Efficient active learning of sparse halfspaces

Chicheng Zhang
Microsoft Research
641 6th Avenue
New York, NY, 10011
USA

Abstract

We study the problem of efficient PAC active learning of homogeneous linear classifiers (halfspaces) in $\mathbb{R}^d$, where the goal is to learn a halfspace with low error, using as few label queries as possible. Given the extra assumption that there is a $t$-sparse halfspace that performs well on the data ($t \ll d$), we would like our active learning algorithm to be attribute efficient, i.e. to have label requirements sublinear in $d$. In this paper, we provide a computationally efficient algorithm that achieves this goal. Under certain distributional assumptions on the data, our algorithm achieves a label complexity of $O(t \text{polylog}(d, \frac{1}{\varepsilon}))$. In contrast, existing algorithms in this setting are either computationally inefficient, or subject to label requirements polynomial in $d$ or $\frac{1}{\varepsilon}$.

1. Introduction

Active learning is a machine learning paradigm that aims at reducing label requirements through interacting with labeling oracles (Settles, 2010). The learner is given a distribution from which it can draw unlabeled examples, and a labeling oracle from which it can query labels interactively. This is in contrast with passive learning, where labeled examples are drawn from distributions directly. Using the ability to adaptively query labels, an active learning algorithm can avoid querying the labels it has known before, thus substantially reducing label requirements. In the PAC active learning model (Valiant, 1984; Kearns et al., 1994), the performance of an active learner is measured by its label complexity, i.e. the number of label requests to satisfy an error requirement $\varepsilon$ with high probability.

There have been many exciting works on active halfspace learning in the literature. In this setting, the instances are in $\mathbb{R}^d$, and the labels are from $\{-1, +1\}$. The goal is to learn a classifier from $\mathcal{H} = \{\text{sign}(w \cdot x) : w \in \mathbb{R}^d\}$, the class of homogeneous linear classifiers, to predict labels from instances. Efficient active halfspace learning algorithms that work under different distributional assumptions have been proposed. Some of these algorithms are computationally efficient, and enjoy information theoretically optimal label complexities (Dasgupta et al., 2005; Balcan et al., 2007; Awasthi et al., 2017; Hanneke et al., 2015; Awasthi et al., 2015; Yan and Zhang, 2017), that is, $O(d \ln \frac{1}{\varepsilon})$ in terms of $d$ and $\varepsilon$ (See also e.g. Kulkarni et al., 1993, for an $\Omega(d \ln \frac{1}{\varepsilon})$ lower bound).

On the other hand, a line of work on attribute efficient learning (Blum, 1990) shows that one can in fact learn faster when the target classifier is sparse, i.e. it depends on only a few of the input features. In the setting of active learning, one can straightforwardly apply existing results to achieve attribute efficiency. For instance, consider running the algorithm of Zhang and Chaudhuri (2014) with concept class $\mathcal{H}_t$, the set of $t$-sparse linear classifiers. Under certain distributional assump-
tions, Zhang and Chaudhuri (2014)’s algorithm achieves label complexities of order $O(t \ln d \ln \frac{1}{\epsilon})$. However, such algorithms are computationally inefficient: they require solving empirical 0-1 loss minimization with respect to $H_t$, which is NP-hard even in the realizable setting (Natarajan, 1995).

The results above raise the following question: are there active learning algorithms that learn linear classifiers in an attribute and computationally efficient manner? A line of work, namely one-bit compressed sensing (Boufounos and Baraniuk, 2008), partially answers this question. They show that when the learning algorithm is allowed to synthesize instances to query their labels (also known as the membership query model (Angluin, 1988), abbrev. MQ), it is possible to approximately recover the target halfspace using a near-optimal number of $\tilde{O}(t (\ln d + \ln \frac{1}{\epsilon}))$ queries (Haupt and Baraniuk, 2011). However, when applied to active learning in the PAC model, these results have strong distributional requirements. For instance, the algorithm of Haupt and Baraniuk (2011) requires the unla-
beled distribution to have a constant probability to observe elements in the discrete set $\{-1, 0, +1\}^d$.

In the PAC setting, recent work of Awasthi et al. (2016) proposes attribute and computationally efficient active halfspace learning algorithms, under the assumption that the unlabeled distribution is isotropic log-concave (Lovász and Vempala, 2007). In the $t$-sparse $\Omega(\epsilon)$-adversarial noise setting, where all but an $\Omega(\epsilon)$ fraction of examples agree with some $t$-sparse linear classifier (see also Definition 1), their algorithm has a label complexity of $\tilde{O}(\frac{\epsilon}{t^2})$. In the $t$-sparse $\eta$-bounded noise setting, where each label is generated by some underlying $t$-sparse linear classifier and then flipped with probability at most a constant $\eta \in [0, \frac{1}{2})$ (see also Definition 2), their algorithm has a label complexity of $\tilde{O}(\frac{1}{t^3})^{O(1)}$. Compared to those achieved by inefficient algorithms (e.g. Zhang and Chaudhuri (2014) discussed above), these label complexity bounds are suboptimal, in that they do not have a logarithmic dependence on $\frac{1}{\epsilon}$.

In this paper, we give an algorithm that combines the advantages of Zhang and Chaudhuri (2014) and Awasthi et al. (2016), achieving computational efficiency and $\tilde{O}(t \text{polylog}(d, \frac{1}{\epsilon}))$ label complexity simultaneously, under certain distributional assumptions on the data. Specifically, our algorithm works if the unlabeled distribution is isotropic log-concave, and has the following guarantee. If one of the two conditions below is true:

1. The $t$-sparse $\Omega(\epsilon)$-adversarial noise condition holds (see Definition 1).
2. The $t$-sparse $\eta$-bounded noise condition holds (see Definition 2), where $\eta$ is at most $c$ for some numerical constant $c > 0$.

then, with high probability, the algorithm outputs a halfspace with excess error at most $\epsilon$, and queries at most $O(t (\ln d + \ln \frac{1}{\epsilon})^3 \ln \frac{1}{\epsilon})$ labels. As a corollary, if there is a $t$-sparse linear classifier that agrees with all the labeled examples drawn from the distribution (see Definition 3), the algorithm also achieves a label complexity of $O(t (\ln d + \ln \frac{1}{\epsilon})^3 \ln \frac{1}{\epsilon})$. In the next section, we give a detailed comparison between our results and related results in the literature.

From a technical perspective, our algorithm combines the margin-based framework of Balcan et al. (2007); Balcan and Long (2013) with iterative hard thresholding (Blumensath and Davies, 2009; Garg and Khandekar, 2009), a technique well-studied in compressed sensing (Candes and Tao, 2006; Donoho, 2006). Our analysis is based on sharp uniform concentration bounds of hinge losses over linear predictors in $\ell_1$ balls in the label query regions, which is in turn built upon classical Rademacher complexity bounds for linear prediction (Kakade et al., 2009).
2. Related work

Attribute efficient active learning of halfspaces. There is a rich body of theoretical literature on active learning of general concept classes in the PAC setting (Dasgupta, 2011; Hanneke, 2014). For the problem of active halfspace learning, sharp distribution-dependent label complexity results are known, in terms of e.g. the splitting index (Dasgupta, 2005), or the disagreement coefficient (Hanneke, 2007). A direct application of these results (without taking advantage of sparsity assumptions) yields label complexity bounds at least $\Omega(d)$ (Kulkarni et al., 1993). To make these algorithms attribute efficient, a natural modification is to consider concept class $\mathcal{H}_t$, the set of $t$-sparse linear classifiers. It is well known that $\mathcal{H}_t$ has VC dimension $O(t \ln d)$. In conjunction with existing results in the active learning literature, this observation immediately yields attribute efficient active learning algorithms. For example, when the unlabeled distribution is isotropic log-concave, an application of Zhang and Chaudhuri (2014)’s algorithm with $\mathcal{H}_t$ yields a label complexity of $O(t \ln d \ln \frac{1}{t})$ in the $t$-sparse realizable setting, and gives $O(t \ln d \cdot (\ln \frac{1}{t} + \nu^2 \ln \frac{1}{t}))$ and $O\left(\frac{t \ln d}{(1-2\eta)^2} \ln \frac{1}{t}\right)$ label complexities in the $t$-sparse $\nu$-adversarial noise and $t$-sparse $\eta$-bounded noise settings.1 However, these algorithms require solving empirical 0-1 loss minimization problems, which are computationally intractable in general. The only attribute and computationally efficient PAC active learning algorithms we are aware of are in Awasthi et al. (2016). Specifically, under the $t$-sparse $\Omega(\epsilon)$-adversarial noise setting, Awasthi et al. (2016) gives an algorithm with label complexity $\tilde{O}\left(\frac{d}{\epsilon^2} \text{polylog}(d, \frac{1}{t})\right)$. Under the $t$-sparse $\eta$-bounded noise setting, Awasthi et al. (2016) gives an efficient algorithm with label complexity $\tilde{O}\left(\frac{d t}{\eta^2} \text{polylog}(d, \frac{1}{t})\right)$.

The notion of attribute efficient learning algorithms is initially studied in the pioneering works of Littlestone (1987); Blum (1990). Littlestone (1987) considers attribute efficient online learning of linear classifiers, with an application to learning disjunctions that depends on only $t$ attributes. The algorithm incurs a mistake bound of $O(t \ln d)$, which can be of substantially lower order than $O(d)$ when $t$ is small. Blum (1990) considers an online learning model where the feature space is infinite dimensional, and each instance shown has a bounded number of nonzero attributes. It gives efficient algorithms that learn $k$-CNFs and disjunctions with finite mistake bounds in this setting. Servedio (2000); Klivans and Servedio (2006); Servedio et al. (2012) study attribute efficient learning of decision lists and analyzes the tradeoff between running time and mistake bound. Long and Servedio (2007) shows that, if the unlabeled distribution is unconcentrated over $\{-1, 1\}^d$, then there is an algorithm that learns $t$-sparse linear classifiers with a sample complexity of $\text{poly}(t, \ln d, 2^{O(\epsilon^{-2})})$. Feldman (2007) gives algorithms for attribute efficient learning parity and DNFs in the membership query model.

One-bit compressed sensing. The line of work on one-bit compressed sensing (Boufounos and Baraniuk, 2008) is closely related to our problem setup. In this setting, there is a unknown $t$-sparse vector $u \in \mathbb{R}^d$, and the algorithm can make measurements of $u$ using vectors $x \in \mathbb{R}^d$ and receives (possibly noisy) values of $\text{sign}(u \cdot x)$. Note that different from classical compressed sensing, the measurement results of one-bit compressed sensing are quantized versions of $(u \cdot x)$’s (i.e. they lie in $\{-1, +1\}$ as opposed to $\mathbb{R}$). The goal is to approximately recover $u$ up to scaling with a few (ideally, $O(t \ln d)$) measurements. In the non-adaptive setting, the measurement vector $x$’s are chosen at the beginning, while in the adaptive setting, the measurement vector $x$’s can be chosen sequen-

1. To see this, note that the $\phi(\cdot, \cdot)$ function defined in Zhang and Chaudhuri (2014) with respect to $\mathcal{H}_t$ can be bounded as: $\phi(r, \xi) \leq O(r \ln \frac{1}{\xi})$, as $\mathcal{H}_t$ is a subset of $\mathcal{H}$. Theorem 4 of Zhang and Chaudhuri (2014) now applies.
tially, based on past observations. The problem of adaptive one-bit compressed sensing is therefore
equivalent to attribute efficient active halfspace learning in the membership query model (Angluin, 1988). We remark that active learning in the PAC model is more challenging than in the membership
model, as the learner has to query the labels of the unlabeled examples it has drawn.

Jacques et al. (2013) gives an algorithm that has robust recovery guarantees, however it is
based on computationally-intractable \( \ell_0 \) minimization. Inspired by the count sketch data struc-
ture (Charikar et al., 2002), Haupt and Baraniuk (2011) proposes an efficient procedure that recov-
ers the support of \( u \) using \( O(t \ln d) \) queries, and has strong noise tolerance properties. In con-
junction with efficient full-dimensional active halfspace learning algorithms (Dasgupta et al., 2005;
Awasthi et al., 2017; Chen et al., 2017; Yan and Zhang, 2017), this procedure yields efficient algo-
rithms that have label complexities of \( O(t(\ln d + \ln \frac{1}{t})) \) (resp. \( O(t(\ln d + \ln \frac{1}{t})) \), \( O(\frac{t}{(1-2\eta)^2}(\ln d + \ln \frac{1}{t})) \)) in the \( t \)-sparse realizable setting (resp. \( t \)-sparse \( \Omega(\epsilon) \)-adversarial noise setting, \( t \)-sparse \( \eta \)-bounded noise setting). Gopi et al. (2013); Acharya et al. (2017) gives upper and lower bounds for univer-
sal one-bit compressed sensing, that is, the same set of measurements can be used to approxi-
mately recover any underlying \( t \)-sparse signal. In this setting, Acharya et al. (2017) shows that,
perhaps surprisingly, the number of measurements necessary and sufficient for support recovery is
\( \tilde{O}(t^2 \ln d) \), as opposed to \( \Theta(t \ln d) \) in the non-universal setting. Plan and Vershynin (2013a)
proposes a linear programming based algorithm that works in the \( t \)-sparse realizable setting, and has a
measurement complexity of \( \tilde{O}(\frac{1}{\epsilon^2}) \), based on a new tool named random hyperplane tessellations. Li
(2016) gives a support recovery algorithm that tolerates bounded noise, using \( \alpha \)-stable random
projections. Plan and Vershynin (2013b) proposes a convex programming based algorithm that works
in the \( t \)-sparse \( \Omega(\epsilon^2) \)-adversarial noise model, and has a measurement complexity of \( \tilde{O}(\frac{1}{\epsilon^2}) \).

Works on one-bit compressed sensing under the symmetric noise condition has been studied in the
literature (Plan and Vershynin, 2013b; Zhang et al., 2014; Chen and Banerjee, 2015; Zhu and Gu,
2015); in this model, it is assumed that there is a known function \( g \), such that for all \( x \), \( \mathbb{E}[y|x] =
g(u \cdot x) \). This assumption captures some realistic scenarios, but is nevertheless strong: it requires
any two examples that have the same projection on \( u \) to have the same conditional label distribution.
In contrast, the \( t \)-sparse adversarial noise and the \( t \)-sparse bounded noise conditions allow hetero-
gegeneous noise levels, even among examples that have the same projection on \( u \). In this setting, the
state of the art result of Zhang et al. (2014) gives a nonadaptive algorithm with \( O(\frac{t \ln d}{\epsilon^2}) \). It also
proposes an adaptive algorithm that works in same setting, achieving a label complexity bound of
\( O(\min(\frac{\ln d}{\epsilon^2}, \sqrt{\frac{\ln d}{\epsilon}})) \), which is sometimes lower than that of the nonadaptive algorithm. The
special case of Gaussian noise before quantization has been studied extensively, i.e. given \( x \), the label
\( y \) is generated by the formula \( \text{sign}(u \cdot x + n) \), where \( n \) is a Gaussian random variable. Gupta et al.
(2010) shows that when \( u \) has a large dynamic range (the absolute value of the ratio between \( u \)'s
largest and smallest nonzero elements in magnitude), adaptive approaches require fewer measure-
ments to identify the support of \( u \) than nonadaptive approaches.

We provide a detailed comparison between our work and the results most closely related to ours
in tables 1, 2, and 3.

3. Preliminaries

We consider active learning in the PAC model (Valiant, 1984; Kearns et al., 1994). Denote by
\( \mathcal{X} := \mathbb{R}^d \) the instance space, and \( \mathcal{Y} := \{-1, +1\} \) the label space. The learning algorithm is given
a data distribution \( D \) over \( \mathcal{X} \times \mathcal{Y} \). Denote by \( D_X \) the marginal distribution of \( D \) over \( X \), and


Table 1: A comparison of algorithms for active learning of halfspaces in the $t$-sparse realizable setting (Definition 3); all the PAC algorithms above work under isotropic log-concave distributions.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Model</th>
<th>Label complexity</th>
<th>Efficient?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haupt and Baraniuk (2011)</td>
<td>MQ</td>
<td>$\tilde{O}(t(\ln d + \ln \frac{1}{\epsilon}))$</td>
<td>Yes</td>
</tr>
<tr>
<td>with Dasgupta et al. (2005)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dasgupta (2005)</td>
<td>PAC</td>
<td>$\tilde{O}(t(\ln d + \ln \frac{1}{\epsilon}))$</td>
<td>No</td>
</tr>
<tr>
<td>Awasthi et al. (2016)</td>
<td>PAC</td>
<td>$\tilde{O}(\frac{t}{\epsilon^2} \text{ polylog}(d, \frac{1}{\epsilon}))$</td>
<td>Yes</td>
</tr>
<tr>
<td>Our work</td>
<td>PAC</td>
<td>$\tilde{O}(t \text{ polylog}(d, \frac{1}{\epsilon}))$</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 2: A comparison of algorithms for active learning of halfspaces in the $t$-sparse $\nu$-adversarial noise setting (Definition 1); all the PAC algorithms above work under isotropic log-concave distributions.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Model</th>
<th>Noise tolerance</th>
<th>Label complexity</th>
<th>Efficient?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haupt and Baraniuk (2011)</td>
<td>MQ</td>
<td>$\nu = \Omega(\epsilon)$</td>
<td>$\tilde{O}(t(\ln d + \ln \frac{1}{\epsilon}))$</td>
<td>Yes</td>
</tr>
<tr>
<td>with Awasthi et al. (2017)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang and Chaudhuri (2014)</td>
<td>PAC</td>
<td>$\nu = \Omega(\epsilon)$</td>
<td>$\tilde{O}(t \ln d \ln \frac{1}{\epsilon})$</td>
<td>No</td>
</tr>
<tr>
<td>Plan and Vershynin (2013b)</td>
<td>PAC</td>
<td>$\nu = \Omega(\epsilon^2)$</td>
<td>$\tilde{O}(\frac{\ln d}{\epsilon^2})$</td>
<td>Yes</td>
</tr>
<tr>
<td>Awasthi et al. (2016)</td>
<td>PAC</td>
<td>$\nu = \Omega(\epsilon)$</td>
<td>$\tilde{O}(\frac{t}{\epsilon^2} \text{ polylog}(d, \frac{1}{\epsilon}))$</td>
<td>Yes</td>
</tr>
<tr>
<td>Our work</td>
<td>PAC</td>
<td>$\nu = \Omega(\epsilon)$</td>
<td>$\tilde{O}(t \text{ polylog}(d, \frac{1}{\epsilon}))$</td>
<td>Yes</td>
</tr>
</tbody>
</table>

$D_{Y \mid X}$ the conditional distribution of $y$ given $x$. The learning algorithm is also given a concept class, the set of homogeneous linear classifiers (halfspaces) $\mathcal{H} := \{\text{sign}(w \cdot x) : w \in \mathbb{R}^d\}$. For any classifier $h : \mathcal{X} \to \mathcal{Y}$, we denote by $err(h) := \mathbb{P}_D(h(x) \neq y)$ the error rate of $h$. Denote by $h^*$ the optimal classifier in $\mathcal{H}$: $h^* := \arg\min_{h \in \mathcal{H}} err(h)$. The excess error of classifier $h$ is defined as $err(h) - err(h^*)$; in words, it is the difference between $h$’s error and the best error in $\mathcal{H}$. A vector $w$ corresponds to a linear classifier $h_w := \text{sign}(w \cdot x)$ whose decision boundary has $w$ as its normal; define $w^*$ as the unit vector $w$ such that $h_w = h^*$. We define the angle between two vectors $w, w'$ in $\mathbb{R}^d$ as $\theta(w, w') = \arccos(\frac{w \cdot w'}{\|w\|_2 \|w'\|_2})$. Balcan and Long (2013) shows that there exist numerical constants $C_1, C_2 > 0$, such that if $D_X$ is isotropic log-concave, then for all $w, w'$ in $\mathbb{R}^d$,

$$C_1 \mathbb{P}_D(h_w(x) \neq h_{w'}(x)) \leq \theta(w, w') \leq C_2 \mathbb{P}_D(h_w(x) \neq h_{w'}(x)).$$

(1)

In active learning, the algorithm has the ability to draw unlabeled examples from $D_X$ and perform adaptive label queries to a labeling oracle $O$. The oracle $O$ takes as input an unlabeled example $x$, and returns a label $y \sim D_{Y \mid X=x}$. Given a random variable $z$ whose distribution is $\Delta$ over $\mathcal{Z}$ and a set $T \subset \mathcal{Z}$, denote by $\Delta|_T$ the conditional distribution of $z$ given that $z$ is in $T$. An active learning algorithm is said to $(\epsilon, \delta)$-PAC learn $\mathcal{H}$ and $D$ with label complexity $n(\epsilon, \delta)$, if with probability $1 - \delta$, it performs at most $n(\epsilon, \delta)$ label queries to $O$, and returns a classifier $\hat{h}$ that has excess error at most $\epsilon$. 

5
Table 3: A comparison of algorithms for active learning of halfspaces in the \(t\)-sparse \(\eta\)-bounded noise setting (Definition 2); all the PAC algorithms above work under isotropic log-concave distributions.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Model</th>
<th>Noise tolerance</th>
<th>Label complexity</th>
<th>Efficient?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haupt and Baraniuk (2011)</td>
<td>MQ</td>
<td>(\eta \in [0, \frac{1}{2}])</td>
<td>(\tilde{O}(\frac{t}{(1-2\eta^2)}(\ln d + \ln \frac{1}{\epsilon})))</td>
<td>Yes</td>
</tr>
<tr>
<td>with Chen et al. (2017)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang and Chaudhuri (2014)</td>
<td>PAC</td>
<td>(\eta \in [0, \frac{1}{2}])</td>
<td>(\tilde{O}(\frac{t}{(1-2\eta^2)} \ln d \ln \frac{1}{\epsilon}))</td>
<td>No</td>
</tr>
<tr>
<td>Awasthi et al. (2016)</td>
<td>PAC</td>
<td>(\eta \in [0, \frac{1}{2}])</td>
<td>(\tilde{O}(\frac{t}{(1-2\eta^2)}))</td>
<td>Yes</td>
</tr>
<tr>
<td>Our work</td>
<td>PAC</td>
<td>(\eta \in [0, \Omega(1)])</td>
<td>(\tilde{O}(t \text{ polylog}(d, \frac{1}{\epsilon})))</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Given a vector \(w\) and example \((x, y)\), the \(\tau\)-hinge loss \(\ell_{\tau}(w, (x, y))\) is defined as \(1 - \frac{w \cdot x}{\tau}_+\), where \((z)_+ := \max(0, z)\). Denote by \(I(\cdot)\) the indicator function, that is, \(I(\mathcal{A}) = 1\) if predicate \(\mathcal{A}\) is true, is 0 if \(\mathcal{A}\) is false. A vector \(v\) in \(\mathbb{R}^d\) is said to be \(s\)-sparse, if it has at most \(s\) nonzero entries. For an integer \(s \in \{1, 2, \ldots, d\}\), define \(P_s(\cdot)\) as the hard thresholding operation that takes a vector \(v\) in \(\mathbb{R}^d\) as input, and outputs a vector that keeps \(v\)'s \(s\) largest entries in absolute value (breaking ties lexicographically), and setting all other entries to zero (Blumensath and Davies, 2009).

In this paper, we focus on the setting where there is a sparse halfspace that performs well under \(D\). Specifically, denote by \(\mathcal{H}_t := \{\text{sign}(w \cdot x) : w \in \mathbb{R}^d, \|w\|_0 \leq t\}\) the set of \(t\)-sparse halfspaces. We consider the following two conditions regarding \(D\):

**Definition 1** A distribution \(D\) over \(\mathcal{X} \times \mathcal{Y}\) is said to satisfy the \(t\)-sparse \(\nu\)-adversarial noise condition for \(\nu \in (0, 1)\) and \(t \in \{1, \ldots, d\}\), if there is a \(t\)-sparse unit vector \(u\), such that \(\mathbb{P}_D(\text{sign}(u \cdot x) \neq y) \leq \nu\).

Observe that under this condition, \(h_u\) is not necessarily the optimal classifier in \(\mathcal{H}\); in fact, it may not even be the optimal classifier in \(\mathcal{H}_t\). Nevertheless, by triangle inequality and Equation (1), the angle between \(u\) and \(w^*\) is at most \(O(\nu)\). It can be readily seen that if \(t\) and \(\nu\) are larger, the learning problem becomes more difficult. When \(t = d\), the condition becomes the \(\nu\)-adversarial noise condition with respect to \(\mathcal{H}\) (Awasthi et al., 2017).

**Definition 2** A distribution \(D\) over \(\mathcal{X} \times \mathcal{Y}\) is said to satisfy the \(t\)-sparse \(\eta\)-bounded noise condition for \(\eta \in [0, \frac{1}{2}]\) and \(t \in \{1, \ldots, d\}\), if there is a \(t\)-sparse unit vector \(u\), such that for every \(x \in \mathcal{X}\), \(\mathbb{P}_D(\text{sign}(u \cdot x) \neq y|x) \leq \eta\).

Under this condition, it can be seen that \(h_u\) is the Bayes optimal classifier, therefore \(u\) coincides with \(w^*\). It can be readily seen that if \(t\) and \(\eta\) are larger, the learning problem becomes more difficult. When \(t = d\), the condition becomes the \(\eta\)-bounded noise condition with respect to \(\mathcal{H}\) (Massart and Nédélec, 2006).

Note that the above two conditions characterize different aspects of the data distribution \(D\). The \(t\)-sparse \(\nu\)-adversarial noise condition only requires an upper bound on the total label flipping probability. On the other hand, the \(t\)-sparse \(\eta\)-bounded noise condition characterizes \(D_{\mathcal{Y}|X}\) everywhere in \(\mathcal{X}\): for every instance \(x\), the expected label \(\mathbb{E}[y|x]\) has the same sign as \(u \cdot x\). The following condition is a special case of the above two conditions by setting \(\nu = 0\) or \(\eta = 0\):
Definition 3 A distribution $D$ over $\mathcal{X} \times \mathcal{Y}$ is said to satisfy the $t$-sparse realizable condition, for $t \in \{1, 2, \ldots, d\}$, if there is a $t$-sparse unit vector $u$, such that $\mathbb{P}(\text{sign}(u \cdot x) \neq y) = 0$.

4. Algorithm

We present our main algorithm, namely Algorithm 1 in this section. We defer the exact settings of constants $c_1, c_2, c_3$ to Appendix A. Our algorithm uses the margin-based active learning framework, initially proposed by Balcan et al. (2007). It proceeds in epochs, where at each epoch $k$, it draws a sample $S_k$ from region $B_k$, queries their labels, and updates its iterate $w_k$ based on the sample. Due to technical reasons, at the first epoch ($k = 0$), the sampling region $B_0$ and the constraint set $W_0$ are different from those in subsequent epochs. Throughout the process, the algorithm maintains the invariant that at each epoch $k$, $w_k$ is a $t$-sparse unit vector.

At each epoch $k \geq 1$, it focuses on sampling in a “small-margin” band $B_k := \{x : |w_{k-1} \cdot x| \leq b_k\}$, with bandwidth $b_k$ decrasing exponentially in $k$. Then it performs constrained empirical hinge loss minimization over $S_k$, getting a linear classifier $w'_k$. The constraint set $W_k$ is the intersection between an $\ell_1$ ball and an $\ell_2$ ball, both centered at $w_{k-1}$ with different radii ($p_k$ and $r_k$). This is similar to the approach in Plan and Vershynin (2013b) for tackling the symmetric noise setting, where a linear optimization problem with a similar shaped constraint set is proposed. The construction of $W_k$’s is inspired by version space constructions in the PAC active learning literature (Cohn et al., 1994; Balcan et al., 2009; Hanneke, 2014). Essentially, we would like to ensure $W_k$ to satisfy the following two properties with high probability: first, $u$ lie in all the $W_k$’s; second, the $W_k$’s are shrinking in size. In addition, the hinge loss used in epoch $k$ is parameterized by $\tau_k$, which also decreases exponentially in $k$.

Observe that $w'_k$ may not be a sparse vector; therefore, we perform a hard thresholding step (applying $P_\ell$), to ensure that our learned halfspace at the end of round $k$, is $t$-sparse. Hard thresholding has been widely used in the (unquantized) compressed sensing literature (See e.g. Blumensath and Davies, 2009; Garg and Khandekar, 2009), however its utility in one-bit compressed sensing is not yet well-understood. For example, Jacques et al. (2013) proposes an algorithm named BiHT (binary iterative thresholding) that has strong empirical performance, but its convergence properties are unknown. To the best of our knowledge, our work is the first that establishes convergence guarantees for iterative hard thresholding style algorithms for one-bit compressed sensing. We then perform a $\ell_2$ normalization step to ensure that our iterate $w_k$ is an unit vector, which has a scale similar to $u$.

Finally, we remark that Algorithm 1 admits a computationally efficient implementation. First, the sampling regions $B_k$’s can be shown to have probability masses at least $\Omega(\epsilon)$ in $D_X$ for all $k$ in $\{0, 1, \ldots, k_0\}$, which makes rejection sampling from $D_X|_{B_k}$ take $O(\frac{1}{\epsilon})$ time per example. Second, optimization problem (2) is convex, and can be approximately solved by e.g. stochastic gradient descent (See e.g. Shamir and Zhang, 2013, Theorem 2) efficiently.

5. Performance guarantees

In this section, we prove Theorem 4, the main result of this paper.

Theorem 4 There exist numerical constants $\mu_1, \mu_2 \in (0, \frac{1}{2})$ such that the following holds. Suppose $D_X$ is isotropic log-concave, and one of the following two conditions hold:

2. We refer the reader to Lemma 6 for a formal statement.
Algorithm 1 Attribute and computationally efficient active learning of halfspaces

**input:** sparsity parameter \( t \), target error \( \epsilon \), failure probability \( \delta \).

**output:** learned halfspace \( \hat{w} \).

1: Initialization: \( k_0 \leftarrow \lceil \log_2 \frac{1}{t \epsilon} \rceil \), where \( C_1 \) is defined in Equation (1).
2: for \( k = 0, 1, 2, \ldots, k_0 \) do
3: \( S_k \leftarrow \) sample \( n_k = c_1 t (\ln d + \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta_k})^3 \) examples from \( D_X|_{B_k} \) and query their labels, where
\[
B_k := \begin{cases} \mathbb{R}^d, & k = 0, \\ \{ x : |w_{k-1} \cdot x| \leq b_k \}, & k \geq 1, \end{cases}
\]
\( \delta_k = \frac{\delta}{(k+1)(k+2)} \) and \( b_k = c_2 \cdot 2^{-k} \).
4: Solve the following optimization problem:
\[
w_k^* \leftarrow \arg\min_{w \in W_k} \sum_{(x,y) \in S_k} \ell_{\tau_k}(w, (x,y)),
\]

where
\[
W_k = \begin{cases} \{ w \in \mathbb{R}^d : \|w\|_2 \leq 1 \text{ and } \|w\|_1 \leq \sqrt{t} \}, & k = 0, \\ \{ w \in \mathbb{R}^d : \|w - w_{k-1}\|_2 \leq r_k \text{ and } \|w - w_{k-1}\|_1 \leq \rho_k \}, & k \geq 1, \end{cases}
\]
\( r_k = 2^{-k-3}, \rho_k = \sqrt{2t} \cdot 2^{-k-3}, \) and \( \tau_k = c_3 \cdot 2^{-k} \).
5: Let \( w_k \leftarrow \frac{\text{Pr}(w_k)}{\|\text{Pr}(w_k)\|_2} \).
6: end for
7: return \( w_{k_0} \).

1. \( D \) satisfies the \( t \)-sparse \( \mu_1 \epsilon \)-adversarial noise condition;
2. \( D \) satisfies the \( t \)-sparse \( \mu_2 \)-bounded noise condition.

In addition, Algorithm 1 is run with sparsity parameter \( t \), target error \( \epsilon \) and failure probability \( \delta \). Then, with probability \( 1 - \delta \), the output halfspace \( \hat{w} \) is such that \( \text{err}(\hat{h}_\hat{w}) - \text{err}(h^*) \leq \epsilon \), and the total number of label queries is \( O(t \cdot (\ln d + \ln \frac{1}{\epsilon})^3 \cdot \ln \frac{1}{\epsilon}) \).

As the \( t \)-sparse realizable setting is a special case of the \( t \)-sparse adversarial noise setting (by setting \( \nu = 0 \)), Theorem 4 immediately implies the following corollary:

**Corollary 5** Suppose \( D_X \) is isotropic log-concave, and the \( t \)-sparse realizable condition holds. In addition, Algorithm 1 is run with sparsity parameter \( t \), target error \( \epsilon \) and failure probability \( \delta \). Then, with probability \( 1 - \delta \), the output halfspace \( \hat{w} \) is such that \( \text{err}(\hat{h}_\hat{w}) - \text{err}(h^*) \leq \epsilon \), and the total number of label queries is \( O(t \cdot (\ln d + \ln \frac{1}{\epsilon})^3 \cdot \ln \frac{1}{\epsilon}) \).

Theorem 4 and Corollary 5 imply that, under the respective noise conditions defined above, Algorithm 1 has a label complexity of \( O(t \text{ polylog}(d, \frac{1}{\epsilon})) \). To the best of our knowledge, this is the first efficient PAC active learning algorithm that has a label complexity linear in \( t \), and polylogarithmic in \( d \) and \( \frac{1}{\epsilon} \). Previous works either need to sacrifice computational efficiency to achieve such guarantee (Dasgupta, 2005; Zhang and Chaudhuri, 2014), or have label complexities polynomial in...
d or $\frac{1}{d}$ (Awasthi et al., 2017, 2016). We remark that in the membership query model (Angluin, 1988; Boufounos and Baraniuk, 2008), efficient algorithms with $O(t \text{polylog}(d, \frac{1}{\delta}))$ label complexities are implicit in the literature (e.g. by combining Haupt and Baraniuk (2011)'s support recovery algorithm with efficient full-dimensional active halfspace learning algorithms (Dasgupta et al., 2005; Awasthi et al., 2017; Chen et al., 2017; Yan and Zhang, 2017)). In contrast, the focus of this paper is on the more challenging PAC setting, and it is unclear how to modify a membership query algorithm to make it work in the PAC setting.

5.1. Proof of Theorem 4

Recall that $\delta_k = \frac{\delta}{(k+1)(k+2)}$; note that $\sum_{t=0}^{k_0} \delta_k \leq \delta$. To prove Theorem 4, we give exact settings of constants $\mu_1, \mu_2 \in (0, \frac{1}{2})$ in Appendix A, such that under either the $t$-sparse $\mu_1\epsilon$-adversarial noise condition or the $t$-sparse $\mu_2$-bounded noise condition, the following lemma holds:

**Lemma 6** For every $k \in \{0, 1, \ldots, k_0\}$, there is an event $E_k$ with probability $1 - \sum_{l=0}^{k} \delta_l$, on which $u$ is in $W_{k+1}$.

The proof of Lemma 6 relies on the following two supporting lemmas. The first lemma (Lemma 7) shows that, $w_{k}'$ produced in the hinge loss minimization step (line 4) has a small angle with $u$. Specifically, the upper bound on $\theta(w_{k}', u)$ is halved compared to that of $\theta(w_{k-1}', u)$, using constrained hinge loss minimization over a fresh set of $n_k = O(t \text{polylog}(d, \frac{1}{\delta}))$ labeled examples. This relies on two ideas: first, as is standard in the margin-based active learning framework (See e.g. Balcan et al., 2007; Balcan and Long, 2013), it suffices to let $w_{k-1}'$ achieve constant error with respect to the sampling distribution at epoch $k$; second, to ensure that the setting of $n_k$ ensures that $w_{k-1}'$ indeed has a constant error under the sampling distribution, we use a novel uniform concentration bound of hinge losses of $W_k$ over $S_k$ tighter than all prior works (Awasthi et al., 2017, 2016). Specifically, our concentration bound of hinge losses is of order $\tilde{O}(\sqrt{\frac{\ln d}{n_k}})$, which can be substantially tighter than $\tilde{O}(\sqrt{\frac{d}{n_k}})$ used in Awasthi et al. (2017); Hanneke et al. (2015) and $\tilde{O}(\sqrt{\frac{\ln d}{n_k}})$ used in Awasthi et al. (2016).

**Lemma 7** For every $k \in \{0, 1, \ldots, k_0\}$, if $u$ is in $W_k$, then with probability $1 - \delta_k$, $\theta(w_{k}', u) \leq 2^{-k-8\pi}$.

The second lemma (Lemma 8) shows that, performing a hard thresholding operation followed by $\ell_2$ normalization on $w_k'$ (line 5) yields a $t$-sparse unit vector $w_k$ that is close to $u$ in terms of both $\ell_1$ and $\ell_2$ distances. This ensures that $W_{k+1}$, the constraint set of the optimization problem at the next epoch, contains $u$.

**Lemma 8** For every $k \in \{0, 1, \ldots, k_0\}$, if $\theta(w_{k}', u) \leq 2^{-k-8\pi}$, then $u$ is in $W_{k+1}$.

We are now ready to prove Lemma 6.

**Proof** [Proof of Lemma 6] We prove the lemma by induction.

**Base case.** In the case of $k = 0$, observe that as $u$ has unit $\ell_2$ norm and $u$ is $t$-sparse, by Cauchy-Schwarz, $\|u\|_1 \leq \sqrt{t}$. Therefore, $u$ belongs to the set $W_0$ deterministically. Lemma 7 with $k = 0$ shows that there is an event $E_0$ with probability $1 - \delta_0$, conditioned on which we have that $\theta(w_0', u) \leq 2^{-8\pi}$. By Lemma 8, we get that $u$ is in $W_1$. 

**Inductive case.** For $k \geq 1$, suppose the inductive hypothesis holds. That is, there is an event $E_{k-1}$ with probability $1 - \sum_{l=0}^{k-1} \delta_l$, such that on $E_{k-1}$, $u$ is in $W_k$. By Lemma 7, there is an event $F_k$ such that $\Pr(F_k|E_{k-1}) \geq 1 - \delta_k$, conditioned on which $\theta(u', u) \leq 2^{-k-8}\pi$.

Define event $E_k := E_{k-1} \cap F_k$. By straightforward calculations, $\Pr(E_k) = \Pr(E_{k-1})\Pr(F_k|E_{k-1}) \geq 1 - \sum_{l=0}^{k} \delta_l$. Now, on event $E_k$, Lemma 8 implies that $u$ is in $W_{k+1}$. This completes the induction.

Theorem 4 is now an immediate consequence of Lemma 6; we give its proof below.

**Proof [Proof of Theorem 4]** From Lemma 6 and the fact that the output $\hat{w}$ is $w_{k_0}$, we have that with probability $1 - \sum_{l=0}^{k_0} \delta_l \geq 1 - \delta$, $u$ is in $W_{k_0+1}$. By the definition of $W_k$,

$$||u - w_{k_0}||_2 \leq r_{k_0+1} = 2^{-k_0-4}.$$  

By Lemma 10 in the Appendix and the fact that $||u||_2 = 1$, we know that $\theta(w_{k_0}, u) \leq \pi ||u - w_{k_0}||_2 \leq 2^{-k_0-2} \leq \frac{C\epsilon}{2}$. By the first inequality of Equation (1), we have that $\Pr_D(h_{w_{k_0}}(x) \neq h_u(x)) \leq \frac{\epsilon}{2}$. Therefore, by triangle inequality and the fact that the output $\hat{w}$ is $w_{k_0}$,

$$\text{err}(h_{\hat{w}}) - \text{err}(h_u) \leq \frac{\epsilon}{2}$$

We now consider two separate cases regarding the two different noise conditions:

1. In the $\mu_1 \epsilon$-adversarial noise setting, we know that $\text{err}(h_u) \leq \mu_1 \epsilon \leq \frac{\epsilon}{2}$. Therefore,

$$\text{err}(h_{\hat{w}}) - \text{err}(h^*) \leq \text{err}(h_{\hat{w}}) \leq \text{err}(h_u) + \frac{\epsilon}{2} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \leq \epsilon.$$

2. In the $\mu_2$-bounded noise setting, as $h_u$ and $h^*$ are identical, it immediately follows that $\text{err}(h_{\hat{w}}) - \text{err}(h^*) \leq \frac{\epsilon}{2} \leq \epsilon$.

We now bound the label complexity of Algorithm 1. The total number of labels queried is $\sum_{k=0}^{k_0} n_k$, where $n_k \leq c_1 \cdot t (\ln d + \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta_0})^3$, and $k_0 = O(\ln \frac{1}{\epsilon})$. As a consequence, the total number of label queries is $O(t \cdot (\ln d + \ln \frac{1}{\epsilon})^3 \cdot \ln \frac{1}{\epsilon})$ in terms of $t$, $d$ and $\epsilon$. The theorem follows.

**6. Conclusions and future work**

We give a computationally efficient PAC active halfspace learning algorithm that enjoys sharp attribute efficient label complexity bounds. It combines the margin-based framework of Balcan et al. (2007); Balcan and Long (2013) with iterative hard thresholding (Blumensath and Davies, 2009; Garg and Khandekar, 2009). The main novel technical component in our analysis is a uniform concentration bound of hinge losses over shrinking $\ell_1$ balls in the sampling regions. We outline several promising directions of future research:

- Can we extend our algorithm to work under $\eta$-bounded noise, when $\eta$ is arbitrarily close to $\frac{1}{2}$? Recall that the results of Zhang and Chaudhuri (2014) imply a computationally inefficient algorithm with a label complexity of $O\left(\frac{\ln d}{(1-2\eta)^2} \ln \frac{1}{\epsilon}\right)$ in this setting, which state of the art efficient algorithms (e.g. Awasthi et al., 2016) cannot achieve.
Can we design attribute and computationally efficient active learning algorithms that work under broader distributions? Existing results in the active learning and one-bit compressed sensing literature have made substantial progress on settings when the unlabeled distribution is $\alpha$-stable (Li, 2016), subgaussian (Ai et al., 2014; Chen and Banerjee, 2015), or s-concave (Balcan and Zhang, 2017); an attribute and computationally efficient, statistically consistent recovery algorithm under any of the above settings would be a step forward.

In one-bit compressed sensing, under the symmetric noise assumption (Plan and Vershynin, 2013b), algorithms with sample complexity polynomial in $\frac{1}{\epsilon}$ have been proposed (Plan and Vershynin, 2013b; Zhang et al., 2014; Zhu and Gu, 2015). Can we develop adaptive one-bit compressed sensing algorithms with $O(t \text{polylog}(d, \frac{1}{\epsilon}))$ measurement complexity in this setting?

Acknowledgments

I am grateful to Daniel Hsu for suggesting this research direction to me, and many insightful discussions along this line. I would also like to thank Pranjal Awasthi, Jie Shen and Hongyang Zhang for helpful initial conversations about the results in this paper. I thank the anonymous reviewers for their thoughtful comments. Special thanks to Yue Liu, who provided unconditional support throughout this research project.

References


Appendix A. Detailed choices of learning and problem parameters

In this section, we give the exact settings of $c_1, c_2, c_3$ that appears in Algorithm 1, and $\mu_1, \mu_2$, the noise rates that can be tolerated by Algorithm 1 under the two noise conditions.

Define $D_k$ as the distribution $D$ over $(x, y)$ conditioned on that $x$ lies in $B_k$. Although cannot be observed directly, for analysis purposes, we define $\tilde{D}$ as the joint distribution of $(x, \text{sign}(u \cdot x))$, and $\tilde{D}_k$ as the distribution of $\tilde{D}$ conditioned on that $x$ lies in $B_k$. Let $\lambda > 0$. We defer the exact setting of $\lambda$ to the end of this section. Given $\lambda$, we define $c_2 := c_2(\lambda)$ such that:

1. $c_2(\lambda) = O(\ln \frac{1}{\lambda})$.
2. For all $w$ such that $\theta(w, w_{k-1}) \leq 2^{-k-3}\pi$,

\[
\mathbb{P}_D(\text{sign}(w \cdot x) \neq \text{sign}(w_{k-1} \cdot x), |w_{k-1} \cdot x| \geq c_2(\lambda) \cdot 2^{-k}) \leq \lambda \cdot 2^{-k}.
\] (3)

The existence of such function $c_2(\cdot)$ is guaranteed by Theorem 21 of Balcan and Long (2013).

In addition, given $\lambda > 0$, define $c_3(\lambda) := \lambda \min(C_3/81, C_3c_2(\lambda)/9)$ (where $C_3$ is a numerical constant defined in Lemma 18), such that $\tau_k = c_32^{-k}$. Under this setting of $\tau_k$, we have that for all $k$ in $\{0, 1, \ldots, k_0\}$:

\[
\mathbb{E}_{D_k} \ell_{\tau_k}(u, x, y) \leq \mathbb{P}_{D_k}(|u \cdot x| \leq \tau_k) \leq \frac{\mathbb{P}_{D_X}(|u \cdot x| \leq \tau_k)}{\mathbb{P}_{D_X}(x \in B_k)} \leq \frac{9\tau_k}{\min(C_3/9, C_3b_k)} \leq \lambda.
\] (4)

where the second inequality uses the fact that $\mathbb{P}(A \mid B) \leq \frac{\mathbb{P}(A)}{\mathbb{P}(B)}$ for any two events $A$, $B$, and the third inequality uses Lemma 18, and the fact that $\mathbb{P}_{D_X}(x \in B_k) = \min(C_3/9, C_3b_k)$ for all $k$ in $\{0, 1, \ldots, k_0\}$.

Recall that $n_k := c_1t(\ln d + \ln \frac{1}{\epsilon} + \ln \frac{1}{6b})^3$. Given $\lambda > 0$ and $c_2(\lambda)$, $c_3(\lambda)$, we set $c_1 := c_1(\lambda)$ such that by Lemmas 13, for all $k$ in $\{0, 1, \ldots, k_0\}$, for all $w$ in $W_k$,

\[
|\mathbb{E}_{S_k} \ell_{\tau_k}(w, (x, y)) - \mathbb{E}_{D_k} \ell_{\tau_k}(w, (x, y))| \leq \lambda.
\] (5)

Given $\lambda$ and $c_2(\lambda)$, $c_3(\lambda)$, we also choose $\mu_1 = \mu_1(\lambda)$, $\mu_2 = \mu_2(\lambda) \in (0, \frac{1}{2})$ such that under the respective noise condition, for all $k$ in $\{0, 1, \ldots, k_0\}$, for all $w$ in $W_k$,

\[
|\mathbb{E}_{D_k} \ell_{\tau_k}(w, (x, y)) - \mathbb{E}_{D_k} \ell_{\tau_k}(w, (x, y))| \leq \lambda.
\] (6)

The existences of $\mu_1(\lambda)$ and $\mu_2(\lambda)$ are guaranteed in light of Lemmas 21 and 22.

Define $\lambda = \min\{\lambda' : C_2(45c_2(\lambda')\lambda' + 5\lambda') \geq 2^{-8}\pi\} > 0$. The existence of $\lambda$ can be seen by noting that $c_2(\lambda)\lambda$ goes to zero as $\lambda$ goes down to zero.

Appendix B. Learning guarantee at each epoch

In this section, we prove two key lemmas, namely Lemmas 7 and 8, both of which serve as the basis of Lemma 6.
B.1. Proof of Lemma 7

The proof of Lemma 7 is based on a uniform concentration bound on the \(\tau_k\)-hinge loss over \(W_k\) in the sampling region \(B_k\), namely Lemma 13. By Lemma 13, the difference between the empirical hinge losses and the expected hinge losses for all \(w\) in \(W_k\) with respect to \(D_k\) is uniformly bounded by \(\tilde{O}(\sqrt{\frac{(ln d + ln \frac{1}{\epsilon})^3}{n_k}})\). As will be seen in the analysis, only a constant concentration error \(\lambda\) is required in the hinge loss minimization step (see Equation (5)). The setting of \(n_k = O(t(ln d + ln \frac{1}{\epsilon})^3)\) indeed fulfills this requirement.

**Proof** [Proof of Lemma 7] We consider the cases of \(k = 0\) and \(k \geq 1\) separately.

**Case 1:** \(k = 0\). By Lemma 9 below and the fact that \(D = D_0\), \(\mathbb{P}_D(\text{sign}(w_0' \cdot x) \neq \text{sign}(u \cdot x)) \leq 5\lambda\) holds. In addition, by the definition of \(\lambda\), and the second inequality of of Equation (1) \((\theta(w_0', u) \leq C_2\mathbb{P}_D(\text{sign}(w_0' \cdot x) \neq \text{sign}(u \cdot x)))\), we have that \(\theta(w_0', u) \leq 2^{-8\pi}\). The lemma follows.

**Case 2:** \(k \geq 1\). By Lemma 9 below, \(\mathbb{P}_{D_k}(\text{sign}(w_k' \cdot x) \neq \text{sign}(u \cdot x)) \leq 5\lambda\) holds. We now show that the above fact implies that the angle between \(w_k'\) and \(u\) is at most \(2^{-k-8\pi}\); this implication is well known in the margin-based active learning literature (Balcan et al., 2007; Balcan and Long, 2013); we provide the proof here for completeness.

By Lemma 18, the above fact implies that

\[
\mathbb{P}_D(\text{sign}(w_k' \cdot x) \neq \text{sign}(u \cdot x), x \in B_k) = \mathbb{P}_{D_k}(\text{sign}(w_k' \cdot x) \neq \text{sign}(u \cdot x)) \cdot \mathbb{P}_{D_{\pi}}(x \in B_k) \leq 5\lambda \cdot 9c_2(\lambda)2^{-k} \leq 45c_2(\lambda)2^{-k}. \tag{7}
\]

On the other hand, observe that for all \(w\) in \(W_k\), \(\|w - w_{k-1}\|_2 \leq 2^{-k-3}\). Using Lemma 10 and the fact that \(w_{k-1}^\perp\) is a unit vector, we get that for all \(w\) in \(W_k\), \(\theta(w, w_{k-1}) \leq 2^{-k-3\pi}\). Specifically, by Equation (3), we have that

\[
\mathbb{P}_D(\text{sign}(w \cdot x) \neq \text{sign}(w_{k-1} \cdot x), x \notin B_k) \leq \lambda 2^{-k}
\]
holds for \(w \in \{u, w_k'\} \subset W_k\) respectively. Therefore, by triangle inequality,

\[
\begin{align*}
\mathbb{P}_D(\text{sign}(w_k' \cdot x) & \neq \text{sign}(u \cdot x), x \notin B_k) \\
& \leq \mathbb{P}_D(\text{sign}(w_k' \cdot x) \neq \text{sign}(w_{k-1} \cdot x), x \notin B_k) + \mathbb{P}_D(\text{sign}(u \cdot x) \neq \text{sign}(w_{k-1} \cdot x), x \notin B_k) \\
& \leq 2\lambda 2^{-k}. \tag{8}
\end{align*}
\]

Combining Equations (7) and (8), we have that

\[
\mathbb{P}_D(\text{sign}(w_k' \cdot x) \neq \text{sign}(u \cdot x)) \leq (45c_2(\lambda)\lambda + 2\lambda)2^{-k}.
\]

Applying the second inequality of Equation (1) gives that

\[
\theta(w_k', u) \leq C_2(45c_2(\lambda) + 2)\lambda 2^{-k}.
\]

By the definition of \(\lambda\), the above is at most \(2^{-k-8\pi}\). The lemma follows.

**Lemma 9** For every \(k\) in \(\{0, 1, \ldots, k_0\}\), if \(u\) is in \(W_k\), then

\[
\mathbb{P}_{D_k}(\text{sign}(w_k' \cdot x) \neq \text{sign}(u \cdot x)) \leq 5\lambda.
\]
Proof If \( u \) is in \( W_k \), then we have the following chain of inequalities:

\[
\mathbb{P}_{D_k}(\text{sign}(w'_k \cdot x) \neq \text{sign}(u \cdot x)) = \mathbb{P}_{\tilde{D}_k}(\text{sign}(w'_k \cdot x) \neq y)
\leq \mathbb{E}_{\tilde{D}_k} \ell_{\tau_k}(w'_k, x, y)
\leq \mathbb{E}_{D_k} \ell_{\tau_k}(w'_k, x, y) + \lambda
\leq \mathbb{E}_{S_k} \ell_{\tau_k}(w'_k, x, y) + 2\lambda
\leq \mathbb{E}_{D_k} \ell_{\tau_k}(u, x, y) + 3\lambda
\leq \mathbb{E}_{S_k} \ell_{\tau_k}(u, x, y) + 4\lambda
\leq \lambda + 4\lambda = 5\lambda,
\]

where the first inequality is from the fact that the \( \tau_k \)-hinge loss is an upper bound of the 0-1 loss; the second inequality is from Equation (6) and that \( w'_k \in W_k \); the third inequality is from Equation (5) and that \( w'_k \in W_k \); the fourth inequality is by the optimality of \( w'_k \) in optimization problem (2) and that \( u \in W_k \); the fifth inequality is from Equation (5) and that \( u \in W_k \); the sixth inequality is from Equation (6) and that \( u \in W_k \); the last inequality is from Equation (4).

The following lemma is used in the proof of Lemma 7; it establishes a connection between angle and \( \ell_2 \) distance of two vectors, when one of them has unit \( \ell_2 \) norm.

**Lemma 10** Suppose \( v \) is an unit vector in \( \mathbb{R}^d \) (that is, \( \|v\|_2 = 1 \)). Then, for any vector \( w \) in \( \mathbb{R}^d \),

\[
\theta(w, v) \leq \pi \|w - v\|_2.
\]

**Proof** Denote by \( \hat{w} \) the \( \ell_2 \) normalized version of \( w \), i.e. \( \hat{w} = \frac{w}{\|w\|_2} \). Lemma 11 below implies that

\[
\|\hat{w} - v\|_2 \leq 2\|w - v\|_2.
\]  \(\text{(9)}\)

Consequently,

\[
\theta(w, v) \leq \frac{\pi}{2} \cdot 2 \sin \left( \frac{\theta(w, v)}{2} \right) = \frac{\pi}{2} \cdot 2 \sin \left( \frac{\theta(\hat{w}, v)}{2} \right) = \frac{\pi}{2} \|\hat{w} - v\|_2 \leq \pi \|w - v\|_2.
\]

where the first inequality is from the elementary inequality that \( \phi \leq \frac{\pi}{2} \sin \phi \) for \( \phi \in [0, \frac{\pi}{2}] \) (by taking \( \phi = \frac{\theta(w,v)}{2} \)), the second inequality is from the identity that \( \|\hat{w} - v\|_2 = 2 \sin \left( \frac{\theta(\hat{w},v)}{2} \right) \) as both \( \hat{w} \) and \( v \) are unit vectors, and the last inequality is from Equation (9).

The following lemma is used in the proof of Lemma 10; it shows that \( \ell_2 \) normalization is roughly a projection onto the unit sphere.

**Lemma 11** Suppose \( v \) is an unit vector in \( \mathbb{R}^d \) (that is, \( \|v\|_2 = 1 \)). Then, for any vector \( w \) in \( \mathbb{R}^d \),

\[
\|\frac{w}{\|w\|_2} - v\|_2 \leq 2\|w - v\|_2.
\]
\textbf{Proof} Denote by $\hat{w}$ the $\ell_2$ normalized version of $w$, i.e. $\hat{w} = \frac{w}{\|w\|_2}$. We have that by triangle inequality,
\[ \|\hat{w} - w\|_2 = \left\| \left( \frac{1}{\|w\|_2} - 1 \right) w \right\|_2 = \|w\|_2 - 1 = \|w\|_2 - 2 \leq \|w - v\|_2. \]
Again by triangle inequality,
\[ \|\hat{w} - v\|_2 \leq \|\hat{w} - w\|_2 + \|w - v\|_2 \leq 2\|w - v\|_2. \]
\[ \qed \]

\textbf{B.2. Proof of Lemma 8}

The proof of Lemma 8 is based on the insight that the hard thresholding operation $P_t$ is effectively a projection onto the $\ell_0$-ball $\{ w \in \mathbb{R}^d : \|w\|_0 \leq t \}$; see Lemma 12 for a formal description.

\textbf{Proof} [Proof of Lemma 8] Denote by $\hat{w}'_k$ the $\ell_2$ normalized version of $w'_k$: $\hat{w}'_k := \frac{w'_k}{\|w'_k\|_2}$. Under the condition that $\theta(w'_k, u) \leq 2^{-k-8}\pi$, as $\hat{w}'_k$ and $u$ are both unit vectors, we have
\[ \|\hat{w}'_k - u\|_2 = 2 \sin \frac{\theta(w'_k, u)}{2} \leq \theta(w'_k, u) \leq 2^{-k-8}\pi \leq 2^{-k-6}. \]
Now, by Lemma 12 below, we have that $\|\hat{w}'_k - P_t(\hat{w}'_k)\|_2 \leq \|\hat{w}'_k - u\|_2 \leq 2^{-k-6}$. By triangle inequality of $\ell_2$ distance, we have that
\[ \| P_t(\hat{w}'_k) - u \|_2 \leq \|\hat{w}'_k - w_k\|_2 + \|\hat{w}'_k - u\|_2 \leq 2^{-k-5}. \]
Observe that as $w_k$ and $\hat{w}'_k$ are equal up to scaling, $w_k := \frac{P_t(\hat{w}'_k)}{\|P_t(\hat{w}'_k)\|_2}$ is identically $\frac{P_t(\hat{w}'_k)}{\|P_t(\hat{w}'_k)\|_2}$. Applying Lemma 11 with $w = P_t(\hat{w}'_k)$ and $\hat{w}'_k = u$, we get that
\[ \|w_k - u\|_2 \leq 2\| P_t(\hat{w}'_k) - u \|_2 \leq 2^{-k-4} = r_{k+1}. \]
In addition, as $w_k$ and $u$ are both $t$-sparse, $w_k - u$ is $2t$-sparse. Therefore, by Cauchy-Schwarz, $\|w_k - u\|_1 \leq \sqrt{2t}\|w_k - u\|_2 \leq \sqrt{2t}r_{k+1} = \rho_{k+1}$. Therefore, $u$ is in the set $\{ w \in \mathbb{R}^d : \|w-w_k\|_2 \leq r_{k+1} \text{ and } \|w-w_k\|_1 \leq \rho_{k+1} \}$, namely $W_{k+1}$.
\[ \qed \]

\textbf{Lemma 12} Suppose $w$ is a vector in $\mathbb{R}^d$. Then, for any $t$-sparse vector $v$ in $\mathbb{R}^d$,
\[ \| P_t(w) - w \|_2 \leq \|v - w\|_2. \]
In other words, $P_t(w)$ is the best $t$-sparse approximation to $w$, measured in $\ell_2$ distance.

\textbf{Proof} Denote by $w(1), w(2), \ldots, w(d)$ the $d$ entries of $w$ in descending order in magnitude. We have that
\[ \| P_t(w) - w \|_2^2 = \sum_{i=t+1}^d w(i)^2. \]
On the other hand, for any $t$-sparse vector $v$, denote by $S$ its support. We have that

$$
\|v - w\|_2^2 \geq \sum_{i \in \{1, \ldots, d\} \setminus S} w_i^2 \geq \sum_{i=t+1}^d w_i^2,
$$

where the second inequality is from that the sum of squares of any $d-t$ entries in $w$ must be greater than that of the bottom $d-t$ entries. The lemma follows.

\[\]

**Appendix C. The uniform concentration of hinge losses in label query regions**

In contrast to Awasthi et al. (2016) where the constraint set of the hinge loss minimization problem at epoch $k$ is the intersection of an $\ell_2$ ball of $O(2^{-k})$ radius and an $\ell_1$ ball of $O(\sqrt{k})$ radius, Algorithm 1 defines the constraint set $W_k$ to be the intersection of an $\ell_2$ ball of $O(2^{-k})$ radius and an $\ell_1$ ball of $O(\sqrt{T2^{-k}})$ radius. The following key lemma, namely Lemma 13, gives justification on the rationale behind our construction of $W_k$. Specifically, it establishes a sharp uniform concentration of hinge losses $\ell_{r_k}$ over $W_k$, with respect to sample $S_k$ drawn from distribution $D_k$. Observe that the concentration bound is $\tilde{O}(\sqrt{\frac{t(ln d + \ln \frac{1}{2}))}{n_k}^{k/2})$; if one were to use the constraint set in Awasthi et al. (2016), one can only get concentration bounds of order $\tilde{O}(\sqrt{\frac{tln d.2^n}{n_k}})$, which is exponentially looser in $k$.

**Lemma 13** For any $c_2, c_3 > 0$, there exists a constant $C_6 > 0$ such that the following holds. Given $k$ in $\{0, 1, \ldots, k_0\}$, suppose $S_k$ is a sample of size $n_k$ drawn from distribution $D_k$. Then with probability $1 - \delta_k$, for all $w \in W_k$, we have:

$$
|\mathbb{E}_{S_k} \ell_{r_k}(w, (x, y)) - \mathbb{E}_{D_k} \ell_{r_k}(w, (x, y))| \leq C_6 \frac{n_k d}{e \delta_k} \cdot \sqrt{\frac{t(ln d + \ln \frac{2}{\delta_k})}{n_k}}.
$$

**Proof** First we show that there is an event $E$ that has probability at least $1 - \delta_k/2$, conditioned on which all the unlabeled examples in $S_k$ have uniformly bounded $\ell_\infty$ norms. Define:

$$
E := \{ \text{for all } (x, y) \text{ in } S_k, (x, y) \text{ is in } T_k \},
$$

where $T_k := \{(x, y) : \|x\|_\infty \leq R_k \}$. Here, $R_k = C_7 \frac{2n_k d}{\delta_k} \max(\frac{9}{C_3}, \frac{1}{C_3 b_k})$ for some large enough constant $C_7$ such that $\mathbb{P}_{D_k}((x, y) \not\in T_k) \leq \min(C_3/9, C_3 b_k) \delta_k/2n_k$ holds. The existence of such $C_7$ is guaranteed by Lemma 20 of Awasthi et al. (2016).

Under the above definitions, for each individual $(x, y)$ in $S_k$ drawn from $D_k$,

$$
\mathbb{P}_{D_k}((x, y) \not\in T_k) \leq \frac{\mathbb{P}_{D_k}((x, y) \not\in T_k)}{\mathbb{P}_{D_k}(x \in B_k)} \leq \frac{\min(C_3/9, C_3 b_k) \delta_k/2n_k}{\min(C_3/9, C_3 b_k)} \leq \frac{\delta_k}{2n_k},
$$

therefore, by union bound, $\mathbb{P}(E) \geq 1 - \delta_k/2$. 

19
By Lemma 14 below, there is an event $F$ such that $\mathbb{P}[F|E] \geq 1 - \delta_k/2$, and on event $F$,

$$\left| \mathbb{E}_{S_k} \ell_{\tau_k}(w, (x,y)) - \mathbb{E}_{D_k\mid \tau_k} \ell_{\tau_k}(w, (x,y)) \right| \leq C_8 \cdot \ln \frac{n_k d}{\epsilon \delta_k} \cdot \sqrt{\frac{t(2 \ln d + \ln \frac{d}{\delta_k})}{n_k}}. \quad (11)$$

for some constant $C_8$ defined in Lemma 14.

Note that $\mathbb{P}(E \cap F) \geq (1 - \delta_k/2)^2 \geq 1 - \delta_k$. We henceforth condition on $E \cap F$ happening. Using Lemma 16, we get that for all $w$ in $W_k$,

$$\left| \mathbb{E}_{S_k} \ell_{\tau_k}(w, (x,y)) - \mathbb{E}_{D_k} \ell_{\tau_k}(w, (x,y)) \right| \leq C_9 \sqrt{\frac{1}{n_k} (1 + \frac{b_k^2 + r_k^2}{\tau_k^2})}, \quad (12)$$

for some constant $C_9$ defined in Lemma 16.

Combining Equations (11) and (12), we conclude that there is a constant $C_6$ such that on event $E \cap F$,

$$\left| \mathbb{E}_{S_k} \ell_{\tau_k}(w, (x,y)) - \mathbb{E}_{D_k} \ell_{\tau_k}(w, (x,y)) \right| \leq C_6 \ln \frac{n_k d}{\epsilon \delta_k} \cdot \sqrt{\frac{t(2 \ln d + \ln \frac{d}{\delta_k})}{n_k}}.$$

This proves the lemma. $\blacksquare$

**Lemma 14** For every $k$ in $\{0, 1, \ldots, k_0\}$, suppose event $E$ is defined as in Equation (10). Then there is an event $F$ such that $\mathbb{P}[F|E] \geq 1 - \delta_k/2$, and on event $F$, for all $w$ in $W_k$,

$$\left| \mathbb{E}_{S_k} \ell_{\tau_k}(w, (x,y)) - \mathbb{E}_{D_k\mid \tau_k} \ell_{\tau_k}(w, (x,y)) \right| \leq C_8 \cdot \ln \frac{n_k d}{\epsilon \delta_k} \cdot \sqrt{\frac{t(2 \ln d + \ln \frac{d}{\delta_k})}{n_k}},$$

for some constant $C_8 > 0$ that depends on $c_2$ and $c_3$.

**Proof** Conditioned on event $E$, sample $S_k$ can be seen as drawn iid from $D_k\mid \tau_k$. We consider the cases of $k = 0$ and $k \geq 1$ separately.

**Case 1:** $k = 0$. Using Corollary 4 of Kakade et al. (2009) with $\ell = \pm \ell_{\tau_0}$, $L_\ell = \frac{1}{\tau_0}$, $X = c \cdot \log \frac{n_0 d}{\epsilon \delta_0}$, and $W_1 = \sqrt{t}$ in the notations therein, we get that there is an event $F$, such that $\mathbb{P}[F|E] \geq 1 - \delta_0/2$, on which for all $w$ in $W_k$,

$$\left| \mathbb{E}_{S_0} \ell_{\tau_0}(w, (x,y)) - \mathbb{E}_{D\mid \tau_0} \ell_{\tau_0}(w, (x,y)) \right| \leq \frac{c}{\tau_0} \ln \frac{n_0 d}{C_3 \delta_0} \cdot \sqrt{\frac{32t(2 \ln d + \ln \frac{d}{\delta_0})}{n_0}}.$$

**Case 2:** $k \geq 1$. From Lemma 15 below, we have that there is an event $F$, such that $\mathbb{P}[F|E] \geq 1 - \delta_k/2$, on which for some constant $C_{10} > 0$ and for all $w$ in $W_k$,

$$\left| \mathbb{E}_{S_k} \ell_{\tau_k}(w, (x,y)) - \mathbb{E}_{D_k\mid \tau_k} \ell_{\tau_k}(w, (x,y)) \right| \leq \sqrt{\frac{2009}{\delta_0}} \cdot \sqrt{\frac{t(2 \ln d + \ln \frac{d}{\delta_k})}{n_k}}.$$
where the second inequality is by observing that \( \frac{b_k}{\tau_k} = \frac{c_k}{c_i} \) and \( \frac{\rho}{\tau_k} = \frac{\sqrt{\gamma}}{c_i} \) and recalling that \( R_k = C_7 \ln(\frac{2n_d}{\delta_k}) \max(\frac{9}{C_9}, \frac{1}{c_3\delta_k}) \).

Combining the above two cases, we can find a large enough constant \( C_8 > 0 \) such that the lemma statement holds. \[ \square \]

We next show Lemma 15, a key concentration result used in Lemma 14.

**Lemma 15** Given \( k \in \{1, \ldots, k_0\} \), suppose \( S_k \) is a set of \( n_k \) iid samples drawn from \( D_k|_{T_k} \). We have that with probability \( 1 - \delta_k/2 \), for all \( w \) in \( W_k \),

\[
|E_{S_k} \ell_{T_k}(w, (x, y)) - E_{D_k} \ell_{T_k}(w, (x, y))| \leq (1 + \frac{b_k}{\tau_k} + \frac{\rho_k R_k}{\tau_k}) \sqrt{\frac{2 \ln d + \ln \frac{2}{\delta_k}}{n_k}}.
\]

**Proof** First, for all \( w \) in \( W_k \), \((x, y) \in T_k\), the instantaneous hinge loss \( \ell_{T_k}(w, (x, y)) \) is at most \((1 + |w - x|) \leq (1 + |w_{k-1} - x| + |w - w_{k-1} - x|) \leq (1 + \frac{b_k}{\tau_k} + \frac{\rho_k R_k}{\tau_k})\). By standard symmetrization arguments (see Theorem 8 of Bartlett and Mendelson (2002)), we have that with probability \( 1 - \delta_k/2 \), for all \( w \) in \( W_k \),

\[
|E_{S_k} \ell_{T_k}(w, (x, y)) - E_{D_k} \ell_{T_k}(w, (x, y))| \leq (1 + \frac{b_k}{\tau_k} + \frac{\rho_k R_k}{\tau_k}) \sqrt{\frac{\ln \frac{2}{\delta_k}}{2n_k}} + R_n(\mathcal{F}), \tag{13}
\]

where \( R_n(\cdot) \) denotes the Rademacher complexity over the examples in \( S_k \), \( \mathcal{F} \) is the set of functions \( \{(x, y) \rightarrow (1 - \frac{\alpha w - x}{\tau_k})_+ : w \in W_k\} \). Note that \( \mathcal{F} \) can be written as the composition of \( \phi_k(\alpha) := (1 - \frac{\alpha}{\tau_k})_+ \) and function class \( \mathcal{G} := \{(x, y) \rightarrow yw \cdot x : w \in W_k\} \).

By the contraction inequality of Rademacher complexity (see Theorem 12 of Bartlett and Mendelson (2002)) and the \( \frac{1}{\tau_k} \) Lipschitzness of \( \phi_k \), \( R_n(\mathcal{F}) \) is at most \( \frac{1}{\tau_k} R_n(\mathcal{G}) \). We now focus on bounding \( R_n(\mathcal{G}) \). First, denote by \((x_i, y_i), i = 1, \ldots, n_k \) the elements of \( S_k \). By the definition of Rademacher complexity,

\[
R_n(\mathcal{G}) = \frac{1}{n_k} E_{\sigma} \sup_{w \in W_k} \sum_{i=1}^{n_k} \sigma_i y_i w \cdot x_i,
\]

where \( \sigma = (\sigma_1, \ldots, \sigma_{n_k}) \), \( \sigma_i \)'s are iid random variables that take values uniformly in \( \{-1, +1\} \).

It can be easily seen that \( \sigma \) has the same distribution as \((\sigma_1 y_1, \ldots, \sigma_{n_k} y_{n_k})\). Hence, \( R_n(\mathcal{G}) \) can be simplified to

\[
R_n(\mathcal{G}) = \frac{1}{n_k} E_{\sigma} \sup_{w \in W_k} \sum_{i=1}^{n_k} \sigma_i w \cdot x_i.
\]

We bound \( R_n(\mathcal{G}) \) as follows:

\[
R_n(\mathcal{G}) \leq \frac{1}{n_k} E_{\sigma} \sup_{w : ||w - w_{k-1}||_1 \leq \rho_k} \sum_{i=1}^{n_k} \sigma_i w \cdot x_i
\]

\[
= \frac{1}{n_k} E_{\sigma} \sup_{v : ||v||_1 \leq \rho_k} \sum_{i=1}^{n_k} \sigma_i (w_{k-1} \cdot x_i + v \cdot x_i)
\]

\[
= \frac{1}{n_k} E_{\sigma} \sum_{i=1}^{n_k} \sigma_i v \cdot x_i + \frac{1}{n_k} E_{\sigma} \sum_{i=1}^{n_k} \sigma_i w_{k-1} \cdot x_i,
\]
where the inequality uses the fact that all $w$’s in $W_k$ satisfies that $\|w - w_{k-1}\|_1 \leq \rho_k$.

As all $x_i$’s have $\ell_\infty$ norm at most $R_k$, by Theorem 1, Example 2 of Kakade et al. (2009), the first term is bounded by $\rho_k \cdot R_k \cdot \sqrt{\frac{2 \ln d}{n_k}}$. In addition, as all $(x_i, y_i)$’s are sampled from $D_k$, for all $i$, $|w_{k-1} \cdot x_i| \leq b_k$. Therefore, the second term can be bounded by:

$$
\frac{1}{n_k} \mathbb{E}_{\sigma} \left[ \sum_{i=1}^{n_k} \sigma_i w_{k-1} \cdot x_i \right] \leq \frac{1}{n_k} \mathbb{E}_{\sigma} \left[ \left( \sum_{i=1}^{n_k} \sigma_i w_{k-1} \cdot x_i \right)^2 \right] \leq b_k \sqrt{\frac{1}{n_k}}.
$$

Summing the two bounds up, we have that $R_n(G) \leq (b_k + \rho_k R_k) \sqrt{\frac{2 \ln d}{n_k}}$. Therefore,

$$
R_n(F) \leq (\frac{b_k}{r_k} + \frac{\rho_k}{r_k} R_k) \sqrt{\frac{2 \ln d}{n_k}}.
$$

Combining this inequality with Equation (13), along with some algebraic calculations, we get the lemma as stated.

**Lemma 16** There is a constant $C_0 > 0$ such that for all $k$ in $\{0, 1, \ldots, k_0\}$, $w$ in $W_k$,

$$
|\mathbb{E}_{D_k \mid T_k} \ell_{r_k}(w, (x, y)) - \mathbb{E}_{D_k} \ell_{r_k}(w, (x, y))| \leq C_0 \sqrt{\frac{1}{n_k}} (1 + \frac{b_k^2 + r_k^2}{\tau_k^2}).
$$

**Proof** We consider the cases of $k = 0$ and $k \geq 1$ separately.

**Case 1:** $k = 0$. Observe that $\mathbb{P}_D((x, y) \notin T_0) \leq \frac{\delta_0}{2n_0 \leq 1}$, and $\mathbb{E}_D(w \cdot x)^2 \leq 1$ for $w$ in $W_0$ as $D$ is isotropic. Using Lemma 17, this implies that

$$
\left| \mathbb{E}_{D \mid T_0} \ell_{r_0}(w, (x, y)) - \mathbb{E}_D \ell_{r_0}(w, (x, y)) \right| \leq \frac{6}{n_0}.
$$

**Case 2:** $k \geq 1$. Observe that by Lemma 19, there is a constant $C_4$ such that for all $w$ in $W_k \subset \{w \in \mathbb{R}^d : \|w - w_{k-1}\|_2 \leq r_k\}$, $\mathbb{E}_{D_k}(w \cdot x)^2 \leq C_4(b_k^2 + r_k^2)$. In addition, $\mathbb{P}_{D_k}((x, y) \notin T_k) \leq \frac{1}{n_k}$. Therefore, by Lemma 17,

$$
\left| \mathbb{E}_{D_k \mid T_k} \ell_{r_k}(w, (x, y)) - \mathbb{E}_{D_k} \ell_{r_k}(w, (x, y)) \right| \leq 6 \sqrt{\frac{1}{n_k} (1 + \frac{C_4(b_k^2 + r_k^2)}{\tau_k^2})}.
$$

Combining the above two cases, we can find a large enough constant $C_0 > 0$ such that the lemma statement holds.

In the proof of Lemma 16, we use the following lemma to bound the difference between $\mathbb{E}_{D_k \mid T_k} \ell_{r_k}(w, (x, y))$ and $\mathbb{E}_{D_k} \ell_{r_k}(w, (x, y))$ in terms of $T_k$’s probability mass in $D_k$ and $D_k$’s second moments.
Lemma 17 For \( k \) in \( \{0, 1, \ldots, k_0\} \), if \( \mathbb{P}_{D_k}(x, y) \notin T_k \leq \frac{\delta_k}{2n^k} \), then the following inequality holds for all \( w \) in \( W_k \):

\[
|E_{D_k\mid T_k} \ell_{\tau_k}(w, (x, y)) - E_{D_k} \ell_{\tau_k}(w, (x, y))| \leq 6 \sqrt{\mathbb{P}_{D_k}(x, y) \notin T_k} \cdot \sqrt{1 + \frac{E_{D_k}(w \cdot x)^2}{\tau_k^2}}.
\]

Proof First, observe that

\[
E_{D_k}(w, (x, y)) = E_{D_k\mid T_k} \ell_{\tau_k}(w, (x, y)) \mathbb{P}_{D_k}(x, y) \in T_k \cdot \mathbb{P}_{D_k}(x, y) \notin T_k).
\]

Therefore,

\[
\begin{align*}
&|E_{D_k\mid T_k} \ell_{\tau_k}(w, (x, y)) - E_{D_k} \ell_{\tau_k}(w, (x, y))| \\
&= \frac{\mathbb{P}_{D_k}(x, y) \notin T_k}{\mathbb{P}_{D_k}(x, y) \in T_k} E_{D_k}(w, (x, y)) - \frac{E_{D_k}(x, y) \notin T_k}{E_{D_k}(x, y) \in T_k} E_{D_k}(w, (x, y))I((x, y) \notin T_k) \\
&\leq 2\mathbb{P}_{D_k}(x, y) \notin T_k)E_{D_k}(1 + \frac{|w \cdot x|}{\tau_k}) + 2E_{D_k}(w, (x, y))I((x, y) \notin T_k) \\
&\leq 2\mathbb{P}_{D_k}(x, y) \notin T_k)E_{D_k}(1 + \frac{|w \cdot x|}{\tau_k}) + 2E_{D_k}(1 + \frac{|w \cdot x|}{\tau_k})I((x, y) \notin T_k) \\
&\leq 2\mathbb{P}_{D_k}(x, y) \notin T_k)E_{D_k}(1 + \frac{E_{D_k}(w \cdot x)^2}{\tau_k^2}) + 2\mathbb{P}_{D_k}(x, y) \notin T_k)E_{D_k}(1 + \frac{(w \cdot x)^2}{\tau_k}) \\
&\leq 6\sqrt{\mathbb{P}_{D_k}(x, y) \notin T_k) \cdot \sqrt{1 + \frac{E_{D_k}(w \cdot x)^2}{\tau_k^2}}.
\end{align*}
\]

where the equality is from Equation (14) and algebra; the first inequality is from that \( \mathbb{P}_{D_k}(x, y) \in T_k \geq 1 - \frac{\delta_k}{2n^k} \geq \frac{1}{2} \); the second inequality is from that \( \ell_{\tau_k}(w, (x, y)) \leq (1 + \frac{|w \cdot x|}{\tau_k}) \); the third inequality is by Cauchy-Schwarz on both terms, and the last inequality is by algebra (using the following elementary inequalities: \( \sqrt{a} + \sqrt{b} \leq \sqrt{2(a + b)} \), \( \mathbb{P}_{D_k}(x, y) \notin T_k \) \leq 1 and \( (a + b)^2 \leq 2(a^2 + b^2) \)).

Appendix D. Auxiliary lemmas

The lemmas in this section are known and used in previous works (See e.g. Awasthi et al., 2017, 2016); we collect them here for completeness.

The following lemma characterizes the density of an isotropic log-concave distribution projected onto one dimension.

Lemma 18 (Lovász and Vempala (2007)) There exists a numerical constant \( C_3 \in (0, 1) \) such that

\[
\min(C_3/9, C_3b) \leq \mathbb{P}_{D_X}(|v \cdot x| \leq b) \leq 9b.
\]
The following technical lemma bounds the second moments of $D|_B$ with band $B$, along directions almost perpendicular to $B$.

**Lemma 19 (Awasthi et al. (2017))** There is a numerical constant $C_4 > 0$, such that the following holds. If $w$ is a unit vector, and $B = \{w : |w \cdot x| \leq b\}$ is a band of width $b > 0$ along $w$. Then, for all $w' \in B_2(w, r)$, we have

$$E_{D_X|_B}(w' \cdot x)^2 \leq C_4(r^2 + b^2).$$

Recall that $D$ (resp. $\tilde{D}$) is the joint distribution over $(x, y)$ (resp. $(x, \text{sign}(u \cdot x))$). The following lemma shows that under certain “local low noise” conditions on $D$, the expected $\tau_k$-hinge loss over $D_k$ is close to that over $\tilde{D}_k$. In Lemmas 21 and 22, we will justify that the above low noise condition is satisfied under the $t$-sparse $\mu_1\epsilon$-adversarial noise condition and $t$-sparse $\mu_2$-bounded noise condition for sufficiently small $\mu_1$ and $\mu_2$.

**Lemma 20** For any choice of $c_2, c_3 > 0$, there exists a constant $C_5 > 0$ such that the following holds. For every $k \in \{0, 1, \ldots, k_0\}$, suppose $P_{D_k}(y \neq \text{sign}(u \cdot x)) \leq \xi_k$, then for every $w \in W_k$,

$$|E_{D_k} \ell_{\tau_k}(w, (x, y)) - E_{\tilde{D}_k} \ell_{\tau_k}(w, (x, y))| \leq \sqrt{C_5} \xi_k. \quad (15)$$

**Proof** We first bound the difference as follows:

$$|E_{D_k} \ell_{\tau_k}(w, (x, y)) - E_{\tilde{D}_k} \ell_{\tau_k}(w, (x, y))| = |E_{D_k} [\ell_{\tau_k}(w, (x, y)) - \ell_{\tau_k}(w, (x, \text{sign}(u \cdot x))))]|$$

$$= |E_{D_k} I(y \neq \text{sign}(u \cdot x)) \cdot (\ell_{\tau_k}(w, (x, y)) - \ell_{\tau_k}(w, (x, \text{sign}(u \cdot x))))|$$

$$\leq E_{D_k} I(y \neq \text{sign}(u \cdot x)) \cdot 2 \frac{|w \cdot x|}{\tau_k}$$

$$\leq 2 \sqrt{P_{D_k}(y \neq \text{sign}(u \cdot x)) E_{D_k}(w \cdot x)^2} \frac{|w \cdot x|}{\tau_k} \quad (16)$$

where the first inequality is from that an example $(x, y)$ drawn from $\tilde{D}_k$ satisfies $y = \text{sign}(u \cdot x)$ with probability 1; the second inequality is by decomposing 1 to $I(y \neq \text{sign}(u \cdot x)) + I(y = \text{sign}(u \cdot x))$; the first inequality is by that $|1 + \frac{|w \cdot x|}{\tau_k} - (1 - \frac{|w \cdot x|}{\tau_k})_+| \leq 2 \frac{|w \cdot x|}{\tau_k}$; the second inequality is by Cauchy-Schwarz. We now consider the cases of $k = 0$ and $k \geq 1$ respectively.

**Case 1:** $k = 0$. In this case, $W_0$ is a subset of $\{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$ and $D_0 = D$. Therefore, for all $w$ in $W_0$, $E_{D_0}(w \cdot x)^2 \leq 1$ as $D$ is isotropic log-concave. Continuing Equation (16), we get that

$$|E_{D_0} \ell_{\tau_0}(w, (x, y)) - E_{D_0} \ell_{\tau_0}(w, (x, y))| \leq 2 \sqrt{\frac{\xi_0}{\tau_0}} = 2 \sqrt{\frac{\xi_0}{\epsilon_3^2}}.$$
Case 2: $k \geq 1$. In this case, $W_k$ is a subset of $\{w \in \mathbb{R}^d : \|w - w_{k-1}\|_2 \leq r_k\}$. By Lemma 19, and the choices of $b_k$ and $r_k$, we have that for all $w \in W_k$,

$$\mathbb{E}_{D_k}(w \cdot x)^2 \leq C_4(r_k^2 + b_k^2). \tag{17}$$

By the definitions of $r_k$, $b_k$, and $\tau_k$, we can find a constant $C_5 \geq \frac{1}{r_k^2}$ (that depends on $c_2$, $c_3$) such that for all $k \geq 1$, the right hand side of Equation (17) is at most $\frac{C_5 \tau_k^2}{4}$. Continuing Equation (16), we get that

$$|\mathbb{E}_{D_k} \ell_{\tau_k}(w, (x, y)) - \mathbb{E}_{\widetilde{D}_k} \ell_{\tau_k}(w, (x, y))| \leq \sqrt{C_5 \xi_k}.$$

Combining the above two cases and the choice of $C_5$, we conclude that Equation (15) holds for all $k \in \{0, 1, \ldots, k_0\}$.

Applying the above lemma to the two noise settings respectively, we have:

Lemma 21 For any $\lambda > 0$ and $c_2, c_3 > 0$, there exists a constant $\mu_1 > 0$ such that the following holds. Suppose $D$ satisfies the $t$-sparse $\mu_1 \epsilon$-bounded noise condition. For every $k \in \{0, \ldots, k_0\}$, and $w$ in $W_k$,

$$|\mathbb{E}_{D_k} \ell_{\tau_k}(w, (x, y)) - \mathbb{E}_{\widetilde{D}_k} \ell_{\tau_k}(w, (x, y))| \leq \lambda.$$

Proof By Lemma 20, it suffices to check that for all $k \in \{0, 1, \ldots, k_0\}$, $\mathbb{P}_{D_k}(y \neq \text{sign}(u \cdot x)) \leq \frac{\lambda^2}{C_5^2}$.

Observe that

$$\mathbb{P}_{D_k}(y \neq \text{sign}(u \cdot x)) \leq \frac{\mathbb{P}_D(y \neq \text{sign}(u \cdot x))}{\mathbb{P}_D(x \in B_k)}.$$

by the fact that $\mathbb{P}(A|B) \leq \frac{\mathbb{P}(A)}{\mathbb{P}(B)}$ for any two events $A, B$. We now consider the cases of $k = 0$ and $k \geq 1$ respectively.

Case 1: $k = 0$. In this case, $B_k = \mathbb{R}^d$, hence $\mathbb{P}_D(x \in B_k) = 1$. It suffices to set $\mu_1 \leq \frac{\lambda^2}{C_5^2}$.

Case 2: $k \geq 1$. In this case, $\mathbb{P}_D(x \in B_k) \geq \min\{C_3/9, C_3 b_k\} \geq \min\{C_3/9, C_3 c_2 C_1 \epsilon/2\}$, where the first inequality is from Lemma 18; the second inequality is from the definition of $b_k$ and $k \leq k_0$. Therefore, for sufficiently small $\mu_1$, if $\mathbb{P}_D(y \neq \text{sign}(u \cdot x)) \leq \mu_1 \epsilon$, then $\mathbb{P}_{D_k}(y \neq \text{sign}(u \cdot x)) \leq \frac{2 \mu_1}{\min\{C_3/9, C_3 c_2 C_1 \epsilon/2\}} \leq \frac{\lambda^2}{C_5^2}$.

Combining the above two cases, we can pick a sufficiently small $\mu_1$ such that the requirements on $\mu_1$ in both cases are satisfied. This completes the proof.

Lemma 22 For any $\lambda > 0$, there exists a constant $\mu_2 > 0$ such that the following holds. Suppose $D$ satisfies the $t$-sparse $\mu_2 \epsilon$-bounded noise condition. For every $k \in \{0, \ldots, k_0\}$, and $w$ in $W_k$,

$$|\mathbb{E}_{D_k} \ell_{\tau_k}(w, (x, y)) - \mathbb{E}_{\widetilde{D}_k} \ell_{\tau_k}(w, (x, y))| \leq \lambda.$$

Proof By Lemma 20, it suffices to check that for all $k \in \{0, 1, \ldots, k_0\}$, $\mathbb{P}_{D_k}(y \neq \text{sign}(u \cdot x)) \leq \frac{\lambda^2}{C_5^2}$.

This can indeed be satisfied by setting $\mu_2 = \frac{\epsilon^2}{C_5^2}$, which immediately implies $\mathbb{P}_{D_k}(y \neq \text{sign}(u \cdot x)) \leq \mu_2 \leq \frac{\lambda^2}{C_5^2}$. \hfill \blacksquare