not be unique, but which still have margin, determined by β. Furthermore, our consideration of general multi-class, together with non-unique Bayes, brings in a bit of added technicality due largely to additional book-keeping. In particular, while in Locatelli et al. (2017); Minsker (2012), 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.

3.3 Rates for General Densities

For general P_X, on the other hand, Algorithm 2 has an excess risk rate $\widetilde{O}(n^{-(\alpha(\beta+1))/(2\alpha+d)})$, 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 23. *Let* $1 \le L^*_{\min} \le L^*_{\max} \le L$*,* $\lambda, C_\beta > 0$*,* $\alpha \in [0, 1), \beta \ge 0$*, and* $\alpha\beta' \le d$*. Let* $\hat{h}_n$ *denote the classifier returned by Algorithm 1 with input* $n > 0$*,* $\lambda > 0$ *and* $0 < \delta < 1$*. Suppose that* $P_{X,Y} \in \mathcal{P}_1(\lambda, \alpha, C_\beta, \beta, L^*_{\min}, L^*_{\max})$*, then with probability at least* $1 - \delta$*,*

$$\mathcal{E}\left(\hat{h}_n\right) \le C_3 \left(\frac{\log L^*_{\max}}{L^*_{\min}}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}} \left(\frac{\log^3(n)\lambda^{\frac{d}{\alpha}} \log\left(\frac{8\lambda^2 n}{\delta}\right)}{n}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}}$$

for some constant $C_3 > 0$ *that does not depend on* $n, \delta, \lambda, L, L^*_{\min}, L^*_{\max}$*.*

The proof ideas follow similar outlines as for Theorem 21, though more direct.

4. Analysis

4.1 Proof of Theorem 11

4.1.1 Construction of the Difficult Distributions

We operate over a dyadic partition $\mathscr{C}_r$ of the unit cubes $[0, 1]^d$. Let $r = (c_1 \log L^*/(nL^*))^{\frac{1}{2\alpha+d}}$, where $c_1 = \frac{8}{9\lambda^2}$. Without loss of generality, we assume that $-\log_2 r \in \mathbb{N}$ and $2 \le L^* \le L-1$. The case where $L^* = L$ can be done using a similar proving strategy with minor adjustments. 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 Lebesgue 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 $\boldsymbol{\sigma} = (\sigma_{\mathcal{C}})_{\mathcal{C}\in\mathscr{C}_r} \in [L^*-1]^{|\mathscr{C}_r|}$. Define the regression function $\boldsymbol{\eta}_{\boldsymbol{z},\boldsymbol{\sigma}}(x) = \left(\eta^1_{\boldsymbol{z},\boldsymbol{\sigma}}(x), \ldots, \eta^L_{\boldsymbol{z},\boldsymbol{\sigma}}(x)\right)$, with

$$\eta^y_{\boldsymbol{z},\boldsymbol{\sigma}}(x) \doteq \begin{cases} \kappa/L^* + c_\eta \sum_{\mathcal{C}\in\mathscr{C}_r} z_{\mathcal{C}} \left(\mathbb{1}(y=\sigma_{\mathcal{C}}) - \mathbb{1}(y=L^*)\right)\phi_{\mathcal{C}}(x) & \text{for } y \in [L^*]\,, \\ (1-\kappa)/(L-L^*) & \text{for } y \in [L]\backslash[L^*]\,. \end{cases}$$

where $c_\eta = \lambda/8$, $\max\{3L^*/(L+2L^*), 1/2\} \le \kappa < 1$, and

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

Here, we adopt the notation $v_+ \doteq \max(0, v)$, $\forall v \in \mathbb{R}$.

Note that $\{\eta^y_{\boldsymbol{z},\boldsymbol{\sigma}}\}_{y\in[L]}$ indeed defines a proper regression function with only the first L^* classes being effective classes. For all $x \in [0,1]^d$, and $n \ge \lambda^{(1/\alpha-2)(2\alpha+d)} \cdot (L^*)^{(\alpha+d)/\alpha} \cdot \log L^*$, we have $\phi_{\mathcal{C}}(x) \le r^\alpha = \left(8/(9\lambda^2)\right)^\alpha \left(\log L^*/(nL^*)\right)^{\alpha/(2\alpha+d)} \le 1/(\lambda L^*)$, and hence

$$\eta^{L^*}_{\boldsymbol{z},\boldsymbol{\sigma}}(x) \ge \kappa/L^* - c_\eta/(\lambda L^*) > 3\kappa/(4L^*); \quad \eta^{\sigma_{\mathcal{C}}}_{\boldsymbol{z},\boldsymbol{\sigma}}(x) \ge \kappa/L^* + c_\eta/(\lambda L^*) < 5\kappa/(4L^*).$$

Also, we have by the choice of κ that $0 < (1-\kappa)/(L-L^*) \le \kappa/(3L^*)$. Therefore, for any $y^* \in [L^*-1]\backslash\{\sigma_{\mathcal{C}}\}$ and $y \in [L]\backslash[L^*]$,

$$0 < \eta^y_{\boldsymbol{z},\boldsymbol{\sigma}}(x) < \eta^{\sigma_{\mathcal{C}}}_{\boldsymbol{z},\boldsymbol{\sigma}}(x)/2 \le \eta^{L^*}_{\boldsymbol{z},\boldsymbol{\sigma}}(x) \le \eta^{y^*}_{\boldsymbol{z},\boldsymbol{\sigma}}(x) \le \eta^{\sigma_{\mathcal{C}}}_{\boldsymbol{z},\boldsymbol{\sigma}}(x)\,.$$

On the other hand, one can easily verify that $\sum_{y\in[L]} \eta^y_{\boldsymbol{z},\boldsymbol{\sigma}}(x) = 1$ for any $\boldsymbol{z}$,$\boldsymbol{\sigma}$ and x, by noticing $\sum_{y\in[L]} \mathbb{1}(y=\sigma_{\mathcal{C}}) - \mathbb{1}(y=L^*) = 0$.

For each pair $(\boldsymbol{z},\boldsymbol{\sigma})$, one can define a joint probability distribution $P_{\boldsymbol{z},\boldsymbol{\sigma}}$ characterized by P_X and $\mathbb{P}[Y=y|X=x] = \eta^y_{\boldsymbol{z},\boldsymbol{\sigma}}(x)$. See Figure 2 for an example of $P_{\boldsymbol{z},\boldsymbol{\sigma}}$ for $L^*=3$, and $d=2$. 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 $z_{\mathcal{C}} = 1$, there is a small bump in the regression function $\eta^{\sigma_{\mathcal{C}}}_{\boldsymbol{z},\boldsymbol{\sigma}}$ of size $c_\eta r^\alpha$. By construction, $\eta_{\boldsymbol{z},\boldsymbol{\sigma}}$ is always a constant in the intersection of any single cell $\mathcal{C}$ and the support of P_X.

Let $\Sigma \doteq \{P_{\boldsymbol{z},\boldsymbol{\sigma}} : (\boldsymbol{z},\boldsymbol{\sigma}) \in \{0,1\}^{|\mathscr{C}_r|} \times [L]^{|\mathscr{C}_r|}\}$, and $\Sigma_\beta \doteq \{P_{\boldsymbol{z},\boldsymbol{\sigma}} : (\boldsymbol{z},\boldsymbol{\sigma}) \in \Theta_\beta\}$ with

$$\Theta_\beta \doteq \{(\boldsymbol{z},\boldsymbol{\sigma}) : \forall \tau > 0, P_X(\{x : \mathcal{M}_{\boldsymbol{z},\boldsymbol{\sigma}}(x) \le \tau\}) \le C_\beta \tau^\beta\},$$

where $\mathcal{M}_{\boldsymbol{z},\boldsymbol{\sigma}}(x)$ is the margin at x with the regression function being $\boldsymbol{\eta}_{\boldsymbol{z},\boldsymbol{\sigma}}(x)$. For simplicity, we use the short notation,

$$\Xi \doteq (c_d, \lambda, \alpha, C_\beta, \beta, L^*, \gamma)\,,$$

to represent all of the parameters for the distribution class from Definition 9.

4.1.2 Establishing the Lower-bound

In this section, we will show that no active learners $\hat{h}$ has excess risk rate faster than

$$C \cdot \left(\frac{\log L^*}{L^*}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}} \cdot \left(\frac{1}{n}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}},$$

with respect to n and L^*.

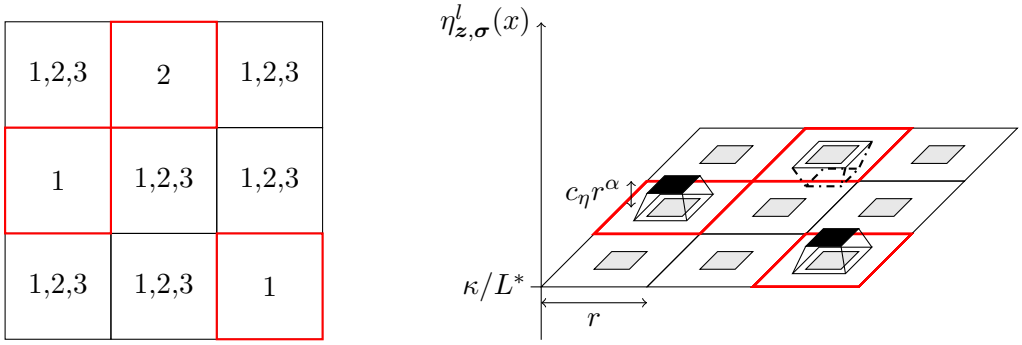

Figure 2: The lower-bound construction of Theorem 11, illustrated in dimension $d = 2$, and number of effective classes $L^* = 3 \leq L$ for simplicity. A key insight is that we require a further layer of randomization than those in usual constructions. Namely, we randomize not only the identity of the Bayes label in fixed regions $\mathcal{C}$ of space ($\mathcal{C}$ is modeled as cells of a partition $\mathscr{C}_r$), but also whether the Bayes is unique in $\mathcal{C}$. **LEFT:** Two sources of randomness $\boldsymbol{z} \in \{0,1\}^{|\mathscr{C}_r|}, \boldsymbol{\sigma} \in [L^*-1]^{|\mathscr{C}_r|}$ are carefully introduced in the construction. The random source $\boldsymbol{z}$ determines the cells in which the Bayes label is unique (highlighted as red squares), $\boldsymbol{\sigma}$ within each of those cells one unique Bayes label uniformly from L^* effective classes (Bayes labels are indicated in each cell). **RIGHT:** A typical regression function $\eta^l_{\boldsymbol{z},\boldsymbol{\sigma}}$ for an effective class $l \in [L^*]$. It takes values from $\{\kappa/L^*, \kappa/L^* \pm c_\eta r^\alpha\}$ and is carefully designed to satisfy the Hölder condition.

First, we show that Σ_β, i.e., the subset of distributions in Σ that satisfies the TMC with parameters (β, C_β), is contained in the distribution class $\mathcal{P}(\Xi)$ as defined in Definition 9. In other words, Σ_β is the non-empty intersection of Σ and $\mathcal{P}(\Xi)$. Therefore, a minimax lower-bound for the class Σ_β is also a minimax lower-bound for the class $\mathcal{P}(\Xi)$ as required in Theorem 11.

Proposition 24. *$\Sigma_\beta \subset \mathcal{P}(\Xi)$. 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 definition. It can be easily verified that strong density condition holds for $c_d = 1$. Clearly, only the first L^* classes are effective classes, and all classes have positive probabilities. It is left to show that $\eta_{\boldsymbol{z},\boldsymbol{\sigma}}$ is (λ, α)-Hölder. In fact, this hold for all $P_{\boldsymbol{z},\boldsymbol{\sigma}} \in \Sigma_\beta$.*

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

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

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, for all $y \in [L]$:

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

Therefore, $\eta_{\boldsymbol{z},\boldsymbol{\sigma}}$ is $(\lambda,\alpha)-$Hölder and we can conclude the proof. ■

Next, we "randomize" the distribution class Σ by letting $\boldsymbol{z} \in \{0,1\}^{|\mathscr{C}_r|} \overset{\text{i.i.d}}{\sim} \text{Ber}(r^{\alpha\beta})$, and $\boldsymbol{\sigma} \in [L^*]^{|\mathscr{C}_r|} \overset{\text{i.i.d}}{\sim} \text{Unif}([L^*])$, $\boldsymbol{z} \perp\!\!\!\perp \boldsymbol{\sigma}$. The following proposition shows that Σ concentrates on Σ_β under such randomness. Consequently, the mean risk (w.r.t. the randomness of $\boldsymbol{z},\boldsymbol{\sigma}$) over the larger distribution class Σ can only be larger than the worst-case risk over the subclass Σ_β by a negligibly small quantity.

Proposition 25. *Let $\hat{h}$ be any active learner. Then,*

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

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 the data distribution $P_{\boldsymbol{z},\boldsymbol{\sigma}}$ and active sampling strategy $\hat{h}$ jointly, and $\mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}}(\cdot)$ is the expectation taken over $P_{\boldsymbol{z},\boldsymbol{\sigma}}$.

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

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

where $c_2 = (C_\beta (tc_\eta)^\beta - 1)^2/3$. 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

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

■

Note that the uncertainties in an active classifier are in both its sampling decision and label prediction. These two types of uncertainties can be de-coupled by considering one single optimal label prediction rule given any sampling decision, if such an optimal rule exists. Formally, we introduce a class of learners with a certain labelling rule in Definition 26 and show that they are indeed optimal in Proposition 27.

Definition 26. *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 $\mathcal{C}$,*

$$\hat{h}_n^*(x) = \mathop{\text{argmax}}_{\sigma\in[L^*]} \prod_{i=1}^{n_{\mathcal{C}}} P_{z_{\mathcal{C}}=1,\sigma_{\mathcal{C}}=\sigma}(Y_i^{\mathcal{C}} | X_i^{\mathcal{C}}),$$

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 27. *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 Let $\hat{h}_n$ be the classifier returned by active learner $\hat{h}$, we can decompose its excess risk as:

$$\mathcal{E}(\hat{h}_n) = \sum_{\mathcal{C}\in\mathscr{C}_r} \mathcal{E}_\mathcal{C}(\hat{h}_n); \tag{6}$$

with $\mathcal{E}_\mathcal{C}(\hat{h}_n) \doteq \int_{\mathcal{C}\cap\{\hat{h}_n\neq\sigma_\mathcal{C}\}} \left[\eta^{\sigma_\mathcal{C}}_{\boldsymbol{z},\boldsymbol{\sigma}}(x) - \eta^{\hat{h}_n(x)}_{\boldsymbol{z},\boldsymbol{\sigma}}(x)\right] dP_X(x)$. Thus, we only need to show that for any $\mathcal{C} \in \mathscr{C}_r$,

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

where $\mathbb{E}_{S|\hat{h}}$ is the expectation taken over the distribution of sample S given active sampling strategy $\hat{h}$ 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 $\hat{h}$. Note that:

$$\mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}|S,\hat{h}} \left[\mathcal{E}_\mathcal{C}(\hat{h}_n)|z_\mathcal{C}=0,\sigma_\mathcal{C}\right] = \left(\frac{\kappa r^d}{L^*} - \frac{1-\kappa}{L-L^*}\right) \mathbb{1}(\hat{h}_n > L^*); \text{ and}$$

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

Clearly, an optimal learner should never predict labels that are larger than or equal to L^*. For those learners with $\hat{h}_n \in [L^*-1]$,

$$\mathop{\mathbb{E}}_{S|\hat{h}} \mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}|S,\hat{h}} \mathcal{E}_\mathcal{C}(\hat{h}_n) = c_\eta r^{\alpha+d} \mathop{\mathbb{E}}_{S|\hat{h}} \left[\sum_{\sigma\in[L^*-1]} \mathbb{1}\left(\hat{h}_n = \sigma\right) \mathbb{P}(z_\mathcal{C}=1, \sigma_\mathcal{C} \neq \sigma|S)\right]$$

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

$$\frac{\mathbb{P}(z_\mathcal{C}=1,\sigma_\mathcal{C}=\sigma|S,\hat{h})}{\mathbb{P}(z_\mathcal{C}=1,\sigma_\mathcal{C}=\sigma'|S,\hat{h})} \geq 1,$$

for any $\sigma' \neq \sigma$. Finally, notice that

$$\frac{\mathbb{P}(z_\mathcal{C}=1,\sigma_\mathcal{C}=\sigma|S,\hat{h})}{\mathbb{P}(z_\mathcal{C}=1,\sigma_\mathcal{C}=\sigma'|S,\hat{h})} = \frac{dP_{S|z_\mathcal{C}=1,\sigma_\mathcal{C}=\sigma,\hat{h}}(S)}{dP_{S|z_\mathcal{C}=1,\sigma_\mathcal{C}=\sigma',\hat{h}}(S)}$$
$$= \frac{\prod_{i=1}^{n_\mathcal{C}} P_{z_\mathcal{C}=1,\sigma_\mathcal{C}=\sigma}(Y_i^\mathcal{C}|X_i^\mathcal{C})}{\prod_{i=1}^{n_\mathcal{C}} P_{z_\mathcal{C}=1,\sigma_\mathcal{C}=\sigma'}(Y_i^\mathcal{C}|X_i^\mathcal{C})}.$$

The last step is clear from the definition

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

where $\boldsymbol{z}_{(\mathcal{C})}, \boldsymbol{\sigma}_{(\mathcal{C})}$ are the $\boldsymbol{z}, \boldsymbol{\sigma}$ vectors with $z_{\mathcal{C}}, \sigma_{\mathcal{C}}$ removed, and $\pi_{\hat{h}}$ is the sampling distribution according to active learner $\hat{h}$. Therefore, the labeling decision of $\hat{h}^*$ minimizes $\mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}} \mathcal{E}_{\mathcal{C}}(\hat{h})$ for each $\mathcal{C}$, hence also minimizes $\mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}} \mathcal{E}(\hat{h})$. ■

By Proposition 27, a conditional Neyman-Pearson learner always has as small an error rate as any other active learner with the same sampling rule. Therefore, we can conclude the proof of Theorem 11 by establishing a lower-bound for the class of all conditional Neyman-Pearson learners.

Proposition 28. *Let $\hat{h}^*$ be any conditional Neyman-Pearson learner. Then,*

$$\mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}^*} \mathcal{E}(\hat{h}_n^*) \geq C_1 \left(\frac{\log L^*}{nL^*}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}}.$$

for some $C_1 > 0$ independent of n, L^*.

Proof *Since $\mathcal{E}_{\mathcal{C}}(\hat{h})$ is a function of $z_{\mathcal{C}}, \sigma_{\mathcal{C}}$ and $S_{\mathcal{C}}$, we have*

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

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 $P_{Y|X,z_{\mathcal{C}},\sigma_{\mathcal{C}}}$:

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

By (6), we have

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

Let $m \doteq 2r^d n \equiv (c_\eta r^\alpha)^{-2} \log L^*/(36L^*)$. By the choice of m, there are at $r^{-d}/2$ cells in $\mathscr{C}_r$ with less than m labeled samples in them. Next, we will establish a lower-bound of the total excess risk in these cells. Note that,

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

where the last inequality follows from Lemma 31. Furthermore,

$$\sum_{\mathcal{C}\in\mathscr{C}_r} \mathop{\mathbb{E}}_{z_\mathcal{C},\sigma_\mathcal{C}} \mathop{\mathbb{P}}_{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}} (z_\mathcal{C}=1; |S_\mathcal{C}|\le m) = \sum_{\mathcal{C}\in\mathscr{C}_r} \mathbb{P}(|S_\mathcal{C}|\le m)\,\mathbb{P}\left(z_\mathcal{C}=1||S_\mathcal{C}|\le m\right))$$
$$\ge \frac{r^{\alpha\beta}}{1+c_4} \sum_{\mathcal{C}\in\mathscr{C}_r} \mathbb{P}(|S_\mathcal{C}|\le m) \ge \frac{r^{\alpha\beta-d}}{2(1+c_4)},$$

where the second last inequality is due to Lemma 32, and the last inequality is from the choice of m and a union bound. Finally,

$$\mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{S|\boldsymbol{z},\boldsymbol{\sigma},\hat{h}} \mathcal{E}(\hat{h}_n^*) = \sum_{\mathcal{C}\in\mathscr{C}_r} \mathop{\mathbb{E}}_{z_\mathcal{C},\sigma_\mathcal{C}} \mathop{\mathbb{E}}_{S_\mathcal{C}|z_\mathcal{C},\sigma_\mathcal{C},\hat{h}} \mathcal{E}_\mathcal{C}(\hat{h}_n^*)$$
$$= (c_3 r^{d+\alpha}) \left(\frac{r^{\alpha\beta-d}}{2(1+c_4)}\right) \ge C_1 \left(\frac{\log L^*}{nL^*}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}},$$

where $C_1 = \frac{c_3(8\lambda^{-2}/9)^{\frac{\alpha(\beta+1)}{2\alpha+d}}}{2(1+c_4)} > 0$. ■

4.1.3 Supporting lemmas

In this section, we present some technical lemmas that have been used in the proof of Proposition 28. The following Lemma 29 shows that the labels in a cell $\mathcal{C}$ are independently and identically distributed given $z_\mathcal{C}, \sigma_\mathcal{C}$, i.e., no information leak among the cells.

Lemma 29. *Conditional on $z_\mathcal{C}$, $\sigma_\mathcal{C}$ and $|S_\mathcal{C}| = n_\mathcal{C}$, $\mathbf{Y}_\mathcal{C} = \{Y_j^\mathcal{C}\}_{j=1}^{n_\mathcal{C}}$ are independently and identically distributed as $P_{Y|X,z_\mathcal{C},\sigma_\mathcal{C}}$.*

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

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

which concludes the proof. ■

The next lemma Lemma 30 is an anti-concentration inequality for multi-classes. It shows that the labeling rule of conditional Neyman-Pearson learners fails to identify the Bayes class in a cell $\mathcal{C}$ with a positive probability bounded away from zero, under an insufficient labelling budget.

Lemma 30 (Anti-concentration). *Let $l \in [L^* - 1]$, $Y_1, \ldots, Y_m$ i.i.d. with*

$$P_l(Y_1 = y) = \begin{cases} \kappa/L^* + c_\eta r^\alpha & \text{if } y = l\,; \\ \kappa/L^* & \text{if } y < L^* \text{ and } y \neq l\,; \\ \kappa/L^* - c_\eta r^\alpha & \text{if } y = L^*\,; \\ (1-\kappa)/(L - L^*) & \text{if } y > L^*\,. \end{cases}$$

where $1/2 \leq \kappa \leq 1$. If $m \leq (c_\eta^{-2} r^{-2\alpha} \log L^*)/(36 L^*)$, then

$$P_l\left(\exists l' \neq l, m_{l'} > m_l\right) \geq c_5,$$

for some absolute constant $c_5 > 0$, where $m_l \doteq \sum_{i=1}^m \mathbb{1}(Y_i = l)$.

Proof *For the case $L^* = 2$, we omit the proof as it follows from the same spirit of binary anti-concentration inequality, e.g., Theorem 2 (ii) of Mousavi (2010). For $L^* > 2$, consider a finite collection of distributions $\{P_0, P_1, \ldots, P_{L^*-1}\}$ where*

$$P_0(Y = y) = \begin{cases} \kappa/L^* & \text{if } y \leq L^*\,; \\ (1-\kappa)/(L - L^*) & \text{if } y > L^*\,. \end{cases}$$

and $\{P_l\}_{l \in [L^*-1]}$ is as defined in the statement. We first find an upper bound for the KL divergence between P_0 and P_l. Let $\varepsilon := c_\eta r^\alpha < \kappa/(8L^*)$ for the choice of large n. By definition of KL divergence,

$$\begin{aligned} KL(P_0|P_l) &= \frac{\kappa}{L^*} \cdot \log\left(\frac{\kappa/L^*}{\kappa/L^* + \varepsilon}\right) + \kappa/L^* \cdot \log\left(\frac{\kappa/L^*}{\kappa/L^* - \varepsilon}\right) \\ &= \frac{\kappa}{L^*} \cdot \left(\log\left(1 + \frac{\varepsilon}{\kappa/L^* - \varepsilon}\right) - \log\left(1 + \frac{\varepsilon}{\kappa/L^*}\right)\right) \\ &= \frac{\kappa}{L^*} \cdot \left(\frac{\varepsilon}{\kappa/L^* - \varepsilon} - \frac{\varepsilon}{\kappa/L^*}\right) + o(\varepsilon^2) \\ &= \frac{\varepsilon^2}{\kappa/L^* - \varepsilon} + o(\varepsilon^2) \\ &\leq 2L^* \varepsilon = 2L^* c_\eta^2 r^{2\alpha}. \end{aligned}$$

Therefore, the product measures P_l^m and P_0^m satisfies $KL(P_0^m|P_l^m) \leq 2mL^* c_\eta^2 r^{2\alpha}$, for $1 \leq l \leq L^* - 1$. Let $\hat{P}_m = P_{\hat{l}_m}$, with $\hat{l}_m = \operatorname{argmax}_l m_l$ being the maximum likelihood estimator of l. Define the discrete metric d, with $d(P, P') = 0$ if $P \equiv P'$ and $d(P, P') = 1$ otherwise. By Theorem 2.5 of Tsybakov (2009), if $m \leq (c_\eta^{-2} r^{-2\alpha} \log L^*)/(36L^*)$, then $2mL^* c_\eta^2 r^{2\alpha} \leq \log L^*/18 \leq \log(L^* - 1)/10$ when $L^* > 2$, and we have

$$\inf_{\hat{P}} \sup_{P \in \{P_1, \ldots, P_{L^*-1}\}} \mathbb{P}(d(P, \hat{P}) = 1) \geq c_5 > 0,$$

for some $c_5 > 0$. By symmetry, $\mathbb{P}(d(P, \hat{P}_m) = 1)$ is the same for all the choice of $P \in \{P_1, \ldots, P_{L^*-1}\}$. Therefore, for any $l \in [L^* - 1]$,

$$\mathbb{P}_{P_l}(d(P_l, \hat{P}_m) = 1) \geq c_5.$$

Hence, the proof is complete by noticing $\{\exists l', m_{l'} > m_l\} = \{d(P_l, \hat{P}_m) = 1\}$. ■

The following Lemma 31 is an immediate corollary of Lemma 30 which provides the excess risk in a cell, under an insufficient labelling budget.

Lemma 31. *Let $n_{\mathcal{C}} \le m = (c_\eta^{-2} r^{-2\alpha} \log L^*)/(36L^*)$ and $\hat{h}^*$ be a conditional Neyman-Pearson learner. Then, in cell $\mathcal{C}$, for any combination of $(z_{\mathcal{C}}, \sigma_{\mathcal{C}})$,*

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

for some $c_3 > 0$.

Proof *When $z_{\mathcal{C}} = 0$, the inequality holds trivially. When $z_{\mathcal{C}} = 1$, let $n_{\mathcal{C}}^\sigma \doteq \sum_{j=1}^{n_{\mathcal{C}}} \mathbb{1}\left(Y_j^{\mathcal{C}} = \sigma\right)$, then*

$$\mathcal{E}_{\mathcal{C}}(\hat{h}_n^*) = c_\eta r^{d+\alpha} \mathbb{1}\left(\sigma_{\mathcal{C}} \neq \mathop{\mathrm{argmax}}_{\sigma \in [L^*-1]} n_{\mathcal{C}}^\sigma\right).$$

Therefore,

$$\begin{aligned}\mathop{\mathbb{E}}_{S_{\mathcal{C}}|z_{\mathcal{C}},\sigma_{\mathcal{C}},\hat{h}}\left[\mathcal{E}_{\mathcal{C}}(\hat{h}_n^*)\middle||S_{\mathcal{C}}| = n_{\mathcal{C}}\right] =& c_\eta r^{d+\alpha} \mathop{\mathbb{P}}_{S_{\mathcal{C}}|z_{\mathcal{C}},\sigma_{\mathcal{C}},\hat{h}}(\exists \sigma, n_{\mathcal{C}}^{\sigma_{\mathcal{C}}} < n_{\mathcal{C}}^\sigma \mid |S_{\mathcal{C}}| = n_{\mathcal{C}}) \\ \ge& c_3 r^{d+\alpha}.\end{aligned}$$

where the last equality holds by Lemma 29 and 30, with $c_3 = c_5 c_\eta$. ■

Finally, Lemma 32 is a technical lemma that shows the distribution of $z_{\mathcal{C}}$, conditioned on small sample size, is not far from the unconditional distribution.

Lemma 32. *Let $S_{\mathcal{C}} = (X_j^{\mathcal{C}}, Y_j^{\mathcal{C}})_j$ be the sample falls in $\mathcal{C}$. Then,*

$$\frac{\int_{|S_{\mathcal{C}}|=n_{\mathcal{C}}} dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=0}(S_{\mathcal{C}})}{\int_{|S_{\mathcal{C}}|=n_{\mathcal{C}}} dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=1}(S_{\mathcal{C}})} \le c_4,$$

for $n_{\mathcal{C}} \le m$ and 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,*

$$\begin{aligned}&dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=0}(S_{\mathcal{C}}) \Big/ \prod_{j=1}^{n_{\mathcal{C}}} dP_{X|\hat{h}}^{\mathcal{C}}\left(X_j^{\mathcal{C}} | (X_i^{\mathcal{C}}, Y_i^{\mathcal{C}})_{i<j}\right) = \left(\frac{\kappa}{L^*}\right)^{\sum_{i=1}^{n_{\mathcal{C}}} \mathbb{1}(Y_i^{\mathcal{C}} \le L^*)} \left(\frac{1-\kappa}{L-L^*}\right)^{\sum_{i=1}^{n_{\mathcal{C}}} \mathbb{1}(Y_i^{\mathcal{C}} > L^*)} \\ &dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=1}(S_{\mathcal{C}}) \Big/ \prod_{j=1}^{n_{\mathcal{C}}} dP_{X|\hat{h}}^{\mathcal{C}}\left(X_j^{\mathcal{C}} | (X_i^{\mathcal{C}}, Y_i^{\mathcal{C}})_{i<j}\right) \\ = \quad & \frac{1}{L^*-1}\left(\sum_{\sigma=1}^{L^*-1} \left(\frac{\kappa}{L^*} + c_\eta r^\alpha\right)^{n_{\mathcal{C}}^\sigma} \left(\frac{\kappa}{L^*} - c_\eta r^\alpha\right)^{n_{\mathcal{C}}^{L^*}} \left(\frac{\kappa}{L^*}\right)^{\sum_{i=1}^{n_{\mathcal{C}}} \mathbb{1}(Y_i^{\mathcal{C}} < L^*, Y_i^{\mathcal{C}} \neq \sigma)} \left(\frac{1-\kappa}{L-L^*}\right)^{\sum_{i=1}^{n_{\mathcal{C}}} \mathbb{1}(Y_i^{\mathcal{C}} > L^*)}\right)\end{aligned}$$

For a fixed $S_{\mathcal{C}}$,

$$\frac{dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=0}(S_{\mathcal{C}})}{dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=1}(S_{\mathcal{C}})} = \frac{L^*-1}{\sum_{\sigma=1}^{L^*-1}\left(1-\frac{L^*c_\eta r^\alpha}{\kappa}\right)^{n_{\mathcal{C}}^{L^*}}\left(1+\frac{L^*c_\eta r^\alpha}{\kappa}\right)^{n_{\mathcal{C}}^{\sigma}}}$$
$$\leq \left[\left(1-\frac{L^*c_\eta r^\alpha}{\kappa}\right)^{n_{\mathcal{C}}^{L^*}}\left(1+\frac{L^*c_\eta r^\alpha}{\kappa}\right)^{\sum_{\sigma=1}^{L^*-1} n_{\mathcal{C}}^{\sigma}/(L^*-1)}\right]^{-1}$$

where the last step is by Jensen's inequality. Therefore,

$$\frac{\int_{|S_{\mathcal{C}}|=n_{\mathcal{C}}} dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=0}(S_{\mathcal{C}})}{\int_{|S_{\mathcal{C}}|=n_{\mathcal{C}}} dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=1}(S_{\mathcal{C}})} \leq 2\cdot\frac{\int_{|S_{\mathcal{C}}|=n_{\mathcal{C}},\sum_{\sigma=1}^{L^*-1} n_{\mathcal{C}}^{\sigma}/(L^*-1)\geq n_{\mathcal{C}}^{L^*}} dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=0}(S_{\mathcal{C}})}{\int_{|S_{\mathcal{C}}|=n_{\mathcal{C}},\sum_{\sigma=1}^{L^*-1} n_{\mathcal{C}}^{\sigma}/(L^*-1)\geq n_{\mathcal{C}}^{L^*}} dP_{S_{\mathcal{C}}|z_{\mathcal{C}}=1}(S_{\mathcal{C}})}$$
$$\leq 2\left(1-\frac{(L^*)^2c_\eta^2 r^{2\alpha}}{\kappa^2}\right)^{-n_{\mathcal{C}}^{L^*}}$$
$$\leq 2\left(1-\frac{L^*}{9m}\right)^{-m/L^*} \leq c_4,$$

for $c_4 = 2\exp(1/9)$. Consequently,

$$\mathbb{P}\left(z_{\mathcal{C}}=1||S_{\mathcal{C}}|\leq m\right) = \frac{\mathbb{P}(z_{\mathcal{C}}=1,|S_{\mathcal{C}}|\leq m)}{\mathbb{P}(|S_{\mathcal{C}}|\leq m)}$$
$$= \frac{\mathbb{P}(z_{\mathcal{C}}=1,|S_{\mathcal{C}}|\leq m)}{\mathbb{P}(z_{\mathcal{C}}=1,|S_{\mathcal{C}}|\leq m)+\mathbb{P}(z_{\mathcal{C}}=0,|S_{\mathcal{C}}|\leq m)}$$
$$= \frac{\mathbb{P}(|S_{\mathcal{C}}|\leq m|z_{\mathcal{C}}=1)\,\mathbb{P}(z_{\mathcal{C}}=1)}{\mathbb{P}(|S_{\mathcal{C}}|\leq m|z_{\mathcal{C}}=1)\,\mathbb{P}(z_{\mathcal{C}}=1)+\mathbb{P}(|S_{\mathcal{C}}|\leq m|z_{\mathcal{C}}=0)\,\mathbb{P}(z_{\mathcal{C}}=0)}$$
$$\geq \frac{r^{\alpha\beta}}{1+c_4}.$$

■

4.2 Proof of Upper-bounds

4.2.1 Technical lemmas

First, we define some quantities and notions that will be used in the lemmas.

Definition 33. *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 $\{(X_j^A, Y_j^A)\}_{j=1}^{n_A}$ in A, an unbiased estimator of $\eta_y(A)$ is

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

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

$$\mathcal{I} \doteq \{2^{-k} : k \in \mathbb{N}\},$$

and

$$\xi_{\mathcal{C},r,\alpha} \doteq \bigcap_{y\in\mathcal{L}_{\mathcal{C}}^{\alpha}} \{|\hat{\eta}_y(\mathcal{C}) - \eta_y(\mathcal{C})| \le \lambda r^\alpha\}$$

where $\mathcal{L}_{\mathcal{C}}^{\alpha}$ is the remaining candidate labels for cell $\mathcal{C}$ **before elimination** as defined in Algorithm 2. The following lemma shows that ξ_α is indeed a high probability event.

Lemma 34. $\mathbb{P}(\xi_\alpha) \ge 1 - \delta_0$.

Proof *First, we show that* $\mathbb{P}(\xi_{\mathcal{C},r,\alpha}^C) \le \delta_0 r^{d+1}$. *By union bound and small deviation version of Chernoff bound (Lemma 20),*

$$\begin{aligned}\mathbb{P}\left(\xi_{\mathcal{C},r,\alpha}^C\right) &\le |\mathcal{L}_{\mathcal{C}}^{\alpha}|\,\mathbb{P}(|\hat{\eta}_y(\mathcal{C}) - \eta_y(\mathcal{C})| > \lambda r^\alpha)\\ &\le |\mathcal{L}_{\mathcal{C}}^{\alpha}|\exp[-2n_{r,\alpha}(\lambda r^\alpha)^2/p_{\max}]\end{aligned}$$

where $p_{\max} \doteq \min\{1, 1/|\mathcal{L}_{\mathcal{C}}^{\alpha}| + \tau_{2r,\alpha}\}$ *is an upper bound for* $\max_{y\in\mathcal{L}_{\mathcal{C}}^{\alpha}} \eta_y(\mathcal{C})$. *Hence, by the choice of* $n_{r,\alpha}$*:*

$$\mathbb{P}\left(\xi_{\mathcal{C},r,\alpha}^C\right) \le 2|\mathcal{L}_{\mathcal{C}}^{\alpha}|\exp(-2n_{r,\alpha}(\lambda r^\alpha)^2/p_{\max}) \le \delta_0 r^{d+1}.$$

Another application of union bound yields $\mathbb{P}(\xi_\alpha) \ge 1 - \sum_{r\in\mathcal{I}} r^{-d}\delta_0 r^{d+1} \ge 1 - \delta_0$. ■

Next, we show some desired properties of Algorithm 2 on the favorable event ξ_α. In particular, Lemma 35 shows that Algorithm 2 never eliminates Bayes classes; Lemma 36 shows that Algorithm 2 predicts only Bayes classes in the area where the soft margin is large enough; Lemma 37 shows that the algorithm will at least reach some certain level $r_{\min}$ of partition.

Lemma 35. *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 eliminates Bayes classes.*

Proof *Let* y^* *be a Bayes class and* $y^* \in \mathcal{L}_{\mathcal{C}}^{\alpha}$ *before elimination. By definition of* ξ_α *and smoothness assumption, we have*

$$\hat{\eta}_{y^*}(\mathcal{C}) \ge \eta_{y^*}(x) - |\eta_{y^*}(x) - \eta_{y^*}(\mathcal{C})| - |\hat{\eta}_{y^*}(\mathcal{C}) - \eta_{y^*}(\mathcal{C})| \ge \eta_{y^*}(x) - 2\lambda r^\alpha\,,$$

Similarly

$$\max_{y\in[L]} \hat{\eta}_y(\mathcal{C}) \le \max_{y\in[L]} \{\eta_y(x) + |\eta_y(\mathcal{C}) - \eta_y(x)| + |\hat{\eta}_y(\mathcal{C}) - \eta_y(\mathcal{C})|\} \le \eta_{y^*}(x) + 2\lambda r^\alpha\,.$$

Therefore, $\hat{\eta}_{(1)}(x) - \hat{\eta}_y(x) \ge 6\lambda r^\alpha$. *Therefore,* $\max_{y\in[L]} \hat{\eta}_y(\mathcal{C}) - \hat{\eta}_{y^*}(\mathcal{C}) < \tau_{r,\alpha}$ *and* y^* *will not be elimiated.* ■

Lemma 36. *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 $\mathcal{M}(x) > \Delta_r$, $\mathcal{L}_\mathcal{C}^\alpha$ contains only Bayes classes.*

Proof *By smoothness assumption, $\eta_{(1)}(\mathcal{C}) - \eta_y(\mathcal{C}) \geq \eta_{(1)}(x) - \eta_y(x) - 2\lambda r^\alpha = 8\lambda r^\alpha$. Let y be any label in $\mathcal{L}_\mathcal{C}^\alpha$ before elimination, 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 37. *On the event ξ_α,*

i) If $P_{X,Y} \in \mathcal{P}_1(\lambda, \alpha, C_\beta, \beta, L^*_{\min}, L^*_{\max})$, then the finest partition Algorithm 2 can reach satisfies

$$r_{\min} \leq \left(\frac{c_6}{n_0\lambda^2 L^*_{\min}} \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)\right)^{1/(2\alpha+d)};$$

for some $c_6 > 0$;

ii) If $P_{X,Y} \in \mathcal{P}_2(c_d, \lambda, \alpha, \varepsilon_0, C_\beta, \beta, \beta', L^*_{\min}, L^*_{\max})$, then

$$r_{\min} \leq \max\left\{\left(\frac{c_7\varepsilon_0 \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0\lambda^2 L^*_{\min}}\right)^{\frac{1}{2\alpha+d}}, \left(\frac{c_7\lambda^{\beta'} \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0\lambda^2 L^*_{\min}}\right)^{\frac{1}{2\alpha+d-\alpha\beta'}}\right\},$$

for some $c_7 > 0$.

Proof *i) Without loss of generality, we assume that n is large enough so that we can at least reach the level r s.t. $1/(2L) \geq \tau_{r,\alpha} = 6\lambda r^\alpha$. Note that the probability gap between the Bayes class and any non-effective label is at least $1/(2L)$, therefore $\mathcal{L}_\mathcal{C}^\alpha$ only contains effective class labels and $L^*_{\min} \leq |\mathcal{L}_\mathcal{C}^\alpha| \leq L^*_{\max}$. Hence,*

$$n_{r,\alpha} \leq \frac{1}{L^*_{\min}} \log\left(\frac{8L^*_{\max}}{\delta_0 r^{d+1}}\right) \Big/ (\lambda r^\alpha)^2.$$

Suppose $r_{\min}$ is the finest partition the algorithm can reach, i.e., the total budget is not sufficient for length $r_{\min}/2$. Then,

$$\begin{aligned} n_0 &\leq \sum_{r\in\mathcal{I}:\ r\geq r_{\min}/2} |\mathcal{A}_r| n_r \\ &\leq \sum_{r\in\mathcal{I}:\ r\geq r_{\min}/2} \frac{r^{-d}}{L^*_{\min}} \log\left(\frac{8L^*_{\max}}{\delta_0 r^{d+1}}\right) \Big/ (\lambda r^\alpha)^2 \\ &\leq \frac{\log\left(8L^*_{\max}/(\delta_0 r^{d+1}_{\min})\right)}{\lambda^2 L^*_{\min}} \sum_{r\in\mathcal{I}:\ r\geq r_{\min}/2} r^{-(2\alpha+d)} \\ &\leq \frac{c_8 \log\left(8L^*_{\max}/(\delta_0 r^{d+1}_{\min})\right)}{\lambda^2 L^*_{\min}} r_{\min}^{-(2\alpha+d)} \end{aligned}$$

where $c_8 > 0$ is independent of $r_{\min}, \delta_0, L^*_{\min}, L^*_{\max}, \lambda$ and n_0, the last inequality is due to the geometric growth of the summands as $r < 1$. We now prove an upper bound for $\log(8L^*_{\max}/(\delta_0 r^{d+1}_{\min}))$. Use the trivial bound

$$\frac{r_{\min}^{-d}}{L^*_{\min}} \cdot \frac{\log\left(8L^*_{\max}/(\delta_0 r^{d+1}_{\min})\right)}{\lambda^2 r^{2\alpha}_{\min}} = n_{r_{\min},\alpha} \leq n_0,$$

and $\frac{r_{\min}^{-d}}{L^*_{\min}} > 1$, $8L^*_{\max}/(\delta_0 r^{d+1}_{\min}) > 8/\delta_0 > 1$, we have $(\lambda^2 r^{2\alpha}_{\min})^{-1} \leq n_0$, which implies $r_{\min} \geq \left(\lambda^2 n_0\right)^{-1/2\alpha}$, and therefore

$$\log(8L^*_{\max}/(\delta_0 r^{d+1}_{\min})) \leq \log\left(\frac{8L^*_{\max}}{\delta_0}\left(\lambda^2 n_0\right)^{(d+1)/2\alpha}\right) \leq \frac{d+1}{2\alpha}\log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right). \tag{7}$$

where the latter step is due to $(d+1)/2\alpha > 1$. With this upper bound (7), we now proceed to upper bound $r_{\min}$. Clearly,

$$n_0 \leq \frac{c_6}{\lambda^2 L^*_{\min}} \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right) r_{\min}^{-(2\alpha+d)}$$

where $c_6 = \frac{c_8(d+1)}{2\alpha}$. Therefore,

$$r_{\min} \leq \left(\frac{c_6}{n_0 \lambda^2 L^*_{\min}} \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)\right)^{1/(2\alpha+d)}.$$

ii) From the strong density condition and Lemma 36, 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 a similar argument as in i), we have

$$\begin{aligned} n_0 &\leq \sum_{r\in\mathcal{I}: r\geq r_{\min}/2} |\mathcal{A}_r| n_r \\ &\leq \sum_{r\in\mathcal{I}: r\geq r_{\min}/2} \frac{\varepsilon_0 + C_\beta (6\lambda r^\alpha)^{\beta'}}{c_d r^d} \cdot \frac{r^{-d}\log(8L^*_{\max}/(\delta_0 r^{d+1}))}{L^*_{\min}\lambda^2 r^{2\alpha}} \\ &\leq \frac{c_7}{\lambda^2 L^*_{\min}} \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right) \max\left\{\varepsilon_0 r_{\min}^{-(2\alpha+d)}, \lambda^{\beta'} r_{\min}^{-(2\alpha+d-\alpha\beta')}\right\}. \end{aligned}$$

where $c_7 > 0$ is independent of $r_{\min}, \delta_0, L^*_{\min}, L^*_{\max}, \lambda$ and n_0. Therefore,

$$r_{\min} \leq \max\left\{ \left(\frac{c_7\varepsilon_0 \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0\lambda^2 L^*_{\min}}\right)^{\frac{1}{2\alpha+d}}, \left(\frac{c_7\lambda^{\beta'}\log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0\lambda^2 L^*_{\min}}\right)^{\frac{1}{2\alpha+d-\alpha\beta'}} \right\}.$$

■

Now we establish an upper-bound for the excess risk rate of the non-adaptive subroutine.

Proposition 38 (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 $\hat{h}_{n_0,\alpha}(x) \in S_{\mathcal{C}}, \forall x \in \mathcal{C} \in \mathscr{C}_{r_0}$.*

i) Suppose that $P_{X,Y} \in \mathcal{P}_1(\lambda, \alpha, C_\beta, \beta, L^*_{\min}, L^*_{\max})$. With probability at least $1-\delta_0$,

$$\mathcal{E}\left(\hat{h}_{n_0,\alpha}\right) \leq C_4 \left(\frac{\lambda^{\frac{d}{\alpha}}\log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0 L^*_{\min}}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}}$$

ii) Suppose that $P_{X,Y} \in \mathcal{P}_2(c_d, \lambda, \alpha, \varepsilon_0, C_\beta, \beta, \beta', L^*_{\min}, L^*_{\max})$. With probability at least $1-\delta_0$,

$$\mathcal{E}\left(\hat{h}_{n_0,\alpha}\right) \leq C_5 \left(\varepsilon_0^{\frac{\alpha(\beta+1)}{2\alpha+d}} \left(\frac{\lambda^{\frac{d}{\alpha}}\log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0 L^*_{\min}}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}} + \left(\frac{\lambda^{\frac{d}{\alpha}}\log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0 L^*_{\min}}\right)^{\frac{\alpha(\beta'+1)}{2\alpha+d-\alpha\beta'}}\right)$$

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

Proof *(Proof of Proposition 38)*

i) On the event ξ_α with probability at least $1-\delta_0$, we have by Part i) of Lemma 37,

$$\Delta_{r_{\min}} = 10\lambda r_{\min}^\alpha \leq 10\lambda \left(\frac{c_6\lambda^{-2}\log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0 L^*_{\min}}\right)^{\frac{\alpha}{2\alpha+d}} \leq 10\left(\frac{c_6\lambda^{\frac{d}{\alpha}}\log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0 L^*_{\min}}\right)^{\frac{\alpha}{2\alpha+d}}.$$

By Lemma 36, the classifier $\hat{h}_{n_0,\alpha}$ makes no error at $\{x : \mathcal{M}(x) > \Delta_{r_{\min}}\}$, and thus

$$\mathcal{E}(\hat{h}_{n_0,\alpha}) \leq \mathbb{P}(\mathcal{M}(X) \leq \Delta_{r_{\min}}) \cdot \Delta_{r_{\min}} \leq C_\beta \Delta_{r_{\min}}^{\beta+1} \leq C_4 \left(\frac{\lambda^{\frac{d}{\alpha}} \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0 L^*_{\min}} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}},$$

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 37,

$$\begin{aligned}\Delta_{r_{\min}} &\leq 10 \max\left\{ \varepsilon_0^{\frac{\alpha}{2\alpha+d}} \left(\frac{c_7 \lambda^{\frac{d}{\alpha}} \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0 L^*_{\min}} \right)^{\frac{\alpha}{2\alpha+d}}, \left(\frac{c_7 \lambda^{\frac{d}{\alpha}} \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0 L^*_{\min}} \right)^{\frac{\alpha}{2\alpha+d-\alpha\beta'}} \right\} \\ &\doteq 10 \max\{Q_1, Q_2\}.\end{aligned}$$

Case 1: $Q_1 \leq Q_2$

Under this case, it is clear that $\varepsilon_0 \leq c_9 \Delta_{r_{\min}}^{\beta'}$ for some $c_9 > 0$. Therefore,

$$\begin{aligned}\mathcal{E}(\hat{h}_{n,\alpha}) &\leq \mathbb{P}(\mathcal{M}(X) \leq \Delta_{r_{\min}}) \Delta_{r_{\min}} \\ &\leq \mathbb{P}(\mathcal{M}'(X) \leq \Delta_{r_{\min}}) \Delta_{r_{\min}} \\ &\leq C_\beta (\varepsilon_0 + \Delta_{r_{\min}}^{\beta'}) \Delta_{r_{\min}} \\ &\leq C_\beta (c_9 + 1) \Delta_{r_{\min}}^{\beta'+1} \\ &\leq C_5' \left(\frac{\lambda^{\frac{d}{\alpha}} \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0 L^*_{\min}} \right)^{\frac{\alpha(\beta'+1)}{2\alpha+d-\alpha\beta'}},\end{aligned}$$

where $C_5' = C_\beta (c_9+1) 10^{\beta'+1} c_7^{\frac{\alpha(\beta'+1)}{2\alpha+d-\alpha\beta'}}$.

Case 2: $Q_1 > Q_2$

Under this case,

$$\mathcal{E}(\hat{h}_{n_0,\alpha}) \leq \mathbb{P}(\mathcal{M}(x) \leq \Delta_{r_{\min}}) \Delta_{r_{\min}} \leq C_\beta \Delta_{r_{\min}}^{\beta+1} \leq C_5'' \varepsilon_0^{\frac{\alpha(\beta+1)}{2\alpha+d}} \left(\frac{\lambda^{\frac{d}{\alpha}} \log\left(\frac{16\lambda^2 n_0}{p_{\min}\delta_0}\right)}{n_0 L^*_{\min}} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}},$$

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

4.2.2 Proof of the main theorems

Proof (Proof of Theorem 21 and 23). Due to their similarity, we only prove Theorem 21, and omit the proof of Theorem 23. 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 $\delta_0 = \delta/\left(\lfloor\log(n)\rfloor^3\right)$ 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 34 and 35, on ξ_{α_i} with probability at least $1-\delta_0$, we have

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

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 $1-\delta$,

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

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 $\mathcal{C} \in \boldsymbol{\mathscr{C}}_{r_0}$. By proposition 38 and the fact that budget for each α_i is $n_0 = \frac{n}{\lfloor\log(n)\rfloor^3}$, we have

$$\mathcal{E}\left(\hat{h}_n\right) \leq C_5 \left(\varepsilon_0^{\frac{\alpha_{i^*}(\beta+1)}{2\alpha_{i^*}+d}} \left(\frac{\lambda^{\frac{d}{\alpha_{i^*}}} \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0 L^*_{\min}} \right)^{\frac{\alpha_{i^*}(\beta+1)}{2\alpha_{i^*}+d}} + \left(\frac{\lambda^{\frac{d}{\alpha_{i^*}}} \log\left(\frac{8L^*_{\max}\lambda^2 n_0}{\delta_0}\right)}{n_0 L^*_{\min}} \right)^{\frac{\alpha_{i^*}(\beta'+1)}{2\alpha_{i^*}+d-\alpha_{i^*}\beta'}} \right)$$

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,

$$\left(\frac{\log^3(n)\lambda^{\frac{d}{\alpha_{i^*}}} \log\left(\left(\frac{8L^*_{\max}\lambda^2 n}{\delta}\right)\right)}{nL^*_{\min}} \right)^{-\frac{\alpha(1+\beta)}{2\alpha+d}+\frac{\alpha_{i^*}(1+\beta)}{2\alpha_{i^*}+d}} \leq 2e^{\frac{1+\beta}{2\log(n)}},$$

$$\left(\frac{\log^3(n)\lambda^{\frac{d}{\alpha_{i^*}}} \log\left(\frac{8L^*_{\max}\lambda^2 n}{\delta}\right)}{nL^*_{\min}} \right)^{-\frac{\alpha(1+\beta')}{2\alpha+d-\alpha\beta'}+\frac{\alpha_{i^*}(1+\beta')}{2\alpha_{i^*}+d-\alpha_{i^*}\beta'}} \leq 2e^{(1+\beta')(2+d)/4}$$

and hence Theorem 21 holds with for $C_2 = 2e^{(1+\beta')(2+d)}C_5$. ■

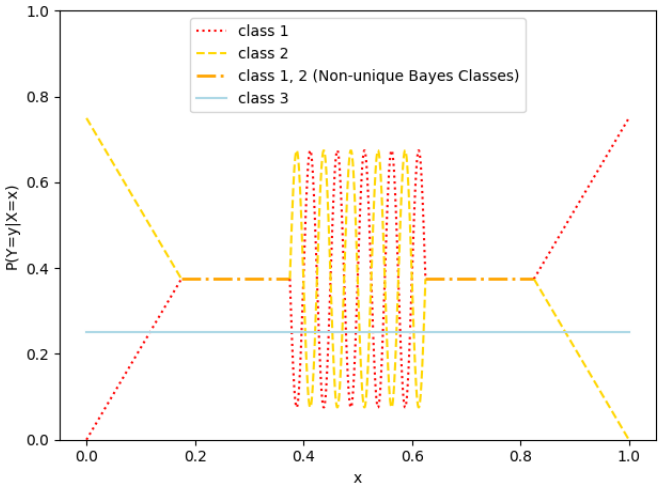

Figure 3: The plot of the regression function. The classification task is relatively easy when x is near the two endpoints of the $[0, 1]$ interval, where the Bayes class is unique with a large margin. The task is more challenging when x is around the center of the interval, where the regression function oscillates rapidly. The Bayes classes are not unique on the two orange dash-dot intervals, where both Class 1 and 2 are Bayes. The total mass of the region where Bayes classes are non-unique is controlled by the parameter ε_0.

5. Experiments

In this section, we demonstrate through a simulation study how non-unique Bayes classes affect the gain in active learning over passive learning.

Data Distribution The joint distribution $P_{X,Y}$ is supported on $[0, 1] \times \{1, 2, 3\}$, characterized by the marginal distribution $P_X \sim \text{Unif}(0, 1)$, and the regression function:

$$\eta_1(x) = \begin{cases} 3x/(3 - 4\varepsilon_0), & 0 \le x \le (3 - 4\varepsilon_0)/8 \\ 3/8, & (3 - 4\varepsilon_0)/8 < x \le 3/8 \\ 3/8, + \sin(40\pi x), & 3/8 < x \le 5/8 \\ 3/8, & 5/8 < x \le (5 + 4\varepsilon_0)/8 \\ 1 + 3(x - 1)/(3 - 4\varepsilon_0), & (5 + 4\varepsilon_0)/8 < x \le 1 \end{cases}$$

$\eta_2(x) = 3/4 - \eta_1(x)$ and $\eta_3(x) = 1/4$. Here, one can easily verify that the parameter $\varepsilon_0 = P_X(\eta_1(X) = \eta_2(X) > \eta_3(X))$ is the mass of region where the Bayes classes are non-unique. See Figure 3 for the plot of the regression function.

Classifiers We compare the performance of the non-adaptive active learner defined as in Algorithm 2 and its passive counterpart. The passive learner samples uniformly on the $[0, 1]$ interval, then partitions the interval into $n^{1/(2\alpha+1)}$ sub-intervals and takes majority votes as its prediction within each partition. The choice of the number of partitions is known to be optimal, see Györfi et al. (2002); Audibert and Tsybakov (2007). Throughout, we assume that both the active and passive learners know the smoothness parameters $\alpha = 1$, $\lambda = 15\pi$.

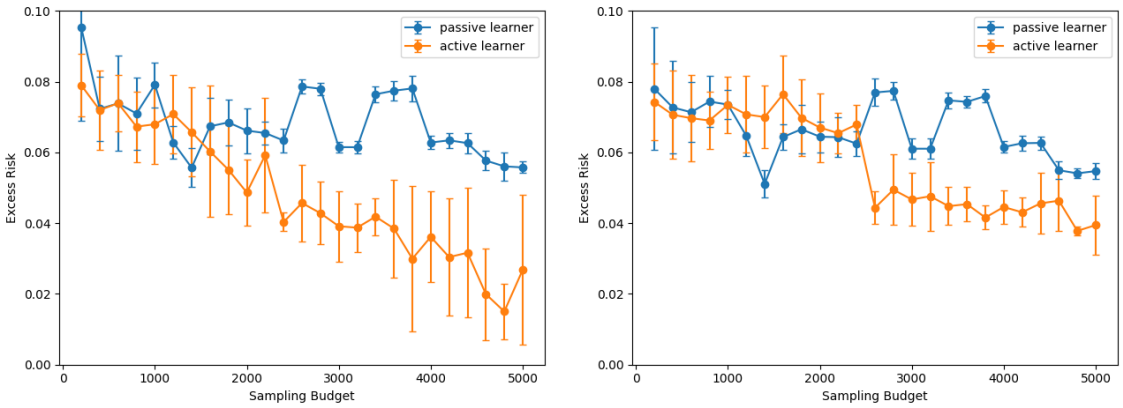

Figure 4: The empirical excess risk of the active learner and passive learner versus the sampling budget, for $\varepsilon_0 = 0$ (left) and $\varepsilon_0 = 0.6$ (right). Each dot represents the mean empirical excess risk over 10 replications, and the error bars stand for the standard deviation.

Evaluation and Results The classifiers are trained with different sampling budgets under multiple levels of ε_0. A test dataset of size $100,000$ is generated and reserved for the evaluation of the classifiers.

Figure 4 shows how the gain of the active learner evolves with changing sampling budgets under two extreme choices of ε_0. When $\varepsilon_0 = 0$, we observe a much faster downward trend for the empirical excess risk of the active learner than that of the passive learner. In this case, the active learner can quickly decide the Bayes class in those regions with large margin and use the majority of the budget where the regression function is highly-oscillated. When $\varepsilon_0 = 0.6$, the gain of the active learner is significantly reduced. The active learner cannot differentiate between the real difficult region and the region where the Bayes classes are non-unique, and hence fails to save budget as efficiently. Nonetheless, there is still limited improvement in the small n regime, which agrees with Theorem 21.

Figure 5 presents the effect of ε_0 on the gain of the active learner on a finer scale. For each different value of ε_0, ranging from 0 to 0.6, we calculate the ratios of the active empirical excess risk to the passive one. We observe an upward trend in the ratio with respect to ε_0. From the level around $\varepsilon_0 = 0.5$ and above, the active learner has almost no advantage over the passive learner.

6. Conclusion

In this paper, we show that classic Tsybakov's margin condition does not guarantee a gain in active learning over passive learning when the marginal distribution P_X is nearly uniform: in particular, there is no gain in the minimax rate whenever the margin condition allows for non-unique Bayes classifiers (up to positive measure). We then propose a refined margin condition that allows for improved active learning rates over passive rates by accounting for the mass of regions with non-unique Bayes classes.

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

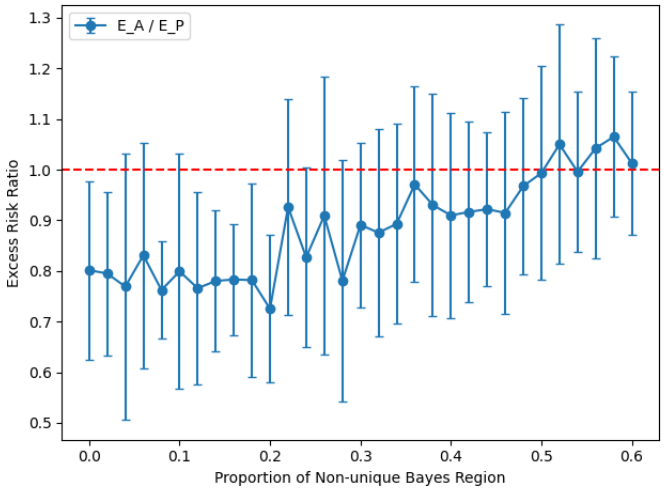

Figure 5: The performance ratio of the active learner and passive learner versus ε_0. Each dot represent the mean performance ratio over 10 replications, and the error bars stand for the standard deviation. The horizontal reference line of level 1 represents the case where the active learner has absolutely no gain over the passive learner.

to gain under the cases with a unique Bayes class throughout the space (Hanneke, 2011; Koltchinskii, 2010; Wang and Singh, 2016). Also, all of the results of this work hold under excess 0-1 risk, different findings might emerge under different performance measures.

Acknowledgments

Samory Kpotufe acknowledges support under a Sloan fellowship, and NSF Grant ID 1739809. He is also a visiting faculty at Google AI Princeton.

References

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

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

László Györfi, Kohler Michael, Adam Krzyżak, and Walk Harro. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

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

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

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

Samory Kpotufe, Gan Yuan, and Yunfan Zhao. Nuances in margin conditions determine gains in active learning. In *International Conference on Artificial Intelligence and Statistics*, pages 8112–8126. PMLR, 2022.

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

Andrea Locatelli, Alexandra Carpentier, and Samory Kpotufe. An adaptive strategy for active learning with smooth decision boundary. In *Algorithmic Learning Theory*, pages 547–571. PMLR, 2018.

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

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

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

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

H. W. J. Reeve and G. Brown. Minimax rates for cost-sensitive learning on manifolds with approximate nearest neighbours. *Proceedings of Machine Learning Research*, 1:1–45, 2017.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

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

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

Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning from imperfect labelers. *Advances in Neural Information Processing Systems*, 29:2128–2136, 2016.