On the Power of Localized Perceptron for Label-Optimal Learning of Halfspaces with Adversarial Noise

Jie Shen [1]

Abstract

We study *online* active learning of homogeneous halfspaces in $\mathbb{R}^d$ with adversarial noise where the overall probability of a noisy label is constrained to be at most ν. Our main contribution is a Perceptron-like online active learning algorithm that runs in polynomial time, and under the conditions that the marginal distribution is isotropic log-concave and $\nu = \Omega(\epsilon)$, where $\epsilon \in (0,1)$ is the target error rate, our algorithm PAC learns the underlying halfspace with near-optimal label complexity of $\tilde{O}\big(d \cdot \mathrm{polylog}(\frac{1}{\epsilon})\big)$ and sample complexity of $\tilde{O}\big(\frac{d}{\epsilon}\big)$.[1] Prior to this work, existing online algorithms designed for tolerating the adversarial noise are subject to either label complexity polynomial in $\frac{1}{\epsilon}$, or suboptimal noise tolerance, or restrictive marginal distributions. With the additional prior knowledge that the underlying halfspace is s-sparse, we obtain attribute-efficient label complexity of $\tilde{O}\big(s \cdot \mathrm{polylog}(d, \frac{1}{\epsilon})\big)$ and sample complexity of $\tilde{O}\big(\frac{s}{\epsilon} \cdot \mathrm{polylog}(d)\big)$. As an immediate corollary, we show that under the agnostic model where no assumption is made on the noise rate ν, our active learner achieves an error rate of $O(\mathrm{OPT}) + \epsilon$ with the same running time and label and sample complexity, where OPT is the best possible error rate achievable by any homogeneous halfspace.

1. Introduction

In many practical applications, there are massive amounts of unlabeled data but labeling is expensive. This distinction has driven the study of active learning (Cohn et al., 1994; Balcan et al., 2007; Dasgupta, 2011; Hanneke, 2014; Awasthi et al., 2017), where labels are initially hidden and the learner must pay for each label it wishes to be revealed. The goal is to design querying strategies to avoid less informative labeling requests, e.g. the labels that can be inferred from previously seen samples. Parallel to active learning, online learning concerns the scenario where the learner observes a stream of samples and makes real-time model updating in order to compete with the best model obtained by seeing all the history data in a batch (Rosenblatt, 1958; Littlestone & Warmuth, 1989; Cesa-Bianchi et al., 1996; Zinkevich, 2003; Shalev-Shwartz, 2012; Hazan, 2019). Online learning algorithms have also been broadly investigated in machine learning, and have found various successful applications owing to its potential of savings in memory cost, low computational cost per sample, and its generalization ability (Cesa-Bianchi et al., 2004; Kakade & Tewari, 2008).

In this paper, we study the important problem of active learning of homogeneous halfspaces in the online setting, where the learner observes a stream of unlabeled data and makes spot decision of whether or not to query the labels. The goal is to achieve the best of the two worlds: label efficiency from active learning and computational efficiency from online learning. In this spectrum, there are a number of early works that share the same merit with this paper. For example, Freund et al. (1997) proposed a query-by-committee learning algorithm and Dasgupta et al. (2005) developed a Perceptron-like algorithm, both of which are implemented in an online fashion and enjoy a near-optimal label complexity bound of $\tilde{O}(d \log \frac{1}{\epsilon})$ where d is the dimension of the instance[2] and $\epsilon \in (0,1)$ is the target error rate. However, there are two crucial assumptions made by these works that seem too stringent: 1) the marginal distribution over the unlabeled data is uniform on the unit sphere in $\mathbb{R}^d$; and 2) there exists a perfect halfspace that incurs zero error rate with respect to the underlying distribution. In this regard, a natural question is: *can we design an online active learner, that provably works under a significantly more general family of marginal distributions while achieving arbitrarily small error rate without the realizability condition?*

To be more concrete, we are interested in designing an online active learning algorithm that PAC learns some underlying halfspace (Valiant, 1984) when the instances are drawn

[1]Stevens Institute of Technology, Hoboken, New Jersey, USA. Correspondence to: Jie Shen <jie.shen@stevens.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]We use the notation $\tilde{O}(f) := O(f \cdot \log f)$, $\tilde{\Omega}(f) := \Omega(f/\log f)$, and $\tilde{\Theta}(f)$ that is between $\tilde{\Omega}(f)$ and $\tilde{O}(f)$.

[2]We will interchangeably use "instance" and "unlabeled data".

from an isotropic log-concave distribution (Lovász & Vempala, 2007) and the labels are corrupted by the adversarial noise (Haussler, 1992; Kearns et al., 1992). It is worth mentioning that the family of isotropic log-concave distributions is a significant generalization of the uniform distribution since it includes a variety of prevalent distributions such as Gaussian, exponential, logistic. It is also known that establishing performance guarantees under this family is often technically subtle compared to that of uniform distribution due to the asymmetricity nature of log-concave distributions (Vempala, 2010; Balcan & Long, 2013; Diakonikolas et al., 2018). Returning to the noise model, we note that the adversarial noise, where the adversary may choose an *arbitrary* joint distribution such that the unlabeled data distribution is isotropic log-concave and the overall probability of a noisy label is constrained to be at most ν, is a realistic yet remarkably challenging regime, as suggested by many hardness results (Feldman et al., 2006; Guruswami & Raghavendra, 2009; Daniely, 2016; Diakonikolas et al., 2020a; Balcan & Haghtalab, 2020).

Under different assumptions on the adversarial noise rate (which might be suboptimal), a large body of works have established PAC guarantees for online learning with adversarial noise. Unfortunately, none of them resolves the aforementioned question in full. For example, under uniform marginal distributions, Theorem 3 of Kalai et al. (2005) showed that a surprisingly simple averaging scheme already is able to tolerate noise rate $\nu = \tilde{\Omega}(\epsilon)$, though the label and sample complexity are both $\tilde{O}(d^2/\epsilon^2)$.[3] The same algorithm was then revisited by Klivans et al. (2009) under isotropic log-concave distributions, with a worse noise tolerance of $\nu = \tilde{\Omega}(\epsilon^3)$. Very recently, Diakonikolas et al. (2020b) proposed a novel objective function by optimizing which with projected online gradient descent, one is guaranteed to tolerate the adversarial noise of $\nu = \tilde{\Omega}(\epsilon)$. Notably, their analysis applies to marginal distributions that are more general than isotropic log-concave. Compared to these passive learning algorithms which have label and sample complexity polynomial in $\frac{1}{\epsilon}$, the work of Yan & Zhang (2017) is more in line with this paper in the sense that they considered the active learning setting. Hence, their label complexity bound has an exponentially better dependence on $\frac{1}{\epsilon}$. However, the noise tolerance of Yan & Zhang (2017) reads as $\nu = \tilde{\Omega}(\epsilon/\log d)$ and their analysis is applicable only to uniform distributions due to the crucial need of the symmetricity of marginal distributions. As we can see in Table 1, none of the prior works subsumes others.

[3] Kalai et al. (2005) analyzed two algorithms: a (batch) polynomial regression and an online averaging. Here we are referring to the online averaging approach. We defer the comparison with the former to Section 4.1 when we are in the position to discuss the connection between adversarial noise and agnostic learning.

1.1. Main results

The main contribution of the paper is a novel online active learning algorithm that improves upon the state-of-the-art online algorithms. We introduce a few useful notations and informally describe our main results in this section; readers are referred to Section 4 for a precise statement.

Let $\mathcal{C}$ be the given concept class of halfspaces, D be the joint distribution over $\mathbb{R}^d \times \{-1, 1\}$, and for any $w \in \mathcal{C}$ define its error rate as $\mathrm{err}_D(w) = \Pr_{(x,y)\sim D}(\mathrm{sign}(w \cdot x) \neq y)$. Our analysis hinges on the following distributional assumptions.

Assumption 1. The unlabeled data distribution is isotropic log-concave, i.e. it has zero mean and unit covariance matrix, and the logarithm of its density function is concave.

Assumption 2. There exists an underlying halfspace $u \in \mathcal{C}$, such that $\mathrm{err}_D(u) \leq \nu$ for some noise rate $\nu \geq 0$.

Let $\epsilon \in (0, 1)$ be the target error rate given to the learner. We have the following theorem.

Theorem 1 (Informal)**.** *If Assumptions 1 and 2 are satisfied and $\nu \leq O(\epsilon)$, then there is an efficient online active learner that outputs a halfspace $\tilde{u} \in \mathcal{C}$ with $\Pr_{(x,y)\sim D}(\mathrm{sign}(\tilde{u} \cdot x) \neq \mathrm{sign}(u \cdot x)) \leq \epsilon$, with label complexity bound of $\tilde{O}\big(d \cdot \mathrm{polylog}(\frac{1}{\epsilon})\big)$ and sample complexity bound of $\tilde{O}\big(\frac{d}{\epsilon}\big)$.*

We compare with state-of-the-art online algorithms of Kalai et al. (2005); Yan & Zhang (2017); Diakonikolas et al. (2020b) that are tolerant to adversarial noise. Yan & Zhang (2017) presented an active learner and obtained analogous label and sample complexity to this work. However, their noise tolerance reads as $\nu = \tilde{\Omega}(\epsilon/\log d)$ while our algorithm is able to tolerate $\nu = \Omega(\epsilon)$; in addition, our results apply to significantly broader marginal distributions. Since Kalai et al. (2005); Diakonikolas et al. (2020b) considered passive learning, our active learning algorithm naturally enjoys label complexity that has exponentially better dependence on ϵ. Even for the sample complexity, we obtain improved dependence on d and ϵ. On the other side, it is worth mentioning that all these three algorithms run faster than our algorithm. Also, the analysis of Diakonikolas et al. (2020b) works under more general marginal distributions, in particular, distributions satisfying concentration, anti-concentration, anti-anti-concentration. Though our results can be generalized to their setting as well (see e.g. Zhang & Li (2021) for the treatment), we do not pursue it in the paper for the sake of clean presentation.

In addition to the properties aforementioned, we show that our algorithm can essentially incorporate attribute efficiency (Littlestone, 1987). That is, when the concept class consists of s-sparse halfspaces, the obtained label and sample complexity scale as $\tilde{O}(s \cdot \mathrm{polylog}(d))$. This characteristic is especially useful when there is limited availability of samples, a problem that has been studied for decades in

Table 1. Comparison to state-of-the-art online algorithms that are robust to adversarial noise. Prior online algorithms cannot even incorporate attribute efficiency. Even in the non-sparse case (i.e. $s = d$), our algorithm (Theorem 1) has better noise tolerance and works under more general distributions than Yan & Zhang (2017), and has improved label and sample complexity compared to Kalai et al. (2005) and Diakonikolas et al. (2020b).

Work	Log-concave?	Label complexity	Sample complexity	Noise tolerance
Theorem 3 of Kalai et al. (2005)	✗	$\tilde{O}(d^2/\epsilon^2)$	$\tilde{O}(d^2/\epsilon^2)$	$\nu = \tilde{\Omega}(\epsilon)$
Yan & Zhang (2017)	✗	$\tilde{O}(d \log \frac{1}{\epsilon})$	$\tilde{O}(d/\epsilon)$	$\nu = \tilde{\Omega}(\epsilon/\log d)$
Diakonikolas et al. (2020b)	✓	$\tilde{O}(d/\epsilon^4)$	$\tilde{O}(d/\epsilon^4)$	$\nu = \tilde{\Omega}(\epsilon)$
This work (Theorem 1)	✓	$\tilde{O}(d \cdot \mathrm{polylog}(\frac{1}{\epsilon}))$	$\tilde{O}(d/\epsilon)$	$\nu = \Omega(\epsilon)$
This work (Theorem 2)	✓	$\tilde{O}(s \cdot \mathrm{polylog}(d, \frac{1}{\epsilon}))$	$\tilde{O}(\frac{s}{\epsilon} \cdot \mathrm{polylog}(d))$	$\nu = \Omega(\epsilon)$

machine learning and statistics; see, e.g. Chen et al. (1998); Tibshirani (1996); Klivans & Servedio (2004); Candès & Tao (2005); Feldman (2007); Plan et al. (2017).

We have the following result for learning sparse halfspaces.

Theorem 2 (Informal)**.** *If Assumptions 1 and 2 are satisfied, $\nu \le O(\epsilon)$, and u is s-sparse, then there is an efficient online active learner that outputs an s-sparse halfspace $\tilde{u} \in \mathcal{C}$ with $\Pr_{(x,y)\sim D}(\mathrm{sign}(\tilde{u} \cdot x) \neq \mathrm{sign}(u \cdot x)) \le \epsilon$, with label complexity bound of $\tilde{O}\big(s \cdot \mathrm{polylog}(d, \frac{1}{\epsilon})\big)$ and sample complexity bound of $\tilde{O}\big(\frac{s}{\epsilon} \cdot \mathrm{polylog}(d)\big)$.*

Observe that Theorem 1 is a special case of the above by setting $s = d$. We note that the state-of-the-art online algorithms of Kalai et al. (2005); Yan & Zhang (2017); Diakonikolas et al. (2020b) do not enjoy attribute efficiency – it is yet nontrivial for them to encompass this property. Hence we have exponentially better dependence on the dimension d in label and sample complexity. See Table 1 for a summary of the comparison.

Finally, through an interesting observation made in Awasthi et al. (2017), it is possible to translate our main results for the adversarial noise model to the agnostic model of Haussler (1992); Kearns et al. (1992) where no assumption is made on the noise rate ν. Let $\mathrm{OPT} := \min_{w\in\mathcal{C}} \mathrm{err}_D(w)$.

Theorem 3 (Informal)**.** *If Assumption 1 is satisfied, then the algorithm tolerating the adversarial noise outputs a halfspace $\tilde{u}$ with $\mathrm{err}_D(\tilde{u}) \le O(\mathrm{OPT}) + \epsilon$, with same label and sample complexity as in Theorem 2.*

1.2. Overview of our techniques

Our algorithm is inspired in part by Zhang et al. (2020). We present an overview of our techniques below, and highlight the algorithmic connection to them as well as the novelty.

1) Active learning via stagewise online mirror descent. The first ingredient in our algorithm is a novel perspective of approaching active learning of halfspaces via stagewise online learning, recently utilized by Zhang et al. (2020) for learning halfspaces with benign noise. In each phase, given an initial halfspace $w_0 \in \mathbb{R}^d$, regardless of how the samples are generated, standard regret bound established for online mirror descent with linear loss $\langle w, \alpha g_t\rangle$ and ℓ_p-norm regularizer $\Phi(w)$ implies that the produced sequence of iterates $\{w_{t-1}\}_{t=1}^T$ must satisfy the following with certainty:

$$\frac{1}{T}\sum_{t=1}^{T}\langle u, -g_t\rangle \le \frac{1}{T}\sum_{t=1}^{T}\langle w_{t-1}, -g_t\rangle + \frac{\mathcal{B}_\Phi(u; w_0)}{\alpha T} + \frac{\alpha}{T}\sum_{t=1}^{T}\|g_t\|_q^2, \quad (1)$$

where $u \in \mathbb{R}^d$ is the underlying halfspace we aim to approximate, $\mathcal{B}_\Phi(\cdot;\cdot)$ denotes Bregman divergence induced by Φ, and $q \in (0, 1)$ is such that $\frac{1}{p} + \frac{1}{q} = 1$. Our goal is threefold: attribute efficiency, label efficiency, and small error rate. We will first specify $p \approx 1$ to achieve attribute efficiency which is a well-known technique in online learning (Grove et al., 2001; Gentile, 2003). In order to reduce the error rate of the initial halfspace w_0 with a few label queries, we need to design suitable gradients g_t and choose proper step size α and iteration number T such that a) the right-hand side of (1) is as small as angle $O(\theta(w_0, u))$; and b) the left-hand side is bounded from below by $\theta(\bar{w}, u)$ for certain halfspace $\bar{w}$ that depends on the sequence $\{w_{t-1}\}_{t=1}^T$. It is then possible to show that $\theta(\bar{w}, u) \le \frac{1}{2} \cdot \theta(w_0, u)$, and we can use $\bar{w}$ as the initial iterate for the next phase of online mirror descent to reduce the distance to u with geometric rate. Therefore, the crucial challenges lie in the design of g_t to accommodate specific noise model and an associated sampling scheme (since g_t depends on the sample). These are also the key technical differences between our work and Zhang et al. (2020), which we elaborate on below.

2) A semi-random gradient update for tolerating adversarial noise. In Zhang et al. (2020), the gradient g_t is heavily tailored to the bounded noise condition (Massart & Nédélec, 2006). We find it technically hard to reuse it for the adversarial noise since it is well known that the latter is a more involved noise model that will al-

ways violate the conditions assumed in the former.[4] Therefore, we consider an alternative yet fairly natural candidate: we choose g_t as the original gradient used in the Perceptron algorithm. That is, given the currently learned halfspace w_{t-1} and a new labeled sample (x_t, y_t), we set $g_t = -y_t x_t \cdot \mathbf{1}_{\{y_t \neq \mathrm{sign}(w_{t-1}\cdot x_t)\}}$. It remains to develop a plausible sampling scheme so that a) $-g_t$ will have non-trivial correlation with the underlying halfspace u; and b) only most informative instances are sampled for labeling. To this end, we propose a new sampling region $X_{\hat{w}_{t-1},b} := \{x \in \mathbb{R}^d : 0 < \hat{w}_{t-1} \cdot x \leq b\}$, where $\hat{w}_{t-1} = \frac{w_{t-1}}{\|w_{t-1}\|}$. Such time-varying region is different from active learning using empirical risk minimization (Balcan et al., 2007; Awasthi et al., 2017; Zhang, 2018) as in these works x is sampled from the full band $|\hat{w}_0 \cdot x| \leq b$. On one hand, using our sampling region leads to a linear loss $\langle w_{t-1}, g_t\rangle$ at most $O(b)$, while the band used by ERM would result in a loss of $O(b\sqrt{s}\log d)$. Observe that a tighter control on the loss implies tighter upper bound in (1). On the other hand, we discover that by restricting on querying the label of instances in $X_{\hat{w}_{t-1},b}$, we are reducing the randomness of model updating because now $g_t = x_t \cdot \mathbf{1}_{\{y_t \neq 1\}}$, i.e. we update the model only when the returned label $y_t = -1$. It turns out that such *semi-randomness* facilitates our control of the correlation between each $-g_t$ and the underlying halfspace u. We note that the semi-random updating rule is inspired by Yan & Zhang (2017), where they used a much narrower sampling region $\left\{x : \frac{b}{2\sqrt{d}} \leq \hat{w}_{t-1}\cdot x \leq \frac{b}{\sqrt{d}}\right\}$ and a carefully rescaled Perceptron gradient $g_t = (\hat{w}_{t-1}\cdot x_t)\cdot x_t \cdot \mathbf{1}_{\{y_t \neq 1\}}$ to accommodate their projection-free algorithm for learning under the uniform marginal distribution. In contrast, we incorporate different sampling strategy and gradients into (projected) mirror descent for PAC learning under isotropic log-concave marginal distributions.

3) A new characterization of the correlation between gradient and the underlying halfspace. Our last ingredient is applying localization in the concept space (Awasthi et al., 2017). Roughly speaking, before running online mirror descent, it is possible to construct an ℓ_2-ball where the underlying halfspace u resides in. Such trust region will be serving as the convex constraint set for online minimization. Using a novel analysis, we show that this interesting observation in allusion to the dedicated design of gradients implies that under the adversarial noise model,

$$\mathbb{E}_{(x_t,y_t)\sim D_{\hat{w}_{t-1},b}}[\langle u, -g_t\rangle] \geq f_{u,b}(w_{t-1}) - \beta\cdot\theta(w_0,u), \quad (2)$$

where we have the potential function

$$f_{u,b}(w_{t-1}) := \mathbb{E}_{(x_t,y_t)\sim D_{\hat{w}_{t-1},b}}\left[\,|u\cdot x_t|\cdot\mathbf{1}_{\{u\cdot x_t<0\}}\right], \quad (3)$$

and $\beta > 0$ is some quantity to be controlled. We argue that the function $f_{u,b}(w_{t-1})$ serves almost as a measure of $\theta(w_{t-1}, u)$; hence combining it with (1) we have that the average of $\theta(w_{t-1}, u)$ is upper bounded by $\frac{1}{2}\cdot\theta(w_0, u)$. This observation, in conjunction with a non-standard online-to-batch conversion, results in the desired halfspace $\bar{w}$. We note two different aspects compared to Zhang et al. (2020). First, our potential function $f_{u,b}$ is slightly distinct since we are considering a smaller sampling region. Second and more importantly, when deriving the lower bound for the correlation of u and $-g_t$, we carry out a more involved analysis but still incur a negative penalty $-\beta\cdot\theta(w_0,u)$ resulted from the adversarial noise model. In contrast, this term does not appear in Zhang et al. (2020) (since the bounded noise model they studied is more benign). Manipulating the penalty turns out to be subtle since if the factor β is large, there is no hope to upper bound the average of $f_{u,b}(w_{t-1})$ by $\frac{1}{2}\cdot\theta(w_0,u)$. We circumvent the technical issue by showing that β is dominated by ν/b which is small as soon as $\nu \leq c_0\epsilon$ for sufficiently small constant c_0 and b is carefully chosen to be greater than ϵ; see Lemma 13 in the appendix.

1.3. Related works

Label-efficient learning has also been broadly studied since gathering high quality labels is often expensive (Cohn et al., 1994; Dasgupta, 2005; 2011). The prominent approaches include disagreement-based active learning (Hanneke, 2011; 2014), margin-based active learning (Balcan et al., 2007; Balcan & Long, 2013; Awasthi et al., 2015), selective sampling (Cavallanti et al., 2011; Dekel et al., 2012), and adaptive one-bit compressed sensing (Zhang et al., 2014; Baraniuk et al., 2017). There are also a number of interesting works that appeal to extra information to mitigate the labeling cost, such as comparison (Xu et al., 2017; Kane et al., 2017; Hopkins et al., 2020; Shen & Zeng, 2020) and search (Beygelzimer et al., 2016).

Adversarial noise is closely related to the agnostic model, which was studied in Haussler (1992) and then coined out by Kearns et al. (1992). Under the uniform marginal distributions, Kalai et al. (2005) obtained the best error rate (see Section 4.1 for a precise statement), though the running time and sample complexity is $O(d^{1/\epsilon^4})$. This bound has been proved almost best possible in very recent works under the statistical query model (Diakonikolas et al., 2020a; Goel et al., 2020). Interestingly, Daniely (2015) characterized the tradeoff between the error rate and running time under the uniform marginal distribution by combining the techniques of polynomial regression (Kalai et al., 2005) and localization (Awasthi et al., 2017). By comprising on the error rate,

[4] In the bounded noise model, the adversary is constrained to flip the label of each given instance with probability at most $\eta \in [0, 1/2)$, which dramatically limits its power. In the adversarial noise, nevertheless, the adversary has the freedom to choose *any* joint distribution over the instance and label space.

Klivans et al. (2009) presented an averaging-based algorithm and showed how to boost it to tolerate an adversarial noise rate $\nu = \tilde{\Omega}(\epsilon^3)$ in polynomial time when the marginal distribution is isotropic log-concave. Such noise tolerance has been improved by a series of recent ERM-based works (Balcan et al., 2009; Beygelzimer et al., 2010; Zhang & Chaudhuri, 2014; Awasthi et al., 2016; 2017; Zhang, 2018; Diakonikolas et al., 2018), among which $\nu = \Omega(\epsilon)$ is the best known noise tolerance. However, solving an ERM often requires more memory storage and computational cost per sample than online methods (Shalev-Shwartz, 2007).

Achieving attribute efficiency has been a long-standing goal in machine learning and statistics (Blum, 1990; Blum et al., 1995), and has been pursued in online classification (Littlestone, 1987), learning decision lists (Servedio, 1999; Klivans & Servedio, 2004; Long & Servedio, 2006), compressed sensing (Donoho, 2006; Candès & Wakin, 2008; Tropp & Wright, 2010; Shen & Li, 2018), one-bit compressed sensing (Boufounos & Baraniuk, 2008; Plan & Vershynin, 2016), and variable selection (Fan & Li, 2001; Fan & Fan, 2008; Zhang, 2010; Shen & Li, 2017a;b; Wang et al., 2018). It is worth mentioning that Awasthi et al. (2016) gave a label-inefficient algorithm for uniformly learning sparse halfspaces with adversarial noise while this work and the closely related works consider non-uniform learning.

Roadmap. In Section 2, we give preliminaries and collect the notations used in the paper. In Section 3, we elaborate on our main algorithms. A theoretical analysis is given in Section 4, along with a proof sketch of the main results. We conclude this paper in Section 5, and defer the proof details to the appendix.

2. Preliminaries

We study PAC learning of sparse homogeneous halfspaces with adversarial noise, where the instance space is $\mathbb{R}^d$, the label space is $\{-1, 1\}$, and the concept class is $\mathcal{C} := \{x \mapsto \mathrm{sign}(w \cdot x) : w \in \mathbb{R}^d,\ \|w\| = 1,\ \|w\|_0 \le s\}$. Here, $\|w\|$ denotes the ℓ_2-norm and $\|w\|_0$ counts the number of non-zero elements in w. Observe that we say a halfspace is non-sparse if $s = d$. An adversary EX with adversarial noise works as follows: it first chooses an arbitrary joint distribution D over $\mathbb{R}^d \times \{-1, 1\}$; the distribution D is then fixed throughout learning. Let D_X denote the marginal distribution over the instance space, which is promised to belong to a family of well-behaved distributions $\mathcal{D}_{\mathcal{X}}$; in this paper it is assumed to be isotropic log-concave (Assumption 1).

A learner is given the instance and label space, the concept class $\mathcal{C}$, the family of distributions $\mathcal{D}_{\mathcal{X}}$ (but not D_X), a target error rate $\epsilon \in (0, 1)$ and a failure confidence $\delta \in (0, 1)$, and the goal is to output in polynomial time a halfspace $\tilde{u} \in \mathcal{C}$ such that with probability at least $1 - \delta$, $\Pr_{(x,y)\sim D}(\mathrm{sign}(\tilde{u} \cdot x) \neq \mathrm{sign}(u \cdot x)) \le \epsilon$. In the passive learning setting, the learner is given access to a sample generation oracle EX which returns a labeled sample $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ randomly drawn from the distribution D. Since we want to optimize the label complexity, we will consider a natural extension: when the learner makes a call to EX, a labeled sample (x, y) is randomly drawn but only the instance x is returned. The learner must make a separate call to a label revealing oracle EX_y to obtain the label y. We refer to the total number of calls to EX as the sample complexity of the algorithm, and that of EX_y as the label complexity.

In our active learning algorithm, we will often want to draw instances from D_X conditioned on a region $X_{\hat{w},b} := \{x \in \mathbb{R}^d : 0 < \hat{w} \cdot x \le b\}$ where $\hat{w} \in \mathbb{R}^d$ and $b > 0$ are given; this can be done by rejection sampling, where we repeatedly call EX until seeing an instance x that falls in $X_{\hat{w},b}$. We will refer to $X_{\hat{w},b}$ as sampling region, and b is called band width. We denote by $D_{X|\hat{w},b}$ (respectively $D_{\hat{w},b}$) the distribution D_X (respectively D) conditioned on the event that $x \in X_{\hat{w},b}$.

Let w be a vector in $\mathbb{R}^d$. We will frequently use $\hat{w}$ to denote its ℓ_2-normalization $\frac{w}{\|w\|}$. For a scalar $\gamma \ge 1$, we denote by $\|w\|_\gamma$ the ℓ_γ-norm of w. Let $s > 0$ be an integer less than d. The hard thresholding operation $\mathcal{H}_s(w)$ zeros out all but the s largest (in magnitude) entries in w. For two vectors w and w', we write $\theta(w, w')$ for the angle between them.

We reserve p and q for specific values: $p = \frac{\ln(8d)}{\ln(8d)-1}$ and $q = \ln(8d)$ (note that $\frac{1}{p} + \frac{1}{q} = 1$). We will use the ℓ_p-norm regularizer, that is, $\Phi(w) = \frac{1}{2(p-1)}\|w - v\|_p^2$ for some given vector v. It is known that $\Phi(w)$ is 1-strongly convex with respect to the ℓ_p-norm (Shalev-Shwartz, 2007, Lemma 17). We denote the Bregman divergence induced by $\Phi(w)$ by $\mathcal{B}_\Phi(w; w') := \Phi(w) - \Phi(w') - \langle \nabla\Phi(w'), w - w' \rangle$. Observe that $\mathcal{B}_\Phi(w; v) = \Phi(w)$ where v is the reference vector appearing in $\Phi(w)$.

We will sometimes phrase our theoretical guarantee in terms of angles between two halfspaces. The following lemma, due to Balcan & Long (2013), is useful to convert the guarantee of angles to that of error rate.

Lemma 4. *There exists an absolute constant $\bar{c} > 0$ such that the following holds. Let D_X be an isotropic log-concave distribution. For any two vectors w and $w' \in \mathbb{R}^d$, $\Pr_{x\sim D_X}(\mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w' \cdot x)) \le \bar{c} \cdot \theta(w, w')$.*

We remark that a closely related noise model is the agnostic model (Haussler, 1992; Kearns et al., 1992; Kalai et al., 2005), where only Assumption 1 is satisfied and the goal is to output a halfspace that approximates the best halfspace in $\mathcal{C}$. The results from our theorems under the adversarial noise model translate immediately into PAC guarantees for the agnostic model; see Section 4.1 for more details.

3. Main Algorithm

We present our online active learning algorithm in Algorithm 1, which consists of two major stages: initialization and refinement. In the initialization stage, the goal is to find a halfspace v_0 that has a constant acute angle with the underlying halfspace u. It will then be used as a warm start for the refinement stage, where the procedure REFINE is repeatedly invoked to cut off the angle with u by half in each phase k. Therefore, after K phases of refinement, we will obtain a halfspace v_K satisfying $\theta(v_K, u) = 2^{-K} = O(\epsilon)$, which by Lemma 4 implies that v_K has small error rate with respect to the underlying halfspace u defined in Assumption 2.

Since INITIALIZE invokes REFINE as well, we will introduce the latter first. Generally speaking, the REFINE algorithm, i.e. Algorithm 3, belongs to the family of online mirror descent algorithms with Perceptron gradient and ℓ_p-norm regularization. The crucial ingredients that make it attribute and label efficient are a carefully crafted constraint set, a time-varying sampling region, and semi-random gradients, as we described in Section 1.2. In particular, the constraint set $\mathcal{K}$ is constructed in such a way that the underlying halfspace u is guaranteed to stay in it (with overwhelming probability). Thus, it serves as a trust region into which all the iterates w_t are projected back. It is worth mentioning that we did not put an ℓ_1-norm constraint in $\mathcal{K}$; this is because we already have utilized the ℓ_p-norm regularization to simultaneously guarantee attribute efficiency (Grove et al., 2001; Gentile, 2003) and the stability of online minimization (Shalev-Shwartz, 2012; Orabona, 2019). The second component in REFINE is the time-varying sampling region $X_{\hat{w}_{t-1},b} = \{x \in \mathbb{R}^d : 0 < \hat{w}_{t-1} \cdot x \le b\}$, which results in a linear loss as small as $O(b)$ in each iteration. In contrast, a naive online approach to simulate the ERM algorithm of Zhang (2018) would lead to a loss as large as $O(b\sqrt{s}\log d)$. The idea of using time-varying sampling paradigm has appeared in a few online active learning algorithms (Yan & Zhang, 2017; Zhang et al., 2020). Ours is less restrictive than the one in Yan & Zhang (2017), and is more dedicated to the much more challenging adversarial noise compared to the bounded noise model considered in Zhang et al. (2020). Along with the new sampling region is a semi-random Perceptron gradient. Recall that the original gradient used in Perceptron is given by $g_t = -y_t x_t \cdot \mathbf{1}_{\{y_t \neq \mathrm{sign}(w_{t-1}\cdot x_t)\}}$. Since $x_t \in X_{\hat{w}_{t-1},b}$, we update the model only when the label y_t returned by the adversary equals -1, thus inducing the gradient displayed in Algorithm 3. Finally, after running mirror descent for T iterations, we perform an averaging scheme followed by hard thresholding, to ensure that the output $\tilde{w}$ belongs the to concept class. We remark that the running time of REFINE is polynomial in d, since in each iteration t, updating the model requires solving a convex program. Regarding the label complexity, it is exactly equal to the total iteration number T. We also remark that obtaining x_t can be done by calling EX for $O(1/b)$ times since the probability mass of $X_{\hat{w}_{t-1},b}$ on D_X is $\Theta(b)$; see Lemma 29.

Algorithm 1 Main Algorithm

Require: Target error rate $\epsilon \in (0,1)$, failure probability $\delta \in (0,1)$, sparsity s.

Ensure: Halfspace $\tilde{u} \in \mathbb{R}^d$ such that $\Pr_{(x,y)\sim D}(\mathrm{sign}(\tilde{u}\cdot x) \neq \mathrm{sign}(u\cdot x)) \le \epsilon$.

1: $v_0 \leftarrow$ INITIALIZE$(\frac{\delta}{2}, s)$.
2: $K \leftarrow \lceil \log \frac{\bar{c}\pi}{8\epsilon} \rceil$ where $\bar{c}$ is defined in Lemma 4.
3: **for** $k = 1, 2, \ldots, K$ **do**
4: $v_k \leftarrow$ REFINE$(v_{k-1}, \frac{\delta}{2k(k+1)}, s, \alpha_k, b_k, \mathcal{K}_k, \Phi_k, T_k)$, where step size $\alpha_k = \tilde{\Theta}\left(2^{-k}\cdot\left(\log\frac{d\cdot k^2\cdot 2^k}{\delta}\right)^{-2}\right)$, band width $b_k = \Theta\left(2^{-k}\right)$, constraint set
$$\mathcal{K}_k = \left\{w : \|w - v_{k-1}\| \le \pi\cdot 2^{-k-2},\ \|w\| \le 1\right\},$$
regularizer $\Phi_k(w) = \frac{1}{2(p-1)}\|w - v_{k-1}\|_p^2$, number of iterations $T_k = \tilde{O}\left(s\log d\cdot\left(\log\frac{d\cdot k^2\cdot 2^k}{\delta}\right)^2\right)$.
5: **end for**
6: **return** $\tilde{u} \leftarrow v_K$.

Algorithm 2 INITIALIZE

Require: Failure probability δ', sparsity s.

Ensure: An s-sparse halfspace v_0 such that $\theta(v_0, u) \le \frac{\pi}{8}$.

1: $(x_1, y_1), \ldots, (x_m, y_m) \leftarrow$ call EX to draw m instances, and query EX$_y$ for their labels, where $m = O(s\log\frac{d}{\delta'})$.
2: Compute $w_{\mathrm{avg}} = \frac{1}{m}\sum_{i=1}^m y_i x_i$.
3: Let $w^\sharp = \frac{\mathcal{H}_{\tilde{s}}(w_{\mathrm{avg}})}{\|\mathcal{H}_{\tilde{s}}(w_{\mathrm{avg}})\|}$, where $\tilde{s} = 81\cdot 2^{40}s$.
4: Let $\mathcal{K} = \left\{w \in \mathbb{R}^d : \|w\| \le 1,\ w\cdot w^\sharp \ge \frac{1}{9\cdot 2^{20}}\right\}$ and find a point $w_0 \in \mathcal{K} \cap \left\{w \in \mathbb{R}^d : \|w\|_1 \le \sqrt{s}\right\}$.
5: **return** $v_0 \leftarrow$ REFINE$(w_0, \frac{\delta'}{2}, s, \alpha, b, \mathcal{K}, \Phi, T)$, where step size $\alpha = \tilde{\Theta}\left(\log^{-2}\frac{d}{\delta'}\right)$, band width $b = \frac{1}{81\cdot 2^{22}}$, regularizer $\Phi(w) = \frac{1}{2(p-1)}\|w - w_0\|_p^2$, and number of iterations $T = \tilde{O}\left(s\log d\cdot\left(\log\frac{d}{\delta'}\right)^2\right)$.

Now we elaborate on the INITIALIZE algorithm, namely Algorithm 2. Technically speaking, one important condition for the success of our analysis is that an overwhelming portion of the iterates must have acute angles with the underlying halfspace u. Therefore, the hypothesis testing approach proposed in Awasthi et al. (2017) does not work out in our case since we will lose control of the intermediate iterates. To circumvent the technical challenge, we tailor the averaging-based initialization scheme of Zhang et al. (2020) to the adversarial noise model. It is possible to show that as far as the noise rate ν is low, w_{avg} has a positive correlation with u, and performing hard thresholding almost preserves it, i.e. $u \cdot w^\sharp = \Omega(1)$. Therefore, we obtain a

Algorithm 3 REFINE

Require: Initial s-sparse halfspace w_0, failure probability δ', sparsity s, step size α, band width b, convex constraint set $\mathcal{K}$, regularization function $\Phi : \mathbb{R}^d \to [0, +\infty)$, number of iterations T.

Ensure: Refined s-sparse halfspace $\tilde{w}$ such that $\theta(\tilde{w}, u) \leq \frac{1}{2} \cdot \theta(w_0, u)$.

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Call EX to obtain an instance x_t in $X_{\hat{w}_{t-1},b}$, and query EX_y for its label y_t (recall that $\hat{w}_{t-1}$ is the ℓ_2-normalization of w_{t-1}).
3: $w_t \leftarrow \arg\min_{w \in \mathcal{K}} \langle w, \alpha g_t \rangle + \mathcal{B}_\Phi(w; w_{t-1})$, where $g_t = x_t \cdot \mathbf{1}_{\{y_t = -1\}}$.
4: **end for**
5: $\bar{w} \leftarrow \frac{1}{T} \sum_{t=1}^{T} \hat{w}_t$.
6: **return** $\tilde{w} \leftarrow \frac{\mathcal{H}_s(\bar{w})}{\|\mathcal{H}_s(\bar{w})\|}$.

good reference vector $w^\sharp$ which is guaranteed to have an acute angle with u. Based on an enhanced constraint set that takes the correlation into consideration, we are able to show that most of the iterates are admissible, and hence our analysis of the REFINE algorithm can be reused to show that the output v_0 will have a small acute angle with u. Note that we are not making efforts to optimize the constants that appear in the INITIALIZE algorithm; in practice, we believe that our algorithm works under reasonable constants. It is also worth mentioning that Zhang & Li (2021) recently developed a simpler initialization scheme for the problem of learning halfspaces with Massart or Tsybakov noise; it will be interesting to adapt their approach to the adversarial noise model as a future work.

4. Performance Guarantee

We are now in the position to state our main theorem, which is a formal statement of Theorem 2.

Theorem 5 (Main result)**.** *Suppose that Assumptions 1 and 2 are satisfied. If $\nu \leq c_0 \epsilon$ for some small absolute constant $c_0 > 0$, then with probability $1-\delta$, the output of Algorithm 1, $\tilde{u}$, satisfies $\Pr_{(x,y)\sim D}\big(\mathrm{sign}(\tilde{u} \cdot x) \neq \mathrm{sign}(u \cdot x)\big) \leq \epsilon$. Moreover, the running time is $\mathrm{poly}(d, \frac{1}{\epsilon}, \log\frac{1}{\delta})$, the label complexity is $\tilde{O}\big(s \cdot \mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})\big)$, and the sample complexity is $\tilde{O}\big(\frac{s}{\epsilon} \cdot \mathrm{polylog}(d, \frac{1}{\delta})\big)$.*

Remark 6. A more concrete label and sample complexity reads as $\tilde{O}(s \log d \cdot \log^3 \frac{d}{\epsilon\delta})$ and $\tilde{O}\big(\frac{s}{\epsilon} \cdot \log^4 \frac{d}{\delta}\big)$, respectively; see Theorem 18 and Theorem 24 in the appendix.

Remark 7 (Excess risk)**.** By the triangle inequality, we have $\mathrm{err}_D(\tilde{u}) - \mathrm{err}_D(u) \leq \Pr_{(x,y)\sim D}(\mathrm{sign}(\tilde{u} \cdot x) \neq \mathrm{sign}(u \cdot x)) \leq \epsilon$. Namely, the excess risk of $\tilde{u}$ with respect to u is at most ϵ over the underlying distribution.

Remark 8 (Implications to passive learning)**.** It is possible to convert our algorithm to an online passive learner, where there is only one oracle EX that always returns a labeled instance upon request. To this end, observe that the active learner interacts with the oracle exclusively in Step 2 of REFINE. Therefore, we only need to modify this step as follows in the passive setting: repeatedly call EX to obtain a sequence of labeled instances $\{x_i, y_i\}_{i \geq 1}$ until seeing a pair (x_t, y_t) such that $x_t \in X_{\hat{w}_{t-1},b}$. Then we use (x_t, y_t) to update the classifier in Step 3 (rather than using all the labeled instances that are drawn from EX). It is easy to see that the label and sample complexity of the passive learner are both $\tilde{O}\big(\frac{s}{\epsilon} \cdot \mathrm{polylog}(d, \frac{1}{\delta})\big)$.

The following corollary, which is a formal statement of Theorem 1, concerns learning of non-sparse halfspaces and is an immediate application of Theorem 5 by setting $s = d$.

Corollary 9. *Assume same conditions as in Theorem 5. With probability $1 - \delta$, $\Pr_{(x,y)\sim D}\big(\mathrm{sign}(\tilde{u} \cdot x) \neq \mathrm{sign}(u \cdot x)\big) \leq \epsilon$. Moreover, the running time is $\mathrm{poly}(d, \frac{1}{\epsilon}, \log\frac{1}{\delta})$, the label complexity is $\tilde{O}\big(d \cdot \mathrm{polylog}(\frac{1}{\epsilon}, \frac{1}{\delta})\big)$, and the sample complexity is $\tilde{O}\big(\frac{d}{\epsilon} \cdot \mathrm{polylog}(\frac{1}{\delta})\big)$.*

4.1. Implications to agnostic learning

In the agnostic model (Haussler, 1992; Kearns et al., 1992), the adversary chooses a joint distribution D over $\mathbb{R}^d \times \{-1, 1\}$ and fixes it throughout the learning process. Let $\mathrm{OPT} = \min_{w \in \mathcal{C}} \mathrm{err}_D(w)$. The goal of the learner is to output a hypothesis $\tilde{u}$ such that $\mathrm{err}_D(\tilde{u}) \leq c \cdot \mathrm{OPT} + \epsilon$ for some approximation factor $c \geq 1$. The crucial difference from the adversarial noise model is that now Assumption 2 may not be satisfied (in other words, OPT can be very large compared to the target error rate ϵ).

Kalai et al. (2005) developed a polynomial regression algorithm that achieves approximation guarantee with $c = 1$, where the computational and sample complexity are both $O(d^{2^{\mathrm{poly}(1/\epsilon)}})$ for learning under isotropic log-concave distributions and are $O(d^{1/\epsilon^4})$ for uniform distributions.

On the other side, Kalai et al. (2005) and a number of recent works also obtained weaker (yet still quite nontrivial) approximation guarantee of $O(\mathrm{OPT}) + \epsilon$ with running time and sample complexity polynomial in d and $\frac{1}{\epsilon}$; see, for example, Awasthi et al. (2017); Zhang (2018); Diakonikolas et al. (2018; 2020b). We follow this line, and remark that our main result, Theorem 5, can be immediately translated into the constant approximation guarantee under the agnostic model. In fact, Lemma C.1 of Awasthi et al. (2017) made an interesting observation that if an algorithm can tolerate adversarial noise of $\nu = \Omega(\epsilon)$, then it essentially achieves error rate of $O(\mathrm{OPT}) + \epsilon$ in the agnostic model, with the same running time, label complexity, and sample complexity (up to a constant multiplicative factor). We therefore have the following result, which is a formal statement of

Theorem 3, by combining Theorem 5 we established and Lemma C.1 of Awasthi et al. (2017); we omit the proof since it is fairly straightforward.

Corollary 10. *Suppose that Assumption 1 is satisfied. Then with probability* $1-\delta$, $\mathrm{err}_D(\tilde{u}) \le c \cdot \mathrm{OPT} + \epsilon$ *for some absolute constant* $c > 1$. *Moreover, the running time is* $\mathrm{poly}(d, \frac{1}{\epsilon})$, *the label complexity is* $\tilde{O}\big(s \cdot \mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})\big)$, *and the sample complexity is* $\tilde{O}\big(\frac{s}{\epsilon} \cdot \mathrm{polylog}(d, \frac{1}{\delta})\big)$.

4.2. Proof of Theorem 5

Theorem 5 hinges on the following two important results, characterizing the performance of INITIALIZE and REFINE respectively. A more precise statement and a detailed proof can be found in Appendix A in the supplementary material.

Theorem 11. *Consider the* INITIALIZE *algorithm. If Assumptions 1 and 2 are satisfied and* $\nu \le c_o\epsilon$, *then with probability* $1-\delta'$, *the output of* INITIALIZE, v_0, *is such that* $\theta(v_0, u) \le \frac{\pi}{8}$. *The running time is* $\mathrm{poly}(d, \log\frac{1}{\delta'})$, *the label complexity is* $\tilde{O}\big(s \cdot \mathrm{polylog}(d, \frac{1}{\delta'})\big)$, *and the sample complexity is* $\tilde{O}\big(s \cdot \mathrm{polylog}(d)\big)$.

Theorem 12. *Consider the* REFINE *algorithm. Suppose that Assumptions 1 and 2 are satisfied and* $\nu \le c_o\epsilon$. *Then with probability* $1-\delta'$, *the output of the* REFINE *algorithm,* $\tilde{w}$, *satisfies* $\theta(\tilde{w}, u) \le \frac{1}{2} \cdot \theta(w_0, u)$. *The running time of* REFINE *is* $T \cdot \mathrm{poly}(d, \frac{1}{b}, \log\frac{1}{\delta'})$, *the label complexity is* T, *and the sample complexity is* $O(T/b + T\log\frac{T}{\delta'})$ *where* T *and* b *are the inputs to* REFINE.

We first explain the results in Theorem 12. The label complexity of REFINE is straightforward since we request one label per iteration, and the total number of iterations is T. We start with the analysis of the sample complexity of REFINE, i.e. the number of calls to EX in Step 3 therein. Since the marginal distribution is assumed to be isotropic log-concave, Lemma 29 shows that $\Pr_{x_t\sim D_X}(x_t \in X_{\hat{w}_{t-1},b}) \ge c_2 b$ for some absolute constant $c_2 > 0$. Thus, by Chernoff bound, we need to call EX for $O(\frac{1}{b} + \log\frac{T}{\delta'})$ times in order to obtain one x_t in the band with probability $1 - \frac{\delta'}{2T}$. Thus, by union bound over the T iterations in REFINE, with probability $1-\frac{\delta'}{2}$, the total number of calls to EX is $O(\frac{T}{b} + T\log\frac{T}{\delta'})$. Note that this is also the computational cost for sampling. On the other side, updating the iterates, i.e. Step 3, involves solving a convex program in $\mathbb{R}^d$, which has a running time polynomial in d per iteration. Thus, the overall computational cost of REFINE is $T \cdot \mathrm{poly}(d, \frac{1}{b}, \log\frac{1}{\delta'})$.

Now we explain the results in Theorem 11. Generally speaking, the INITIALIZE algorithm consists of two major steps, one for constructing the constraint set $\mathcal{K}$ and one for obtaining a good initial halfspace v_0 based on $\mathcal{K}$. To obtain $\mathcal{K}$, it consumes m labeled instances and the computational cost for sampling them is $O(m)$ since rejection sampling is not needed. Then it aims to find a point w_0 in a convex set, which is polynomial-time solvable; in fact, we can set w_0 to the zero vector and then project it onto $\mathcal{K}$, corresponding to solving a convex program. The second major step is to invoke REFINE, for which we have just analyzed. Combining these observations and the parameters specified in the INITIALIZE algorithm, we obtain the announced results.

Proof of Theorem 5. First, Theorem 11 implies $\theta(v_0, u) \le \frac{\pi}{8}$ with probability $1-\frac{\delta}{2}$. In addition, for any phase k in Algorithm 1, we specify in Theorem 12 that $w_0 = v_{k-1}$ and $\tilde{w} = v_k$, and obtain that $\theta(v_k, u) \le \frac{1}{2} \cdot \theta(v_{k-1}, u)$ with probability $1 - \frac{\delta}{2k(k+1)}$. By telescoping, we get $\theta(v_K, u) \le 2^{-K} \cdot \frac{\pi}{8} \le \epsilon/\bar{c}$ in light of our setting of K; this inequality holds with probability $1 - \frac{\delta}{2} - \sum_{k=1}^{K} \frac{\delta}{2k(k+1)} \ge 1-\delta$ by union bound. This in allusion to Lemma 4 gives the desired error rate of $\tilde{u} = v_K$ with respect to u.

The running time of Algorithm 1 follow from those we analyzed for INITIALIZE and REFINE, and from the hyperparameter settings on b_k, T_k, and δ_k in each phase k. In particular, observe that $b_k \ge \epsilon$ for all $k \le K$. Therefore, the running time is given by $\mathrm{poly}(d, \log\frac{1}{\delta}) + \sum_{k=1}^{K} T_k \cdot \mathrm{poly}(d, \frac{1}{b_k}, \log\frac{k^2}{\delta}) = \mathrm{poly}(d, \frac{1}{\epsilon}, \log\frac{1}{\delta})$.

Likewise, for label complexity and sample complexity, we can add up the cost in the initialization stage and that of the K phases of refinement to obtain the bounds as claimed. □

5. Conclusion and Future Works

This paper studies the fundamental problem of learning halfspaces with adversarial noise. We have presented the first attribute-efficient, label-efficient, and noise-tolerant algorithm in the online setting, under the general isotropic log-concave marginal distributions. Prior to this work, existing online learners are either subject to label inefficiency or suboptimal noise tolerance, or work under restrictive marginal distributions. We have shown that our label and sample complexity are near-optimal, and the learner achieves PAC guarantee in polynomial time. Prior to this work, such performance guarantee is only achieved by a very recent batch algorithm. Our results also have immediate implications to the agnostic model, and match the best known results obtained by polynomial-time batch algorithms.

We discuss a few important directions for future investigation. First, it is interesting to develop online PAC algorithms with $\mathrm{OPT}+\epsilon$ approximation error under the agnostic model, by leveraging, for example, the polynomial regression technique (Kalai et al., 2005) into the online mirror descent framework. Second, it is useful to extend the analysis to more general concept classes such as intersections of halfspaces (Klivans et al., 2002; Diakonikolas et al., 2018). It will also be important to design PAC algorithms that leverage additional types of queries such as pairwise comparison (Kane

et al., 2017; Xu et al., 2017) in the scenario where labels are extremely demanding (e.g. medical data), or to develop new projection-free algorithms for even faster computation.

Acknowledgements

We thank Jing Wang for valuable feedback on the merit of the work, and thank the anonymous reviewers for helpful suggestions on improving the presentation. This work is supported by NSF-IIS-1948133 and the startup funding of Stevens Institute of Technology.

References

Awasthi, P., Balcan, M., Haghtalab, N., and Urner, R. Efficient learning of linear separators under bounded noise. In *Proceedings of the 28th Annual Conference on Learning Theory*, pp. 167–190, 2015.

Awasthi, P., Balcan, M., Haghtalab, N., and Zhang, H. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of the 29th Annual Conference on Learning Theory*, pp. 152–192, 2016.

Awasthi, P., Balcan, M., and Long, P. M. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM*, 63(6):50:1–50:27, 2017.

Balcan, M. and Haghtalab, N. Noise in classification. In Roughgarden, T. (ed.), *Beyond the Worst-Case Analysis of Algorithms*, pp. 361–381. Cambridge University Press, 2020.

Balcan, M. and Long, P. M. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the 26th Annual Conference on Learning Theory*, pp. 288–316, 2013.

Balcan, M., Broder, A. Z., and Zhang, T. Margin based active learning. In *Proceedings of the 20th Annual Conference on Learning Theory*, pp. 35–50, 2007.

Balcan, M., Beygelzimer, A., and Langford, J. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

Baraniuk, R. G., Foucart, S., Needell, D., Plan, Y., and Wootters, M. Exponential decay of reconstruction error from binary measurements of sparse signals. *IEEE Transactions on Information Theory*, 63(6):3368–3385, 2017.

Beygelzimer, A., Hsu, D. J., Langford, J., and Zhang, T. Agnostic active learning without constraints. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, pp. 199–207, 2010.

Beygelzimer, A., Hsu, D. J., Langford, J., and Zhang, C. Search improves label for active learning. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, pp. 3342–3350, 2016.

Blum, A. Learning boolean functions in an infinite attribute space. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing*, pp. 64–72, 1990.

Blum, A., Hellerstein, L., and Littlestone, N. Learning in the presence of finitely or infinitely many irrelevant attributes. *Journal of Computer and System Sciences*, 50 (1):32–40, 1995.

Boufounos, P. and Baraniuk, R. G. 1-bit compressive sensing. In *Proceedings of the 42nd Annual Conference on Information Sciences and Systems*, pp. 16–21, 2008.

Candès, E. J. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Candès, E. J. and Wakin, M. B. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2): 21–30, 2008.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. Learning noisy linear classifiers via adaptive and selective sampling. *Machine Learning*, 83(1):71–102, 2011.

Cesa-Bianchi, N., Long, P. M., and Warmuth, M. K. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7:604–619, 1996.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Chen, S. S., Donoho, D. L., and Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

Cohn, D., Atlas, L., and Ladner, R. Improving generalization with active learning. *Machine Learning*, 15(2): 201–221, 1994.

Daniely, A. A PTAS for agnostically learning halfspaces. In *Proceedings of the 28th Annual Conference on Learning Theory*, volume 40, pp. 484–502, 2015.

Daniely, A. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pp. 105–117, 2016.

Dasgupta, S. Coarse sample complexity bounds for active learning. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, pp. 235–242, 2005.

Dasgupta, S. Active learning. *Encyclopedia of Machine Learning*, 2011.

Dasgupta, S., Kalai, A. T., and Monteleoni, C. Analysis of Perceptron-based active learning. In *Proceedings of the 18th Annual Conference on Learning Theory*, pp. 249–263, 2005.

Dekel, O., Gentile, C., and Sridharan, K. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, 13:2655–2697, 2012.

Diakonikolas, I., Kane, D. M., and Stewart, A. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM Symposium on Theory of Computing*, pp. 1061–1073, 2018.

Diakonikolas, I., Kane, D., and Zarifis, N. Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under gaussian marginals. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, pp. 13586–13596, 2020a.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Non-convex SGD learns halfspaces with adversarial label noise. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, pp. 18540–18549, 2020b.

Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Fan, J. and Fan, Y. High dimensional classification using features annealed independence rules. *Annals of Statistics*, 36(6):2605–2637, 2008.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Feldman, V. Attribute-efficient and non-adaptive learning of parities and DNF expressions. *Journal of Machine Learning Research*, 8:1431–1460, 2007.

Feldman, V., Gopalan, P., Khot, S., and Ponnuswami, A. K. New results for learning noisy parities and halfspaces. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pp. 563–574, 2006.

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

Gentile, C. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

Goel, S., Gollakota, A., and Klivans, A. R. Statistical-query lower bounds via functional gradients. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020.

Grove, A. J., Littlestone, N., and Schuurmans, D. General convergence results for linear discriminant updates. *Machine Learning*, 43(3):173–210, 2001.

Guruswami, V. and Raghavendra, P. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39 (2):742–765, 2009.

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

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

Haussler, D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

Hazan, E. Introduction to online convex optimization. *CoRR*, abs/1909.05207, 2019.

Hopkins, M., Kane, D., Lovett, S., and Mahajan, G. Noise-tolerant, reliable active classification with comparison queries. In *Proceedings of the 33rd Annual Conference on Learning Theory*, pp. 1957–2006, 2020.

Kakade, S. M. and Tewari, A. On the generalization ability of online strongly convex programming algorithms. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, pp. 801–808, 2008.

Kalai, A. T., Klivans, A. R., Mansour, Y., and Servedio, R. A. Agnostically learning halfspaces. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pp. 11–20, 2005.

Kane, D. M., Lovett, S., Moran, S., and Zhang, J. Active classification with comparison queries. In Umans, C. (ed.), *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, pp. 355–366, 2017.

Kearns, M. J., Schapire, R. E., and Sellie, L. Toward efficient agnostic learning. In Haussler, D. (ed.), *Proceedings of the 5th Annual Conference on Computational Learning Theory*, pp. 341–352, 1992.

Klivans, A. R. and Servedio, R. A. Toward attribute efficient learning of decision lists and parities. In *Proceedings of the 17th Annual Conference on Learning Theory*, pp. 224–238, 2004.

Klivans, A. R., O'Donnell, R., and Servedio, R. A. Learning intersections and thresholds of halfspaces. In *Proceedings fo the 43rd Annual IEEE Symposium on Foundations of Computer Science*, pp. 177–186, 2002.

Klivans, A. R., Long, P. M., and Servedio, R. A. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10:2715–2740, 2009.

Littlestone, N. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. In *Proceedings of the 28th Annual IEEE Symposium on Foundations of Computer Science*, pp. 68–77, 1987.

Littlestone, N. and Warmuth, M. K. The weighted majority algorithm. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pp. 256–261, 1989.

Long, P. M. and Servedio, R. A. Attribute-efficient learning of decision lists and linear threshold functions under unconcentrated distributions. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, pp. 921–928, 2006.

Lovász, L. and Vempala, S. S. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3):307–358, 2007.

Massart, P. and Nédélec, É. Risk bounds for statistical learning. *The Annals of Statistics*, pp. 2326–2366, 2006.

Orabona, F. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019.

Plan, Y. and Vershynin, R. The generalized lasso with nonlinear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, 2016.

Plan, Y., Vershynin, R., and Yudovina, E. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

Rosenblatt, F. The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958.

Servedio, R. A. Computational sample complexity and attribute-efficient learning. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pp. 701–710, 1999.

Shalev-Shwartz, S. *Online learning: Theory, algorithms, and applications*. PhD thesis, Hebrew University, 2007.

Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Shen, J. and Li, P. On the iteration complexity of support recovery via hard thresholding pursuit. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3115–3124, 2017a.

Shen, J. and Li, P. Partial hard thresholding: Towards a principled analysis of support recovery. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pp. 3127–3137, 2017b.

Shen, J. and Li, P. A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18(208):1–42, 2018.

Shen, J. and Zeng, S. Learning from the crowd with pairwise comparisons. *CoRR*, abs/2011.01104, 2020.

Shen, J. and Zhang, C. Attribute-efficient learning of halfspaces with malicious noise: Near-optimal label complexity and noise tolerance. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pp. 1072–1113, 2021.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tropp, J. A. and Wright, S. J. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.

Valiant, L. G. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Vempala, S. S. A random-sampling-based algorithm for learning intersections of halfspaces. *Journal of the ACM*, 57(6):32:1–32:14, 2010.

Wang, J., Shen, J., and Li, P. Provable variable selection for streaming features. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5158–5166, 2018.

Xu, Y., Zhang, H., Singh, A., Dubrawski, A., and Miller, K. Noise-tolerant interactive learning using pairwise comparisons. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pp. 2431–2440, 2017.

Yan, S. and Zhang, C. Revisiting Perceptron: Efficient and label-optimal learning of halfspaces. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pp. 1056–1066, 2017.

Zhang, C. Efficient active learning of sparse halfspaces. In *Proceedings of the 31st Annual Conference On Learning Theory*, pp. 1856–1880, 2018.

Zhang, C. and Chaudhuri, K. Beyond disagreement-based agnostic active learning. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, pp. 442–450, 2014.

Zhang, C. and Li, Y. Improved algorithms for efficient active learning halfspaces with Massart and Tsybakov noise. *CoRR*, abs/2102.05312, 2021.

Zhang, C., Shen, J., and Awasthi, P. Efficient active learning of sparse halfspaces with arbitrary bounded noise. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, pp. 7184–7197, 2020.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pp. 894–942, 2010.

Zhang, L., Yi, J., and Jin, R. Efficient algorithms for robust one-bit compressive sensing. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 820–828, 2014.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 928–936, 2003.