Revisiting Perceptron: Efficient and Label-Optimal Learning of Halfspaces

Songbai Yan
UC San Diego
La Jolla, CA
yansongbai@ucsd.edu

Chicheng Zhang[*]
Microsoft Research
New York, NY
chicheng.zhang@microsoft.com

Abstract

It has been a long-standing problem to efficiently learn a halfspace using as few labels as possible in the presence of noise. In this work, we propose an efficient Perceptron-based algorithm for actively learning homogeneous halfspaces under the uniform distribution over the unit sphere. Under the bounded noise condition [49], where each label is flipped with probability at most $\eta < \frac{1}{2}$, our algorithm achieves a near-optimal label complexity of $\tilde{O}\left(\frac{d}{(1-2\eta)^2}\ln\frac{1}{\epsilon}\right)^2$ in time $\tilde{O}\left(\frac{d^2}{\epsilon(1-2\eta)^3}\right)$. Under the adversarial noise condition [6, 45, 42], where at most a $\tilde{\Omega}(\epsilon)$ fraction of labels can be flipped, our algorithm achieves a near-optimal label complexity of $\tilde{O}\left(d\ln\frac{1}{\epsilon}\right)$ in time $\tilde{O}\left(\frac{d^2}{\epsilon}\right)$. Furthermore, we show that our active learning algorithm can be converted to an efficient passive learning algorithm that has near-optimal sample complexities with respect to ϵ and d.

1 Introduction

We study the problem of designing efficient noise-tolerant algorithms for actively learning homogeneous halfspaces in the streaming setting. We are given access to a data distribution from which we can draw unlabeled examples, and a noisy labeling oracle $\mathcal{O}$ that we can query for labels. The goal is to find a computationally efficient algorithm to learn a halfspace that best classifies the data while making as few queries to the labeling oracle as possible.

Active learning arises naturally in many machine learning applications where unlabeled examples are abundant and cheap, but labeling requires human effort and is expensive. For those applications, one natural question is whether we can learn an accurate classifier using as few labels as possible. Active learning addresses this question by allowing the learning algorithm to sequentially select examples to query for labels, and avoid requesting labels which are less informative, or can be inferred from previously-observed examples.

There has been a large body of work on the theory of active learning, showing sharp distribution-dependent label complexity bounds [21, 11, 34, 27, 35, 46, 60, 41]. However, most of these general active learning algorithms rely on solving empirical risk minimization problems, which are computationally hard in the presence of noise [5].

On the other hand, existing computationally efficient algorithms for learning halfspaces [17, 29, 42, 45, 6, 23, 7, 8] are not optimal in terms of label requirements. These algorithms have different degrees of noise tolerance (e.g. adversarial noise [6], malicious noise [43], random classification noise [3],

[*] Work done while at UC San Diego.

[2] We use $\tilde{O}(f(\cdot)) := O(f(\cdot)\ln f(\cdot))$, and $\tilde{\Omega}(f(\cdot)) := \Omega(f(\cdot)/\ln f(\cdot))$. We say $f(\cdot) = \tilde{\Theta}(g(\cdot))$ if $f(\cdot) = \tilde{O}(g(\cdot))$ and $f(\cdot) = \tilde{\Omega}\left(g(\cdot)\right)$

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

bounded noise [49], etc), and run in time polynomial in $\frac{1}{\epsilon}$ and d. Some of them naturally exploit the utility of active learning [6, 7, 8], but they do not achieve the sharpest label complexity bounds in contrast to those computationally-inefficient active learning algorithms [10, 9, 60].

Therefore, a natural question is: is there any active learning halfspace algorithm that is computationally efficient, and has a minimum label requirement? This has been posed as an open problem in [50]. In the realizable setting, [26, 10, 9, 56] give efficient algorithms that have optimal label complexity of $\tilde{O}(d \ln \frac{1}{\epsilon})$ under some distributional assumptions. However, the challenge still remains open in the nonrealizable setting. It has been shown that learning halfspaces with agnostic noise even under Gaussian unlabeled distribution is hard [44]. Nonetheless, we give an affirmative answer to this question under two moderate noise settings: bounded noise and adversarial noise.

1.1 Our Results

We propose a Perceptron-based algorithm, ACTIVE-PERCEPTRON, for actively learning homogeneous halfspaces under the uniform distribution over the unit sphere. It works under two noise settings: bounded noise and adversarial noise. Our work answers an open question by [26] on whether Perceptron-based active learning algorithms can be modified to tolerate label noise.

In the *η-bounded noise setting* (also known as the Massart noise model [49]), the label of an example $x \in \mathbb{R}^d$ is generated by $\mathrm{sign}(u \cdot x)$ for some underlying halfspace u, and flipped with probability $\eta(x) \le \eta < \frac{1}{2}$. Our algorithm runs in time $\tilde{O}\left(\frac{d^2}{(1-2\eta)^3\epsilon}\right)$, and requires $\tilde{O}\left(\frac{d}{(1-2\eta)^2} \cdot \ln \frac{1}{\epsilon}\right)$ labels. We show that this label complexity is *nearly optimal* by providing an almost matching information-theoretic lower bound of $\Omega\left(\frac{d}{(1-2\eta)^2} \cdot \ln \frac{1}{\epsilon}\right)$. Our time and label complexities substantially improve over the state of the art result of [8], which runs in time $\tilde{O}(d^{O(\frac{1}{(1-2\eta)^4})}\frac{1}{\epsilon})$ and requires $\tilde{O}(d^{O(\frac{1}{(1-2\eta)^4})} \ln \frac{1}{\epsilon})$ labels.

Our main theorem on learning under bounded noise is as follows:

Theorem 2 (Informal). *Suppose the labeling oracle $\mathcal{O}$ satisfies the η-bounded noise condition with respect to u, then for* ACTIVE-PERCEPTRON, *with probability at least $1-\delta$: (1) The output halfspace v is such that $\mathbb{P}[\mathrm{sign}(v \cdot X) \ne \mathrm{sign}(u \cdot X)] \le \epsilon$; (2) The number of label queries to oracle $\mathcal{O}$ is at most $\tilde{O}\left(\frac{d}{(1-2\eta)^2} \cdot \ln \frac{1}{\epsilon}\right)$; (3) The number of unlabeled examples drawn is at most $\tilde{O}\left(\frac{d}{(1-2\eta)^3\epsilon}\right)$; (4) The algorithm runs in time $\tilde{O}\left(\frac{d^2}{(1-2\eta)^3\epsilon}\right)$.*

In addition, we show that our algorithm also works in a more challenging setting, the *ν-adversarial noise setting* [6, 42, 45].[3] In this setting, the examples still come iid from a distribution, but the assumption on the labels is just that $\mathbb{P}[\mathrm{sign}(u \cdot X) \ne Y] \le \nu$ for some halfspace u. Under this assumption, the Bayes classifier may not be a halfspace. We show that our algorithm achieves an error of ϵ while tolerating a noise level of $\nu = \Omega\left(\frac{\epsilon}{\ln \frac{d}{\delta} + \ln \ln \frac{1}{\epsilon}}\right)$. It runs in time $\tilde{O}\left(\frac{d^2}{\epsilon}\right)$, and requires only $\tilde{O}\left(d \cdot \ln \frac{1}{\epsilon}\right)$ labels which is near-optimal. ACTIVE-PERCEPTRON has a label complexity bound that matches the state of the art result of [39][4], while having a lower running time.

Our main theorem on learning under adversarial noise is as follows:

Theorem 3 (Informal). *Suppose the labeling oracle $\mathcal{O}$ satisfies the ν-adversarial noise condition with respect to u, where $\nu < \Theta(\frac{\epsilon}{\ln \frac{d}{\delta} + \ln \ln \frac{1}{\epsilon}})$. Then for* ACTIVE-PERCEPTRON, *with probability at least $1-\delta$: (1) The output halfspace v is such that $\mathbb{P}[\mathrm{sign}(v \cdot X) \ne \mathrm{sign}(u \cdot X)] \le \epsilon$; (2) The number of label queries to oracle $\mathcal{O}$ is at most $\tilde{O}\left(d \cdot \ln \frac{1}{\epsilon}\right)$; (3) The number of unlabeled examples drawn is at most $\tilde{O}\left(\frac{d}{\epsilon}\right)$; (4) The algorithm runs in time $\tilde{O}\left(\frac{d^2}{\epsilon}\right)$.*

[3] Note that the adversarial noise model is not the same as that in online learning [18], where each example can be chosen adversarially.

[4] The label complexity bound is implicit in [39] by a refined analysis of the algorithm of [6] (See their Lemma 8 for details).

Table 1: A comparison of algorithms for active learning of halfspaces under the uniform distribution, in the η-bounded noise model.

Algorithm	Label Complexity	Time Complexity
[10, 9, 60]	$\tilde{O}(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon})$	superpoly$(d, \frac{1}{\epsilon})$ [5]
[8]	$\tilde{O}(d^{O(\frac{1}{(1-2\eta)^4})} \cdot \ln \frac{1}{\epsilon})$	$\tilde{O}(d^{O(\frac{1}{(1-2\eta)^4})} \cdot \frac{1}{\epsilon})$
Our Work	$\tilde{O}(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon})$	$\tilde{O}\left(\frac{d^2}{(1-2\eta)^3} \frac{1}{\epsilon}\right)$

Table 2: A comparison of algorithms for active learning of halfspaces under the uniform distribution, in the ν-adversarial noise model.

Algorithm	Noise Tolerance	Label Complexity	Time Complexity
[60]	$\nu = \Omega(\epsilon)$	$\tilde{O}(d \ln \frac{1}{\epsilon})$	superpoly$(d, \frac{1}{\epsilon})$
[39]	$\nu = \Omega(\epsilon)$	$\tilde{O}(d \ln \frac{1}{\epsilon})$	poly$(d, \frac{1}{\epsilon})$
Our Work	$\nu = \Omega(\frac{\epsilon}{\ln d + \ln \ln \frac{1}{\epsilon}})$	$\tilde{O}(d \ln \frac{1}{\epsilon})$	$\tilde{O}\left(d^2 \cdot \frac{1}{\epsilon}\right)$

Throughout the paper, ACTIVE-PERCEPTRON is shown to work if the unlabeled examples are drawn uniformly from the unit sphere. The algorithm and analysis can be easily generalized to any spherical symmetrical distributions, for example, isotropic Gaussian distributions. They can also be generalized to distributions whose densities with respect to uniform distribution are bounded away from 0.

In addition, we show in Section 6 that ACTIVE-PERCEPTRON can be converted to a passive learning algorithm, PASSIVE-PERCEPTRON, that has near optimal sample complexities with respect to ϵ and d under the two noise settings. We defer the discussion to the end of the paper.

2 Related Work

Active Learning. The recent decades have seen much success in both theory and practice of active learning; see the excellent surveys by [54, 37, 25]. On the theory side, many label-efficient active learning algorithms have been proposed and analyzed. An incomplete list includes [21, 11, 34, 27, 35, 46, 60, 41]. Most algorithms relies on solving empirical risk minimization problems, which are computationally hard in the presence of noise [5].

Computational Hardness of Learning Halfspaces. Efficient learning of halfspaces is one of the central problems in machine learning [22]. In the realizable case, it is well known that linear programming will find a consistent hypothesis over data efficiently. In the nonrealizable setting, however, the problem is much more challenging.

A series of papers have shown the hardness of learning halfspaces with agnostic noise [5, 30, 33, 44, 23]. The state of the art result [23] shows that under standard complexity-theoretic assumptions, there exists a data distribution, such that the best linear classifier has error $o(1)$, but no polynomial time algorithms can achieve an error at most $\frac{1}{2} - \frac{1}{d^c}$ for every $c > 0$, even with improper learning. [44] shows that under standard assumptions, even if the unlabeled distribution is Gaussian, any agnostic halfspace learning algorithm must run in time $(\frac{1}{\epsilon})^{\Omega(\ln d)}$ to achieve an excess error of ϵ. These results indicate that, to have nontrivial guarantees on learning halfspaces with noise in polynomial time, one has to make additional assumptions on the data distribution over instances and labels.

Efficient Active Learning of Halfspaces. Despite considerable efforts, there are only a few halfspace learning algorithms that are both computationally-efficient and label-efficient even under the uniform distribution. In the realizable setting, [26, 10, 9] propose computationally efficient active learning algorithms which have an optimal label complexity of $\tilde{O}(d \ln \frac{1}{\epsilon})$.

Since it is believed to be hard for learning halfspaces in the general agnostic setting, it is natural to consider algorithms that work under more moderate noise conditions. Under the bounded noise

[5] The algorithm needs to minimize 0-1 loss, the best known method for which requires superpolynomial time.

setting [49], the only known algorithms that are both label-efficient and computationally-efficient are [7, 8]. [7] uses a margin-based framework which queries the labels of examples near the decision boundary. To achieve computational efficiency, it adaptively chooses a sequence of hinge loss minimization problems to optimize as opposed to directly optimizing the 0-1 loss. It works only when the label flipping probability upper bound η is small ($\eta \leq 1.8 \times 10^{-6}$). [8] improves over [7] by adapting a polynomial regression procedure into the margin-based framework. It works for any $\eta < 1/2$, but its label complexity is $O(d^{O(\frac{1}{(1-2\eta)^4})} \ln \frac{1}{\epsilon})$, which is far worse than the information-theoretic lower bound $\Omega(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon})$. Recently [20] gives an efficient algorithm with a near-optimal label complexity under the membership query model where the learner can query on synthesized points. In contrast, in our stream-based model, the learner can only query on points drawn from the data distribution. We note that learning in the stream-based model is harder than in the membership query model, and it is unclear how to transform the DC algorithm in [20] into a computationally efficient stream-based active learning algorithm.

Under the more challenging ν-adversarial noise setting, [6] proposes a margin-based algorithm that reduces the problem to a sequence of hinge loss minimization problems. Their algorithm achieves an error of ϵ in polynomial time when $\nu = \Omega(\epsilon)$, but requires $\tilde{O}(d^2 \ln \frac{1}{\epsilon})$ labels. Later, [39] performs a refined analysis to achieve a near-optimal label complexity of $\tilde{O}(d \ln \frac{1}{\epsilon})$, but the time complexity of the algorithm is still an unspecified high order polynomial.

Tables 1 and 2 present comparisons between our results and results most closely related to ours in the literature. Due to space limitations, discussions of additional related work are deferred to Appendix A.

3 Definitions and Settings

We consider learning homogeneous halfspaces under uniform distribution. The instance space $\mathcal{X}$ is the unit sphere in $\mathbb{R}^d$, which we denote by $\mathbb{S}^{d-1} := \left\{x \in \mathbb{R}^d : \|x\| = 1\right\}$. We assume $d \geq 3$ throughout this paper. The label space $\mathcal{Y} = \{+1, -1\}$. We assume all data points (x, y) are drawn i.i.d. from an underlying distribution D over $\mathcal{X} \times \mathcal{Y}$. We denote by $D_{\mathcal{X}}$ the marginal of D over $\mathcal{X}$ (which is uniform over $\mathbb{S}^{d-1}$), and $D_{Y|X}$ the conditional distribution of Y given X. Our algorithm is allowed to draw unlabeled examples $x \in \mathcal{X}$ from $D_{\mathcal{X}}$, and to make queries to a labeling oracle $\mathcal{O}$ for labels. Upon query x, $\mathcal{O}$ returns a label y drawn from $D_{Y|X=x}$. The hypothesis class of interest is the set of homogeneous halfspaces $\mathcal{H} := \left\{h_w(x) = \operatorname{sign}(w \cdot x) \mid w \in \mathbb{S}^{d-1}\right\}$. For any hypothesis $h \in \mathcal{H}$, we define its error rate $\operatorname{err}(h) := \mathbb{P}_D[h(X) \neq Y]$. We will drop the subscript D in $\mathbb{P}_D$ when it is clear from the context. Given a dataset $S = \left\{(X_1, Y_1), \ldots, (X_m, Y_m)\right\}$, we define the empirical error rate of h over S as $\operatorname{err}_S(h) := \frac{1}{m} \sum_{i=1}^m \mathbb{1}\left\{h(x_i) \neq y_i\right\}$.

Definition 1 (Bounded Noise [49])**.** *We say that the labeling oracle $\mathcal{O}$ satisfies the* η-bounded noise condition *for some $\eta \in [0, 1/2)$ with respect to u, if for any x, $\mathbb{P}[Y \neq \operatorname{sign}(u \cdot x) \mid X = x] \leq \eta$.*

It can be seen that under η-bounded noise condition, h_u is the Bayes classifier.

Definition 2 (Adversarial Noise [6])**.** *We say that the labeling oracle $\mathcal{O}$ satisfies the* ν-adversarial noise condition *for some $\nu \in [0, 1]$ with respect to u, if $\mathbb{P}[Y \neq \operatorname{sign}(u \cdot X)] \leq \nu$.*

For two unit vectors v_1, v_2, denote by $\theta(v_1, v_2) = \arccos(v_1 \cdot v_2)$ the angle between them. The following lemma gives relationships between errors and angles (see also Lemma 1 in [8]).

Lemma 1. *For any $v_1, v_2 \in \mathbb{S}^{d-1}$, $\left|\operatorname{err}(h_{v_1}) - \operatorname{err}(h_{v_2})\right| \leq \mathbb{P}\left[h_{v_1}(X) \neq h_{v_2}(X)\right] = \frac{\theta(v_1, v_2)}{\pi}$. Additionally, if the labeling oracle satisfies the η-bounded noise condition with respect to u, then for any vector v, $\left|\operatorname{err}(h_v) - \operatorname{err}(h_u)\right| \geq (1 - 2\eta)\mathbb{P}\left[h_v(X) \neq h_u(X)\right] = \frac{1-2\eta}{\pi}\theta(v, u)$.*

Given access to unlabeled examples drawn from $D_{\mathcal{X}}$ and a labeling oracle $\mathcal{O}$, our goal is to find a polynomial time algorithm $\mathcal{A}$ such that with probability at least $1 - \delta$, $\mathcal{A}$ outputs a halfspace $h_v \in \mathcal{H}$ with $\mathbb{P}[\operatorname{sign}(v \cdot X) \neq \operatorname{sign}(u \cdot X)] \leq \epsilon$ for some target accuracy ϵ and confidence δ. (By Lemma 1, this guarantees that the excess error of h_v is at most ϵ, namely, $\operatorname{err}(h_v) - \operatorname{err}(h_u) \leq \epsilon$.) The desired algorithm should make as few queries to the labeling oracle $\mathcal{O}$ as possible.

We say an algorithm $\mathcal{A}$ achieves a *label complexity* of $\Lambda(\epsilon, \delta)$, if for any target halfspace $h_u \in \mathcal{H}$, with probability at least $1 - \delta$, $\mathcal{A}$ outputs a halfspace $h_v \in \mathcal{H}$ such that $\mathrm{err}(h_v) \leq \mathrm{err}(h_u) + \epsilon$, and requests at most $\Lambda(\epsilon, \delta)$ labels from oracle $\mathcal{O}$.

4 Main Algorithm

Our main algorithm, ACTIVE-PERCEPTRON (Algorithm 1), works in epochs. It works under the bounded and the adversarial noise models, if its sample schedule $\{m_k\}$ and band width $\{b_k\}$ are set appropriately with respect to each noise model. At the beginning of each epoch k, it assumes an upper bound of $\frac{\pi}{2^k}$ on $\theta(v_{k-1}, u)$, the angle between current iterate v_{k-1} and the underlying halfspace u. As we will see, this can be shown to hold with high probability inductively. Then, it calls procedure MODIFIED-PERCEPTRON (Algorithm 2) to find an new iterate v_k, which can be shown to have an angle with u at most $\frac{\pi}{2^{k+1}}$ with high probability. The algorithm ends when a total of $k_0 = \lceil \log_2 \frac{1}{\epsilon} \rceil$ epochs have passed.

For simplicity, we assume for the rest of the paper that the angle between the initial halfspace v_0 and the underlying halfspace u is acute, that is, $\theta(v_0, u) \leq \frac{\pi}{2}$; Appendix F shows that this assumption can be removed with a constant overhead in terms of label and time complexities.

Algorithm 1 ACTIVE-PERCEPTRON

```
Input: Labeling oracle $\mathcal{O}$, initial halfspace $v_0$, target error $\epsilon$, confidence $\delta$, sample schedule $\{m_k\}$,
    band width $\{b_k\}$.
Output: learned halfspace $v$.
Let $k_0 = \lceil \log_2 \frac{1}{\epsilon} \rceil$.
for $k = 1, 2, \ldots, k_0$ do
    $v_k \leftarrow$ MODIFIED-PERCEPTRON($\mathcal{O}, v_{k-1}, \frac{\pi}{2^k}, \frac{\delta}{k(k+1)}, m_k, b_k$).
end for
return $v_{k_0}$.
```

Procedure MODIFIED-PERCEPTRON (Algorithm 2) is the core component of ACTIVE-PERCEPTRON. It sequentially performs a modified Perceptron update rule on the selected new examples (x_t, y_t) [51, 17, 26]:

$$w_{t+1} \leftarrow w_t - 2\mathbb{1}\left\{y_t w_t \cdot x_t < 0\right\}(w_t \cdot x_t) \cdot x_t \tag{1}$$

Define $\theta_t := \theta(w_t, u)$. Update rule (1) implies the following relationship between θ_{t+1} and θ_t (See Lemma 8 in Appendix E for its proof):

$$\cos\theta_{t+1} - \cos\theta_t = -2\mathbb{1}\left\{y_t w_t \cdot x_t < 0\right\}(w_t \cdot x_t) \cdot (u \cdot x_t) \tag{2}$$

This motivates us to take $\cos\theta_t$ as our measure of progress; we would like to drive $\cos\theta_t$ up to 1(so that θ_t goes down to 0) as fast as possible.

To this end, MODIFIED-PERCEPTRON samples new points x_t under time-varying distributions $D_{\mathcal{X}}|_{R_t}$ and query for their labels, where $R_t = \left\{x \in \mathbb{S}^{d-1} : \frac{b}{2} \leq w_t \cdot x \leq b\right\}$ is a band inside the unit sphere. The rationale behind the choice of R_t is twofold:

1. We set R_t to have a probability mass of $\tilde{\Omega}(\epsilon)$, so that the time complexity of rejection sampling is at most $\tilde{O}(\frac{1}{\epsilon})$ per example. Moreover, in the adversarial noise setting, we set R_t large enough to dominate the noise of magnitude $\nu = \tilde{\Omega}(\epsilon)$.
2. Unlike the active Perceptron algorithm in [26] or other margin-based approaches (for example [55, 10]) where examples with small margin are queried, we query the label of the examples with a range of margin $[\frac{b}{2}, b]$. From a technical perspective, this ensures that θ_t decreases by a decent amount in expectation (see Lemmas 9 and 10 for details).

Following the insight of [32], we remark that the modified Perceptron update (1) on distribution $D_{\mathcal{X}}|_{R_t}$ can be alternatively viewed as performing stochastic gradient descent on a special non-convex loss function $\ell(w, (x, y)) = \min(1, \max(0, -1 - \frac{2}{b} y w \cdot x))$. It is an interesting open question whether optimizing this new loss function can lead to improved empirical results for learning halfspaces.

Algorithm 2 MODIFIED-PERCEPTRON

```
Input: Labeling oracle $\mathcal{O}$, initial halfspace $w_0$, angle upper bound $\theta$, confidence $\delta$, number of
    iterations $m$, band width $b$.
Output: Improved halfspace $w_m$.
for $t = 0, 1, 2, \ldots, m-1$ do
    Define region $R_t = \left\{ x \in \mathbb{S}^{d-1} : \frac{b}{2} \le w_t \cdot x \le b \right\}$.
    Rejection sample $x_t \sim D_{\mathcal{X}}|_{R_t}$. In other words, draw $x_t$ from $D_{\mathcal{X}}$ until $x_t$ is in $R_t$. Query $\mathcal{O}$
        for its label $y_t$.
    $w_{t+1} \leftarrow w_t - 2\mathbb{1}\{y_t w_t \cdot x_t < 0\} \cdot (w_t \cdot x_t) \cdot x_t$.
end for
return $w_m$.
```

5 Performance Guarantees

We show that ACTIVE-PERCEPTRON works in the bounded and the adversarial noise models, achieving computational efficiency and near-optimal label complexities. To this end, we first give a lower bound on the label complexity under bounded noise, and then give computational and label complexity upper bounds under the two noise conditions respectively. We defer all proofs to the Appendix.

5.1 A Lower Bound under Bounded Noise

We first present an information-theoretic lower bound on the label complexity in the bounded noise setting under uniform distribution. This extends the distribution-free lower bounds of [53, 37], and generalizes the realizable-case lower bound of [47] to the bounded noise setting. Our lower bound can also be viewed as an extension of [59]'s Theorem 3; specifically it addresses the hardness under the α-Tsybakov noise condition where $\alpha = 0$ (while [59]'s Theorem 3 provides lower boundes when $\alpha \in (0, 1)$).

Theorem 1. *For any $d > 4$, $0 \le \eta < \frac{1}{2}$, $0 < \epsilon \le \frac{1}{4\pi}$, $0 < \delta \le \frac{1}{4}$, for any active learning algorithm $\mathcal{A}$, there is a $u \in \mathbb{S}^{d-1}$, and a labeling oracle $\mathcal{O}$ that satisfies η-bounded noise condition with respect to u, such that if with probability at least $1-\delta$, $\mathcal{A}$ makes at most n queries of labels to $\mathcal{O}$ and outputs $v \in \mathbb{S}^{d-1}$ such that $\mathbb{P}[\operatorname{sign}(v \cdot X) \neq \operatorname{sign}(u \cdot X)] \le \epsilon$, then $n \ge \Omega\left(\frac{d \log \frac{1}{\epsilon}}{(1-2\eta)^2} + \frac{\eta \log \frac{1}{\delta}}{(1-2\eta)^2}\right)$.*

5.2 Bounded Noise

We establish Theorem 2 in the bounded noise setting. The theorem implies that, with appropriate settings of input parameters, ACTIVE-PERCEPTRON efficiently learns a halfspace of excess error at most ϵ with probability at least $1-\delta$, under the assumption that $D_{\mathcal{X}}$ is uniform over the unit sphere and $\mathcal{O}$ has bounded noise. In addition, it queries at most $\tilde{O}(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon})$ labels. This matches the lower bound of Theorem 1, and improves over the state of the art result of [8], where a label complexity of $\tilde{O}(d^{O(\frac{1}{(1-2\eta)^4})} \ln \frac{1}{\epsilon})$ is shown using a different algorithm.

The proof and the precise setting of parameters (m_k and b_k) are given in Appendix C.

Theorem 2 (ACTIVE-PERCEPTRON under Bounded Noise)**.** *Suppose Algorithm 1 has inputs labeling oracle $\mathcal{O}$ that satisfies η-bounded noise condition with respect to halfspace u, initial halfspace v_0 such that $\theta(v_0, u) \in [0, \frac{\pi}{2}]$, target error ϵ, confidence δ, sample schedule $\{m_k\}$ where $m_k = \Theta\left(\frac{d}{(1-2\eta)^2}(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{k}{\delta})\right)$, band width $\{b_k\}$ where $b_k = \Theta\left(\frac{2^{-k}(1-2\eta)}{\sqrt{d}\ln(km_k/\delta)}\right)$. Then with probability at least $1-\delta$:*

1. *The output halfspace v is such that $\mathbb{P}[\operatorname{sign}(v \cdot X) \neq \operatorname{sign}(u \cdot X)] \le \epsilon$.*
2. *The number of label queries is $O\left(\frac{d}{(1-2\eta)^2} \cdot \ln \frac{1}{\epsilon} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\epsilon}\right)\right)$.*

3. *The number of unlabeled examples drawn is* $O\left(\frac{d}{(1-2\eta)^3} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta} + \ln\ln \frac{1}{\epsilon}\right)^2 \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon}\right)$.

4. *The algorithm runs in time* $O\left(\frac{d^2}{(1-2\eta)^3} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta} + \ln\ln \frac{1}{\epsilon}\right)^2 \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon}\right)$.

The theorem follows from Lemma 2 below. The key ingredient of the lemma is a delicate analysis of the dynamics of the angles $\{\theta_t\}_{t=0}^m$, where $\theta_t = \theta(w_t, u)$ is the angle between the iterate w_t and the halfspace u. Since x_t is randomly sampled and y_t is noisy, we are only able to show that θ_t decreases by a decent amount *in expectation.* To remedy the stochastic fluctuations, we apply martingale concentration inequalities to carefully control the upper envelope of sequence $\{\theta_t\}_{t=0}^m$.

Lemma 2 (MODIFIED-PERCEPTRON under Bounded Noise)**.** *Suppose Algorithm 2 has inputs labeling oracle $\mathcal{O}$ that satisfies η-bounded noise condition with respect to halfspace u, initial halfspace w_0 and angle upper bound $\theta \in (0, \frac{\pi}{2}]$ such that $\theta(w_0, u) \leq \theta$, confidence δ, number of iterations $m = \Theta(\frac{d}{(1-2\eta)^2}(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta}))$, band width $b = \Theta\left(\frac{\theta(1-2\eta)}{\sqrt{d}\ln(m/\delta)}\right)$. Then with probability at least $1-\delta$:*

1. *The output halfspace w_m is such that $\theta(w_m, u) \leq \frac{\theta}{2}$.*

2. *The number of label queries is* $O\left(\frac{d}{(1-2\eta)^2}\left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta}\right)\right)$.

3. *The number of unlabeled examples drawn is* $O\left(\frac{d}{(1-2\eta)^3} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta}\right)^2 \cdot \frac{1}{\theta}\right)$.

4. *The algorithm runs in time* $O\left(\frac{d^2}{(1-2\eta)^3} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta}\right)^2 \cdot \frac{1}{\theta}\right)$.

5.3 Adversarial Noise

We establish Theorem 3 in the adversarial noise setting. The theorem implies that, with appropriate settings of input parameters, ACTIVE-PERCEPTRON efficiently learns a halfspace of excess error at most ϵ with probability at least $1-\delta$, under the assumption that $D_{\mathcal{X}}$ is uniform over the unit sphere and $\mathcal{O}$ has an adversarial noise of magnitude $\nu = \Omega(\frac{\epsilon}{\ln d + \ln\ln \frac{1}{\epsilon}})$. In addition, it queries at most $\tilde{O}(d \ln \frac{1}{\epsilon})$ labels. Our label complexity bound is information-theoretically optimal [47], and matches the state of the art result of [39]. The benefit of our approach is computational: it has a running time of $\tilde{O}(\frac{d^2}{\epsilon})$, while [39] needs to solve a convex optimization problem whose running time is some polynomial over d and $\frac{1}{\epsilon}$ with an unspecified degree.

The proof and the precise setting of parameters (m_k and b_k) are given in Appendix C.

Theorem 3 (ACTIVE-PERCEPTRON under Adversarial Noise)**.** *Suppose Algorithm 1 has inputs labeling oracle $\mathcal{O}$ that satisfies ν-adversarial noise condition with respect to halfspace u, initial halfspace v_0 such that $\theta(v_0, u) \leq \frac{\pi}{2}$, target error ϵ, confidence δ, sample schedule $\{m_k\}$ where $m_k = \Theta(d(\ln d + \ln \frac{k}{\delta}))$, band width $\{b_k\}$ where $b_k = \Theta\left(\frac{2^{-k}}{\sqrt{d}\ln(km_k/\delta)}\right)$. Additionally $\nu \leq \Omega(\frac{\epsilon}{\ln \frac{d}{\delta} + \ln\ln \frac{1}{\epsilon}})$. Then with probability at least $1-\delta$:*

1. *The output halfspace v is such that $\mathbb{P}[\mathrm{sign}(v \cdot X) \neq \mathrm{sign}(u \cdot X)] \leq \epsilon$.*

2. *The number of label queries is* $O\left(d \cdot \ln \frac{1}{\epsilon} \cdot \left(\ln d + \ln \frac{1}{\delta} + \ln\ln \frac{1}{\epsilon}\right)\right)$.

3. *The number of unlabeled examples drawn is* $O\left(d \cdot \left(\ln d + \ln \frac{1}{\delta} + \ln\ln \frac{1}{\epsilon}\right)^2 \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon}\right)$.

4. *The algorithm runs in time* $O\left(d^2 \cdot \left(\ln d + \ln \frac{1}{\delta} + \ln\ln \frac{1}{\epsilon}\right)^2 \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon}\right)$.

The theorem follows from Lemma 3 below, whose proof is similar to Lemma 2.

Lemma 3 (Modified-Perceptron under Adversarial Noise)**.** *Suppose Algorithm 2 has inputs labeling oracle $\mathcal{O}$ that satisfies ν-adversarial noise condition with respect to halfspace u, initial halfspace w_0 and angle upper bound $\theta \in (0, \frac{\pi}{2}]$ such that $\theta(w_0, u) \le \theta$, confidence δ, number of iterations $m = \Theta(d(\ln d + \ln \frac{1}{\delta}))$, band width $b = \Theta\left(\frac{\theta}{\sqrt{d}\ln(m/\delta)}\right)$. Additionally $\nu \le \Omega(\frac{\theta}{\ln(m/\delta))})$. Then with probability at least $1 - \delta$:*

1. *The output halfspace w_m is such that $\theta(w_m, u) \le \frac{\theta}{2}$.*
2. *The number of label queries is $O\left(d \cdot \left(\ln d + \ln \frac{1}{\delta}\right)\right)$.*
3. *The number of unlabeled examples drawn is $O\left(d \cdot \left(\ln d + \ln \frac{1}{\delta}\right)^2 \cdot \frac{1}{\theta}\right)$*
4. *The algorithm runs in time $O\left(d^2 \cdot \left(\ln d + \ln \frac{1}{\delta}\right)^2 \cdot \frac{1}{\theta}\right)$.*

6 Implications to Passive Learning

Active-Perceptron can be converted to a passive learning algorithm, Passive-Perceptron, for learning homogeneous halfspaces under the uniform distribution over the unit sphere. Passive-Perceptron has PAC sample complexities close to the lower bounds under the two noise models. We give a formal description of Passive-Perceptron in Appendix B. We give its formal guarantees in the corollaries below, which are immediate consequences of Theorems 2 and 3.

In the η-bounded noise model, the sample complexity of Passive-Perceptron improves over the state of the art result of [8], where a sample complexity of $\tilde{O}(\frac{d^{O(\frac{1}{(1-2\eta)^4})}}{\epsilon})$ is obtained. The bound has the same dependency on ϵ and d as the minimax upper bound of $\tilde{\Theta}(\frac{d}{\epsilon(1-2\eta)})$ by [49], which is achieved by a computationally inefficient ERM algorithm.

Corollary 1 (Passive-Perceptron under Bounded Noise)**.** *Suppose* Passive-Perceptron *has inputs distribution D that satisfies η-bounded noise condition with respect to u, initial halfspace v_0, target error ϵ, confidence δ, sample schedule $\{m_k\}$ where $m_k = \Theta\left(\frac{d}{(1-2\eta)^2}(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{k}{\delta})\right)$, band width $\{b_k\}$ where $b_k = \Theta\left(\frac{2^{-k}(1-2\eta)}{\sqrt{d}\ln(km_k/\delta)}\right)$. Then with probability at least $1 - \delta$: (1) The output halfspace v is such that $\mathrm{err}(h_v) \le \mathrm{err}(h_u) + \epsilon$; (2) The number of labeled examples drawn is $\tilde{O}\left(\frac{d}{(1-2\eta)^3\epsilon}\right)$. (3) The algorithm runs in time $\tilde{O}\left(\frac{d^2}{(1-2\eta)^3\epsilon}\right)$.*

In the ν-adversarial noise model, the sample complexity of Passive-Perceptron matches the minimax optimal sample complexity upper bound of $\tilde{\Theta}(\frac{d}{\epsilon})$ obtained in [39]. Same as in active learning, our algorithm has a faster running time than [39].

Corollary 2 (Passive-Perceptron under Adversarial Noise)**.** *Suppose* Passive-Perceptron *has inputs distribution D that satisfies ν-adversarial noise condition with respect to u, initial halfspace v_0, target error ϵ, confidence δ, sample schedule $\{m_k\}$ where $m_k = \Theta\left(d(\ln d + \ln \frac{k}{\delta})\right)$, band width $\{b_k\}$ where $b_k = \Theta\left(\frac{2^{-k}}{\sqrt{d}\ln(km_k/\delta)}\right)$. Furthermore $\nu = \Omega(\frac{\epsilon}{\ln\ln\frac{1}{\epsilon} + \ln\frac{d}{\delta}})$. Then with probability at least $1 - \delta$: (1) The output halfspace v is such that $\mathrm{err}(h_v) \le \mathrm{err}(h_u) + \epsilon$; (2) The number of labeled examples drawn is $\tilde{O}\left(\frac{d}{\epsilon}\right)$. (3) The algorithm runs in time $\tilde{O}\left(\frac{d^2}{\epsilon}\right)$.*

Tables 3 and 4 present comparisons between our results and results most closely related to ours.

Acknowledgments. The authors thank Kamalika Chaudhuri for help and support, Hongyang Zhang for thought-provoking initial conversations, Jiapeng Zhang for helpful discussions, and the anonymous reviewers for their insightful feedback. Much of this work is supported by NSF IIS-1167157 and 1162581.

Table 3: A comparison of algorithms for PAC learning halfspaces under the uniform distribution, in the η-bounded noise model.

Algorithm	Sample Complexity	Time Complexity
[8]	$\tilde{O}(\frac{d^{O(\frac{1}{(1-2\eta)^4})}}{\epsilon})$	$\tilde{O}(\frac{d^{O(\frac{1}{(1-2\eta)^4})}}{\epsilon})$
ERM [49]	$\tilde{O}(\frac{d}{(1-2\eta)\epsilon})$	superpoly$(d, \frac{1}{\epsilon})$
Our Work	$\tilde{O}(\frac{d}{(1-2\eta)^3\epsilon})$	$\tilde{O}(\frac{d^2}{(1-2\eta)^3} \cdot \frac{1}{\epsilon})$

Table 4: A comparison of algorithms for PAC learning halfspaces under the uniform distribution, in the ν-adversarial noise model where $\nu = \Omega(\frac{\epsilon}{\ln\ln\frac{1}{\epsilon}+\ln d})$.

Algorithm	Sample Complexity	Time Complexity
[39]	$\tilde{O}(\frac{d}{\epsilon})$	poly$(d, \frac{1}{\epsilon})$
ERM [57]	$\tilde{O}(\frac{d}{\epsilon})$	superpoly$(d, \frac{1}{\epsilon})$
Our Work	$\tilde{O}(\frac{d}{\epsilon})$	$\tilde{O}(\frac{d^2}{\epsilon})$

References

[1] Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. *ICML (3)*, 28: 1220–1228, 2013.

[2] Nir Ailon, Ron Begleiter, and Esther Ezra. Active learning using smooth relative regret approximations with applications. *Journal of Machine Learning Research*, 15(1):885–920, 2014.

[3] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, Apr 1988. ISSN 1573-0565. doi: 10.1023/A:1022873112823. URL `https://doi.org/10.1023/A:1022873112823`.

[4] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

[5] Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 724–733. IEEE, 1993.

[6] Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 449–458. ACM, 2014.

[7] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In *COLT*, pages 167–190, 2015.

[8] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of The 28th Conference on Learning Theory, COLT 2016*, 2016.

[9] M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *COLT*, 2013.

[10] M.-F. Balcan, A. Z. Broder, and T. Zhang. Margin based active learning. In *COLT*, 2007.

[11] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *J. Comput. Syst. Sci.*, 75(1): 78–89, 2009.

[12] Maria-Florina Balcan and Vitaly Feldman. Statistical active learning algorithms. In *NIPS*, pages 1295–1303, 2013.

[13] Maria-Florina Balcan and Hongyang Zhang. S-concave distributions: Towards broader distributions for noise-tolerant and sample-efficient learning algorithms. *arXiv preprint arXiv:1703.07758*, 2017.

[14] Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The true sample complexity of active learning. *Machine learning*, 80(2-3):111–139, 2010.

[15] A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *NIPS*, 2010.

[16] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Twenty-Sixth International Conference on Machine Learning*, 2009.

[17] Avrim Blum, Alan M. Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1998.

[18] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[19] Nicolò Cesa-Bianchi, Claudio Gentile, and erancesco Orabona. Robust bounds for classification via selective sampling. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 121–128, 2009.

[20] Lin Chen, Hamed Hassani, and Amin Karbasi. Near-optimal active learning of halfspaces via query synthesis in the noisy setting. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[21] David A. Cohn, Les E. Atlas, and Richard E. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[22] Nello Cristianini and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. 2000.

[23] Amit Daniely. Complexity theoretic limitations on learning halfspaces. *arXiv preprint arXiv:1505.05800*, 2015.

[24] S. Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.

[25] Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781, 2011.

[26] Sanjoy Dasgupta, Adam Tauman Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. In *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, pages 249–263, 2005.

[27] Sanjoy Dasgupta, Daniel Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems 20*, 2007.

[28] Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, 13(Sep):2655–2697, 2012.

[29] John Dunagan and Santosh Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 315–320. ACM, 2004.

[30] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 563–574. IEEE, 2006.

[31] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[32] Andrew Guillory, Erick Chastain, and Jeff Bilmes. Active learning as non-convex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 201–208, 2009.

[33] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[34] S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

[35] S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Carnegie Mellon University, 2009.

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

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

[38] Steve Hanneke and Liu Yang. Surrogate losses in passive and active learning. *arXiv preprint arXiv:1207.3772*, 2012.

[39] Steve Hanneke, Varun Kanade, and Liu Yang. Learning with a drifting target concept. In *International Conference on Algorithmic Learning Theory*, pages 149–164. Springer, 2015.

[40] D. Hsu. *Algorithms for Active Learning*. PhD thesis, UC San Diego, 2010.

[41] Tzu-Kuo Huang, Alekh Agarwal, Daniel Hsu, John Langford, and Robert E. Schapire. Efficient and parsimonious agnostic active learning. *CoRR*, abs/1506.08669, 2015.

[42] Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[43] Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[44] Adam Klivans and Pravesh Kothari. Embedding Hard Learning Problems Into Gaussian Space. In *APPROX/RANDOM 2014*, pages 793–809, 2014.

[45] Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(Dec):2715–2740, 2009.

[46] V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *JMLR*, 2010.

[47] Sanjeev R Kulkarni, Sanjoy K Mitter, and John N Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11(1):23–35, 1993.

[48] Philip M Long. On the sample complexity of pac learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

[49] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, pages 2326–2366, 2006.

[50] Claire Monteleoni. Efficient algorithms for general active learning. In *International Conference on Computational Learning Theory*, pages 650–652. Springer, 2006.

[51] TS Motzkin and IJ Schoenberg. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6(3):393–404, 1954.

[52] Francesco Orabona and Nicolo Cesa-Bianchi. Better algorithms for selective sampling. In *Proceedings of the 28th international conference on Machine learning (ICML-11)*, pages 433–440, 2011.

[53] Maxim Raginsky and Alexander Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems*, pages 1026–1034, 2011.

[54] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[55] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[56] Christopher Tosh and Sanjoy Dasgupta. Diameter-based active learning. In *ICML*, pages 3444–3452, 2017.

[57] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

[58] Liwei Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12(Jul):2269–2292, 2011.

[59] Yining Wang and Aarti Singh. Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition. In *AAAI*, 2016.

[60] Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 442–450, 2014.

[61] Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *COLT*, pages 1980–2022, 2017.