Journal of Machine Learning Research 12 (2011) 2269-2292 Submitted 6/10; Revised 4/11; Published 7/11

Smoothness, Disagreement Coefficient, and the Label Complexity of Agnostic Active Learning

Liwei Wang WANGLW@CIS.PKU.EDU.CN

Key Laboratory of Machine Perception, MOE
School of Electronics Engineering and Computer Science
Peking University
Beijing, 100871, P.R.China

Editor: Rocco Servedio

Abstract

We study pool-based active learning in the presence of noise, that is, the agnostic setting. It is known that the effectiveness of agnostic active learning depends on the learning problem and the hypothesis space. Although there are many cases on which active learning is very useful, it is also easy to construct examples that no active learning algorithm can have an advantage. Previous works have shown that the label complexity of active learning relies on the *disagreement coefficient* which often characterizes the intrinsic difficulty of the learning problem. In this paper, we study the disagreement coefficient of classification problems for which the classification boundary is smooth and the data distribution has a density that can be bounded by a smooth function. We prove upper and lower bounds for the disagreement coefficients of both finitely and infinitely smooth problems. Combining with existing results, it shows that active learning is superior to passive supervised learning for smooth problems.

Keywords: active learning, disagreement coefficient, label complexity, smooth function

1. Introduction

Active learning addresses the problem that the algorithm is given a pool of unlabeled data drawn i.i.d. from some underlying distribution; the algorithm can then pay for the label of any example in the pool. The goal is to learn an accurate classifier by requesting as few labels as possible. This is in contrast with the standard passive supervised learning, where the labeled examples are chosen randomly.

The simplest example that demonstrates the potential of active learning is to learn the optimal threshold on an interval. Suppose the instances are uniformly distributed on $[0,1]$, and there exists a perfect threshold separating the two classes (i.e., there is no noise), then binary search needs $O(\log\frac{1}{\varepsilon})$ labels to learn an ε-accurate classifier, while passive learning requires $O(\frac{1}{\varepsilon})$ labels. Another encouraging example is to learn homogeneous linear separators. If the data are distributed on the unit sphere of $\mathbb{R}^d$, and the distribution has a density function upper and lower bounded by λ and $1/\lambda$ respectively, where λ is some constant, then active learning can still give exponential savings in the label complexity (Dasgupta, 2005).

However, there are also very simple problems that active learning does not help. Suppose again that the instances are uniformly distributed on $[0,1]$. But this time the positive class could be any interval on $[0,1]$. In this case, for any active learning algorithm there exists a distribution (i.e., a target classifier) such that the algorithm needs $\Omega(\frac{1}{\varepsilon})$ label requests to learn an ε-accurate classifier (Dasgupta, 2005). Thus there is no improvement over passive learning in the minimax sense. All above are realizable problems. Of more interest and more realistic is the agnostic setting, where the best classifier in the hypothesis space has a non-zero error ν. For agnostic active learning, there is no active learning algorithm that can always reduce label requests due to a lower bound $\Omega(\frac{\nu^2}{\varepsilon^2})$ for the label complexity (Kääriäinen, 2006).

Previous results have shown that whether active learning helps relies crucially on the *disagreement coefficient* of the learning problem (Hanneke, 2007). The disagreement coefficient depends on the distribution of the instance-label pairs and the hypothesis space and often describes the intrinsic difficulty of the active learning problem. In particular, it has been shown that the label complexity of two important agnostic active learning algorithms A^2 (Balcan et al., 2006) and the one due to Dasgupta et al. (2007) (will be referred to as DHM) are characterized by the disagreement coefficient. If the disagreement coefficient is small, active learning usually has smaller label complexity than passive learning.

In this paper, we study the disagreement coefficient for smooth problems. Specifically we analyze the disagreement coefficient for learning problems whose classification boundaries are smooth. Such problems are often referred to as the boundary fragment class (van der Vaart and Wellner, 1996). Under some mild assumptions on the distribution, we show that the magnitude of the disagreement coefficient depends on the order of smoothness. For finite order smoothness, it is polynomially smaller than the largest possible value, and exponentially smaller for infinite smoothness. Combining with known upper bounds on the label complexity in terms of disagreement coefficient, we give sufficient condition under which active learning is strictly superior to passive learning.

1.1 Related Works

Our work is closely related to Castro and Nowak (2008) which proved label complexity bounds for problems with smooth classification boundary under Tsybakov's noise condition (Tsybakov, 2004). Please see Section 3.3 for a detailed discussion on this work.

Another related work is due to Friedman (2009). He introduced a different notion of smoothness. In particular, he considered smooth problems whose hypothesis space is a finite dimensional parametric space (and therefore has finite VC dimension). He gave conditions under which the disagreement coefficient is always bounded from above by a constant. In contrast, the hypothesis space (the boundary fragment class) studied in our work is a nonparametric class and is more expressive than VC classes.

2. Background

Let $\mathcal{X}$ be an instance space, $\mathcal{D}$ a distribution over $\mathcal{X} \times \{-1,1\}$. Let $\mathcal{H}$ be the hypothesis space, a set of classifiers from $\mathcal{X}$ to $\{-1,1\}$. Denote $\mathcal{D}_X$ the marginal of $\mathcal{D}$ over $\mathcal{X}$. In our active learning model, the algorithm has access to a pool of unlabeled examples from $\mathcal{D}_X$. For any unlabeled point x, the algorithm can ask for its label y, which is generated from the conditional distribution at x. The error of a hypothesis h according to $\mathcal{D}$ is $er_{\mathcal{D}}(h) = \Pr_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$. The empirical error on a sample $\mathcal{S}$ of size n is $er_{\mathcal{S}}(h) = \frac{1}{n}\sum_{(x,y)\in\mathcal{S}} \mathbb{I}[h(x) \neq y]$, where $\mathbb{I}$ is the indicator function. We use h^* denote the best classifier in $\mathcal{H}$. That is, $h^* = \arg\min_{h\in\mathcal{H}} er_{\mathcal{D}}(h)$. Let $\nu = er_{\mathcal{D}}(h^*)$. Our goal is to learn a $\hat{h} \in \mathcal{H}$ with error rate at most $\nu + \varepsilon$, where ε is the desired accuracy.

```
Input: unlabeled data pool $(x_1, x_2, \dots, x_m)$ i.i.d. from $\mathcal{D}_X$, hypothesis space $\mathcal{H}$;
Initially: $V \leftarrow \mathcal{H}$, $R \leftarrow DIS(V)$, $Q \leftarrow \emptyset$;
for $t = 1, 2, \dots, m$ do
    if $\Pr(DIS(V)) \leq \frac{1}{2}\Pr(R)$ then
        $R \leftarrow DIS(V)$; $Q \leftarrow \emptyset$;
    end
    Find a new data $x_i$ from the data pool with $x_i$ in $R$;
    Request the label $y_i$ of $x_i$, and let $Q \leftarrow Q \cup \{(x_i, y_i)\}$;
    $V \leftarrow \{h \in V : LB(h, Q, \delta/m) \leq \min_{h' \in V} UB(h', Q, \delta/m)\}$;
    $h_t \leftarrow \arg\min_{h \in V} er_Q(h)$;
    $\beta_t \leftarrow (UB(h_t, Q, \delta/m) - LB(h_t, Q, \delta/m)) \Pr(R)$ ;
end
Return $\hat{h} = h_j$, where $j = \arg\min_{t \in \{1,2,\dots,m\}} \beta_t$.
```

Algorithm 1: The A^2 algorithm

A^2 (Balcan et al., 2006) is the first rigorous agnostic active learning algorithm. It can be viewed as a robust version of the active learning algorithm due to Cohn et al. (1994) for the realizable setting. A description of the algorithm is given in Algorithm 1. It was shown that A^2 is never much worse than passive learning in terms of the label complexity. The key observation that A^2 can be superior to passive learning is that, since our goal is to choose an $\hat{h}$ such that $er_{\mathcal{D}}(\hat{h}) \leq er_{\mathcal{D}}(h^*) + \varepsilon$, we only need to *compare* the errors of hypotheses. Therefore we can just request labels of those x on which the hypotheses under consideration have disagreement.

To do this, the algorithm keeps track of two spaces. One is the current version space V, consisting of hypotheses that with statistical confidence are not too bad compared to h^*; the other is the region of disagreement $DIS(V)$, which is the set of all $x \in \mathcal{X}$ for which there are hypotheses in V that disagree on x. Formally, for any subset $V \subset \mathcal{H}$,

$$DIS(V) = \{x \in \mathcal{X} : \exists h, h' \in V, \;\; h(x) \neq h'(x)\}.$$

To achieve the statistical guarantee that the version space V contains only good hypotheses, the algorithm must be provided with a uniform convergence bound over the hypothesis space. That is, with probability at least $1 - \delta$ over the draw of sample $\mathcal{S}$ according to $\mathcal{D}$ conditioned on $DIS(V)$ for any version spaces V,

$$LB(\mathcal{S}, h, \delta) \leq er_{\mathcal{D}_{|V}}(h) \leq UB(\mathcal{S}, h, \delta),$$

hold simultaneously for all $h \in \mathcal{H}$, where the lower bound $LB(\mathcal{S}, h, \delta)$ and upper bound $UB(\mathcal{S}, h, \delta)$ can be computed from the empirical error $er_{\mathcal{S}}(h)$. Here $\mathcal{D}_{|V}$ is the distribution of $\mathcal{D}$ conditioned on $DIS(V)$. If $\mathcal{H}$ has finite VC dimension $VC(\mathcal{H})$, then $er_{\mathcal{S}}(h) \pm O(\frac{VC(\mathcal{H})}{n})^{-1/2}$ are upper and lower bounds of $er_{\mathcal{D}_{|V}}(h)$ respectively.

We will denote the volume of $DIS(V)$ by $\Delta(V) = \Pr_{X \sim \mathcal{D}_X}(X \in DIS(V))$. Requesting labels of the instances from $DIS(V)$ rather than from the whole space $\mathcal{X}$ allows A^2 require fewer labels than passive learning. Hence the key issue is how fast $\Delta(V)$ reduces. This process, and in turn the label complexity of A^2, are nicely characterized by the disagreement coefficient θ introduced in Hanneke (2007).

```
Input: unlabeled data pool $(x_1, x_2, \ldots, x_m)$ i.i.d. from $\mathcal{D}_X$, hypothesis space $\mathcal{H}$;
Initially: $\mathcal{G}_0 \leftarrow \emptyset$, $\mathcal{T}_0 \leftarrow \emptyset$;
for $t = 1, 2, \ldots, m$ do
    For each $\hat{y} \in \{-1, 1\}$, $h_{\hat{y}} \leftarrow \mathrm{LEARN}_{\mathcal{H}}(\mathcal{G}_{t-1} \cup \{(x_t, \hat{y})\}, \mathcal{T}_{t-1})$;
    if $er_{\mathcal{G}_{t-1} \cup \mathcal{T}_{t-1}}(h_{-\hat{y}}) - er_{\mathcal{G}_{t-1} \cup \mathcal{T}_{t-1}}(h_{\hat{y}}) > \Delta_{t-1}$ for some $\hat{y} \in \{-1, 1\}$ then
        $\mathcal{G}_t \leftarrow \mathcal{G}_{t-1} \cup \{(x_t, \hat{y})\}$; $\mathcal{T}_t \leftarrow \mathcal{T}_{t-1}$;
    end
    else
        Request the true label $y_t$ of $x_t$; $\mathcal{G}_t \leftarrow \mathcal{G}_{t-1}$; $\mathcal{T}_t \leftarrow \mathcal{T}_{t-1} \cup \{(x_t, y_t)\}$;
    end
end
Return $h = \mathrm{LEARN}_{\mathcal{H}}(\mathcal{G}_m, \mathcal{T}_m)$.
```

Algorithm 2: The DHM algorithm

Definition 1 *Let $\rho(\cdot,\cdot)$ be the pseudo-metric on a hypothesis space $\mathcal{H}$ induced by $\mathcal{D}_X$. That is, for $h, h' \in \mathcal{H}$, $\rho(h, h') = \Pr_{X \sim \mathcal{D}_X}(h(X) \neq h'(X))$. Let $B(h, r) = \{h' \in \mathcal{H} \colon \rho(h, h') \leq r\}$. The disagreement coefficient $\theta(\varepsilon)$ is*

$$\theta(\varepsilon) = \sup_{r \geq \varepsilon} \frac{\Delta(B(h^*, r))}{r} = \sup_{r \geq \varepsilon} \frac{\Pr_{X \sim \mathcal{D}_X}(X \in DIS(B(h^*, r)))}{r},$$

where $h^* = \arg\min_{h \in \mathcal{H}} er_{\mathcal{D}}(h)$.

Note that θ depends on $\mathcal{H}$ and $\mathcal{D}$, and $1 \leq \theta(\varepsilon) \leq \frac{1}{\varepsilon}$.[1] The following is an upper bound of the label complexity of A^2 in terms of the disagreement coefficient $\theta(\varepsilon)$ (Hanneke, 2007).

Theorem 2 *Suppose that $\mathcal{H}$ has finite VC dimension $VC(\mathcal{H})$. Then using the definitions given above, the label complexity of A^2 is*

$$O\left((\theta(\nu + \varepsilon))^2 \left(\frac{\nu^2}{\varepsilon^2} + 1\right) \mathrm{polylog}\left(\frac{1}{\varepsilon}\right) \log\left(\frac{1}{\delta}\right) VC(\mathcal{H})\right). \tag{1}$$

In addition, Hanneke (2007) showed that $\tilde{\Omega}(\theta^2 \log \frac{1}{\delta})$ is a lower bound for the A^2 algorithm of any problem with $\nu = 0$, where in $\tilde{\Omega}$ we hide the logrithm terms.

Another important agnostic active learning algorithm is DHM. (Algorithm 2 gives a formal description of the algorithm.) DHM reduces active learning to a series of *constrained* supervised learning. The key idea of the algorithm is that each time we encounter a new unlabeled data x, we test if we can guess the label of x with high confidence, using the information obtained so far. If we can, we put the data and the confidently guessed label $(x, \hat{y})$ into the *guessed* set $\mathcal{G}$; otherwise, we request the true label y of x and put (x, y) into the *true* set $\mathcal{T}$. The criterion of whether we can guess the label of x confidently is as follows. For each $\tilde{y} \in \{-1, +1\}$, we learn a classifier $h_{\tilde{y}} \in \mathcal{H}$ such that $h_{\tilde{y}}(x) = \tilde{y}$, and $h_{\tilde{y}}$ is consistent with all $(x, \hat{y}) \in \mathcal{G}$ and has minimal error on $\mathcal{T}$. (This is the subroutine LEARN in Algorithm 2.) If for some $\tilde{y} \in \{-1, +1\}$ the error rate of $h_{\tilde{y}}$ is smaller than

[1] Here we only consider the nontrivial case that $\Delta(B(h^*, r)) \geq r$ for all r. This condition is satisfied by the smooth problems studied in this paper.

that of $h_{-\tilde{y}}$ by a threshold Δ_t (t is the number of unlabeled data encountered so far), then we guess that $\hat{y} = \tilde{y}$ confidently.

The algorithm DHM relies crucially on a good choice of the threshold function Δ_t. If $\mathcal{H}$ has finite VC dimension $VC(\mathcal{H})$, Dasgupta et al. (2007) suggested to choose Δ_t based on the normalized uniform convergence bound on $\mathcal{H}$ (Vapnik, 1998). They also showed that DHM is never much worse than passive learning and it has label complexity[2]

$$\tilde{O}\left(\theta(\nu+\varepsilon)\left(1+\frac{\nu^2}{\varepsilon^2}\right)\text{polylog}\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\delta}\right)VC(\mathcal{H})\right). \tag{2}$$

From (1) and (2) it can be seen that if $\varepsilon > \nu$, the term $\frac{\nu^2}{\varepsilon^2}$ is upper bounded by 1 and the label complexity of the active learning algorithms crucially depends on the disagreement coefficient θ. However, the asymptotic label complexity as ε tends to 0 (assuming $\nu > 0$) can at best only be upper bounded by $O\left(\frac{\nu^2}{\varepsilon^2}\right)$. In fact, this bound cannot be improved: it is known that given some hypothesis space $\mathcal{H}$, for every active learning algorithm A, there is a learning problem (to be concrete, a h^*) such that the label complexity of A is at least $\Omega(\frac{\nu^2}{\varepsilon^2})$ (Kääriäinen, 2006). Thus $\Omega(\frac{\nu^2}{\varepsilon^2})$ is a minimax lower bound of the label complexity of agnostic active learning algorithms.

Although no active learning algorithm is superior to passive learning in all agnostic settings, it turns out that if the disagreement coefficient is small, active learning does always help under a finer parametrization of the noise distribution, known as Tsybakov's noise condition (Tsybakov, 2004).

Definition 3 *Let $\eta(x) = \Pr(Y = 1|X = x)$. We say that the distribution of the learning problem has noise exponent $\kappa = \frac{a+1}{a}$ ($\kappa \geq 1$) if there exists constant $c > 0$ such that*

$$\Pr\left(\left|\eta(X) - \frac{1}{2}\right| \leq t\right) \leq ct^a, \qquad 0 < a \leq +\infty$$

for all $0 < t \leq t_0$ for some constant t_0.

Tsybakov's noise condition characterizes the behavior of $\eta(x)$ when x crosses the class boundary. If $\kappa = 1$, $\eta(x)$ has a jump from $\frac{1}{2} - t_0$ to $\frac{1}{2} + t_0$. The larger the κ, the more "flat" $\eta(x)$ is.

Under Tsybakov's noise condition, Hanneke (2009, 2011) proved that a variant of the DHM algorithm (by choosing the threshold Δ_t based on local Rademacher complexity (Koltchinskii, 2006)) has the following asymptotic label complexity.

Theorem 4 *Suppose that the learning problem satisfies the Tsybakov's noise condition with noise exponent κ. Assume that the hypothesis space $\mathcal{H}$ and the marginal distribution $\mathcal{D}_X$ satisfies that the entropy with bracketing $H_{[\,]}(\varepsilon, \mathcal{H}, L_2(\mathcal{D}_X)) = O\left(\left(\frac{1}{\varepsilon}\right)^{2p}\right)$ for some $0 < p < 1$. If the Bayes classifier $h_B^* \in \mathcal{H}$, then the label complexity of DHM is*

$$O\left(\theta(\varepsilon_0)\left(\frac{1}{\varepsilon}\right)^{2-\frac{2-p}{\kappa}}\left(\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right)\right), \tag{3}$$

where ε_0 depends on ε, κ, p, δ and the learning problem. In particular, setting $\varepsilon_0 = \varepsilon^{\frac{1}{\kappa}}$ the theorem holds.

[2] Here in $\tilde{O}$ we hide terms like $\log\log(\frac{1}{\varepsilon})$ and $\log\log(\frac{1}{\delta})$.

Inspired by this result, Koltchinskii (2010) further proved that under similar conditions a variant of the A^2 algorithm has label complexity

$$O\left(\theta(\varepsilon^{\frac{1}{\kappa}})\left(\left(\frac{1}{\varepsilon}\right)^{2-\frac{2-p}{\kappa}}+\left(\frac{1}{\varepsilon}\right)^{2-\frac{2}{\kappa}}\left(\log\frac{1}{\delta}+\log\log\frac{1}{\varepsilon}\right)\right)\right). \tag{4}$$

Note that in the last formula $\left(\frac{1}{\varepsilon}\right)^{2-\frac{2-p}{\kappa}}$ dominates over $\left(\frac{1}{\varepsilon}\right)^{2-\frac{2}{\kappa}}$ as $\varepsilon \to 0$ if $p > 0$.

If the hypothesis space $\mathcal{H}$ has finite VC dimension, the entropy with bracketing is

$$H_{[\,]}(\varepsilon, \mathcal{H}, L_2(\mathcal{D}_X)) = O\left(\log\frac{1}{\varepsilon}\right),$$

smaller than $\left(\frac{1}{\varepsilon}\right)^{2p}$ for any $p > 0$. In this case, it can be shown that the above label complexity bounds still hold by just putting $p = 0$ into them.

In contrast, the sample complexity for passive learning under the same conditions is known to be (Tsybakov, 2004)

$$O\left(\left(\frac{1}{\varepsilon}\right)^{2-\frac{1-p}{\kappa}}\left(\log\frac{1}{\delta}+\log\log\frac{1}{\varepsilon}\right)\right), \tag{5}$$

and it is also a minimax lower bound. Comparing (3), (4) and (5) one can see that whether active learning is strictly superior to passive learning entirely depends on how small the disagreement coefficient $\theta(\varepsilon)$ is.

One shortcoming of A^2 and DHM is that they are computationally expensive. This is partially because that they need to minimize the 0-1 loss and need to maintain the version space. Beygelzimer et al. (2009) proposed an importance weighting procedure IWAL which, during learning, minimize a convex surrogate loss and therefore avoid 0-1 minimization. Furthermore, Beygelzimer et al. (2010) developed an active learning algorithm which does not need to keep the version space and therefore is computationally efficient. There are also upper bounds on the label complexity of these two algorithms in terms of the disagreement coefficient.

Finally, for a comprehensive survey of the theoretical research on active learning, please see the excellent tutorial (Dasgupta and Langford, 2009).

3. Main Results

As described in the previous section, whether active learning helps largely depends on the disagreement coefficient which often characterizes the intrinsic difficulty of the learning problem using a given hypothesis space. So it is important to understand if the disagreement coefficient is small for learning problems with practical and theoretical interests. In this section we give bounds on the disagreement coefficient for problems that have smooth classification boundaries, under additional assumptions on the distribution. Such smooth problems are often referred to as boundary fragment class and has been extensively studied in passive learning and especially in empirical processes.

In Section 3.1 we give formal definitions of the smooth problems. Section 3.2 contains the main results, where we establish upper and lower bounds for the disagreement coefficient of smooth problems. In Section 3.3 we provide some discussions on some closely related works.

3.1 Smoothness

Let f be a function defined on $\Omega \subset \mathbb{R}^d$ and $\alpha > 0$ be a real number. Let $\underline{\alpha}$ be the largest integer strictly smaller than α. (Hence $\underline{\alpha} = \alpha - 1$ when α is an integer.) For any vector $\mathbf{k} = (k_1, \cdots, k_d)$ of d nonnegative integers, let $|\mathbf{k}| = \sum_{i=1}^d k_i$. Let

$$D^{\mathbf{k}} = \frac{\partial^{|\mathbf{k}|}}{\partial^{k_1} x^1 \cdots \partial^{k_d} x^d},$$

be the differential operator. Define the α-norm as (van der Vaart and Wellner, 1996)

$$\|f\|_\alpha := \max_{|\mathbf{k}| \le \underline{\alpha}} \sup_x |D^{\mathbf{k}} f(x)| + \max_{|\mathbf{k}| = \underline{\alpha}} \sup_{x,x'} \frac{|D^{\mathbf{k}} f(x) - D^{\mathbf{k}} f(x')|}{\|x - x'\|^{\alpha - \underline{\alpha}}},$$

where the suprema are taken over all x, x' over Ω with $x \neq x'$.

Definition 5 ***(Finite Smooth Functions)*** *A function f is said to be αth order smooth with respect to a constant C, if $\|f\|_\alpha \le C$. The set of αth order smooth functions is defined as*

$$F_C^\alpha := \{f : \|f\|_\alpha \le C\}.$$

Thus αth order smooth functions have uniformly bounded partial derivatives up to order $\underline{\alpha}$, and the $\underline{\alpha}$th order partial derivatives are Hölder continuous. As a special case, note that if f has continuous partial derivatives upper bounded by C up to order m, where m is any positive integer, then $f \in F_C^m$. Also, if $0 < \beta < \alpha$, then $f \in F_C^\alpha$ implies $f \in F_C^\beta$.

Definition 6 ***(Infinitely Smooth Functions)*** *A function f is said to be infinitely smooth with respect to a constant C, if $\|f\|_\alpha \le C$ for all $\alpha > 0$. The set of infinitely smooth functions is denoted by F_C^∞.*

With the definitions of smoothness, we introduce the hypothesis space we use in active learning algorithms.

Definition 7 ***(Hypotheses with Smooth Classification Boundaries)*** *A set of hypotheses $\mathcal{H}_C^\alpha$ defined on $\mathcal{X} = [0,1]^{d+1}$ is said to have αth order smooth classification boundaries, if for every $h \in \mathcal{H}_C^\alpha$, the classification boundary is a αth order smooth function on $[0,1]^d$. To be precise, let $\mathbf{x} = (x^1, x^2, \ldots, x^{d+1}) \in [0,1]^{d+1}$. The classification boundary is the graph of function $x^{d+1} = f(x^1, \ldots, x^d)$, where $f \in F_C^\alpha$. Similarly, a hypothesis space $\mathcal{H}_C^\infty$ is said to have infinitely smooth boundaries, if for every $h \in \mathcal{H}_C^\infty$ the classification boundary is the graph an infinitely smooth function on $[0,1]^d$.*

The first thing we need to guarantee is that smooth problems are learnable, both passively and actively. To be concrete, we must show that the entropy with bracketing of smooth problems satisfies

$$H_{[\,]}(\varepsilon, \mathcal{H}, L_2(\mathcal{D}_X)) = O\left(\left(\frac{1}{\varepsilon}\right)^{2p}\right),$$

for some $p < 1$ (van der Vaart and Wellner, 1996) (see also Theorem 4). For smooth problems, the following proposition is known.

Proposition 8 *(van der Vaart and Wellner, 1996)*

Let the instance space be $[0,1]^{d+1}$ *and the hypothesis space be* $\mathcal{H}_C^\alpha$*. Assume that the marginal distribution* $\mathcal{D}_X$ *has a density upper bounded by a constant. Then*

$$H_{[\,]}(\varepsilon, \mathcal{H}_C^\alpha, L_2(\mathcal{D}_X)) = O\left(\left(\frac{1}{\varepsilon}\right)^{\frac{2d}{\alpha}}\right).$$

The problem is learnable if $\alpha > d$.

In the rest of this paper, we only consider smooth problems such that $\alpha > d$.

3.2 Disagreement Coefficient

The disagreement coefficient θ plays an important role for the label complexity of active learning algorithms. In fact previous negative examples for which active learning does not work are all because of large θ. For instance the interval learning problem, $\theta(\varepsilon) = \frac{1}{\varepsilon}$, which leads to the same label complexity as passive learning. (Recall that $\theta(\varepsilon) \le \frac{1}{\varepsilon}$, so this is the worst case.)

In this section we will show that that the disagreement coefficient $\theta(\varepsilon)$ for smooth problems is small. Especially, we establish both upper bounds (Theorem 9 and Theorem 10) and lower bounds (Theorem 13) for the disagreement coefficient of smooth problems. Finally we will combine our upper bounds on the disagreement coefficient with the label complexity result of Theorem 4 and show that active learning is strictly superior to passive learning for smooth problems.

Theorem 9 *Let the instance space be* $\mathcal{X} = [0,1]^{d+1}$*. Let the hypothesis space be* $\mathcal{H}_C^\alpha$*, where* $d < \alpha < \infty$*. If the marginal distribution* $\mathcal{D}_X$ *has a density* $p(\mathbf{x})$ *on* $[0,1]^{d+1}$ *such that there exists an* α*th order smooth function* $g(\mathbf{x})$ *and two constants* $0 < a \le b$ *such that* $ag(\mathbf{x}) \le p(\mathbf{x}) \le bg(\mathbf{x})$ *for all* $\mathbf{x} \in [0,1]^{d+1}$*, then*[3]

$$\theta(\varepsilon) = O\left(\left(\frac{1}{\varepsilon}\right)^{\frac{d}{\alpha+d}}\right).$$

The key points in the theorem are: the classification boundaries are smooth; and the density is bounded from above and below by constants times a smooth function.[4] Note that the density itself is not necessarily smooth. We merely require the density does not change too rapidly.

The intuition behind the theorem above is as follows. Let $f_{h^*}(x)$ and $f_h(x)$ be the classification boundaries of h^* and h, and suppose $\rho(h,h^*)$ is small, where $\rho(h,h^*) = \Pr_{x\sim\mathcal{D}_X}(h(x) \ne h^*(x))$ is the pseudo metric. If the classification boundaries and the density are all smooth, then the two boundaries have to be close to each other everywhere. That is, $|f_h(x) - f_{f^*}(x)|$ is small uniformly for all x. Hence only the points close to the classification boundary of h^* can be in $DIS(B(h^*,\varepsilon))$, which leads to a small disagreement coefficient.

For infinitely smooth problems, we have the following theorem. Note that the requirement on the density is stronger than finite smoothness problems.

[3] This upper bound was obtained with the help of Yanqi Dai, Kai Fan, Chicheng Zhang and Ziteng Wang. It improves a previous upper bound $O\left(\left(\frac{1}{\varepsilon}\right)^{1-\frac{\alpha^d}{(1+\alpha)^d}}\right)$, which converges to the current bound as $\frac{\alpha}{d} \to \infty$.

[4] These two conditions include a large class of learning problems. For example, the boundary fragment class equipped with most elementary distributions (truncated in $[0,1]^{d+1}$) satisfies these conditions.

Theorem 10 *Let the hypothesis space be $\mathcal{H}_C^\infty$. If the distribution $\mathcal{D}_X$ has a density $p(\mathbf{x})$ such that there exist two constants $0 < a \le b$ such that $a \le p(\mathbf{x}) \le b$ for all $\mathbf{x} \in [0,1]^{d+1}$, then $\theta(\varepsilon) = O(\log^{2d}(\frac{1}{\varepsilon}))$.*

The proofs of Theorem 9 and Theorem 10 rely on the following two lemmas.

Lemma 11 *Let Φ be a function defined on $[0,1]^d$ and is αth order smooth. If*

$$\int_{[0,1]^d} |\Phi(x)|dx \le r,$$

then

$$\|\Phi\|_\infty = O\left(r^{\frac{\alpha}{\alpha+d}}\right) = O\left(r \cdot \left(\frac{1}{r}\right)^{\frac{d}{\alpha+d}}\right),$$

where $\|\Phi\|_\infty = \sup_{x\in[0,1]^d} |\Phi(x)|$.

Lemma 12 *Let Φ be a function defined on $[0,1]^d$ and is infinitely smooth. If*

$$\int_{[0,1]^d} |\Phi(x)|dx \le r,$$

then

$$\|\Phi\|_\infty = O\left(r \cdot \left(\log\frac{1}{r}\right)^{2d}\right).$$

Proof of Theorem 9 First of all, since we focus on binary classification, $DIS(B(h^*,r))$ can be written equivalently as

$$DIS(B(h^*,r)) = \{x \in \mathcal{X},\ \exists h \in B(h^*,r),\ s.t.\ h(x) \ne h^*(x)\}.$$

Consider any $h \in B(h^*,r)$. Let $f_h, f_{h^*} \in F_C^\alpha$ be the corresponding classification boundaries of h and h^* respectively. If r is sufficiently small, we must have

$$\rho(h,h^*) = \Pr_{X\sim\mathcal{D}_X}(h(X) \ne h^*(X)) = \int\limits_{[0,1]^d} dx^1 \dots dx^d \left| \int_{f_{h^*}(x^1,\dots,x^d)}^{f_h(x^1,\dots,x^d)} p(x^1,\dots,x^{d+1})dx^{d+1} \right|.$$

Denote

$$\Phi_h(x^1,\dots,x^d) = \int_{f_{h^*}(x^1,\dots,x^d)}^{f_h(x^1,\dots,x^d)} p(x^1,\dots,x^{d+1})dx^{d+1}.$$

We assert that there is a αth order smooth function $\tilde{\Phi}_h(x^1,\dots,x^d)$ and two constants $0 < a \le b$ such that $a|\tilde{\Phi}_h| \le |\Phi_h| \le b|\tilde{\Phi}_h|$. To see this, remember that f_h and f_{h^*} are αth order smooth functions; and the density p is upper and lower bounded by constants times a αth order smooth function $g(x^1,\dots,x^{d+1})$. Also note that if we define

$$\tilde{\Phi}_h(x^1,\dots,x^d) = \int_{f_{h^*}(x^1,\dots,x^d)}^{f_h(x^1,\dots,x^d)} g(x^1,\dots,x^{d+1})dx^{d+1},$$

$\tilde{\Phi}_h$ is a αth order smooth function, which is easy to check by taking derivatives. Now

$$\int_{[0,1]^d} |\tilde{\Phi}_h(x)|dx \leq \int_{[0,1]^d} \frac{1}{a}|\Phi_h(x)|dx \leq \frac{r}{a}.$$

According to Lemma 11, we have $\|\tilde{\Phi}_h\|_\infty = O(r^{\frac{\alpha}{\alpha+d}})$. Thus $\|\Phi_h\|_\infty \leq b\|\tilde{\Phi}_h\|_\infty = O(r^{\frac{\alpha}{\alpha+d}})$. Because this holds for all $h \in B(h^*, r)$, we have

$$\sup_{h\in B(h^*,r)} \|\Phi_h\|_\infty = O\left(r^{\frac{\alpha}{\alpha+d}}\right).$$

Now consider the region of disagreement of $B(h^*, r)$. Note that

$$DIS(B(h^*,r)) = \cup_{h\in B(h^*,r)}\{x : h(x) \neq h^*(x)\}.$$

Hence

$$\begin{aligned}\Pr_{X\sim\mathcal{D}_X}(x \in DIS(B(h^*,r))) &= \Pr_{X\sim\mathcal{D}_X}\left(x \in \cup_{h\in B(h^*,r)}\{x : h(x) \neq h^*(x)\}\right)\\ &\leq 2\int_{[0,1]^d} \sup_{h\in B(h^*,r)} \|\Phi_h\|_\infty dx^1 \ldots dx^d = O\left(r^{\frac{\alpha}{\alpha+d}}\right) = O\left(r \cdot \left(\frac{1}{r}\right)^{\frac{d}{\alpha+d}}\right).\end{aligned}$$

The theorem follows by the definition of $\theta(\varepsilon)$. ■

Theorem 10 can be proved similarly by using Lemma 12.

In the next theorem, we give lower bounds on the disagreement coefficient for finite smooth problems under the condition that the marginal distribution $\mathcal{D}_X$ is the uniform distribution.[5] Note that in this case the lower bound matches the upper bound in Theorem 9. Thus in general Theorem 9 cannot be improved.

Theorem 13 *Let the hypothesis space be $\mathcal{H}_C^\alpha$ where $\alpha < \infty$. Assume that the marginal distribution $\mathcal{D}_X$ is uniform on $[0,1]^{d+1}$. Then the disagreement coefficient has the following lower bound*[6]

$$\theta(\varepsilon) = \Omega\left(\left(\frac{1}{\varepsilon}\right)^{\frac{d}{\alpha+d}}\right).$$

Proof Without loss of generality, we assume that the classification boundary of the optimal classifier h^* is the graph of function $x^{d+1} = f(x^1, x^2, \ldots, x^d) \equiv 1/2$. That is, the classification boundary of h^* is a hyperplane orthogonal to the $d+1$th axis. We will show that for most points $(x^1, x^2, \ldots, x^{d+1}) \in [0,1]^{d+1}$ that are $\varepsilon \cdot \left(\frac{1}{\varepsilon}\right)^{\frac{d}{\alpha+d}}$-close to the classification boundary of h^*, that is, $|x^{d+1} - \frac{1}{2}| \leq \varepsilon \cdot \left(\frac{1}{\varepsilon}\right)^{\frac{d}{\alpha+d}}$, there is a $h_f \in \mathcal{H}_C^\alpha$ satisfying

$$h_f(x^1, x^2, \ldots, x^{d+1}) \neq h^*(x^1, x^2, \ldots, x^{d+1}),$$

[5] The condition can be relaxed to that $\mathcal{D}_X$ is bounded from above and below by positive constants.

[6] The bound was obtained with the help of Yanqi Dai, Kai Fan, Chicheng Zhang and Ziteng Wang. It improves a previous lower bound $\Omega\left(\left(\frac{1}{\varepsilon}\right)^{\frac{d}{2\alpha+d}}\right)$.

and at the same time

$$\rho(h_f,h^*) = \Pr\left(h_f(X) \neq h^*(X)\right) \leq \varepsilon,$$

and therefore $h_f \in B(h^*,\varepsilon)$. Thus the volume of $DIS(B(h^*,\varepsilon))$ is $\Omega\left(\varepsilon \cdot \left(\frac{1}{\varepsilon}\right)^{\frac{d}{\alpha+d}}\right)$; and consequently

$$\theta(\varepsilon) \geq \frac{\Pr\left(DIS(B(h^*,\varepsilon))\right)}{\varepsilon} = \Omega\left(\left(\frac{1}{\varepsilon}\right)^{\frac{d}{\alpha+d}}\right).$$

For this purpose, fixing $(x^1,x^2,\ldots,x^{d+1}) \in [0,1]^{d+1}$ with

$$0 \leq x^{d+1} - \frac{1}{2} \leq c\varepsilon\left(\frac{1}{\varepsilon}\right)^{\frac{d}{\alpha+d}},$$

for some constant c. We only consider point $(x^1,x^2,\ldots,x^{d+1}) \in [0,1]^{d+1}$ that are not too close to the "boundary" of $[0,1]^{d+1}$. (e.g., $0.1 \leq x^i \leq 0.9$ for all $1 \leq i \leq d$.) We construct h_f whose classification boundary is the graph of the following function f. For convenience, we shift $(x^1,x^2,\ldots,x^d,\frac{1}{2})$ to the origin. Let f be defined on $[0,1]^d$ as

$$f(u_1,u_2,\ldots,u_d) = \begin{cases} \xi^{-\alpha}\left(\xi^2 - \sum_{i=1}^d u_i^2\right)^\alpha & \text{if } \sum_{i=1}^d u_i^2 \leq \xi^2, \\ 0 & \text{otherwise,} \end{cases}$$

where ξ is determined by

$$\int_\Omega |f| d\omega = \varepsilon,$$

that is, $\rho(h_f,h^*) = \varepsilon$, and Ω is the region obtained from $[0,1]^d$ after shifting $(x^1,x^2,\ldots,x^d,\frac{1}{2})$ to the origin.

First, it is not hard to check by calculus that f is αth order smooth. Next, since $\int_\Omega |f| d\omega = \varepsilon$, it is not difficult to calculate that $\xi = c'\varepsilon^{\frac{1}{\alpha+d}}$, for some constant c'. Thus

$$\|f\|_\infty = f(0,0,\ldots,0) = c'\varepsilon^{\frac{\alpha}{\alpha+d}} = c'\varepsilon\left(\frac{1}{\varepsilon}\right)^{\frac{d}{\alpha+d}}.$$

So we have $h_f(0,0,\ldots,0) \neq h^*(0,0,\ldots,0)$ and $\rho(h_f,h^*) = \varepsilon$. This completes the proof. ∎

For infinite smoothness, we do not know any lower bound for the disagreement coefficient larger than the trivial $\Omega(1)$.

3.2.1 LABEL COMPLEXITY FOR SMOOTH PROBLEMS

Now we combine our results (Theorem 9 and Theorem 10) with the label complexity bounds for active learning (Theorem 4 and (4)) and show that active learning is strictly superior to passive learning for smooth problems.

Remember that under Tsybakov's noise conditions the label complexity of active learning is (see Theorem 4 and (4))

$$O\left(\theta(\varepsilon^{\frac{1}{\kappa}})\left(\frac{1}{\varepsilon}\right)^{2-\frac{2-p}{\kappa}}\left(\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right)\right).$$

While for passive learning it is (see (5))

$$O\left(\left(\frac{1}{\varepsilon}\right)^{2-\frac{1-p}{\kappa}}\left(\log\frac{1}{\delta}+\log\log\frac{1}{\varepsilon}\right)\right).$$

We see that if

$$\theta(\varepsilon^{\frac{1}{\kappa}}) = o\left(\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\kappa}}\right),$$

then active learning requires strictly fewer labels than passive learning.

By Theorem 9 and remember that $\alpha > d$ (see Proposition 8) we obtain

$$\theta(\varepsilon^{\frac{1}{\kappa}}) = O\left(\frac{1}{\varepsilon}\right)^{\frac{d}{\kappa(\alpha+d)}} = o\left(\left(\frac{1}{\varepsilon}\right)^{\frac{1}{2\kappa}}\right) = o\left(\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\kappa}}\right).$$

So we have the following conclusion.

Theorem 14 *Assume that the Tsybakov noise exponent κ is finite. Then active learning algorithms A^2 and DHM have label complexity strictly smaller than passive learning for αth order smooth problems whenever $\alpha > d$.*

3.3 Discussion

In this section we discuss and compare our results to a closely related work due to Castro and Nowak (2008), which also studied the label complexity of smooth problems under Tsybakov's noise condition. Castro and Nowak's work is heavily based on their detailed analysis of actively learning a threshold on $[0,1]$ described below.

Consider the learning problem in which the instance space $\mathcal{X} = [0,1]$; the hypothesis space $\mathcal{H}$ contains all threshold functions, that is, $\mathcal{H} = \{\mathbb{I}(x \geq t) : t \in [0,1]\} \cup \{\mathbb{I}(x < t) : t \in [0,1]\}$, where $\mathbb{I}$ is indicator function; and the marginal distribution $\mathcal{D}_X$ is the uniform distribution on $[0,1]$. Suppose that the Bayes classifier $h_B^* \in \mathcal{H}$. Assume that the learning problem satisfies the "geometric" Tsybakov's noise condition

$$\left|\eta(x) - \frac{1}{2}\right| \geq b\,|x - x_B^*|^{\kappa-1}, \tag{6}$$

for some constant $b > 0$ and for all x such that $|\eta(x) - \frac{1}{2}| \leq \tau_0$ with the constant $\tau_0 > 0$. Here x_B^* is the threshold of the Bayes classifier. (One can verity that (6) implies the ordinary Tsybakov's condition with noise exponent κ when $\mathcal{D}_X$ is uniform on [0,1].) In addition, assume that the learning problem satisfies a reverse-sided Tsybakov's condition

$$\left|\eta(x) - \frac{1}{2}\right| \leq B\,|x - x_B^*|^{\kappa-1},$$

for some constant $B > 0$.

Under these assumptions, Castro and Nowak showed that an active learning algorithm they attributed to Burnashev and Zigangirov (1974) (will be referred to as BZ), which is essentially a Bayesian binary search algorithm,[7] has label complexity $\tilde{O}\left(\left(\frac{1}{\varepsilon}\right)^{2-\frac{2}{\kappa}}\right)$. Moreover, due to the

[7] Note that this BZ algorithm can choose any point x from the instance space, not necessarily from the given pool of unlabeled data. This model is called membership query, making stronger assumptions than the pool-based active learning model.

reverse-sided Tsybakov's condition, one can show that with high probability that the threshold $\hat{x}$ returned by the active learning algorithm converges to the Bayes threshold x_B^* exponentially fast with respect to the number of label requests.[8]

Castro and Nowak then generalized this result to smooth problems similar to what we studied in this paper. Let the hypothesis space be $\mathcal{H}_C^\alpha$. Suppose that the Bayes classifier $h_B^* \in \mathcal{H}_C^\alpha$. Assume that on every vertical line segment in $[0,1]^{d+1}$, (that is, for every $\{(x^1,x^2,\ldots,x^d,x^{d+1}) : x^{d+1} \in [0,1]\}$, $(x^1,x^2,\ldots,x^d) \in [0,1]^d$,) the one dimensional distribution $\eta_{x^1,\ldots,x^d}(x^{d+1})$ satisfies the two-sided geometric Tsybakov's condition with noise exponent κ. Based on these assumptions, they proposed the following active learning algorithm: Choosing M vertical line segments in $[0,1]^{d+1}$. Performing one dimensional threshold learning on each line segment as in the one-dimensional case described above. After obtaining the threshold for each line, doing a piecewise polynomial interpolation on these thresholds and return the interpolation function as the classification boundary. They showed that this algorithm has label complexity $\tilde{O}\left(\left(\frac{1}{\varepsilon}\right)^{2-\frac{2-\frac{d}{\alpha}}{\kappa}}\right)$.

In sum, their algorithm makes the following main assumptions:

(A1) On every vertical line, the conditional distribution is two-sided Tsybakov. Thus the distribution $\mathcal{D}_{XY}$ has a uniform "one-dimensional" behavior along the $(d+1)$th axis.

(A2) The algorithm can choose any point from $\mathcal{X} = [0,1]^{d+1}$ and ask for its label.

Comparing the label complexity of this algorithm

$$\tilde{O}\left(\left(\frac{1}{\varepsilon}\right)^{2-\frac{2-\frac{d}{\alpha}}{\kappa}}\right)$$

and that obtained from Theorem 4 and Proposition 8

$$\tilde{O}\left(\theta(\varepsilon_0)\left(\frac{1}{\varepsilon}\right)^{2-\frac{2-\frac{d}{\alpha}}{\kappa}}\right)$$

one sees that their label complexity is smaller by $\theta(\varepsilon_0)$. It seems that the disagreement coefficient of smooth problem does not play a role in their label complexity formula. The reason is that the assumption (A1) in their model assumes that the distribution of the problem has a uniform one-dimensional behavior: on each line segment parallel to the $d+1$th axis, the conditional distribution $\eta_{x^1,\ldots,x^d}(x^{d+1})$ satisfies the two-sided Tsybakov's condition with equal noise exponent κ. Therefore the algorithm can assign equal label budget to each line segment for one-dimensional learning. Recall that the disagreement coefficient of the one-dimensional threshold learning problem is at most 2 for all $\varepsilon > 0$, so there seems no $\theta(\varepsilon)$ term in the final label complexity formula. If, instead of assumption (A1), we assume the ordinary Tsybakov's noise condition, the algorithm has to assign

[8] It needs to be pointed out that the BZ algorithm requires that the noise exponent κ is known and the algorithm takes it as imput. But by Threorem 4 and (4) we know that both A^2 and DHM (with slight modifications) have the same label complexity and the convergence property, since the disagreement coefficient for this threshold problem is $\theta(\varepsilon) = 2$ for all $\varepsilon > 0$.

label budgets according to the “worst” noise on all the line segments, that is, the largest κ of the conditional distributions $\eta_{x^1,\dots,x^d}(x^{d+1})$ over all $(x^1,\dots,x^d)$. But under the ordinary Tsybakov’s condition, the largest κ on lines can be arbitrarily large, resulting a label complexity equal to that of passive learning.

4. Proof of Lemma 11 and 12—Some Generalizations of the Landau-Kolmogorov Type Inequalities

In this section we give proofs of Lemma 11 and Lemma 12. The two lemmas are closely related to the Landau-Kolmogorov type inequalities (Landau, 1913; Kolmogorov, 1938; Schoenberg, 1973) (see also Mitrinović et al., 1991 Chapter I for a comprehensive survey), and specifically the following result due to Gorny (1939).

Theorem 15 *Let $f(x)$ be a function defined on $[0,1]$ and has derivatives up to the nth order. Let $M_k = \|f^{(k)}\|_\infty$, $k = 0,1,\dots,n$. Then*

$$M_k \le 4\left(\frac{n}{k}\right)^k e^k M_0^{1-\frac{k}{n}} M_n'^{\frac{k}{n}},$$

where $M_n' = \max(M_0 n!, M_n)$.

Roughly, for function f defined on a finite interval, the above theorem bounds the ∞-norm of the kth order derivative of f by the ∞-norm of f and its nth derivative.

In order to prove Lemma 11, we give the following generalization of Theorem 15 (in the direction of dimensionality and non-integer smoothness). Our proof is elementary and is simpler than Gorny’s proof of Theorem 15. But note that the definition of M_α' in Theorem 16 is different to that of Theorem 15.

Theorem 16 *Let f be a function defined on $[0,1]^d$. Let $\alpha > 1$ be a real number. Assume that f has partial derivatives up to order $\underline{\alpha}$. For all $1 \le t \le \alpha$, define*

$$M_t = \max_{|\mathbf{k}|=\underline{t}} \sup_{x,x'} \frac{|D^{\mathbf{k}} f(x) - D^{\mathbf{k}} f(x')|}{\|x - x'\|^{t-\underline{t}}},$$

where $D^{\mathbf{k}}$ is the differential operator defined in Section 3.1 and $x, x' \in [0,1]^d$. Also define $M_0 = \sup_{x\in[0,1]^d} f(x)$. Then

$$M_k \le C M_0^{1-\frac{k}{\alpha}} M_\alpha'^{\frac{k}{\alpha}}, \tag{7}$$

where $M_\alpha' = \max(M_0, M_\alpha)$, $k = 1,2,\dots,\underline{\alpha}$, and the constant C depends on k and α but does not depend on M_0 and M_α.

Lemma 11 can be derived from Theorem 16.

Proof of Lemma 11 The proof has two steps. First, we construct a function $\overline{f}$ by scaling Φ and redefine the domain of the function, so that a) the integral of $\overline{f}$ over the unit hypercube is at most 1; b) the $\underline{\alpha}$th order derivatives of $\overline{f}$ is Hölder continuous with the same constant as Φ, and c) $\|\overline{f}\|_\infty = \left(\frac{1}{r}\right)^{\frac{\alpha}{\alpha+d}} \|\Phi\|_\infty$. Next, we use Theorem16 to show that $\|\overline{f}\|_\infty$ can be bounded from above by a constant depending only on α, d, C (recall that $\Phi \in F_C^{\alpha}$) but independent of r. Combining the two steps concludes the theorem.

Now, for the first step assume

$$\int_{[0,1]^d} |\Phi(x)| dx = t \leq r.$$

Let

$$f(x^1, x^2, \ldots, x^d) = \left(\frac{1}{t}\right)^{\frac{\alpha}{\alpha+d}} \Phi(t^{\frac{1}{\alpha+d}} x^1, t^{\frac{1}{\alpha+d}} x^2, \ldots, t^{\frac{1}{\alpha+d}} x^d),$$

where now the domain of f is $(x^1, \ldots, x^d) \in [0, \frac{1}{t}^{\frac{1}{\alpha+d}}]^d$.

First, it is easy to check that

$$\int_{[0, \frac{1}{t}^{\frac{1}{\alpha+d}}]^d} |f(x)| dx = 1;$$

and the $\underline{\alpha}$th order derivatives of f is Hölder continuous with the same constant C as Φ. That is,

$$\max_{|\mathbf{k}|=\underline{\alpha}} \sup_{x,x' \in [0, \frac{1}{t}^{\frac{1}{\alpha+d}}]^d} \frac{|D^{\mathbf{k}} f(x) - D^{\mathbf{k}} f(x')|}{\|x - x'\|^{\alpha - \underline{\alpha}}} = \max_{|\mathbf{k}|=\underline{\alpha}} \sup_{x,x' \in [0,1]^d} \frac{|D^{\mathbf{k}} \Phi(x) - D^{\mathbf{k}} \Phi(x')|}{\|x - x'\|^{\alpha - \underline{\alpha}}} \leq C.$$

In addition, clearly we have

$$\|\Phi\|_\infty = t^{\frac{\alpha}{\alpha+d}} \|f\|_\infty \leq r^{\frac{\alpha}{\alpha+d}} \|f\|_\infty.$$

Thus in order to prove the lemma, we only need to show $\|f\|_\infty$ is bounded from above by a universal constant independent of r.

Note that the domain of f is $[0, \frac{1}{t}^{\frac{1}{\alpha+d}}]^d$, larger than $[0,1]^d$. Assume f achieves its maximum at $(a_1, a_2, \ldots, a_d) \in [0, \frac{1}{t}^{\frac{1}{\alpha+d}}]^d$. Now we truncate the domain of f to a d-dimensional hypercube $[z_1, z_1 + 1] \otimes [z_2, z_2 + 1] \otimes \ldots, \otimes [z_d, z_d + 1]$ so that $(a_1, a_2, \ldots, a_d) \in [z_1, z_1 + 1] \otimes [z_2, z_2 + 1] \otimes \ldots, \otimes [z_d, z_d + 1]$. Let $\overline{f}$ be the function by restricting f on this hypercube $[z_1, z_1 + 1] \otimes [z_2, z_2 + 1] \otimes \ldots, \otimes [z_d, z_d + 1]$. Clearly, we have

$$\|\overline{f}\|_\infty = \|f\|_\infty,$$

where $\|\overline{f}\|_\infty$ is the maximum over the hypercube $[z_1, z_1 + 1] \otimes [z_2, z_2 + 1] \otimes \ldots, \otimes [z_d, z_d + 1]$. Thus we just need to show $\|\overline{f}\|_\infty$ has a universal upper bound.

Now we begin the second step of the proof, where our goal is to show $\overline{f}$ has an upper bound independent of r. Assume $z_i = 0$ for $i = 1, \ldots, d$ by shifting if necessary.

Let

$$g_d(x^1, \ldots, x^d) = \int_0^{x^d} \overline{f}(x^1, \ldots, x^{d-1}, u_d) du_d.$$

For any fixed $x^1, \ldots, x^{d-1}$, consider g_d as a function of the single variable x^d. Since $\overline{f}$ is the first order derivative of g_d, it is easy to check that g_d has derivatives up to order $\underline{\alpha} + 1$ with respect to x^d and its $\underline{\alpha} + 1$ order derivative is Hölder continuous with constant C. Thus according to Theorem16, we have

$$\|\overline{f}\|_\infty \leq \|g_d\|_\infty^{\frac{\alpha}{\alpha+1}} C^{\frac{1}{\alpha+1}}. \tag{8}$$

Similarly, let

$$g_i(x^1,\ldots,x^d)=\int_0^{x^i} g_{i+1}(x^1,\ldots,x^{i-1},u_i,x^{i+1},\ldots,x^d)du_i,\quad i=1,2,\ldots,d-1.$$

For each g_i use the above argument and observe that the $\underline{\alpha}+1$th order derivative of each g_i is bounded from above by C, then it is easy to obtain that

$$\|g_i\|_\infty \le \|g_{i+1}\|_\infty^{\frac{\alpha}{\alpha+1}} C^{\frac{1}{\alpha+1}}. \tag{9}$$

Combining (8) and (9) for all $i=1,2,\ldots,d$ we have

$$\|\overline{f}\|_\infty \le \|g_1\|_\infty^{\left(\frac{\alpha}{\alpha+1}\right)^d} C',$$

where C' is a constant depending on $C,\ \alpha,\ d$. This completes the proof since

$$g_1(x^1,\ldots,x^d)=\int_0^{x^1}\cdots\int_0^{x^d}\overline{f}(u_1,\ldots,u_d)du_1,\ldots,du_d \le 1.$$

■

Proof of Theorem 16 The structure of the proof is as follows: we first deal with the case of $d=1$ and then generalize to $d>1$. For the case of $d=1$, we first show the case $1<\alpha\le 2$, and then prove general α by induction.

Now assume $d=1$ and $1<\alpha\le 2$. Our goal is to show

$$M_1 \le C M_0^{1-\frac{1}{\alpha}} M_\alpha'^{\frac{1}{\alpha}}.$$

For any fixed $x\in[0,1]$, there must be a $y\in[0,1]$ such that $|y-x|=1/2$. We thus have

$$\frac{f(y)-f(x)}{y-x}=f'(x+u), \tag{10}$$

where $|u|\le 1/2$. Since $1<\alpha\le 2$, we know $\underline{\alpha}=1$. By the definition of M_α we have

$$|f'(x+u)-f'(x)|\le M_\alpha |u|^{\alpha-1}. \tag{11}$$

Combining (10) and (11) and recall $|y-x|=1/2$, we obtain

$$\begin{aligned}|f'(x)| &\le \left|\frac{f(y)-f(x)}{y-x}\right| + M_\alpha |u|^{\alpha-1}\\ &\le 4M_0+\left(\frac{1}{2}\right)^{\alpha-1} M_\alpha\\ &\le 4M_0+M_\alpha. \end{aligned} \tag{12}$$

Let $g(x)=f(ax+r)$, where $0<a\le 1$, $r\in[0,1-a]$ and $x\in[0,1]$. Let

$$M_0^g=\sup_{x\in[0,1]}|g(x)|,\qquad M_\alpha^g=\sup_{x,x'\in[0,1]}\frac{|g'(x)-g'(x')|}{|x-x'|^{\alpha-1}}.$$

It is easy to check that $M_0^g \leq M_0$ and $M_\alpha^g \leq a^\alpha M_\alpha$. Applying (12) to $g(x)$, which is defined on $[0,1]$, we obtain that for every $a \in (0,1]$

$$|f'(ax+r)| = \frac{1}{a}|g'(x)| \leq \frac{4M_0^g}{a} + \frac{M_\alpha^g}{a} \leq \frac{4M_0}{a} + a^{\alpha-1}M_\alpha.$$

Taking

$$a = \min\left(\left(\frac{M_0}{M_\alpha}\right)^{\frac{1}{\alpha}}, 1\right),$$

we obtain for all $x \in [r, a+r]$

$$|f'(x)| \leq 5M_0^{1-\frac{1}{\alpha}} M_\alpha'^{\frac{1}{\alpha}},$$

where $M'_\alpha = \max(M_0, M_\alpha)$. Since $r \in [0, 1-a]$ is arbitrary, we have that

$$M_1 = \sup_{x\in[0,1]} |f'(x)| \leq 5M_0^{1-\frac{1}{\alpha}} M_\alpha'^{\frac{1}{\alpha}}. \tag{13}$$

Note that this implies that for all nonnegative integer m and $1 < \alpha \leq 2$, we have

$$M_{m+1} \leq 5M_m^{1-\frac{1}{\alpha}} M_{m+\alpha}'^{\frac{1}{\alpha}}.$$

We next prove the general $\alpha > 1$ case. Let n be a positive integer. By induction, assume for all $1 < \alpha \leq n$ we already have, for $k = 1, 2, \ldots, \underline{\alpha}$

$$M_k \leq CM_0^{1-\frac{k}{\alpha}} \left(\max(M_0, M_\alpha)\right)^{\frac{k}{\alpha}}, \tag{14}$$

where the constant C depends on k and α but does not depend on M_0 and M_α. (In the following the constant C my be different from line to line and even in the same line.) We will prove that (14) is true for $\alpha \in (n, n+1]$. Here we will treat the two cases $1 \leq k < n$ and $k = n$ separately.

For the case $1 \leq k < n$, since $\alpha - k \leq n$, by the assumption of the induction we have

$$M_n \leq CM_k^{1-\frac{n-k}{\alpha-k}} \left(\max(M_k, M_\alpha)\right)^{\frac{n-k}{\alpha-k}}. \tag{15}$$

Combining (14) and (15), and setting $\alpha = n$ in (14). We distinguish three cases.

Case I: $M_0 > M_n$
We have

$$\begin{aligned} M_k &\leq CM_0^{1-\frac{k}{n}} \left(\max(M_0, M_n)\right)^{\frac{k}{n}} \\ &= CM_0 \\ &\leq CM_0^{1-\frac{k}{\alpha}} \left(\max(M_0, M_\alpha)\right)^{\frac{k}{\alpha}}. \end{aligned}$$

Case II: $M_0 \leq M_n$ and $M_k > M_\alpha$
We have

$$M_k \leq CM_0^{1-\frac{k}{n}} M_n^{\frac{k}{n}} \leq CM_0^{1-\frac{k}{n}} M_k^{\frac{k}{n}}.$$

Thus

$$M_k \leq CM_0 \leq CM_0^{1-\frac{k}{\alpha}} \left(\max(M_0, M_\alpha)\right)^{\frac{k}{\alpha}}.$$

Case III: $M_0 \leq M_n$ and $M_k \leq M_\alpha$

We have

$$\begin{aligned} M_k &\leq CM_0^{1-\frac{k}{n}} M_n^{\frac{k}{n}} \\ &\leq CM_0^{1-\frac{k}{n}} \left(M_k^{1-\frac{n-k}{\alpha-k}} M_\alpha^{\frac{n-k}{\alpha-k}} \right)^{\frac{k}{n}} \\ &= CM_0^{1-\frac{k}{n}} M_k^{\frac{k(\alpha-n)}{n(\alpha-k)}} M_\alpha^{\frac{k(n-k)}{n(\alpha-k)}}. \end{aligned}$$

We obtain after some simple calculations that

$$M_k \leq CM_0^{1-\frac{k}{\alpha}} M_\alpha^{\frac{k}{\alpha}}.$$

This completes the proof for $1 \leq k < n$.

For the case $k = n$, note that

$$M_n \leq CM_1^{1-\frac{n-1}{\alpha-1}} \max(M_1, M_\alpha)^{\frac{n-1}{\alpha-1}}, \tag{16}$$

and

$$M_1 \leq CM_0^{1-\frac{1}{n}} \max(M_0, M_n)^{\frac{1}{n}}. \tag{17}$$

We need to distinguish four cases.

Case I: $M_1 > M_\alpha$ and $M_0 > M_n$

Combining (16) and (17), we have

$$M_n \leq CM_1 \leq CM_0 \leq CM_0^{1-\frac{n}{\alpha}} \left(\max(M_0, M_\alpha)\right)^{\frac{n}{\alpha}}.$$

Case II: $M_1 > M_\alpha$ and $M_0 \leq M_n$

We have

$$M_n \leq CM_1 \leq CM_0^{1-\frac{1}{n}} M_n^{\frac{1}{n}}.$$

Thus

$$M_n \leq CM_0 \leq CM_0^{1-\frac{n}{\alpha}} \left(\max(M_0, M_\alpha)\right)^{\frac{n}{\alpha}}.$$

Case III: $M_1 \leq M_\alpha$ and $M_0 > M_n$

We have

$$\begin{aligned} M_n &\leq CM_1^{1-\frac{n-1}{\alpha-1}} M_\alpha^{\frac{n-1}{\alpha-1}} \\ &\leq CM_0^{1-\frac{n-1}{\alpha-1}} M_\alpha^{\frac{n-1}{\alpha-1}}. \end{aligned} \tag{18}$$

If $M_0 \leq M_\alpha$, then from (18) we obtain

$$\begin{aligned} M_n &\leq CM_0^{1-\frac{n}{\alpha}} M_\alpha^{\frac{n}{\alpha}} \left(\frac{M_0}{M_\alpha} \right)^{\frac{n}{\alpha}-\frac{n-1}{\alpha-1}} \\ &\leq CM_0^{1-\frac{n}{\alpha}} \left(\max(M_0, M_\alpha)\right)^{\frac{n}{\alpha}}. \end{aligned}$$

Otherwise, $M_0 > M_\alpha$, by (18)

$$M_n \leq CM_0.$$

Case IV: $M_1 \leq M_\alpha$ and $M_0 \leq M_n$
We have

$$\begin{aligned} M_n &\leq C\left(M_0^{1-\frac{1}{n}} M_n^{\frac{1}{n}}\right)^{1-\frac{n-1}{\alpha-1}} M_\alpha^{\frac{n-1}{\alpha-1}} \\ &= CM_n^{\frac{(n-1)(\alpha-n)}{n(\alpha-1)}} M_\alpha^{\frac{n-1}{\alpha-1}} M_n^{\frac{\alpha-n}{n(\alpha-1)}}. \end{aligned}$$

After some simple calculation, this yields

$$M_n \leq CM_0^{1-\frac{n}{\alpha}} M_\alpha^{\frac{n}{\alpha}}.$$

This completes the proof of the $k = n$ case, and we finished the discussion of the $d = 1$ case.

The above arguments are easy to generalize to the $d \geq 2$ case. Consider the function $f(x^1, \ldots, x^d)$. Again we first look at $1 < \alpha \leq 2$. Assume M_1 is achieved at the jth partial derivative of f. That is,

$$\left\| \frac{\partial f}{\partial x^j} \right\|_\infty = M_1.$$

Fixing $x^1, \ldots, x^{j-1}, x^{j+1}, \ldots, x^d$, consider f as a function of a single variable x^j. By the argument for the $d = 1$ case, we know that

$$M_1 \leq 5M_0^{1-\frac{1}{\alpha}} \tilde{M}^{\frac{1}{\alpha}},$$

where

$$\tilde{M} = \max\left(M_0, \sup_{x,x'} \frac{\left|\frac{\partial f}{\partial x^j}\big|_x - \frac{\partial f}{\partial x^j}\big|_{x'}\right|}{\|x - x'\|^{\alpha-1}}\right).$$

Clearly, $\tilde{M} \leq \max(M_0, M_\alpha) = M'_\alpha$. Hence for $1 < \alpha \leq 2$, we have $M_1 \leq 5M_0^{1-\frac{1}{\alpha}} M_\alpha'^{\frac{1}{\alpha}}$. Finally, using the previous induction argument and noting that it does not depend on the dimensionality d, we obtain the desired result for all $\alpha > 1$. ■

To prove Lemma 12 however, Theorem 16 is not a suitable tool. Note that the M'_α in (7) is $\max(M_0, M_\alpha)$, while in Theorem 15 $M'_\alpha = \max(n!M_0, M_\alpha)$. Therefore the constant in Theorem 16 grows exponentially regarding to α. In the following we give another generalization of Gorny's inequality which will be used to prove Lemma 12. The price however is that it cannot handle the non-integer smoothness.

Theorem 17 *Let $f(x)$ be defined on $[0,1]^d$ and have uniformly bounded partial derivatives up to order n. Let*

$$M_k = \sup_{|\mathbf{k}|=k} \|D^{\mathbf{k}} f\|_\infty, \quad k = 0, 1, 2, \ldots, n.$$

Then

$$M_k \leq Cn^k M_0^{1-\frac{k}{n}} M_n'^{\frac{k}{n}},$$

where $M'_n = \max(n!M_0, M_n)$, and C is a constant depending on k but does not depend on M_0, M_n and n.

This theorem is a straightforward generalization of Gorny's theorem to multidimension. Now we can use this theorem to prove Lemma 12.

Proof of Lemma 12 Similar to the proof of Lemma 11, for $(x^1,\ldots,x^d)\in[0,1]^d$, let

$$f(x^1,\ldots,x^d)=\int_0^{x^1}\int_0^{x^2}\cdots\int_0^{x^d}\Phi(u_1,\ldots,u_d)du_1,\ldots,du_d.$$

It is easy to check that f is infinitely smooth. Let M_t be defined as in Theorem 15 for f. Clearly $M_0\le r$ and $\|\Phi\|_\infty\le M_d$. Since f is infinitely smooth, there is a constant C such that $M_n\le C$ for $n=d+1,d+2,\ldots$.

Now for r sufficiently small, take $n=\frac{\log\frac{1}{r}}{\log\log\frac{1}{r}}$. Let's first look at $n!M_0$. Note that

$$n^n=\left(\frac{\log\frac{1}{r}}{\log\log\frac{1}{r}}\right)^{\frac{\log\frac{1}{r}}{\log\log\frac{1}{r}}}\le\left(\log\frac{1}{r}\right)^{\frac{\log\frac{1}{r}}{\log\log\frac{1}{r}}}=\frac{1}{r}.$$

We have

$$\begin{aligned}M_0n! &\le r\sqrt{2\pi n}n^ne^{-n}\le\sqrt{2\pi\frac{\log\frac{1}{r}}{\log\log\frac{1}{r}}}(r)^{\frac{1}{\log\log\frac{1}{r}}}\\ &\le \sqrt{2\pi}\exp\left(\frac{\log\log\frac{1}{r}-\log\log\log\frac{1}{r}}{2}-\frac{\log\frac{1}{r}}{\log\log\frac{1}{r}}\right),\end{aligned}$$

which tends to zero as $r\to 0$ and therefore

$$M_n'=\max(M_0n!,M_n))\le C.$$

Thus we have, by Theorem 17

$$\begin{aligned}\|\Phi\|_\infty &\le M_d\\ &\le Cn^dM_0^{1-\frac{d}{n}}M_n'^{\frac{d}{n}}\\ &\le Cn^dM_0^{1-\frac{d}{n}}\\ &\le C\left(\frac{\log\frac{1}{r}}{\log\log\frac{1}{r}}\right)^d r\left(\frac{1}{r}\right)^{\frac{d\log\log\frac{1}{r}}{\log\frac{1}{r}}}\\ &\le Cr\left(\log\frac{1}{r}\right)^{2d}.\end{aligned}$$

■

Proof of Theorem 17 We know that the theorem is valid when $d=1$. Now assume $d\ge 2$. Using the same argument in the proof of Theorem 16, we have that for all positive integers n,

$$M_1\le CnM_0^{1-\frac{1}{n}}M_n'^{\frac{1}{n}},$$

and in general

$$M_{m+1} \leq CnM_m^{1-\frac{1}{n}} M'^{\frac{1}{n}}_{m+n}, \tag{19}$$

for any nonnegative integer m.

Now we prove the theorem by induction on k. Assume we have already shown

$$M_{k-1} \leq Cn^{k-1} M_0^{1-\frac{k-1}{n}} \left(\max(M_0 n!, M_n)\right)^{\frac{k-1}{n}}. \tag{20}$$

By (19) we have

$$M_k \leq C(n-k+1) M_{k-1}^{1-\frac{1}{n-k+1}} \left(\max\Big((n-k+1)! M_{k-1}, M_n\Big)\right)^{\frac{1}{n-k+1}}. \tag{21}$$

We consider the following four cases separately. Note that below we will frequently use the fact that for $m = 1, 2, \ldots$

$$\frac{1}{\sqrt{2\pi}} (m!)^{\frac{1}{m}} e \leq m \leq (m!)^{\frac{1}{m}} e^{1+\frac{1}{12}}.$$

To see this, just note that

$$\sqrt{2\pi m} m^m e^{-(m+\frac{1}{12m})} \leq m! \leq \sqrt{2\pi m} m^m e^{-m},$$

and

$$1 \leq \left(\sqrt{2\pi m}\right)^{\frac{1}{m}} \leq \sqrt{2\pi}.$$

Case I: $(n-k+1)! M_{k-1} > M_n$ and $n! M_0 \leq M_n$

From (21) we have

$$M_k \leq C(n-k+1) M_{k-1} \Big((n-k+1)!\Big)^{\frac{1}{n-k+1}} \leq C(n-k+1)^2 M_{k-1} \leq Cn^2 M_{k-1}.$$

Taking into consideration of (20), we have

$$\begin{aligned} M_k &\leq Cn^{k+1} M_0^{1-\frac{k-1}{n}} M_n'^{\frac{k-1}{n}} \\ &\leq Cn^k M_0^{1-\frac{k}{n}} M_n'^{\frac{k}{n}} \left(nM_0^{\frac{1}{n}} M_n'^{-\frac{1}{n}}\right) \\ &\leq Cn^k M_0^{1-\frac{k}{n}} M_n'^{\frac{k}{n}} \left(\frac{n! M_0}{M_n'}\right)^{\frac{1}{n}} \\ &\leq Cn^k M_0^{1-\frac{k}{n}} M_n'^{\frac{k}{n}}. \end{aligned}$$

Case II: $(n-k+1)! M_{k-1} > M_n$ and $n! M_0 > M_n$

By the similar argument as in Case I, we have

$$\begin{aligned} M_k &\leq Cn^{k+1} M_0^{1-\frac{k-1}{n}} M_n'^{\frac{k-1}{n}} \\ &= Cn^{k+1} M_0 (n!)^{\frac{k-1}{n}} \\ &\leq Cn^k M_0 (n!)^{\frac{k}{n}} \\ &= Cn^k M_0^{1-\frac{k}{n}} (n! M_0)^{\frac{k}{n}}. \end{aligned}$$

Case III: $(n-k+1)!M_{k-1} \le M_n$ and $n!M_0 > M_n$

Combining (20) and (21),

$$\begin{aligned} M_k &\le C(n-k+1)\left(n^{k-1}M_0(n!)^{\frac{k-1}{n}}\right)^{1-\frac{1}{n-k+1}} M_n^{\frac{1}{n-k+1}} \\ &\le Cn^k M_0^{1-\frac{1}{n-k+1}}(n!)^{(\frac{k-1}{n})(1-\frac{1}{n-k+1})}(n!M_0)^{\frac{1}{n-k+1}} \\ &\le Cn^k M_0(n!)^{\frac{k}{n}} \\ &\le Cn^k M_0^{1-\frac{k}{n}}(n!M_0)^{\frac{k}{n}}. \end{aligned}$$

Case IV: $(n-k+1)!M_{k-1} \le M_n$ and $n!M_0 \le M_n$ Combining (20) and (21), we obtain

$$\begin{aligned} M_k &\le C(n-k+1)M_{k-1}^{1-\frac{1}{n-k+1}} M_n^{\frac{1}{n-k+1}} \\ &\le Cn\left(n^{k-1}M_0^{1-\frac{k-1}{n}} M_n^{\frac{k-1}{n}}\right)^{1-\frac{1}{n-k+1}} M_n^{\frac{1}{n-k+1}} \\ &\le Cn^k M_0^{1-\frac{k}{n}} M_n^{\frac{k}{n}}. \end{aligned}$$

This completes the proof. ■

5. Conclusion

This paper studies the disagreement coefficient of smooth problems and extends our previous results (Wang, 2009). Comparing to the worst case $\theta(\varepsilon) = \frac{1}{\varepsilon}$ for which active learning has the same label complexity as passive learning, the disagreement coefficient is $\theta(\varepsilon) = O\left(\left(\frac{1}{\varepsilon}\right)^{\frac{d}{\alpha+d}}\right)$ for αth ($\alpha < \infty$) order smooth problems, and is $\theta(\varepsilon) = O\left(\log^{2d}\left(\frac{1}{\varepsilon}\right)\right)$ for infinite order smooth problems. Combining with the bounds on the label complexity in terms of disagreement coefficient, we give sufficient conditions for which active learning algorithm A^2 and DHM are superior to passive learning under Tsybakov's noise condition.

Although we assume that the classification boundary is the graph of a function, our results can be generalized to the case that the boundaries are a finite number of functions. To be precise, consider N (N is even) functions $f_1(\mathbf{x}) \le \cdots \le f_N(\mathbf{x})$, for all $\mathbf{x} \in [0,1]^d$. Let $f_0(\mathbf{x}) \equiv 0$, $f_{N+1}(\mathbf{x}) \equiv 1$. The positive (or negative) set defined by these functions is $\{(\mathbf{x}, x^{d+1}): \; f_{2i}(\mathbf{x}) \le x^{d+1} \le f_{2i+1}(\mathbf{x}), \; i = 0, 1, \ldots, \frac{N}{2}\}$. It is easy to show that our main theorems still hold in this case. Moreover, using the techniques in Dudley (1999, page 259), our results may generalize to the case that the classification boundaries are intrinsically smooth, and not necessarily graphs of smooth functions. This would include a substantially richer class of problems which can be benefit from active learning.

There is an open problems worthy of further study. For infinitely smooth problems we proved that the disagreement coefficient can be upper and lower bounded by $O\left(\log^{2d}\left(\frac{1}{\varepsilon}\right)\right)$ and $\Omega(1)$ respectively. Improving the upper bound and (or) the lower bound would be interesting.

Acknowledgments

The author thank Yanqi Dai, Kai Fan, Chicheng Zhang and Ziteng Wang for very helpful discussions. In particular, they helped to improve Theorem 9 and Theorem 13 over previous versions. The author is grateful to an anonymous reviewer who gave many constructive suggestions. It greatly improves the quality of the paper. This work was supported by NSFC (61075003) and a grant from MOE-Microsoft Key Laboratory of Statistics and Information Technology of Peking University.

References

M.-F. Balcan, A.Beygelzimer, and J. Langford. Agnostic active learning. In *23th International Conference on Machine Learning*, 2006.

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *26th International Conference on Machine Learning*, 2009.

A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems*, 2010.

M. Burnashev and K. Zigangirov. An interval estimation problem for controlled problem. *Problems of Information Transmission*, 10:223–231, 1974.

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

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994.

S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*, 2005.

S. Dasgupta and J. Langford. A tutorial on active learning, 2009.

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*, 2007.

R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

E. Friedman. Active learning for smooth problems. In *22th Annual Conference on Learning Theory*, 2009.

A. Gorny. Contribution a l'etude des fonctions dérivables d'une variable réelle. *Acta Mathematica*, 13:317–358, 1939.

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

S. Hanneke. Adaptive rates of convergence in active learning. In *22th Annual Conference on Learning Theory*, 2009.

S. Hanneke. Rates of convergence in active learning. *Annals of Statistics*, 39:333–361, 2011.

M. Kääriäinen. Active learning in the non-realizable case. In *17th International Conference on Algorithmic Learning Theory*, 2006.

A. Kolmogorov. Une généralisation de l'inégalite de J. Hadamard entre les bornes supérieurs des dérivés successives d'une fonction. *C. R. Académie des Sciences, Paris*, 207:763–765, 1938.

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34:2593–2656, 2006.

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

E. Landau. Einige ungleichungen für zweimal differentiierbare funktionen. *Proceedings of the London Mathematical Society*, 13:43–49, 1913.

D.S. Mitrinović, J.E. Pečarié, and A.M. Fink. *Inequalities Involving Functions and Their Integrals and Derivatives*. Kluwer Academic Publishers, 1991.

I.J. Schoenberg. The elementary cases of landau's problem of inequlities between derivatives. *The American Mathematical Monthly*, 80:121–158, 1973.

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

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes with Application to Statistics*. Springer Verlag, 1996.

V. Vapnik. *Statistical Learning Theory*. John Wiely and Sons, 1998.

L. Wang. Sufficient conditions for agnostic active learnable. In *Advances in Neural Information Processing Systems*, 2009.