Attribute-Efficient Learning of Halfspaces with Malicious Noise: Near-Optimal Label Complexity and Noise Tolerance

Jie Shen JIE.SHEN@STEVENS.EDU
Stevens Institute of Technology
Hoboken, New Jersey, USA

Chicheng Zhang CHICHENGZ@CS.ARIZONA.EDU
University of Arizona
Tucson, Arizona, USA

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

This paper is concerned with computationally efficient learning of homogeneous sparse halfspaces in $\mathbb{R}^d$ under noise. Though recent works have established attribute-efficient learning algorithms under various types of label noise (e.g. bounded noise), it remains an open question of when and how s-sparse halfspaces can be efficiently learned under the challenging *malicious noise* model, where an adversary may corrupt both the unlabeled examples and the labels. We answer this question in the affirmative by designing a computationally efficient active learning algorithm with near-optimal label complexity of $\tilde{O}(s \log^4 \frac{d}{\epsilon})$[1] and noise tolerance $\eta = \Omega(\epsilon)$, where $\epsilon \in (0, 1)$ is the target error rate, under the assumption that the distribution over (uncorrupted) unlabeled examples is isotropic log-concave. Our algorithm can be straightforwardly tailored to the passive learning setting, and we show that its sample complexity is $\tilde{O}(\frac{1}{\epsilon} s^2 \log^5 d)$ which also enjoys attribute efficiency. Our main techniques include attribute-efficient paradigms for soft outlier removal and for empirical risk minimization, and a new analysis of uniform concentration for unbounded instances – all of them crucially take the sparsity structure of the underlying halfspace into account.

Keywords: halfspaces, malicious noise, passive and active learning, attribute efficiency

1. Introduction

This paper investigates the fundamental problem of learning halfspaces under noise (Valiant, 1984, 1985). In the absence of noise, this problem is well understood (Rosenblatt, 1958; Blumer et al., 1989). However, the premise changes immediately when the unlabeled examples[2] or the labels are corrupted by noise. In the last decades, various types of label noise have been extensively studied, and a plethora of polynomial-time algorithms have been developed that are resilient to random classification noise (Blum et al., 1996), bounded noise (Sloan, 1988, 1992; Massart and Nédélec, 2006), and adversarial noise (Kearns et al., 1992; Kalai et al., 2005). Significant progress towards optimal noise tolerance is also witnessed in the past few years (Daniely, 2015; Awasthi et al., 2015; Yan and Zhang, 2017; Diakonikolas et al., 2019, 2020). In this regard, a surge of recent research interest is concentrated on further improvement of the performance guarantees by leveraging the structure of the underlying halfspace into algorithmic design. Of central interest is a property termed attribute efficiency, which proves to be useful when the data lie in a high-dimensional space (Littlestone, 1987), or even in an infinite-dimensional space but with bounded

1. We use the notation $\tilde{O}(f) := O(f \log f)$.
2. We will also refer to unlabeled examples as instances in this paper.

number of effective attributes (Blum, 1990). In the statistics and signal processing community, it is often referred to as sparsity, dating back to the celebrated Lasso estimator (Tibshirani, 1996; Chen et al., 1998; Candès and Tao, 2005; Donoho, 2006). Recently, learning of sparse halfspaces in an attribute-efficient manner was highlighted as an open problem in Feldman (2014), and in a series of recent works (Plan and Vershynin, 2013b; Awasthi et al., 2016; Zhang, 2018; Zhang et al., 2020), this property was carefully explored for label-noise-tolerant learning of halfspaces with improved or even near-optimal sample complexity, label complexity, or generalization error, where the key insight is that such structural constraint effectively controls the complexity of the hypothesis class (Zhang, 2002; Kakade et al., 2008).

Compared to the rich set of positive results on attribute-efficient learning of sparse halfspaces under label noise, less is known when both instances and labels are corrupted. Specifically, under the η-malicious noise model (Valiant, 1985; Kearns and Li, 1988), there is an unknown hypothesis w^* and an unknown instance distribution D selected from a certain family by an adversary. Each time with probability $1-\eta$, the adversary returns an instance x drawn from D and the label $y = \operatorname{sign}(w^* \cdot x)$; with probability η, it instead is allowed to return an arbitrary pair $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ that may depend on the state of the learning algorithm and the history of its outputs. Since this is a much more challenging noise model, only recently has an algorithm with near-optimal noise tolerance been established in Awasthi et al. (2017), although without attribute efficiency. It is worth noting that the problem of learning sparse halfspaces is also closely related to one-bit compressed sensing (Boufounos and Baraniuk, 2008) where one is allowed to utilize any distribution D over measurements for recovering the target hypothesis. However, even with such strong condition, existing theory therein can only handle label noise (Plan and Vershynin, 2013a; Awasthi et al., 2016; Baraniuk et al., 2017). This naturally raises two fundamental questions: 1) can we design attribute-efficient learning algorithms that are capable of tolerating the malicious noise; and 2) can we still obtain near-optimal performance guarantees on the degree of noise tolerance and on the sample complexity.

In this paper, we answer the two questions in the affirmative under a mild distributional assumption that D is chosen from the family of isotropic log-concave distributions (Lovász and Vempala, 2007; Vempala, 2010), which covers prominent distributions such as normal distributions, exponential distributions, and logistic distributions. Moreover, we take label complexity into consideration (Cohn et al., 1994), for which we show that our bound is near-optimal in that aspect. We build our algorithm upon the margin-based active learning framework (Balcan et al., 2007), which queries the label of an instance when it has small "margin" with respect to the currently learned hypothesis.

From a high level, this work can be thought of as extending the best known result of Awasthi et al. (2017) to the high-dimensional regime. However, even under the low-dimensional setting where $s = d$, our bound of label complexity is better than theirs in terms of the dependence on the dimension d: they have a quadratic dependence whereas we have a linear dependence (up to logarithmic factors). Moreover, as we will describe in Section 3, obtaining such algorithmic extension is nontrivial both computationally and statistically. This work can also be viewed as an extension of Zhang (2018) to the malicious noise model. In fact, our construction of empirical risk minimization is inspired by that work. However, they considered only label noise which makes their algorithm and analysis not applicable to our setting: it turns out that when facing malicious noise, a sophisticated design of outlier removal paradigm is crucial for optimal noise tolerance (Klivans et al., 2009).

Also in line with this work is learning with nasty noise (Diakonikolas et al., 2018) and robust sparse functional estimation (Balakrishnan et al., 2017). Both works considered more general setting

in the following sense: Diakonikolas et al. (2018) showed that by properly adapting the techniques in robust mean estimation, some more general concepts, e.g. low-degree polynomial threshold functions and intersections of halfspaces, can be efficiently learned with $\mathrm{poly}\left(d, 1/\epsilon\right)$ sample complexity; Balakrishnan et al. (2017) showed that under proper sparsity assumptions, a sample complexity bound of $\mathrm{poly}\left(s, \log d, 1/\epsilon\right)$ can be achieved for many sparse estimation problems, such as generalized linear models with Lipschitz mapping functions and covariance estimation. However, we remark that neither of them obtained label efficiency. In addition, when adapted to our setting, Theorem 1.5 of Diakonikolas et al. (2018) only handles noise rate $\eta \leq O(\epsilon^c)$ for some constant c that is greater than one, while as to be shown in Section 4, we obtain the near-optimal noise tolerance $\eta \leq O(\epsilon)$. Balakrishnan et al. (2017) achieved near-optimal noise tolerance but their analysis is restricted to the Gaussian marginal distribution and Lipschitz mapping functions. In addition to such fundamental differences, the main techniques we develop are distinct from theirs, which will be described in more detail in Section 3.3.3.

1.1. Main results

We informally present our main results below; readers are referred to Theorem 4 in Section 4 for a precise statement.

Theorem 1 (Informal) *Consider the malicious noise model with noise rate η. If the unlabeled data distribution is isotropic log-concave and the underlying halfspace w^* is s-sparse, then there is an algorithm that for any given target error rate $\epsilon \in (0, 1)$, PAC learns the underlying halfspace in polynomial time provided that $\eta \leq O(\epsilon)$. In addition, the label complexity is $\tilde{O}\left(s \log^4 \frac{d}{\epsilon}\right)$ and the sample complexity is $\tilde{O}\left(\frac{1}{\epsilon} s^2 \log^5 d\right)$.*

First of all, note that the noise tolerance is near-optimal as Kearns and Li (1988) showed that a noise rate greater than $\frac{\epsilon}{1+\epsilon}$ cannot be tolerated by any algorithm regardless of the computational power. The following fact establishes the optimality of our label complexity.

Lemma 2 *Active learning of s-sparse halfspaces under isotropic log-concave distributions in the realizable case has an information-theoretic label complexity lower bound of $\Omega\left(s(\log \frac{1}{\epsilon} + \log \frac{d}{s})\right)$.*

To see this lemma, observe that there exist ϵ-packings of s-sparse halfspaces with sizes $(\frac{1}{\epsilon})^{\Omega(s)}$ (Long, 1995) and $(\frac{d}{s})^{\Omega(s)}$ (Raskutti et al., 2011); applying Theorem 1 of Kulkarni et al. (1993) gives the lower bound.

1.2. Related works

Kearns and Li (1988) presented a general analysis on efficiently learning halfspaces, showing that even without any distributional assumptions, it is possible to tolerate the malicious noise at a rate of $\Omega(\epsilon/d)$, but a noise rate greater than $\frac{\epsilon}{1+\epsilon}$ cannot be tolerated. The noise model was further studied by Schapire (1992); Bshouty (1998); Cesa-Bianchi et al. (1999), and Kalai et al. (2005) obtained a noise tolerance $\Omega(\epsilon/d^{1/4})$ when D is the uniform distribution. Klivans et al. (2009) improved this result to $\Omega(\epsilon^2/\log(d/\epsilon))$ for the uniform distribution, and showed a noise tolerance $\Omega(\epsilon^3/\log^2(d/\epsilon))$ for isotropic log-concave distributions. A near-optimal result of $\Omega(\epsilon)$ was established in Awasthi et al. (2017) for both uniform and isotropic log-concave distributions.

Achieving attribute efficiency has been a long-standing goal in machine learning and statistics (Blum, 1990; Blum et al., 1995), and has found a variety of applications with strong theoretical backend. A partial list includes online classification (Littlestone, 1987), learning decision lists (Servedio, 1999; Klivans and Servedio, 2004; Long and Servedio, 2006), compressed sensing (Donoho, 2006; Candès and Wakin, 2008; Tropp and Wright, 2010; Shen and Li, 2018), one-bit compressed sensing (Boufounos and Baraniuk, 2008; Plan and Vershynin, 2016), and variable selection (Fan and Li, 2001; Fan and Fan, 2008; Shen and Li, 2017a,b).

Label-efficient learning has also been broadly studied since gathering high quality labels is often expensive. The prominent approaches include disagreement-based active learning (Hanneke, 2011, 2014), margin-based active learning (Balcan et al., 2007; Balcan and Long, 2013; Yan and Zhang, 2017), 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) and search (Balcan and Hanneke, 2012; Beygelzimer et al., 2016).

Recent works such as Diakonikolas et al. (2016); Lai et al. (2016) studied mean estimation under a strong noise model where in addition to returning dirty instances, the adversary has also the power of eliminating a few clean instances, similar to the nasty noise model in learning halfspaces (Bshouty et al., 2002). The main technique of robust mean estimation is a novel outlier removal paradigm, which uses the spectral norm of the covariance matrix to detect dirty instances. This is similar in spirit to the idea of Klivans et al. (2009); Awasthi et al. (2017) and the current work. However, there is no direct connection between mean estimation and halfspace learning since the former is an unsupervised problem while the latter is supervised (although any connection would be very interesting). Very recently, such technique was extensively investigated in a variety of problems such as clustering and linear regression; we refer the reader to a comprehensive survey by Diakonikolas and Kane (2019) for more information.

Roadmap. We collect useful notations and formally define the problem in Section 2. In Section 3, we describe our algorithms, followed by a theoretical analysis in Section 4. We conclude this paper in Section 5, and defer all proof details to the appendix.

2. Preliminaries

We study the problem of learning sparse halfspaces in $\mathbb{R}^d$ under the malicious noise model with noise rate $\eta \in [0, 1/2)$ (Valiant, 1985; Kearns and Li, 1988), where an oracle $\mathrm{EX}_\eta(D, w^*)$ (i.e. adversary) first selects a member D from a family of distributions $\mathcal{D}$ and a concept w^* from a concept class $\mathcal{C}$; during the learning process, D and w^* are fixed. Each time the adversary is called, with probability $1-\eta$, a random pair (x, y) is returned to the learner with $x \sim D$ and $y = \mathrm{sign}\,(w^* \cdot x)$, referred to as a clean sample; with probability η, the adversary can return an *arbitrary* pair $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$, referred to as a dirty sample. The adversary is assumed to have unrestricted computational power to search dirty samples that may depend on, e.g. the states of the learning algorithm and the history of its outputs. Formally, we make the following distributional assumptions.

Assumption 1 *Let $\mathcal{D}$ be the family of isotropic log-concave distributions. The underlying distribution D from which clean instances are drawn is chosen from $\mathcal{D}$ by the adversary, and is fixed during the learning process. The learner is given the knowledge of $\mathcal{D}$ but not of D.*

Assumption 2 *With probability* $1-\eta$*, the adversary returns a pair* (x, y) *where* $x \sim D$ *and* $y = \operatorname{sign}(w^* \cdot x)$*; with probability* η*, it may return an arbitrary pair* $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$.

Since we are interested in obtaining a label-efficient algorithm, we will consider a natural extension of such passive learning model. In particular, Awasthi et al. (2017) proposed to consider the following: when a labeled instance (x, y) is generated, the learner only has access to an instance-generation oracle $\mathrm{EX}^x_\eta(D, w^*)$ which returns x, and must make a separate call to a label revealing oracle $\mathrm{EX}^y_\eta(D, w^*)$ to obtain y. We refer to the total number of calls to $\mathrm{EX}^x_\eta(D, w^*)$ as the sample complexity of the learning algorithm, and to that of $\mathrm{EX}^y_\eta(D, w^*)$ as the label complexity.

We will presume that the concept class $\mathcal{C}$ consists of homogeneous halfspaces that have unit ℓ_2-norm and are s-sparse, i.e. the number of non-zero elements of any $w \in \mathcal{C}$ is at most s where $s \in \{1, 2, \ldots, d\}$. The learning algorithm is given this concept class, that is, the set of homogeneous s-sparse halfspaces. For a hypothesis $w \in \mathcal{C}$, we define its error rate on a distribution D as $\mathrm{err}_D(w) = \Pr_{x\sim D}\left(\operatorname{sign}(w \cdot x) \neq \operatorname{sign}(w^* \cdot x)\right)$. The goal of the learner is to find a hypothesis w in polynomial time such that with probability $1-\delta$, $\mathrm{err}_D(w) \leq \epsilon$ for any given failure confidence $\delta \in (0, 1)$ and any error rate $\epsilon \in (0, 1)$, with a few calls to $\mathrm{EX}^x_\eta(D, w^*)$ and $\mathrm{EX}^y_\eta(D, w^*)$.

For a reference vector $u \in \mathbb{R}^d$ and a positive scalar b, we call the region $X_{u,b} := \{x \in \mathbb{R}^d : |u \cdot x| \leq b\}$ as band, and we denote by $D_{u,b}$ the distribution obtained by conditioning D on the event $x \in X_{u,b}$. Given a hypothesis w in $\mathbb{R}^d$, a labeled instance (x, y), and a parameter $\tau > 0$, we define the τ-hinge loss $\ell_\tau(w; x, y) = \max\left\{0, 1 - \frac{1}{\tau} y(w \cdot x)\right\}$. For a labeled set $S = \{(x_i, y_i)\}_{i=1}^n$, we define $\ell_\tau(w; S) = \frac{1}{n}\sum_{i=1}^n \ell_\tau(w; x_i, y_i)$.

For $p \geq 1$, we denote by $B_p(u, r)$ the ℓ_p-ball centering at the point u with radius $r > 0$, i.e. $B_p(u, r) = \{w \in \mathbb{R}^d : \|w - u\|_p \leq r\}$. We will be particularly interested in the cases $p = 1, 2, \infty$. For a vector $u \in \mathbb{R}^d$, the hard thresholding operation $\mathcal{H}_s(u)$ keeps its s largest (in absolute value) elements and sets the remaining to zero. Let $u, v \in \mathbb{R}^d$ be two vectors; we write $\theta(u, v)$ to denote the angle between them, and write $u \cdot v$ to denote their inner product. For a matrix H, we denote by $\|H\|_*$ its trace norm (also known as the nuclear norm), i.e. the sum of its singular values. We will also use $\|H\|_1$ to denote the entrywise ℓ_1-norm of H, i.e. the sum of absolute values of its entries. If H is a symmetric matrix, we use $H \succeq 0$ to denote that it is positive semidefinite.

Throughout this paper, the subscript variants of the lowercase letter c, e.g. c_1 and c_2, are reserved for specific absolute constants that are uniquely determined by the distribution D. We also reserve C_1 and C_2 for specific constants. We remark that the value of all the constants involved in the paper does not depend on the underlying distribution D, but rather on the knowledge of $\mathcal{D}$ given to the learner. We collect all the definitions of these constants in Appendix A.

3. Main Algorithm

We first present an overview of our learning algorithm, followed by specifying all the hyper-parameters used therein. Then we describe in detail the attribute-efficient outlier removal scheme, which is the core technique in the paper.

3.1. Overview

Our main algorithm, namely Algorithm 1, is based on the celebrated margin-based active learning framework (Balcan et al., 2007). The key observation is that a good classifier can be learned by concentrating on fitting only the most informative labeled instances, measured by the closeness to

the current decision boundary (i.e. the closer the more informative). In our algorithm, the sampling region is set as $\mathbb{R}^d$ at phase $k = 1$, and is set as the band $X_{w_{k-1},b_k} = \{x \in \mathbb{R}^d : |w_{k-1} \cdot x| \leq b_k\}$ at phases $k \geq 2$. Once we obtain the instance set $\bar{T}$, we perform a pruning step that removes all instances having large ℓ_∞-norm. This is motivated by our analysis that with high probability, all clean instances in $\bar{T}$ must have small ℓ_∞-norm provided that Assumption 1 is satisfied. Since the oracle $\mathrm{EX}^x_\eta(D, w^*)$ may output dirty instances, we design an attribute-efficient soft outlier removal procedure, which aims to find proper weights for all instances in T, such that the clean instances (i.e. those from D_{w_{k-1},b_k}) have overwhelming weights compared to dirty instances. Equipped with the learned weights, it is possible to minimize the reweighted hinge loss to obtain a refined halfspace. However, this would lead to a suboptimal label complexity since we have to query the label for all instances in T. Our remedy is to randomly sample a few points from T according to their importance, which is crucial for us to obtain near-optimal label complexity.

When minimizing the hinge loss, we carefully construct the constraint set W_k with three properties. First, it has an ℓ_2-norm constraint. As a useful fact of isotropic log-concave distributions, the ℓ_2-distance to the underlying halfspace w^* is of the same order as the error rate. Thus, if we were able to ensure that the target halfspace w^* stays in W_k, we would show that the error rate of w_k is as small as $O(r_k)$, the radius of the ℓ_2-ball. Second, W_k has an ℓ_1-norm constraint, which is well-known for its power to promote sparse solutions and to guarantee attribute-efficient sample complexity (Tibshirani, 1996; Chen et al., 1998; Candès and Tao, 2005; Plan and Vershynin, 2013b). Lastly, the ℓ_2 and ℓ_1 radii of W_k shrinks by a constant factor in each phase; hence, when Algorithm 1 terminates, the radius of the ℓ_2-ball will be as small as $O(\epsilon)$. Notably, Zhang (2018) also utilizes such constraint for active learning of sparse halfspaces, but only under the setting of label noise.

The last step in Algorithm 1 is to perform hard-thresholding $\mathcal{H}_s$ on the solution v_k followed by ℓ_2-normalization. Roughly speaking, these two steps will produce an iterate w_k consistent with the structure of w^* (i.e. w_k is guaranteed to belong to the concept class $\mathcal{C}$), and more importantly, will be useful to show that w^* lies in W_k in all phases.

3.2. Hyper-parameter setting

We elaborate on our hyper-parameter setting that is used in Algorithm 1 and our analysis. Let $g(t) = c_2 \left(2t \exp(-t) + \frac{c_3\pi}{4} \exp\left(-\frac{c_4 t}{4\pi}\right) + 16 \exp(-t)\right)$, where the constants are specified in Appendix A. Observe that there exists an absolute constant $\bar{c} \geq 8\pi/c_4$ satisfying $g(\bar{c}) \leq 2^{-8}\pi$, since the continuous function $g(t) \to 0$ as $t \to +\infty$ and all the involved quantities in $g(t)$ are absolute constants. Given such constant $\bar{c}$, we set $b_k = \bar{c} \cdot 2^{-k-3}$, $\tau_k = c_0 \kappa \cdot \min\{b_k, 1/9\}$, $\delta_k = \frac{\delta}{(k+1)(k+2)}$,

$$r_k = \begin{cases} 1, & k = 1 \\ 2^{-k-3}, & k \geq 2 \end{cases}, \text{ and } \rho_k = \begin{cases} \sqrt{s}, & k = 1 \\ \sqrt{2s} \cdot 2^{-k-3}, & k \geq 2 \end{cases}.$$

We set the constant $\kappa = \exp(-\bar{c})$, and choose $\xi_k = \min\left\{\frac{1}{2}, \frac{\kappa^2}{16}\left(1 + 4\sqrt{C_2} z_k/\tau_k\right)^{-2}\right\}$. Observe that all ξ_k's are lower bounded by the constant $c_6 := \min\left\{\frac{1}{2}, \frac{\kappa^2}{16}\left(1 + \frac{4}{c_0\kappa\bar{c}}\sqrt{C_2\bar{c}^2 + C_2}\right)^{-2}\right\}$. Our theoretical guarantee holds for any noise rate $\eta \leq c_5\epsilon$, where the constant $c_5 := \frac{c_8}{2\pi}\bar{c}c_1c_6$.

We set the total number of phases $k_0 = \left\lceil \log\left(\frac{\pi}{16c_1\epsilon}\right) \right\rceil$ in Algorithm 1. Consider any phase $k \geq 1$. We use $n_k = \tilde{O}\left(s^2 \log^4 \frac{d}{b_k} \cdot \left(\log d + \log^3 \frac{1}{\delta_k}\right)\right)$ as the size of unlabeled instance set $\bar{T}$. We will show that by making $N_k = O\left(n_k/b_k\right)$ calls to $\mathrm{EX}^x_\eta(D, w^*)$, Algorithm 1 is guaranteed to obtain

Algorithm 1 Attribute and Label-Efficient Algorithm Tolerating Malicious Noise

Require: Error rate ϵ, failure probability δ, sparsity parameter s, an instance generation oracle $\mathrm{EX}^x_\eta(D, w^*)$, a label revealing oracle $\mathrm{EX}^y_\eta(D, w^*)$.

Ensure: A halfspace w_{k_0} such that $\mathrm{err}_D(w_{k_0}) \le \epsilon$ with probability $1-\delta$.

1: $k_0 \leftarrow \left\lceil \log\left(\frac{\pi}{16 c_1 \epsilon}\right) \right\rceil$.
2: Initialize w_0 as the zero vector in $\mathbb{R}^d$.
3: **for** phases $k = 1, 2, \ldots, k_0$ **do**
4: Clear the working set $\bar{T}$.
5: If $k = 1$, independently draw n_k instances from $\mathrm{EX}^x_\eta(D, w^*)$ and put them into $\bar{T}$; otherwise, draw n_k instances from $\mathrm{EX}^x_\eta(D, w^*)$ conditioned on $|w_{k-1} \cdot x| \le b_k$ and put into $\bar{T}$.
6: **Pruning:** Remove all instances x in $\bar{T}$ with $\|x\|_\infty > c_9 \log \frac{48 n_k d}{b_k \delta_k}$ to form a set T.
7: **Soft outlier removal:** Apply Algorithm 2 to T with $u \leftarrow w_{k-1}$, $b \leftarrow b_k$, $r \leftarrow r_k$, $\rho \leftarrow \rho_k$, $\xi \leftarrow \xi_k$, $C \leftarrow 2C_2$, and let $q = \{q(x)\}_{x \in T}$ be the returned function. Normalize q to form a probability distribution p over T.
8: **Random sampling:** $S_k \leftarrow$ Independently draw m_k instances (with replacement) from T according to p and query $\mathrm{EX}^y_\eta(D, w^*)$ for their labels.
9: Let $W_k = B_2(w_{k-1}, r_k) \cap B_1(w_{k-1}, \rho_k)$. Find $v_k \in W_k$ such that

$$\ell_{\tau_k}(v_k; S_k) \le \min_{w \in W_k} \ell_{\tau_k}(w; S_k) + \kappa.$$

10: $w_k \leftarrow \frac{\mathcal{H}_s(v_k)}{\|\mathcal{H}_s(v_k)\|_2}$.
11: **end for**
12: **return** w_{k_0}.

such $\bar{T}$ in each phase with high probability. We set $m_k = \tilde{O}\left(s \log^2 \frac{d}{b_k \delta_k} \cdot \log \frac{d}{\delta_k}\right)$ as the size of labeled instance set S_k, which is also the number of calls to $\mathrm{EX}^y_\eta(D, w^*)$. Note that $N := \sum_{k=1}^{k_0} N_k$ is the sample complexity of Algorithm 1, and $m := \sum_{k=1}^{k_0} m_k$ is its label complexity.

3.3. Attribute and computationally efficient soft outlier removal

Our soft outlier removal procedure is inspired by Awasthi et al. (2017). We first briefly describe their main idea. Then we introduce a natural extension of their approach to the high-dimensional regime and show why it fails. Lastly, we present our novel outlier removal scheme.

To ease our discussion, we decompose $T = T_{\mathrm{C}} \cup T_{\mathrm{D}}$ where T_{C} is the set of clean instances in T and T_{D} consists of all dirty instances. Ideally, we would expect to find a function $q : T \to [0, 1]$ such that $q(x) = 1$ for all $x \in T_{\mathrm{C}}$ and $q(x) = 0$ otherwise. Suppose that ξ is the fraction of dirty instances in T. Then one would expect that the total weights $\sum_{x \in T} q(x)$ is as large as $(1 - \xi)\,|T|$ in order to include such ideal function. On the other hand, we must restrict the weights of dirty instances; namely, we need to characterize under what conditions T_{C} can be distinguished from T_{D}. The key observation made in Klivans et al. (2009) and Awasthi et al. (2017) is that if the dirty instances want to deteriorate the hinge loss (which is the purpose of the adversary), they must lead to a variance[3] of $w \cdot x$ orders of magnitude larger than $\Omega(b^2 + r^2)$ on the direction of a particular

3. We follow Awasthi et al. (2017) and slightly abuse the word "variance" without subtracting the squared mean of $w \cdot x$.

Algorithm 2 Attribute-Efficient Localized Soft Outlier Removal

Require: Reference vector u, band width b, radius r for ℓ_2-ball, radius ρ for ℓ_1-ball, empirical noise rate ξ, absolute constant C, a set of unlabeled instances T where for all $x \in T$, $|u \cdot x| \leq b$.

Ensure: A function $q : T \to [0, 1]$.

1: Define the convex set of matrices $\mathcal{M} = \left\{H \in \mathbb{R}^{d\times d} : \ H \succeq 0, \ \|H\|_* \leq r^2, \ \|H\|_1 \leq \rho^2\right\}$.

2: Find a function $q : T \to [0, 1]$ satisfying the following constraints:

1. for all $x \in T, 0 \leq q(x) \leq 1$;
2. $\sum_{x\in T} q(x) \geq (1 - \xi)\,|T|$;
3. $\sup_{H\in\mathcal{M}} \frac{1}{|T|} \sum_{x\in T} q(x) x^\top H x \leq C(b^2 + r^2)$.

3: **return** q.

halfspace. Thus, it suffices to find a proper weight for each instance, such that the reweighted variance $\frac{1}{|T|} \sum_{x\in T} q(x)(w \cdot x)^2$ is as small as $O(b^2 + r^2)$ for all feasible halfspaces $w \in W$. Now it remains to resolve two questions: 1) how many instances do we need to draw in order to guarantee the existence of such function q; and 2) how to find a feasible function q in polynomial time.

If label complexity were our only objective, we could have used the soft outlier removal procedure of Awasthi et al. (2017) directly, i.e. we set $W = B_2(u, r)$, which in conjunction with the ℓ_1-norm constrained hinge loss minimization of Zhang (2018) would result in an $\tilde{O}\left(\frac{d^2}{\epsilon}\right)$ sample complexity and a $\mathrm{poly}\left(s, \log d, \log(1/\epsilon)\right)$ label complexity. However, as we would also like to optimize for the learner's sample complexity by utilizing the sparsity assumption, we need an attribute-efficient outlier removal procedure.

3.3.1. A NATURAL APPROACH AND WHY IT FAILS

It is well-known that incorporating an ℓ_1-norm constraint often leads to a sample complexity sublinear in the dimension (Zhang, 2002; Kakade et al., 2008). Thus, a natural approach for attribute-efficient outlier removal is to set $W = B_2(u, r) \cap B_1(u, \rho)$ for some carefully chosen radius $\rho > 0$. With the new localized concept space, it is possible to show that a sample size of $\mathrm{poly}\,(s, \log d)$ suffices to guarantee the existence of a function q such that the reweighted variance is small over all $w \in W$. However, on the computational side, for a given q, we will have to check the reweighted variance for all $w \in W$, which amounts to finding a global optimum of the following program:

$$\max_{w\in\mathbb{R}^d} \ \frac{1}{|T|} \sum_{x\in T} q(x)(w \cdot x)^2, \ \text{s.t.} \ \|w - u\|_2 \leq r, \ \|w - u\|_1 \leq \rho. \tag{3.1}$$

The above program is closely related to the problem of sparse principal component analysis (PCA) (Zou et al., 2006), and unfortunately it is known that finding a global optimum is NP-hard (Steinberg, 2005; Tillmann and Pfetsch, 2014).

3.3.2. CONVEX RELAXATION OF SPARSE PRINCIPAL COMPONENT ANALYSIS

Our goal is to find a function q such that the objective value in (3.1) is less than $O\left(b^2 + r^2\right)$ for all $w \in W$. To circumvent the computational intractability caused by the non-convexity of the

objective function, we consider an alternative formulation using semidefinite programming (SDP), similar to the approach of d'Aspremont et al. (2007). First, let $v = w - u$. It is not hard to see that $(w \cdot x)^2 \le 2(u \cdot x)^2 + 2(v \cdot x)^2$. Due to our localized sampling scheme, we have $(u \cdot x)^2 \le b^2$ with probability 1. Thus, we only need to examine the maximum value of $\frac{1}{|T|}\sum_{x \in T} q(x)(v \cdot x)^2$ over $v \in B_2(0, r) \cap B_1(0, \rho)$. Now the technique of d'Aspremont et al. (2007) comes in: the rank-one symmetric matrix $vv^\top$ is replaced by a new variable $H \in \mathbb{R}^{d \times d}$ which is positive semidefinite, and the vector ℓ_2 and ℓ_1-norm constraints are relaxed to the matrix trace and ℓ_1-norm constraints respectively as follows:

$$\max_{H \in \mathbb{R}^{d \times d}} \frac{1}{|T|} \sum_{x \in T} q(x) x^\top H x, \text{ s.t. } H \succeq 0, \ \|H\|_* \le r^2, \ \|H\|_1 \le \rho^2. \tag{3.2}$$

The program (3.2) has two salient features: first, it is a semidefinite program that can be optimized efficiently (Boyd and Vandenberghe, 2004); second, if its objective value is upper bounded by $O\left(b^2 + r^2\right)$, we immediately obtain that the reweighted variance is well controlled. This is the theme of the following lemma.

Lemma 3 *Suppose that Assumption 1 and 2 are satisfied, and that $\eta \le c_5\epsilon$. There exists a constant $C_2 > 2$ such that the following holds. For any phase k of Algorithm 1 with $1 \le k \le k_0$, write $\mathcal{M}_k = \{H \in \mathbb{R}^{d \times d} : \ H \succeq 0, \ \|H\|_* \le r_k^2, \ \|H\|_1 \le \rho_k^2\}$. Then with probability $1 - \frac{\delta_k}{24}$ over the draw of T_{C}, we have*

$$\sup_{H \in \mathcal{M}_k} \frac{1}{|T_{\mathrm{C}}|} \sum_{x \in T_{\mathrm{C}}} x^\top H x \le 2C_2(b_k^2 + r_k^2),$$

provided that $|T_{\mathrm{C}}| \ge \tilde{O}\left(s^2 \log^4 \frac{d}{b_k} \cdot \left(\log d + \log^2 \frac{1}{\delta_k}\right)\right)$.

Recall that Algorithm 1 sets $n_k = \tilde{O}\left(s^2 \log^4 \frac{d}{b_k} \cdot \left(\log d + \log^3 \frac{1}{\delta_k}\right)\right)$, which suffices to guarantee the condition on $|T_{\mathrm{C}}|$ holds (see Appendix D.2); therefore, the above concentration bound holds with high probability. As a result, it is not hard to verify that the function $q : T \to [0, 1]$, where $q(x) = 1$ for all $x \in T_{\mathrm{C}}$ and $q(x) = 0$ for all $x \in T_{\mathrm{D}}$, satisfies all three constraints in Algorithm 2. In other words, Lemma 3 establishes the existence of a feasible function q to Algorithm 2. Furthermore, observe that the optimization problem of finding a feasible q in Algorithm 2 is a semi-infinite linear program. For a given candidate q, we can construct an efficient oracle as follows: it checks if q violates the first two constraints; if not, it checks the last constraint by invoking a polynomial-time SDP solver to find the maximum objective value of (3.2). It is well-known that equipped with such separation oracle, Algorithm 2 will return a desired function q in polynomial time by the ellipsoid method (Grötschel et al., 2012, Chapter 3).

3.3.3. COMPARISON TO PRIOR WORKS

We remark that the setting of n_k results in a sample complexity of $\tilde{O}\left(\frac{s^2}{b_k}\right)$ for phase k (see a formal statement in Lemma 6), which implies a total sample complexity of $\tilde{O}\left(\frac{s^2}{\epsilon}\right)$. When $s \ll d$, this substantially improves upon the sample complexity of $\tilde{O}\left(\frac{d^2}{\epsilon}\right)$ when naively applying the soft outlier removal procedure in Awasthi et al. (2017).

We remark three crucial technical differences from Diakonikolas et al. (2018) and Balakrishnan et al. (2017). First, we progressively restrict the variance to identify dirty instances, i.e. the variance

upper bound is set as $O(1)$ at the beginning of Algorithm 1 and progressively decreases to $O(\epsilon^2)$ (see our setting of b_k and r_k), while in Diakonikolas et al. (2018); Balakrishnan et al. (2017) and many of their follow-up works it is typically fixed to $O(\epsilon)$. Second, we control the variance locally, i.e. we only require a small variance over a localized instance space D_{w_{k-1},b_k} and a localized concept space $\mathcal{M}_k$. Third, the small variance is used to robustly estimate the hinge loss in our work, while in Diakonikolas et al. (2018) it was utilized to approximate the Chow parameters. All these problem-specific design of outlier removal are vital for us to obtain the first near-optimal guarantee on attribute efficiency and label efficiency for learning sparse halfspaces.

4. Performance Guarantee

In the following, we always presume that the underlying halfspace is parameterized by w^*, which is s-sparse and has unit ℓ_2-norm. This condition may not be explicitly stated in our analysis.

Our main theorem is as follows. We note that there are two sources of randomness in Algorithm 1: the random draw of instances from $\mathrm{EX}^x_\eta(D, w^*)$, and the random sampling step (i.e. Step 8); the probability is taken over all the randomness in the algorithm.

Theorem 4 *Suppose that Assumptions 1 and 2 are satisfied. There exists an absolute constant c_5 such that for any $\epsilon \in (0,1)$ and $\delta \in (0,1)$, if $\eta \leq c_5\epsilon$, then with probability at least $1-\delta$, $\mathrm{err}_D(w_{k_0}) \leq \epsilon$ where w_{k_0} is the output of Algorithm 1. Furthermore, Algorithm 1 has a sample complexity of $\tilde{O}\big(\frac{1}{\epsilon}s^2\log^4 d\cdot\big(\log d+\log^3\frac{1}{\delta}\big)\big)$, and a label complexity of $\tilde{O}\big(s\log^2\frac{d}{\epsilon\delta}\cdot\log\frac{d}{\delta}\cdot\log\frac{1}{\epsilon}\big)$, and has running time* $\mathrm{poly}\big(d, 1/\epsilon, 1/\delta\big)$.

Algorithm 1 can be straightforwardly modified to work in the passive learning setting, where the learner has direct access to the labeled instance oracle $\mathrm{EX}_\eta(D, w^*)$. The modified algorithm works as follows: it calls $\mathrm{EX}_\eta(D, w^*)$ to obtain a pair of instance and the label whenever Algorithm 1 calls $\mathrm{EX}^x_\eta(D, w^*)$. In particular, for the passive learning algorithm, the working set $\bar{T}$ is always a labeled instance set, and there is no need for it to query $\mathrm{EX}^y_\eta(D, w^*)$ in the random sampling step.

We have the following simple corollary which is an immediate result from Theorem 4.

Corollary 5 *Suppose that Assumptions 1 and 2 are satisfied. There exists a polynomial-time algorithm (that has access to only $\mathrm{EX}_\eta(D, w^*)$) and an absolute constant c_5 such that for any $\epsilon \in (0,1)$ and $\delta \in (0,1)$, if $\eta \leq c_5\epsilon$, then with probability at least $1-\delta$, the algorithm outputs a hypothesis with error at most ϵ, using $\tilde{O}\big(\frac{1}{\epsilon}s^2\log^4 d\cdot\big(\log d+\log^3\frac{1}{\delta}\big)\big)$ labeled instances.*

We need an ensemble of new results to prove Theorem 4. Specifically, we propose new techniques to control the sample and computational complexity of soft outlier removal, and a new analysis of label complexity by making full use of the localization in the instance and concept spaces. We elaborate on them in the following, and sketch the proof of Theorem 4 at the end of this section.

4.1. Localized sampling in the instance space

Localized sampling, also known as margin-based active learning, is a useful technique proposed in Balcan et al. (2007). Interestingly, under isotropic log-concave distributions, Balcan and Long (2013) showed that if the band width b is large enough, the region outside the band, i.e. $\{x \in \mathbb{R}^d : |w\cdot x| > b\}$, can be safely "ignored", in the sense that, if w is close enough to w^*, it is guaranteed to incur a small error rate therein. Motivated by this elegant finding, theoretical analyses in the literature

are often dedicated to bounding the error rate within the band, and it is now well understood that a constant error rate within the band suffices to ensure significant progress in each phase (Awasthi et al., 2015, 2017; Zhang, 2018). We follow this line of reasoning and our technical contribution is to show how to obtain such constant error rate with near-optimal label complexity and noise tolerance.

Our analysis will rely on the condition that $\bar{T}$ has sufficiently many instances. Specifically, in order to collect n_k instances to form the working set $\bar{T}$, we need to call $\mathrm{EX}_\eta^x(D, w^*)$ enough number of times since our sampling is localized within the band $X_k := \{x : |w_{k-1} \cdot x| \le b_k\}$. The following lemma characterizes the sample complexity at phase k.

Lemma 6 *Suppose that Assumption 1 and 2 are satisfied. Further assume $\eta < \frac{1}{2}$. With probability $1 - \frac{\delta_k}{4}$, we will obtain n_k instances that fall into the band $X_k = \{x : |w_{k-1} \cdot x| \le b_k\}$ by making a number of $N_k = O\left(\frac{1}{b_k}\left(n_k + \log\frac{1}{\delta_k}\right)\right)$ calls to the instance generation oracle $\mathrm{EX}_\eta^x(D, w^*)$.*

4.2. Attribute and computationally efficient soft outlier removal

We summarize the performance guarantee of Algorithm 2 in the following proposition.

Proposition 7 *Consider phase k of Algorithm 1 for any $1 \le k \le k_0$. Suppose that Assumption 1 and 2 are satisfied, and that $\eta \le c_5\epsilon$. With the setting of n_k, with probability $1 - \frac{\delta_k}{8}$ over the draw of $\bar{T}$, Algorithm 2 will output a function $q : T \to [0, 1]$ in polynomial time with the following properties: (1) $\frac{1}{|T|}\sum_{x\in T} q(x) \ge 1 - \xi_k$; (2) for all $w \in W_k$, $\frac{1}{|T|}\sum_{x\in T} q(x)(w \cdot x)^2 \le 5C_2\left(b_k^2 + r_k^2\right)$.*

Again, we remind that the key difference between our algorithm and that of Awasthi et al. (2017) is in Constraint 3 of Algorithm 2: we require that the "variance proxy" $\sum_{x\in T} q(x)x^\top Hx$ of the reweighted instances are small for all positive semidefinite H that lies in an intersection of a trace-norm ball and an ℓ_1-norm ball. On the statistical side, this favorable constraint set of H, in conjunction with Adamczak's bound in empirical processes literature (Adamczak, 2008), results in sufficient uniform concentration of the variance proxy $x^\top Hx$ with a sample complexity of $\mathrm{poly}\,(s, \log d)$. This significantly improves the sample complexity of $\mathrm{poly}\,(d)$ established in Awasthi et al. (2017). The detailed proof can be found in Appendix D.3.

Remark 8 *While in some standard settings, a proper ℓ_1-norm constraint suffices to guarantee a desired bound of sample complexity in the high-dimensional regime (Wainwright, 2009; Kakade et al., 2008), we note that in order to establish near-optimal noise tolerance, the ℓ_2-norm constraint of w (hence the trace-norm of H) is vital as well. Though eliminating it eases the search of a feasible function q, this leads to a suboptimal noise tolerance $\eta \le \Omega(\epsilon/s)$. Informally speaking, the per-phase error rate, expected to be a constant, is inherently proportional to the variance $(w \cdot x)^2$ times ξ_k, the noise rate within the band. Now without the trace-norm constraint, the variance would be s times larger than before (since we now have to use $\rho_k^2 = O(sr_k^2)$ as a proxy for the constraint set's radius, measured in trace norm). This implies that we need to set ξ_k a factor $1/s$ of before, which in turn indicates that the noise tolerance η becomes a factor $1/s$ of before since $\eta/\epsilon \approx \xi_k$. We refer the reader to Proposition 33 and Lemma 39 for details.*

Remark 9 *The quantity n_k has a quadratic dependence on the sparsity parameter s. This cannot be improved in some sparse PCA related problems (Berthet and Rigollet, 2013), but it is not clear whether such dependence is optimal in our case. We leave this investigation to future work.*

Next, we describe the statistical property of the distribution p (obtained by normalizing q returned by Algorithm 2). Observe that the noise rate within the band is at most $\eta/b_k \leq O(\eta/\epsilon) \leq \xi_k$ since the probability mass of the band is $\Theta(b_k)$ – an important property of isotropic log-concave distributions. Also, it is possible to show that the variance of clean instances on directions $H \in \mathcal{M}_k$ is $O(b_k^2 + r_k^2)$ (see Lemma 18). Therefore, Algorithm 2 is essentially searching for a weighting such that clean instances have overwhelming weights over dirty instances, and that the variance of the weighted instances is similar to that of the clean instances. Recall that $T_{\mathrm{C}} \subset T$ is the set of clean instances in T. Let $\tilde{T}_{\mathrm{C}} = \{(x, y_x)\}_{x \in T_{\mathrm{C}}}$ be the unrevealed labeled set where each instance is correctly annotated by w^*. The following proposition, which is similar to Lemma 4.7 of Awasthi et al. (2017) but with refinement, states that the reweighted hinge loss $\ell_{\tau_k}(w; p) := \sum_{x \in T} p(x) \ell_{\tau_k}(w; x, y_x)$, is a good proxy for the hinge loss evaluated exclusively on clean labeled instances $\tilde{T}_{\mathrm{C}}$.

Proposition 10 *Suppose Assumption 1 and 2 are satisfied, and $\eta \leq c_5\epsilon$. For any phase k of Algorithm 1, with probability $1 - \frac{\delta_k}{4}$ over the draw of $\bar{T}$, we have $\sup_{w \in W_k} \left| \ell_{\tau_k}(w; \tilde{T}_{\mathrm{C}}) - \ell_{\tau_k}(w; p) \right| \leq \kappa$.*

Note that though this proposition is phrased in terms of the hinge loss on pairs (x, y_x), it is only used in the analysis and our algorithm does not require the knowledge of the labels y_x – the algorithm even does not need to exactly identify the set of clean instances T_{C}. As a result, the size of T_{C} does not count towards our label complexity. Proposition 7 together with Proposition 10 implies that with high probability, Algorithm 2 produces a desired probability distribution in polynomial time, which justifies its computational and statistical efficiency.

In addition, let $L_{\tau_k}(w) := \mathbb{E}_{x \sim D_{w_{k-1}, b_k}} \left[\ell_{\tau_k} \left(w; x, \operatorname{sign}(w^* \cdot x) \right) \right]$ be the expected loss on D_{w_{k-1}, b_k}. The following result links $L_{\tau_k}(w)$ to the empirical hinge loss on clean instances.

Proposition 11 *Under Assumption 1 and 2, and $\eta \leq c_5\epsilon$, for any phase k of Algorithm 1, with probability $1 - \frac{\delta_k}{4}$ over the draw of $\bar{T}$, we have $\sup_{w \in W_k} \left| L_{\tau_k}(w) - \ell_{\tau_k}(w; \tilde{T}_{\mathrm{C}}) \right| \leq \kappa$.*

4.3. Attribute and label-efficient empirical risk minimization

In light of Proposition 10, one may want to find an iterate by minimizing its reweighted hinge loss $\ell_{\tau_k}(w; p)$. This requires collecting labels for all instances in T, which leads to a suboptimal label complexity $O\left(s^2 \cdot \operatorname{polylog}\left(d, 1/\epsilon\right)\right)$. As a remedy, we perform a random sampling process, which draws m_k instances from T according to the distribution p and then query their labels, resulting in the labeled instance set S_k. By standard uniform convergence arguments, it is expected that $\ell_{\tau_k}(w; S_k) \approx \ell_{\tau_k}(w; p)$ provided that m_k is large enough, as is shown in the following proposition.

Proposition 12 *Suppose that Assumption 1 and 2 are satisfied. For any phase k of Algorithm 1, with probability $1 - \frac{\delta_k}{4}$, we have $\sup_{w \in W_k} \left| \ell_{\tau_k}(w; p) - \ell_{\tau_k}(w; S_k) \right| \leq \kappa$.*

We remark that when establishing the performance guarantee, the ℓ_1-norm constraint on the hypothesis space, together with an ℓ_∞-norm upper bound on the localized instance space, leads to a Rademacher complexity that has a linear dependence on the sparsity (up to a logarithmic factor). Technically speaking, our analysis is more involved than that of Awasthi et al. (2017): applying their analysis to the setting of learning sparse halfspaces along with the fact that the VC dimension of the class of s-sparse halfspaces is $O(s \log(d/s))$ would give a label complexity quadratic in s.

4.4. Uniform concentration for unbounded data

Our analysis involves building uniform concentration bounds. The primary issue of applying standard concentration results, e.g. Theorem 1 of Kakade et al. (2008), is that the instances are not contained in a pre-specified ℓ_∞-ball with probability 1 under isotropic log-concave distribution. Awasthi et al. (2017); Zhang (2018) construct a conditional distribution, on which the data are all bounded from above, and then measure the difference between this conditional distribution and the original one. We circumvent such technical complication by using the Adamczak's bound (Adamczak, 2008) in the empirical process literature, which provides a generic way to analyze concentration inequalities for well-behaved distributions with unbounded support. See Appendix C for a concrete treatment.

4.5. Proof sketch of Theorem 4

Proof We first show that error rate of v_k on D_{w_{k-1},b_k} is a constant, and that of w_k follows since hard thresholding and ℓ_2-norm projection can only deviate the error rate by a constant factor. Observe that in light of Proposition 10, Proposition 11, and Proposition 12, we have $\left|\ell_{\tau_k}(w;S_k) - L_{\tau_k}(w)\right| \leq 3\kappa$ for all $w \in W_k$. Therefore, if $w^* \in W_k$, by the optimality of v_k, we have $L_{\tau_k}(v_k) \leq \ell_{\tau_k}(v_k;S_k) + 3\kappa \leq \ell_{\tau_k}(w^*;S_k) + 4\kappa \leq L_{\tau_k}(w^*) + 7\kappa \leq 8\kappa$, where the last inequality is by Lemma 3.7 of Awasthi et al. (2017). Since $L_{\tau_k}(v_k)$ always serves as an upper bound of $\mathrm{err}_{D_{w_{k-1},b_k}}(v_k)$, the constant error rate on D_{w_{k-1},b_k} follows. Next we can use the analysis framework of margin-based active learning to show that such constant error rate ensures that the angle between w_k and w^* is as small as $O(2^{-k})$, which in turn implies $w^* \in W_{k+1}$. It remains to show $w^* \in W_1$; this can be easily seen by the definition of W_1: $W_1 = B_2(0,1) \cap B_1(0,\sqrt{s})$. Hence, we conclude $w^* \in W_k$ for all $1 \leq k \leq k_0$. Observe that the radius of ℓ_2-ball of W_{k_0} is as small as ϵ, which, by a basic property of isotropic log-concave distributions, implies the error rate of w_{k_0} on D is less than ϵ.

The sample and label complexity bounds follow from our setting of N_k and m_k, and the fact that $b_k \in [\epsilon, \bar{c}/16]$ for all $k \leq k_0$. See Appendix D.5 for the full proof. ■

5. Conclusion and Open Questions

We have presented a computationally efficient algorithm for learning sparse halfspaces under the challenging malicious noise model. Our algorithm leverages the well-established margin-based active learning framework, with a particular treatment on attribute efficiency, label complexity, and noise tolerance. We have shown that our theoretical guarantees for label complexity and noise tolerance are near-optimal, and the sample complexity of a passive learning variant of our algorithm is attribute-efficient, thanks to the set of new techniques proposed in this paper.

We raise three open questions for further investigation. First, as we discussed in Section 4.2, the sample complexity for concentration of $x^\top Hx$ has a quadratic dependence on s. It would be interesting to study whether this is a fundamental limit of learning under isotropic log-concave distributions, or it can be improved by a more sophisticated localization scheme in the instance and the concept spaces. Second, while isotropic log-concave distributions bear favorable properties that fit perfectly in the margin-based framework, it would be interesting to examine whether the established results can be extended to heavy-tailed distributions. This may lead to a large error rate within the band that cannot be controlled at a constant level, and new techniques must be developed. Finally, it would be interesting to design computationally more efficient algorithms, e.g. stochastic gradient descent-type algorithms similar to Dasgupta et al. (2005), with comparable statistical guarantees.

Acknowledgments

The authors thank the anonymous reviewers for helpful suggestions. Jie Shen is supported by NSF-IIS-1948133 and the startup funding of Stevens Institute of Technology. Chicheng Zhang acknowledges the startup funding support from the University of Arizona.

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13(34):1000–1034, 2008.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In *Proceedings of the 28th Annual Conference on Learning Theory*, pages 167–190, 2015.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of the 29th Annual Conference on Learning Theory*, pages 152–192, 2016.

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

Sivaraman Balakrishnan, Simon S. Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 30th Annual Conference on Learning Theory*, pages 169–212, 2017.

Maria Florina Balcan and Steve Hanneke. Robust interactive learning. In *Conference on Learning Theory*, pages 20–1, 2012.

Maria-Florina Balcan and Philip M. Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of The 26th Annual Conference on Learning Theory*, pages 288–316, 2013.

Maria-Florina Balcan, Andrei Z. Broder, and Tong Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Learning Theory*, pages 35–50, 2007.

Richard G. Baraniuk, Simon Foucart, Deanna Needell, Yaniv Plan, and Mary Wootters. Exponential decay of reconstruction error from binary measurements of sparse signals. *IEEE Transactions on Information Theory*, 63(6):3368–3385, 2017.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 1046–1066, 2013.

Alina Beygelzimer, Daniel J. Hsu, John Langford, and Chicheng Zhang. Search improves label for active learning. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, pages 3342–3350, 2016.

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

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

Avrim Blum, Alan M. Frieze, Ravi Kannan, and Santosh S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. In *Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science*, pages 330–338, 1996.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

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

Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Nader H. Bshouty. A new composition theorem for learning algorithms. In *Proceedings of the 30th Annual ACM Symposium on the Theory of Computing*, pages 583–589, 1998.

Nader H. Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

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

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

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

Nicolò Cesa-Bianchi, Eli Dichterman, Paul Fischer, Eli Shamir, and Hans Ulrich Simon. Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM*, 46(5):684–719, 1999.

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

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

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

Sanjoy Dasgupta, Adam Tauman Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 249–263, 2005.

Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

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

Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Zheng Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. *CoRR*, abs/1604.06443, 2016.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM Symposium on Theory of Computing*, pages 1061–1073, 2018.

Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent PAC learning of halfspaces with Massart noise. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*, pages 4751–4762, 2019.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning halfspaces with Massart noise under structured distributions. In *Proceedings of the 33rd Annual Conference on Learning Theory*, volume 125, pages 1486–1513, 2020.

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

Richard M. Dudley. *Uniform central limit theorems*, volume 142. Cambridge University Press, 2014.

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

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

Vitaly Feldman. Open problem: The statistical query complexity of learning sparse halfspaces. In *Proceedings of The 27th Annual Conference on Learning Theory*, volume 35, pages 1283–1289, 2014.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

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

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

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, pages 793–800, 2008.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 11–20, 2005.

Daniel M. Kane, Shachar Lovett, Shay Moran, and Jiapeng Zhang. Active classification with comparison queries. In Chris Umans, editor, *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, pages 355–366, 2017.

Michael J. Kearns and Ming Li. Learning in the presence of malicious errors. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, pages 267–280, 1988.

Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. In David Haussler, editor, *Proceedings of the 5th Annual Conference on Computational Learning Theory*, pages 341–352, 1992.

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

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

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

Kevin A. Lai, Anup B. Rao, and Santosh S. Vempala. Agnostic estimation of mean and covariance. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, pages 665–674, 2016.

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

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

Philip M. Long and Rocco A. Servedio. 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*, pages 921–928, 2006.

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

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

Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013a.

Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013b.

Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, 2016.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10): 6976–6994, 2011.

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

Robert E. Schapire. *Design and analysis of efficient learning algorithms*. MIT Press, Cambridge, MA, USA, 1992.

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

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

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

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

Robert H. Sloan. Types of noise in data for concept learning. In *Proceedings of the First Annual Workshop on Computational Learning Theory*, pages 91–96, 1988.

Robert H. Sloan. Corrigendum to types of noise in data for concept learning. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory*, page 450, 1992.

Daureen Steinberg. Computation of matrix norms with applications to robust optimization. *Research thesis, Technion-Israel University of Technology*, 2005.

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

Andreas M. Tillmann and Marc E. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2014.

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

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

Leslie G. Valiant. Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 560–566, 1985.

Sara van de Geer and Johannes Lederer. The Bernstein-Orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157:225–250, 2013.

Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer, 1996.

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

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5): 2183–2202, 2009.

Yichong Xu, Hongyang Zhang, Aarti Singh, Artur Dubrawski, and Kyle Miller. Noise-tolerant interactive learning using pairwise comparisons. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pages 2431–2440, 2017.

Songbai Yan and Chicheng Zhang. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pages 1056–1066, 2017.

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

Chicheng Zhang, Jie Shen, and Pranjal Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. *CoRR*, abs/2002.04840, 2020.

Lijun Zhang, Jinfeng Yi, and Rong Jin. Efficient algorithms for robust one-bit compressive sensing. In *Proceedings of the 31st International Conference on Machine Learning*, pages 820–828, 2014.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

Appendix A. Detailed Choices of Reserved Constants and Additional Notations

Constants. The absolute constants c_0, c_1 and c_2 are specified in Lemma 14, and c_3 and c_4 are specified in Lemma 15. c_5 and c_6 were clarified in Section 3.2. The definition of c_7, c_8, c_9 can be found in Lemma 16, Lemma 19, and Lemma 20 respectively. The absolute constant C_1 acts as an upper bound of all b_k's, and by our choice in Section 3.2, $C_1 = \bar{c}/16$. The absolute constant C_2 is defined in Lemma 18. Other absolute constants, such as C_3, C_4 are not quite crucial to our analysis or algorithmic design. Therefore, we do not track their definitions. The subscript variants of K, e.g. K_1 and K_2, are also absolute constants but their values may change from appearance to appearance. We remark that the value of all these constants does not depend on the underlying distribution D chosen by the adversary, but rather depends on the knowledge of $\mathcal{D}$.

Pruning. Consider Algorithm 1. For each phase k, we sample a working set $\bar{T}$ and remove all instances that have large ℓ_∞-norm to obtain T (Step 6), which is equivalent to intersecting it with the ℓ_∞-ball $B_\infty(0, \nu_k) := \{x : \|x\|_\infty \leq \nu_k\}$ where $\nu_k = c_9 \log \frac{48|\bar{T}|d}{b_k \delta_k}$. This is motivated by Lemma 20, which states that with high probability, all clean instances in $\bar{T}$ are in $B_\infty(0, \nu_k)$. Specifically, Denote by $\bar{T}_{\mathrm{C}}$ (respectively $\bar{T}_{\mathrm{D}}$) the set of clean (respectively dirty) instances in $\bar{T}$. Lemma 20 implies that with probability $1 - \frac{\delta_k}{48}$, $\bar{T}_{\mathrm{C}} \subset B_\infty(0, \nu_k)$. Therefore, with high probability, all the instances in $\bar{T}_{\mathrm{C}}$ are kept in this step and only instances in $\bar{T}_{\mathrm{D}}$ may be removed. Denote by $T_{\mathrm{C}} = \bar{T}_{\mathrm{C}} \cap B_\infty(0, \nu_k)$ and $T_{\mathrm{D}} = \bar{T}_{\mathrm{D}} \cap B_\infty(0, \nu_k)$; we therefore also have the decomposition $T = T_{\mathrm{C}} \cup T_{\mathrm{D}}$. We finally denote by $\hat{T}_{\mathrm{C}}$ the unrevealed labeled set that corresponds to $\bar{T}_{\mathrm{C}}$.

Table 1: Summary of useful notations associated with the working set $\bar{T}$ at each phase k.

$\bar{T}$	instance set obtained by calling $\mathrm{EX}^x_\eta(D, w^*)$ conditioned on $\lvert w_{k-1} \cdot x\rvert \leq b_k$
$\bar{T}_{\mathrm{C}}$	set of instances in $\bar{T}$ that $\mathrm{EX}^x_\eta(D, w^*)$ draws from the distribution D
$\bar{T}_{\mathrm{D}}$	set of dirty instances in $\bar{T}$, i.e. $\bar{T} \backslash \bar{T}_{\mathrm{C}}$
T	set of instances in $\bar{T}$ that lie in $B_\infty(0, \nu_k)$
T_{C}	set of instances in $\bar{T}_{\mathrm{C}}$ that lie in $B_\infty(0, \nu_k)$
T_{D}	set of instances in $\bar{T}_{\mathrm{D}}$ that lie in $B_\infty(0, \nu_k)$
$\hat{T}_{\mathrm{C}}$	unrevealed labeled set of $\bar{T}_{\mathrm{C}}$
$\tilde{T}_{\mathrm{C}}$	unrevealed labeled set of T_{C}

Regularity condition on $D_{u,b}$. We will frequently work with the conditional distribution $D_{u,b}$ obtained by conditioning D on the event that x is in the band $\{x \in \mathbb{R}^d : |u \cdot x| \leq b\}$. We give the following regularity condition to ease our terminology.

Definition 13 *A conditional distribution $D_{u,b}$ is said to satisfy the regularity condition if one of the following holds: 1) the vector $u \in \mathbb{R}^d$ has unit ℓ_2-norm and $0 < b \leq C_1$; 2) the vector u is the zero vector and $b = C_1$.*

In particular, at each phase k of Algorithm 1, u is set to w_{k-1} and b is set to b_k. For $k = 1$, $u = w_0$ is a zero vector, $b = b_1 = C_1$, satisfying the regularity condition. It is worth mentioning that at phase 1 the conditional distribution $D_{u,b}$ boils down to D. For all $k \geq 2$, u is a unit vector and $b \in (0, C_1]$ in view of our construction of b_k. Therefore, for all $k \geq 1$, D_{w_{k-1}, b_k} satisfy the regularity condition.

Appendix B. Useful Properties of Isotropic Log-Concave Distributions

We record some useful properties of isotropic log-concave distributions.

Lemma 14 *There are absolute constants $c_0, c_1, c_2 > 0$, such that the following holds for all isotropic log-concave distributions $D \in \mathcal{D}$. Let f_D be the density function. We have*

1. *Orthogonal projections of D onto subspaces of $\mathbb{R}^d$ are isotropic log-concave;*
2. *If $d = 1$, then $\Pr_{x\sim D}(a \le x \le b) \le |b - a|$;*
3. *If $d = 1$, then $f_D(x) \ge c_0$ for all $x \in [-1/9, 1/9]$;*
4. *For any two vectors $u, v \in \mathbb{R}^d$,*

$$c_1 \cdot \Pr_{x\sim D}\left(\operatorname{sign}(u \cdot x) \neq \operatorname{sign}(v \cdot x)\right) \le \theta(u, v) \le c_2 \cdot \Pr_{x\sim D}\left(\operatorname{sign}(u \cdot x) \neq \operatorname{sign}(v \cdot x)\right);$$

5. $\Pr_{x\sim D}\left(\|x\|_2 \ge t\sqrt{d}\right) \le \exp(-t + 1)$.

We remark that Parts 1, 2, 3, and 5 are due to Lovász and Vempala (2007), and Part 4 is from Vempala (2010); Balcan and Long (2013).

The following lemma is implied by the proof of Theorem 21 of Balcan and Long (2013), which shows that if we choose a proper band width $b > 0$, the error outside the band will be small. This observation is crucial for controlling the error over the distribution D, and has been broadly recognized in the literature (Awasthi et al., 2017; Zhang, 2018).

Lemma 15 (Theorem 21 of Balcan and Long (2013)) *There are absolute constants $c_3, c_4 > 0$ such that the following holds for all isotropic log-concave distributions $D \in \mathcal{D}$. Let u and v be two unit vectors in $\mathbb{R}^d$ and assume that $\theta(u, v) = \alpha < \pi/2$. Then for any $b \ge \frac{4}{c_4}\alpha$, we have*

$$\Pr_{x\sim D}(\operatorname{sign}(u \cdot x) \neq \operatorname{sign}(v \cdot x) \text{ and } |v \cdot x| \ge b) \le c_3\alpha \exp\left(-\frac{c_4 b}{2\alpha}\right).$$

Lemma 16 (Lemma 20 of Awasthi et al. (2016)) *There is an absolute constant $c_7 > 0$ such that the following holds for all isotropic log-concave distributions $D \in \mathcal{D}$. Draw n i.i.d. instances from D to form a set S. Then*

$$\Pr_{S\sim D^n}\left(\max_{x\in S}\|x\|_\infty \ge c_7 \log\frac{|S|\, d}{\delta}\right) \le \delta.$$

Lemma 17 *There is an absolute constant $\bar{C}_2 \ge 1$ such that the following holds for all isotropic log-concave distributions $D \in \mathcal{D}$ and all $D_{u,b}$ that satisfy the regularity condition:*

$$\sup_{w\in B_2(u,r)} \mathbb{E}_{x\sim D_{u,b}}\left[(w \cdot x)^2\right] \le \bar{C}_2(b^2 + r^2).$$

Proof When u is a unit vector, Lemma 3.4 of Awasthi et al. (2017) shows that there exists a constant K_1 such that

$$\sup_{w\in B_2(u,r)} \mathbb{E}_{x\sim D_{u,b}}\left[(w \cdot x)^2\right] \le K_1(b^2 + r^2).$$

When u is a zero vector, $D_{u,b}$ reduces to D and the constraint $w \in B_2(u,r)$ reads as $\|w\|_2 \le r$. Thus we have

$$\mathbb{E}_{x\sim D_{u,b}}\left[(w\cdot x)^2\right] = \|w\|_2^2 \le r^2 < b^2 + r^2.$$

The proof is complete by choosing $\bar{C}_2 = K_1 + 1$. ■

Lemma 18 *There is an absolute constant $C_2 \ge 2$ such that the following holds for all isotropic log-concave distributions $D \in \mathcal{D}$ and all $D_{u,b}$ that satisfy the regularity condition:*

$$\sup_{H\in\mathcal{M}} \mathbb{E}_{x\sim D_{u,b}}\left[x^\top H x\right] \le C_2(b^2+r^2),$$

where $\mathcal{M} := \{H \in \mathbb{R}^{d\times d} :\ H \succeq 0,\ \|H\|_* \le r^2,\ \|H\|_1 \le \rho^2\}$.

Proof Since $H \in \mathcal{M}$ is a positive semidefinite matrix with trace norm at most r^2, it has eigendecomposition $H = \sum_{i=1}^d \lambda_i v_i v_i^\top$, where $\lambda_i \ge 0$ are the eigenvalues such that $\sum_{i=1}^d \lambda_i \le r^2$, and v_i's are orthonormal vectors in $\mathbb{R}^d$. Thus,

$$x^\top H x = \frac{1}{r^2}\sum_{i=1}^d \lambda_i (rv_i\cdot x)^2 \le \frac{2}{r^2}\cdot\sum_{i=1}^d \lambda_i\left[\left((rv_i+u)\cdot x\right)^2 + (u\cdot x)^2\right].$$

Since x is drawn from $D_{u,b}$, we have $(u\cdot x)^2 \le b^2$. Moreover, applying Lemma 17 with the setting of $w = rv + u$ implies that

$$\sup_{v\in B_2(0,1)} \mathbb{E}_{x\sim D_{u,b}}\left[\left((rv+u)\cdot x\right)^2\right] \le \bar{C}_2(b^2+r^2).$$

Therefore,

$$\sup_{H\in\mathcal{M}} \mathbb{E}_{x\sim D_{u,b}}\left[x^\top H x\right] \le \frac{2}{r^2}\cdot\sum_{i=1}^d \lambda_i\left(\bar{C}_2(b^2+r^2)+b^2\right) \le 2(\bar{C}_2+1)(b^2+r^2).$$

The proof is complete by choosing $C_2 = 2(\bar{C}_2+1)$. ■

Lemma 19 *Let $c_8 = \min\left\{2c_0, \frac{2c_0}{9C_1}, \frac{1}{C_1}\right\}$. Then for all isotropic log-concave distributions $D \in \mathcal{D}$ and all $D_{u,b}$ satisfying the regularity condition,*

1. $\Pr_{x\sim D}\left(|u\cdot x| \le b\right) \ge c_8 \cdot b$;
2. $\Pr_{x\sim D_{u,b}}(E) \le \frac{1}{c_8 b}\Pr_{x\sim D}(E)$ *for any event E.*

Proof We first consider the case that u is a unit vector.

For the lower bound, Part 3 of Lemma 14 shows that the density function of the random variable $u\cdot x$ is lower bounded by c_0 when $|u\cdot x| \le 1/9$. Thus

$$\Pr_{x\sim D}\left(|u\cdot x| \le b\right) \ge \Pr_{x\sim D}\left(|u\cdot x| \le \min\{b, 1/9\}\right) \ge 2c_0\min\{b,1/9\} \ge 2c_0\min\left\{1, \frac{1}{9C_1}\right\}\cdot b$$

where in the last inequality we use the condition $b \leq C_1$.

For any event E, we always have

$$\Pr_{x\sim D_{u,b}}(E) \leq \frac{\Pr_{x\sim D}(E)}{\Pr_{x\sim D}(|u\cdot x| \leq b)} \leq \frac{1}{c_8 b}\Pr_{x\sim D}(E).$$

Now we consider the case that u is the zero vector and $b = C_1$. Then $\Pr_{x\sim D}\left(|u\cdot x| \leq b\right) = 1 \geq c_8 \cdot b$ in view of the choice c_8. Thus Part 2 still follows. The proof is complete. ■

Lemma 20 *There exists an absolute constant $c_9 > 0$ such that the following holds for all isotropic log-concave distributions $D \in \mathcal{D}$ and all $D_{u,b}$ that satisfy the regularity condition. Let S be a set of i.i.d. instances drawn from $D_{u,b}$. Then*

$$\Pr_{S\sim D^n_{u,b}}\left(\max_{x\in S}\|x\|_\infty \geq c_9 \log\frac{|S|\,d}{b\delta}\right) \leq \delta.$$

Proof Using Lemma 16 we have

$$\Pr_{S\sim D^n}\left(\max_{x\in S}\|x\|_\infty \geq c_7 \log\frac{|S|\,d}{\delta}\right) \leq \delta.$$

Thus, using Part 2 of Lemma 19 gives

$$\Pr_{S\sim D^n_{u,b}}\left(\max_{x\in S}\|x\|_\infty \geq c_7 \log\frac{|S|\,d}{\delta}\right) \leq \frac{\delta}{c_8 b}.$$

The proof is complete by changing δ to $\delta' = \frac{\delta}{c_8 b}$. ■

Appendix C. Orlicz Norm and Concentration Results using Adamczak's Bound

The following notion of Orlicz norm (van de Geer and Lederer, 2013; Dudley, 2014) is useful in handling random variables that have tails of the form $\exp(-t^\alpha)$ for general α's beyond $\alpha = 2$ (subgaussian) and $\alpha = 1$ (subexponential).

Definition 21 (Orlicz norm) *For any $z \in \mathbb{R}$, let $\psi_\alpha : z \mapsto \exp(z^\alpha) - 1$. Furthermore, for a random variable $Z \in \mathbb{R}$ and $\alpha > 0$, define $\|Z\|_{\psi_\alpha}$, the Orlicz norm of Z with respect to ψ_α, as:*

$$\|Z\|_{\psi_\alpha} = \inf\left\{t > 0 : \mathbb{E}_Z\left[\psi_\alpha\left(|Z|/t\right)\right] \leq 1\right\}.$$

We collect some basic facts about Orlicz norms in the following lemma; they can be found in Section 1.3 of Van Der Vaart and Wellner (1996).

Lemma 22 *Let Z, Z_1, Z_2 be real-valued random variables. Consider the Orlicz norm with respect to ψ_α. We have the following:*

1. *$\|\cdot\|_{\psi_\alpha}$ is a norm. For any $a \in \mathbb{R}$, $\|aZ\|_{\psi_\alpha} = |a|\cdot\|Z\|_{\psi_\alpha}$; $\|Z_1 + Z_2\|_{\psi_\alpha} \leq \|Z_1\|_{\psi_\alpha} + \|Z_2\|_{\psi_\alpha}$.*

2. $\|Z\|_p \le \|Z\|_{\psi_p} \le p! \|Z\|_{\psi_1}$ *where* $\|Z\|_p := \left(\mathbb{E}\left[|Z|^p\right]\right)^{1/p}$.

3. *For any* $p, \alpha > 0$, $\|Z\|_{\psi_p}^{\alpha} = \|Z^{\alpha}\|_{\psi_{p/\alpha}}$.

4. *If* $\Pr\left(|Z| \ge t\right) \le K_1 \exp\left(-K_2 t^{\alpha}\right)$ *for any* $t \ge 0$, *then* $\|Z\|_{\psi_\alpha} \le \left(\frac{2(\ln K_1 + 1)}{K_2}\right)^{1/\alpha}$.

5. *If* $\|Z\|_{\psi_\alpha} \le K$, *then for all* $t \ge 0$, $\Pr\left(|Z| \ge t\right) \le 2\exp\left(-(\frac{t}{K})^{\alpha}\right)$.

The following auxiliary results, tailored to the localized sampling scheme in Algorithm 1, will also be useful in our analysis.

Lemma 23 *There exists an absolute constant* $C_3 > 0$ *such that the following holds for all isotropic log-concave distributions* $D \in \mathcal{D}$ *and all* $D_{u,b}$ *that satisfy the regularity condition. Let* $S = \{x_1, \ldots, x_n\}$ *be a set of* n *instances drawn from* $D_{u,b}$. *Then*

$$\left\| \max_{x \in S} \|x\|_\infty \right\|_{\psi_1} \le C_3 \log \frac{nd}{b}.$$

Consequently,

$$\mathbb{E}_{S \sim D_{u,b}^n}\left[\max_{x \in S} \|x\|_\infty\right] \le C_3 \log \frac{nd}{b}.$$

Proof Let Z be isotropic log-concave random variable in $\mathbb{R}$. Part 5 of Lemma 14 shows that for all $t > 0$,

$$\Pr(|Z| > t) \le \exp(-t + 1).$$

Fix $i \in \{1, \ldots, n\}$ and fix $j \in \{1, \ldots, d\}$. Denote by $x_i^{(j)}$ the j-th coordinate of x_i. Part 1 of Lemma 14 suggests that $x_i^{(j)}$ is isotropic log-concave. Thus, by Part 2 of Lemma 19,

$$\Pr_{x \sim D_{u,b}}\left(\left|x_i^{(j)}\right| > t\right) \le \frac{1}{c_8 b} \Pr_{x \sim D}\left(\left|x_i^{(j)}\right| > t\right) \le \frac{1}{c_8 b} \exp(-t + 1).$$

Taking the union bound over $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, d\}$, we have for all $t > 0$

$$\Pr_{x \sim D_{u,b}}\left(\max_{x \in S} \|x\|_\infty > t\right) \le \frac{nd}{c_8 b} \exp(-t + 1).$$

Now Part 4 of Lemma 22 immediately implies that

$$\left\| \max_{x \in S} \|x\|_\infty \right\|_{\psi_1} \le C_3 \log \frac{nd}{b}$$

for some constant $C_3 > 0$. The second inequality of the lemma is an immediate result by combining the above and Part 2 of Lemma 22. ■

C.1. Adamczak's bound

In this section, we establish the key concentration results that will be used to analyze the performance of soft outlier removal and random sampling in Algorithm 1. Since we are considering the isotropic log-concave distribution, any unlabeled instance x is unbounded. This prevents us from using standard concentration bounds, e.g. Kakade et al. (2008). We henceforth appeal to the following generalization of Talagrand's inequality, due to Adamczak (2008).

Lemma 24 (Adamczak's bound) *For any $\alpha \in (0,1]$, there exists a constant $\Lambda_\alpha > 0$, such that the following holds. Given any function class $\mathcal{F}$, and a function F such that for any $f \in \mathcal{F}$, $\left|f(x)\right| \leq F(x)$, we have with probability at least $1-\delta$ over the draw of a set $S = \{x_1, \ldots, x_n\}$ of i.i.d. instances from D,*

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbb{E}_{x\sim D}\left[f(x)\right]\right| \leq \Lambda_\alpha \left(\mathbb{E}_{S\sim D^n}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbb{E}_{x\sim D}\left[f(x)\right]\right|\right] + \sqrt{\frac{\sup_{f\in\mathcal{F}}\mathbb{E}_{x\sim D}\left[(f(x))^2\right]\ln\frac{1}{\delta}}{n}} + \frac{(\ln\frac{1}{\delta})^{1/\alpha}}{n}\left\|\max_{1\leq i\leq n} F(x_i)\right\|_{\psi_\alpha}\right).$$

We first establish the following result that upper bounds the expected value of Rademacher complexity of linear classes by the Orlicz norm of the random instances.

Lemma 25 *There exists an absolute constant $C_5 > 0$ such that the following holds for all isotropic log-concave distributions $D \in \mathcal{D}$ and all $D_{u,b}$ that satisfy the regularity condition. Let $S = \{x_1, \ldots, x_n\}$ be a set of n i.i.d. unlabeled instances drawn from $D_{u,b}$. Denote $W = B_2(u,r) \cap B_1(u,\rho)$. Let a sequence of random variables $Z = \{z_1, \ldots, z_n\}$ be drawn from a distribution supported on a bounded interval $[-\lambda, \lambda]$ for some $\lambda > 0$. Let $\sigma = \{\sigma_1, \ldots, \sigma_n\}$, where the σ_i's are i.i.d. Rademacher random variables independent of S and Z. We have:*

$$\mathbb{E}_{S,Z,\sigma}\left[\sup_{w\in W}\left|\sum_{i=1}^{n}\sigma_i z_i (w\cdot x_i)\right|\right] \leq \lambda b\sqrt{n} + C_5\rho\lambda\sqrt{n\log d}\cdot\log\frac{nd}{b}.$$

Proof Let $V = B_2(0,r)\cap B_1(0,\rho)$ so that any $w \in W$ can be expressed as $w = u + v$ for some $v \in V$. First, conditioned on S and Z, we have that

$$\mathbb{E}_{\sigma}\left[\sup_{v\in V}\left|\sum_{i=1}^{n}\sigma_i z_i (v\cdot x_i)\right|\right] \leq \rho\sqrt{2n\log(2d)}\cdot\max_{1\leq i\leq n}\|z_i x_i\|_\infty \leq \rho\lambda\sqrt{2n\log(2d)}\cdot\max_{1\leq i\leq n}\|x_i\|_\infty.$$

Thus,

$$\begin{aligned}\mathbb{E}_{S,Z,\sigma}\left[\sup_{v\in V}\left|\sum_{i=1}^{n}\sigma_i z_i (v\cdot x_i)\right|\right] &\leq \rho\lambda\sqrt{2n\log(2d)}\cdot\mathbb{E}_S\left[\max_{1\leq i\leq n}\|x_i\|_\infty\right] \\ &\leq C_5\rho\lambda\sqrt{n\log d}\cdot\log\frac{nd}{b}, \qquad \text{(C.1)}\end{aligned}$$

where the second inequality follows from Lemma 23.

On the other side, using the fact that for any random variable A, $\mathbb{E}[A] \le \left(\mathbb{E}[A^2]\right)^{1/2}$, we have

$$\mathbb{E}_{S,Z,\sigma}\left[\left|\sum_{i=1}^n \sigma_i z_i (u\cdot x_i)\right|\right] \le \sqrt{\mathbb{E}_{S,Z,\sigma}\left[\left(\sum_{i=1}^n \sigma_i z_i (u\cdot x_i)\right)^2\right]}$$
$$= \sqrt{\mathbb{E}_{S,Z}\left[\sum_{i=1}^n z_i^2 (u\cdot x_i)^2\right]} \le \sqrt{nb^2\lambda^2},$$

where in the equality we use the observation that $\mathbb{E}_{S,Z,\sigma}\left[\sigma_i\sigma_j z_i z_j (u\cdot x_i)(u\cdot x_j)\right] = 0$ when $i \neq j$, and in the last inequality we used the condition that x_i is drawn from $D_{u,b}$. Combining the above with (C.1) we obtain the desired result. ■

C.2. Uniform concentration of hinge loss

Proposition 26 *There exists an absolute constant $C_6 > 0$ such that the following holds for all isotropic log-concave distributions $D \in \mathcal{D}$ and all $D_{u,b}$ that satisfy the regularity condition. Let $S = \{x_1, \ldots, x_n\}$ be a set of n i.i.d. unlabeled instances drawn from $D_{u,b}$ which satisfies the regularity condition. Let $y_x = \operatorname{sign}(w^* \cdot x)$ for any $x \sim D_{u,b}$. Denote $W = B_2(u, r) \cap B_1(u, \rho)$ and let $G(w) = \frac{1}{n}\sum_{i=1}^n \ell_\tau(w; x_i, y_{x_i}) - \mathbb{E}_{x\sim D_{u,b}}\left[\ell_\tau(w; x, y_x)\right]$. Then with probability $1-\delta$,*

$$\sup_{w\in W} \left|G(w)\right| \le C_6 \left(\frac{b + \rho\sqrt{\log d}\log\frac{nd}{b}}{\tau\sqrt{n}} + \frac{b+r}{\tau\sqrt{n}}\sqrt{\log\frac{1}{\delta}} + \frac{b+\rho\log\frac{nd}{b}}{\tau n}\log\frac{1}{\delta}\right).$$

In particular, suppose $b = O(r)$, $\rho = O(\sqrt{s}r)$ and $\tau = \Omega(r)$. Then we have: for any $t > 0$, a sample size $n = \tilde{O}\left(\frac{1}{t^2} s \log^2\frac{d}{b}\cdot\log\frac{d}{\delta}\right)$ suffices to guarantee that with probability $1-\delta$, $\sup_{w\in W}\left|G(w)\right| \le t$.

Proof We will use Lemma 24 with function class $\mathcal{F} = \{(x,y) \mapsto \ell_\tau(w; x, y) : w \in W\}$ and the Orlicz norm with respect to ψ_1. We define $F(x,y) = 1 + \frac{b}{\tau} + \frac{\rho}{\tau}\|x\|_\infty$. It can be seen that for every $w \in W$,

$$\left|\ell_\tau(w; x, y)\right| \le 1 + \frac{|w\cdot x|}{\tau} \le 1 + \frac{u\cdot x}{\tau} + \frac{(w-u)\cdot x}{\tau} \le 1 + \frac{b}{\tau} + \frac{\rho}{\tau}\|x\|_\infty = F(x,y).$$

That is, for every f in $\mathcal{F}$, $\left|f(x,y)\right| \le F(x,y)$.

Step 1. We upper bound $\left\|\max_{1\le i\le n} F(x_i, y_{x_i})\right\|_{\psi_1}$. Since $\|\cdot\|_{\psi_1}$ is a norm, we have

$$\left\|\max_{1\le i\le n} F(x_i, y_{x_i})\right\|_{\psi_1} \le \left\|1 + \frac{b}{\tau}\right\|_{\psi_1} + \left\|\frac{\rho}{\tau}\cdot\max_{1\le i\le n}\|x_i\|_\infty\right\|_{\psi_1}$$
$$= 1 + \frac{b}{\tau} + \frac{\rho}{\tau}\cdot\left\|\max_{1\le i\le n}\|x_i\|_\infty\right\|_{\psi_1}$$
$$\le 1 + \frac{b}{\tau} + \frac{C_3\rho}{\tau}\log\frac{nd}{b}, \tag{C.2}$$

where we applied Lemma 23 in the last inequality.

Step 2. Next, we upper bound $\sup_{w\in W}\mathbb{E}_{x\sim D_{u,b}}\left[(\ell_\tau(w;x,y_x))^2\right]$. For all w in W, we have

$$\sup_{w\in W}\mathbb{E}_{x\sim D_{u,b}}\left[(\ell_\tau(w;x,y_x))^2\right]\le 2\cdot\sup_{w\in W}\mathbb{E}_{x\sim D_{u,b}}\left[1+\frac{(w\cdot x)^2}{\tau^2}\right]\le 2+2\bar{C}_2\cdot\frac{r^2+b^2}{\tau^2} \quad \text{(C.3)}$$

where the last inequality uses Lemma 17.

Step 3. Finally, we upper bound $\mathbb{E}_{S\sim D^n_{u,b}}\left[\sup_{w\in W}\left|G(w)\right|\right]$. Let $\sigma=\{\sigma_1,\ldots,\sigma_n\}$ where each σ_i is an i.i.d. draw from the Rademacher distribution. We have

$$\begin{aligned}\mathbb{E}_S\left[\sup_{w\in W}\left|G(w)\right|\right]&\le\frac{2}{n}\mathbb{E}_{S,\sigma}\left[\sup_{w\in W}\left|\sum_{i=1}^n\sigma_i\ell_\tau\left(w;x_i,y_{x_i}\right)\right|\right]\\&\le\frac{2}{\tau n}\mathbb{E}_{S,\sigma}\left[\sup_{w\in W}\left|\sum_{i=1}^n\sigma_i y_{x_i}(w\cdot x_i)\right|\right]\\&\le\frac{2b}{\tau\sqrt{n}}+\frac{2C_5\rho}{\tau}\cdot\sqrt{\frac{\log d}{n}}\cdot\log\frac{nd}{b}.\end{aligned} \quad \text{(C.4)}$$

In the above, the first inequality used standard symmetrization arguments; see, for example, Lemma 26.2 of Shalev-Shwartz and Ben-David (2014). In the second inequality, we used the contraction property of Rademacher complexity and the fact that $\ell_\tau(w;x,y)$ can be seen as a $\frac{1}{\tau}$-Lipschitz function $\phi(a)=\max\left\{0,1-\frac{a}{\tau}\right\}$ applied on input $a=yw\cdot x$. In the last inequality, we applied Lemma 25 with the fact that $\left|y_{x_i}\right|\le 1$.

Putting together. The first inequality of the proposition follows from combining (C.2), (C.3), and (C.4), and using Lemma 24 with $\mathcal{F}$ and ψ_1. Under our choice of (b,r,ρ,τ), with some calculation we obtain the bound of n. ■

C.3. Uniform concentration of relaxed sparse PCA

Proposition 27 *There exists an absolute constant $C_7>0$ such that the following holds for all isotropic log-concave distributions $D\in\mathcal{D}$ and all $D_{u,b}$ that satisfy the regularity condition. Let $S=\{x_1,\ldots,x_n\}$ be a set of n i.i.d. unlabeled instances drawn from $D_{u,b}$. Denote $G(H)=\frac{1}{n}\sum_{i=1}^n x_i^\top Hx_i-\mathbb{E}_{x\sim D_{u,b}}\left[x^\top Hx\right]$. Then with probability $1-\delta$,*

$$\sup_{H\in\mathcal{M}}\left|G(H)\right|\le C_7\rho^2\log^2\frac{nd}{b}\left(\sqrt{\frac{\log d}{n}}+\sqrt{\frac{\log(1/\delta)}{n}}+\frac{\log^2\frac{1}{\delta}}{n}\right).$$

In particular, suppose $\rho=O(\sqrt{s}r)$ and $r=O(b)$. Then we have: for any $t>0$, a sample size

$$n=\tilde{O}\left(\frac{1}{t^2}s^2b^4\log^4\frac{d}{b}\cdot\left(\log d+\log^2\frac{1}{\delta}\right)\right)$$

suffices to guarantee that with probability $1-\delta$, $\sup_{H\in\mathcal{M}}\left|G(H)\right|\le t$.

Proof Recall that $\mathcal{M}=\{H\in\mathbb{R}^{d\times d}:H\succeq 0,\|H\|\le r^2,\|H\|_1\le\rho^2\}$. For any matrix H, we denote by H_{ij} the (i,j)-th entry of the matrix H. For any vector x, we denote by $x^{(i)}$ the i-th coordinate of x.

We will use Lemma 24 with function class $\mathcal{F} = \{x \mapsto x^\top Hx : H \in \mathcal{M}\}$ and the Orlicz norm with respect to $\psi_{0.5}$. Consider the function $f(x) := x^\top Hx$ parameterized by $H \in \mathcal{M}$. First, we wish to find a function $F(x)$ that upper bounds $\left|f(x)\right|$. It is easy to see that

$$\left|x^\top Hx\right| = \left|\sum_{i,j} H_{ij} x^{(i)} x^{(j)}\right| \le \|x\|_\infty^2 \sum_{i,j} \left|H_{ij}\right| \le \rho^2 \|x\|_\infty^2 . \tag{C.5}$$

Thus it suffices to choose $F(x) = \rho^2 \|x\|_\infty^2$.

Step 1. We first bound $\left\|\sqrt{\max_{1\le i\le n} F(x_i)}\right\|_{\psi_1} = \left\|\rho \cdot \max_{1\le i\le n} \|x_i\|_\infty\right\|_{\psi_1} \le C_3 \rho \log \frac{nd}{b}$ by Lemma 23. By Part 3 of Lemma 22, $\left\|\max_{1\le i\le n} F(x)\right\|_{\psi_{0.5}}$ equals $\left\|\sqrt{\max_{1\le i\le n} F(x)}\right\|_{\psi_1}^2$. Thus

$$\left\|\max_{1\le i\le n} F(x)\right\|_{\psi_{0.5}} \le \left(C_3 \rho \log \frac{nd}{b}\right)^2 . \tag{C.6}$$

Step 2. Next we upper bound $\sup_{f\in\mathcal{F}} \mathbb{E}_{x\sim D_{u,b}} \left[(f(x))^2\right]$ where we remark that taking the superum over $f \in \mathcal{F}$ is equivalent to taking that over $H \in \mathcal{M}$. Since $\left|f(x)\right| \le F(x)$, we have

$$(f(x))^2 \le (F(x))^2 \le \rho^4 \|x\|_\infty^4 .$$

In view of Part 2 of Lemma 22, we have

$$\left(\mathbb{E}_{x\sim D_{u,b}} \left[\|x\|_\infty^4\right]\right)^{1/4} \le 24 \left\|\|x\|_\infty\right\|_{\psi_1} \le 24 C_3 \log \frac{d}{b}, \tag{C.7}$$

where the last inequality follows from Lemma 23. Hence,

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{x\sim D_{u,b}} \left[(f(x))^2\right] \le K_1 \rho^4 \log^4 \frac{d}{b} \tag{C.8}$$

for some absolute constant $K_1 > 0$.

Step 3. Finally, we upper bound $\mathbb{E}_{S\sim D^n} \left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}_{x\sim D_{u,b}} \left[f(x)\right]\right|\right]$. Let $\sigma = \{\sigma_1, \ldots, \sigma_n\}$ where σ_i's are independent draw from the Rademacher distribution. By standard symmetrization arguments (see e.g. Lemma 26.2 of Shalev-Shwartz and Ben-David (2014)), we have

$$\mathbb{E}_S \left[\sup_{f\in\mathcal{F}} \left|G(v, H)\right|\right] \le \frac{2}{n} \mathbb{E}_{S,\sigma} \left[\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right] = \frac{2}{n} \mathbb{E}_{S,\sigma} \left[\sup_{H\in\mathcal{M}} \left|\sum_{i=1}^n \sigma_i x_i^\top H x_i\right|\right] . \tag{C.9}$$

We first condition on S and consider the expectation over σ. For a matrix H, we use $\mathrm{vec}(H)$ to denote the vector obtained by concatenating all of the columns of H; likewise for $x_i x_i^\top$. It is crucial to observe that with this notation, for any $H \in \mathcal{M}$, we have $\left\|\mathrm{vec}(H)\right\|_1 = \|H\|_1 \le \rho^2$. It follows that

$$\begin{aligned}
\mathbb{E}_\sigma \left[\left|\sup_{H\in\mathcal{M}} \sum_{i=1}^n \sigma_i x_i^\top H x_i\right|\right] &\le \mathbb{E}_\sigma \left[\sup_{H:\left\|\mathrm{vec}(H)\right\|_1 \le \rho^2} \left|\sum_{i=1}^n \sigma_i \left\langle \mathrm{vec}(H), \mathrm{vec}(x_i x_i^\top)\right\rangle\right|\right] \\
&\le \rho^2 \sqrt{n \ln(2d^2)} \cdot \max_{1\le i\le n} \left\|\mathrm{vec}(x_i x_i^\top)\right\|_\infty \cdot \\
&= \rho^2 \sqrt{n \ln(2d^2)} \cdot \max_{1\le i\le n} \|x_i\|_\infty^2 .
\end{aligned}$$

where the second inequality is from Lemma 43, and the equality is from the observation that $\|\operatorname{vec}(x_i x_i^\top)\|_\infty = \|x_i\|_\infty^2$. Therefore,

$$\begin{aligned}\mathbb{E}_{S,\sigma}\left[\left|\sup_{H\in\mathcal{M}}\sum_{i=1}^n \sigma_i x_i^\top H x_i\right|\right] &\le \rho^2\sqrt{n\ln(2d^2)}\cdot\mathbb{E}_S\left[\max_{1\le i\le n}\|x_i\|_\infty^2\right]\\ &\le \rho^2\sqrt{2n\ln(2d)}\cdot 2\left\|\max_{1\le i\le n}\|x_i\|_\infty\right\|_{\psi_1}^2\\ &\le \rho^2\sqrt{2n\ln(2d)}\cdot C_3^2\log^2\frac{nd}{b},\end{aligned}$$

where the second inequality follows from Part 2 of Lemma 22, and the last inequality follows from Lemma 23. In summary,

$$\mathbb{E}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i x_i^\top H x_i\right|\right] \le K_2\sqrt{n\ln d}\cdot\rho^2\log^2\frac{nd}{b} \tag{C.10}$$

for some constant $K_2 > 0$.

Combining (C.9) and (C.10), we have

$$\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|G(H)\right|\right] \le \frac{K_3\sqrt{\log d}}{\sqrt{n}}\cdot\rho^2\log^2\frac{nd}{b}. \tag{C.11}$$

Putting together. Combining (C.6), (C.8), (C.11), and using Lemma 24 gives the first inequality of the proposition. Under our setting of (b, r, ρ), by some calculation we obtain the bound of n. The proof is complete. ■

Appendix D. Performance Guarantee of Algorithm 1

In this section, we leverage all the tools from previous sections to establish the performance guarantee of Algorithm 1. Our main theorem, Theorem 4, follows from the analysis of each step of the algorithm, as we describe below.

D.1. Analysis of sample complexity

Recall that we refer to the number of calls to $\mathrm{EX}_\eta^x(D, w^*)$ as the sample complexity of Algorithm 1. In order to obtain n_k instances residing the band $X_k := \{x : |w_{k-1}\cdot x| \le b_k\}$, we have to call $\mathrm{EX}_\eta^x(D, w^*)$ sufficient times.

Lemma 28 (Restatement of Lemma 6) *Consider phase k of Algorithm 1 for any $k \ge 1$. Suppose that Assumption 1 and 2 are satisfied. Further assume $\eta < \frac{1}{2}$. By making a number of $N_k = O\left(\frac{1}{b_k}\left(n_k + \log\frac{1}{\delta_k}\right)\right)$ calls to the instance generation oracle $\mathrm{EX}_\eta^x(D, w^*)$, we will obtain n_k instances that fall into X_k with probability $1 - \frac{\delta_k}{4}$.*

Proof By Lemma 19

$$\Pr_{x\sim D}(x \in X_k) \geq c_8 b_k.$$

This implies that

$$\begin{aligned}
&\Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}(x \in X_k \text{ and } x \text{ is clean}) \\
=\ &\Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}(x \in X_k \mid x \text{ is clean}) \cdot \Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}(x \text{ is clean}) \\
\geq\ &c_8 b_k (1-\eta).
\end{aligned}$$

We want to ensure that by drawing N_k instances from $\mathrm{EX}^x_\eta(D, w^*)$, with probability at least $1 - \frac{\delta_k}{4}$, n_k out of them fall into the band X_k. We apply the second inequality of Lemma 42 by letting $Z_i = \mathbf{1}_{\{x_i \in X_k \text{ and } x_i \text{ is clean}\}}$ and $\alpha = 1/2$, and obtain

$$\Pr\left(\left|\bar{T}_{\mathrm{C}}\right| \leq \frac{c_8 b_k (1-\eta)}{2} N_k\right) \leq \exp\left(-\frac{c_8 b_k (1-\eta) N_k}{8}\right),$$

where the probability is taken over the event that we make a number of N_k calls to $\mathrm{EX}^x_\eta(D, w^*)$. Thus, when $N_k \geq \frac{8}{c_8 b_k (1-\eta)}\left(n_k + \ln\frac{4}{\delta_k}\right)$, we are guaranteed that at least n_k samples from $\mathrm{EX}^x_\eta(D, w^*)$ fall into the band X_k with probability $1 - \frac{\delta_k}{4}$. The lemma follows by observing $\eta < \frac{1}{2}$. ■

D.2. Analysis of pruning and the structure of $\bar{T}$

With the instance set $\bar{T}$ on hand, we estimate the empirical noise rate after applying pruning (Step 6) in Algorithm 1. Recall that $n_k = \left|\bar{T}\right|$, i.e. the number of unlabeled instances before pruning.

Lemma 29 *Suppose that Assumption 1 and Assumption 2 are satisfied. Further assume $\eta < \frac{1}{2}$. If $D_{u,b}$ satisfies the regularity condition, we have*

$$\Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}\left(x \text{ is dirty} \mid x \in X_{u,b}\right) \leq \frac{2\eta}{c_8 b}$$

where c_8 was defined in Lemma 19 and $X_{u,b} := \{x \in \mathbb{R}^d : |u \cdot x| \leq b\}$.

Proof For an instance x, we use $\mathrm{tag}_x = 1$ to denote that x is drawn from D, and use $\mathrm{tag}_x = -1$ to denote that x is adversarially generated.

We first calculate the probability that an instance returned by $\mathrm{EX}^x_\eta(D, w^*)$ falls into the band $X_{u,b}$ as follows:

$$\begin{aligned}
&\Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}\left(x \in X_{u,b}\right) \\
=\ &\Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}\left(x \in X_{u,b} \text{ and } \mathrm{tag}_x = 1\right) + \Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}\left(x \in X_{u,b} \text{ and } \mathrm{tag}_x = -1\right) \\
\geq\ &\Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}\left(x \in X_{u,b} \text{ and } \mathrm{tag}_x = 1\right) \\
=\ &\Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}\left(x \in X_{u,b} \mid \mathrm{tag}_x = 1\right) \cdot \Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}\left(\mathrm{tag}_x = 1\right) \\
=\ &\Pr_{x\sim D}\left(x \in X_{u,b}\right) \cdot \Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}\left(\mathrm{tag}_x = 1\right) \\
\overset{\zeta}{\geq}\ &c_8 b \cdot (1-\eta) \\
\geq\ &\frac{1}{2} c_8 b,
\end{aligned}$$

where in the inequality ζ we applied Part 1 of Lemma 19. It is thus easy to see that

$$\Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}\left(\mathrm{tag}_x = -1 \mid x \in X_{u,b}\right) \le \frac{\Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}\left(\mathrm{tag}_x = -1\right)}{\Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}\left(x \in X_{u,b}\right)} \le \frac{2\eta}{c_8 b},$$

which is the desired result. ■

Lemma 30 *Suppose that Assumptions 1 and 2 are satisfied. Further assume $\eta \le c_5\epsilon$. For any $1 \le k \le k_0$, if $n_k \ge \frac{6}{\xi_k}\ln\frac{48}{\delta_k}$, then with probability $1-\frac{\delta_k}{24}$ over the draw of $\bar{T}$, the following results hold simultaneously:*

1. *$T_{\mathrm{C}} = \bar{T}_{\mathrm{C}}$ and hence $\tilde{T}_{\mathrm{C}} = \hat{T}_{\mathrm{C}}$, i.e. all clean instances in $\bar{T}$ are intact after pruning;*
2. *$\frac{|T_{\mathrm{D}}|}{|T|} \le \xi_k$, i.e. the empirical noise rate after pruning is upper bounded by ξ_k;*
3. *$|T_{\mathrm{C}}| \ge (1-\xi_k)n_k$.*

In particular, with the hyper-parameter setting in Section 3.2, $|T_{\mathrm{C}}| \ge \frac{1}{2}n_k$.

Proof Let us write events $E_1 := \{T_{\mathrm{C}} = \bar{T}_{\mathrm{C}}\}$, $E_2 := \{\left|\bar{T}_{\mathrm{D}}\right| \le \xi_k n_k\}$. We bound the probability of the two events over the draw of $\bar{T}$.

Recall that Lemma 20 implies that with probability $1-\frac{\delta_k}{48}$, all instances in $\bar{T}_{\mathrm{C}}$ are in the ℓ_∞-ball $B_\infty(0,\nu_k)$ for $\nu_k = c_9\log\frac{48|\bar{T}|d}{b_k\delta_k}$, which implies $\Pr(E_1) \ge 1-\frac{\delta_k}{48}$.

We next calculate the noise rate within the band $X_k := \{x : |w_{k-1}\cdot x| \le b_k\}$ by Lemma 29:

$$\Pr_{x\sim \mathrm{EX}^x_\eta(D,w^*)}(x \text{ is dirty} \mid x \in X_k) \le \frac{2\eta}{c_8 b_k} = \frac{2\eta}{c_8\bar{c}\cdot 2^{-k-3}} \le \frac{\pi}{c_8\bar{c}c_1}\cdot\frac{\eta}{\epsilon} \le \frac{\pi c_5}{c_8\bar{c}c_1} \le \frac{\xi_k}{2},$$

where the equality applies our setting on b_k, the second inequality uses the condition $k \le k_0$ and the setting $k_0 = \log\left(\frac{\pi}{16c_1\epsilon}\right)$, and the last inequality is guaranteed by our choice of c_5. Now we apply the first inequality of Lemma 42 by specifying $Z_i = \mathbf{1}_{\{x_i \text{ is dirty}\}}$, $\alpha = 1$ therein, which gives

$$\Pr\left(\left|\bar{T}_{\mathrm{D}}\right| \ge \xi_k n_k\right) \le \exp\left(-\frac{\xi_k n_k}{6}\right),$$

where the probability is taken over the draw of $\bar{T}$. This implies $\Pr(E_2) \ge 1-\frac{\delta_k}{48}$ provided that $n_k \ge \frac{6}{\xi_k}\ln\frac{48}{\delta_k}$.

By union bound, we have $\Pr(E_1\cap E_2) \ge 1-\frac{\delta_k}{24}$. We show that on the event $E_1\cap E_2$, the second and third parts of the lemma follow. To see this, we note that it trivially holds that $\frac{|T_{\mathrm{D}}|}{|T|} \le \frac{\left|\bar{T}_{\mathrm{D}}\right|}{n_k}$ since only dirty instances have chance to be removed. This proves the second part. Also, it is easy to see that $|T_{\mathrm{C}}| = \left|\bar{T}_{\mathrm{C}}\right| = \left|\bar{T}\right| - \left|\bar{T}_{\mathrm{D}}\right| \ge (1-\xi_k)\left|\bar{T}\right|$, which is exactly the third part. ■

D.3. Analysis of Algorithm 2

Lemma 31 (Restatement of Lemma 3) *Suppose that Assumption 1 and 2 are satisfied, and that $\eta \le c_5\epsilon$. There exists a constant $C_2 > 2$ such that the following holds. Consider phase k of Algorithm 1 for any $1 \le k \le k_0$. Denote by $\mathcal{M}_k$ the constraint set of (3.2). If $|T_{\mathrm{C}}| = \tilde{O}\left(s^2 \log^4 \frac{d}{b_k} \cdot \left(\log d + \log^2 \frac{1}{\delta_k}\right)\right)$, then with probability $1 - \frac{\delta_k}{24}$ over the draw of T_{C}, we have*

1. $\sup_{H \in \mathcal{M}_k} \frac{1}{|T_{\mathrm{C}}|} \sum_{x \in T_{\mathrm{C}}} x^\top H x \le 2C_2(b_k^2 + r_k^2)$;
2. $\sup_{w \in W_k} \frac{1}{|T_{\mathrm{C}}|} \sum_{x \in T_{\mathrm{C}}} (w \cdot x)^2 \le 5C_2\left(b_k^2 + r_k^2\right)$.

Proof The first part is an immediate result by combining Proposition 27 and Lemma 18, and recognizing our setting of b_k and r_k.

To see the second part, for any $w \in W_k$, we can upper bound $(w \cdot x)^2$ as follows:

$$(w \cdot x)^2 \le 2(w_{k-1} \cdot x)^2 + 2(v \cdot x)^2 \le 2b_k^2 + 2x^\top (vv^\top)x,$$

where $v = w - w_{k-1} \in B_2(0, r_k) \cap B_1(0, \rho_k)$. Hence it is easy to see that $vv^\top$ lies in $\mathcal{M}_k$. This indicates that for any $w \in W_k$, there exists an $H \in \mathcal{M}_k$ such that

$$(w \cdot x)^2 \le 2\left[b_k^2 + x^\top H x\right]. \tag{D.1}$$

Thus,

$$\sup_{w \in W_k} \frac{1}{|T_{\mathrm{C}}|} \sum_{x \in T_{\mathrm{C}}} (w \cdot x)^2 \le 2b_k^2 + 2 \sup_{H \in \mathcal{M}_k} \frac{1}{|T_{\mathrm{C}}|} \sum_{x \in T_{\mathrm{C}}} x^\top H x \le 5C_2(b_k^2 + r_k^2),$$

where the last inequality follows from the fact $C_2 \ge 2$. ■

Proposition 32 (Formal statement of Proposition 7) *Consider phase k of Algorithm 1 for any $1 \le k \le k_0$. Suppose that Assumption 1 and 2 are satisfied, and that $\eta \le c_5\epsilon$. With probability $1 - \frac{\delta_k}{8}$ (over the draw of $\bar{T}$), Algorithm 2 will output a function $q : T \to [0, 1]$ with the following properties:*

1. *for all $x \in T$, $q(x) \in [0, 1]$;*
2. $\frac{1}{|T|} \sum_{x \in T} q(x) \ge 1 - \xi_k$;
3. *for all $w \in W_k$, $\frac{1}{|T|} \sum_{x \in T} q(x)(w \cdot x)^2 \le 5C_2\left(b_k^2 + r_k^2\right)$.*

Furthermore, such function q can be found in polynomial time.

Proof Our choice on n_k satisfies the condition $n_k \ge \frac{6}{\xi_k} \ln \frac{48}{\delta_k}$ since ξ_k is lower bounded by a constant (see Section 3.2 for our parameter setting). Thus by Lemma 30, with probability $1 - \frac{\delta_k}{24}$, $|T_{\mathrm{C}}| \ge (1 - \xi_k)n_k$. We henceforth condition on this happening.

On the other side, Lemma 3 and Proposition 27 together implies that with probability $1 - \frac{\delta_k}{24}$, for all $H \in \mathcal{M}_k$, we have

$$\frac{1}{|T_{\mathrm{C}}|} \sum_{x \in T_{\mathrm{C}}} x^\top H x \le 2C_2(b_k^2 + r_k^2) \tag{D.2}$$

provided that

$$|T_{\mathrm{C}}| = \tilde{O}\left(s^2 \log^4 \frac{d}{b_k} \cdot \left(\log d + \log^2 \frac{1}{\delta_k}\right)\right). \tag{D.3}$$

Note that (D.3) is satisfied in view of the aforementioned event $|T_{\mathrm{C}}| \geq (1-\xi_k)n_k$ along with the setting of n_k and ξ_k. By union bound, the events (D.2) and $|T_{\mathrm{C}}| \geq (1-\xi_k)\,|T|$ hold simultaneously with probability at least $1 - \frac{\delta_k}{8}$.

Now we show that these two events together implies the existence of a feasible function $q(x)$ to Algorithm 2. Consider a particular function $q(x)$ with $q(x) = 0$ for all $x \in T_{\mathrm{D}}$ and $q(x) = 1$ for all $x \in T_{\mathrm{C}}$. We immediately have

$$\frac{1}{|T|}\sum_{x\in T} q(x) = \frac{|T_{\mathrm{C}}|}{|T|} \geq 1 - \xi_k.$$

In addition, for all $H \in \mathcal{M}_k$,

$$\frac{1}{|T|}\sum_{x\in T} q(x)x^\top Hx = \frac{1}{|T|}\sum_{x\in T_{\mathrm{C}}} x^\top Hx \leq \frac{1}{|T_{\mathrm{C}}|}\sum_{x\in T_{\mathrm{C}}} x^\top Hx \leq 2C_2(b_k^2 + r_k^2), \tag{D.4}$$

where the first inequality follows from the fact $|T| \geq |T_{\mathrm{C}}|$ and the second inequality follows from (D.2). Namely, such function $q(x)$ satisfies all the constraints in Algorithm 2. Finally, combining (D.1) and (D.4) gives Part 3.

It remains to show that for a given candidate function q, a separation oracle for Algorithm 2 can be constructed in polynomial time. First, it is straightforward to check whether the first two constraints $q(x) \in [0,1]$ and $\sum_{x\in T} q(x) \geq (1-\xi)\,|T|$ are violated. If not, we just need to further check if there exists an $H \in \mathcal{M}_k$ such that $\frac{1}{|T|}\sum_{x\in T} q(x)x^\top Hx > 2C_2(b_k^2 + r_k^2)$. To this end, we appeal to solving the following program:

$$\max_{H\in\mathcal{M}_k} \frac{1}{|T|}\sum_{x\in T} q(x)x^\top Hx.$$

This is a semidefinite program that can be solved in polynomial time (Boyd and Vandenberghe, 2004). If the maximum objective value is greater than $2C_2(b_k^2 + r_k^2)$, then we conclude that q is not feasible; otherwise we would have found a desired function. ■

The analysis of the following proposition closely follows Awasthi et al. (2017) with a refined treatment. Let $\ell_{\tau_k}(w;p) := \sum_{x\in T} p(x)\ell_{\tau_k}(w;x,y_x)$ where y_x is the unrevealed label of x that the adversary has committed to.

Proposition 33 (Formal statement of Proposition 10) *Consider phase k of Algorithm 1. Suppose that Assumption 1 and 2 are satisfied. Assume that $\eta \leq c_5\epsilon$. Set N_k and ξ_k as in Section 3.2. Denote $z_k := \sqrt{b_k^2 + r_k^2} = \sqrt{\bar{c}^2+1}\cdot 2^{-k-3}$. With probability $1 - \frac{\delta_k}{4}$ over the draw of $\bar{T}$, for all $w \in W_k$*

$$\ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}}) \leq \ell_{\tau_k}(w;p) + 2\xi_k\left(1 + \sqrt{10C_2}\cdot\frac{z_k}{\tau_k}\right) + \sqrt{10C_2\xi_k}\cdot\frac{z_k}{\tau_k},$$
$$\ell_{\tau_k}(w;p) \leq \ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}}) + 2\xi_k + \sqrt{20C_2\xi_k}\cdot\frac{z_k}{\tau_k}.$$

In particular, with our hyper-parameter setting,

$$\left|\ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}}) - \ell_{\tau_k}(w;p)\right| \leq \kappa.$$

Proof The choice of n_k guarantees that Lemma 30 and Proposition 32 hold simultaneously with probability $1 - \frac{\delta_k}{4}$. We thus have for all $w \in W_k$

$$\frac{1}{|T|}\sum_{x\in T} q(x)(w\cdot x)^2 \le 5C_2 z_k^2, \tag{D.5}$$

$$\frac{1}{|T_{\mathrm{C}}|}\sum_{x\in T_{\mathrm{C}}} (w\cdot x)^2 \le 5C_2 z_k^2, \tag{D.6}$$

$$\frac{|T_{\mathrm{D}}|}{|T|} \le \xi_k. \tag{D.7}$$

In the above expression, (D.5) and (D.6) follow from Part 3 and Part 2 of Lemma 31 respectively, (D.7) follows from Lemma 30. It follows from Eq. (D.7) and $\xi_k \le 1/2$ that

$$\frac{|T|}{|T_{\mathrm{C}}|} = \frac{|T|}{|T| - |T_{\mathrm{D}}|} = \frac{1}{1 - |T_{\mathrm{D}}|/|T|} \le \frac{1}{1-\xi_k} \le 2. \tag{D.8}$$

In the following, we condition on the event that all these inequalities are satisfied.

Step 1. First we upper bound $\ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}})$ by $\ell_{\tau_k}(w;p)$.

$$\begin{aligned}
|T_{\mathrm{C}}|\cdot \ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}}) &= \sum_{x\in T_{\mathrm{C}}} \ell(w;x,y_x) \\
&= \sum_{x\in T}\Big[q(x)\ell(w;x,y_x) + \big(\mathbf{1}_{\{x\in T_{\mathrm{C}}\}} - q(x)\big)\ell(w;x,y_x)\Big] \\
&\overset{\zeta_1}{\le} \sum_{x\in T} q(x)\ell(w;x,y_x) + \sum_{x\in T_{\mathrm{C}}}(1-q(x))\ell(w;x,y_x) \\
&\overset{\zeta_2}{\le} \sum_{x\in T} q(x)\ell(w;x,y_x) + \sum_{x\in T_{\mathrm{C}}}(1-q(x))\left(1+\frac{|w\cdot x|}{\tau_k}\right) \\
&\overset{\zeta_3}{\le} \sum_{x\in T} q(x)\ell(w;x,y_x) + \xi_k|T| + \frac{1}{\tau_k}\sum_{x\in T_{\mathrm{C}}}(1-q(x))|w\cdot x| \\
&\overset{\zeta_4}{\le} \sum_{x\in T} q(x)\ell(w;x,y_x) + \xi_k|T| + \frac{1}{\tau_k}\sqrt{\sum_{x\in T_{\mathrm{C}}}(1-q(x))^2}\cdot\sqrt{\sum_{x\in T_{\mathrm{C}}}(w\cdot x)^2} \\
&\overset{\zeta_5}{\le} \sum_{x\in T} q(x)\ell(w;x,y_x) + \xi_k|T| + \frac{1}{\tau_k}\sqrt{\xi_k|T|}\cdot\sqrt{5C_2|T_{\mathrm{C}}|}\cdot z_k,
\end{aligned} \tag{D.9}$$

where ζ_1 follows from the simple fact that

$$\begin{aligned}
\sum_{x\in T}\big(\mathbf{1}_{\{x\in T_{\mathrm{C}}\}} - q(x)\big)\ell(w;x,y_x) &= \sum_{x\in T_{\mathrm{C}}}(1-q(x))\ell(w;x,y_x) + \sum_{x\in T_{\mathrm{D}}}(-q(x))\ell(w;x,y_x) \\
&\le \sum_{x\in T_{\mathrm{C}}}(1-q(x))\ell(w;x,y_x),
\end{aligned}$$

ζ_2 explores the fact that the hinge loss is always upper bounded by $1+\frac{|w\cdot x|}{\tau_k}$ and that $1-q(x)\ge 0$, ζ_3 follows from Part 2 of Proposition 32, ζ_4 applies Cauchy-Schwarz inequality, and ζ_5 uses Eq. (D.6).

In view of Eq. (D.8), we have $\frac{|T|}{|T_{\mathrm{C}}|} \leq 2$. Continuing Eq. (D.9), we obtain

$$\begin{aligned}
\ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}}) &\leq \frac{1}{|T_{\mathrm{C}}|}\sum_{x\in T} q(x)\ell(w;x,y_x) + 2\xi_k + \sqrt{10C_2\xi_k}\cdot\frac{z_k}{\tau_k} \\
&= \frac{\sum_{x\in T} q(x)}{|T_{\mathrm{C}}|}\sum_{x\in T} p(x)\ell(w;x,y_x) + 2\xi_k + \sqrt{10C_2\xi_k}\cdot\frac{z_k}{\tau_k} \\
&= \ell_{\tau_k}(w;p) + \left(\frac{\sum_{x\in T} q(x)}{|T_{\mathrm{C}}|} - 1\right)\sum_{x\in T} p(x)\ell(w;x,y_x) + 2\xi_k + \sqrt{10C_2\xi_k}\cdot\frac{z_k}{\tau_k} \\
&\leq \ell_{\tau_k}(w;p) + \left(\frac{|T|}{|T_{\mathrm{C}}|} - 1\right)\sum_{x\in T} p(x)\ell(w;x,y_x) + 2\xi_k + \sqrt{10C_2\xi_k}\cdot\frac{z_k}{\tau_k} \\
&\leq \ell_{\tau_k}(w;p) + 2\xi_k\sum_{x\in T} p(x)\ell(w;x,y_x) + 2\xi_k + \sqrt{10C_2\xi_k}\cdot\frac{z_k}{\tau_k}, \qquad \text{(D.10)}
\end{aligned}$$

where in the last inequality we use $|T|/|T_{\mathrm{C}}| - 1 = \frac{|T_{\mathrm{D}}|/|T|}{1-|T_{\mathrm{D}}|/|T|} \leq 2|T_{\mathrm{D}}|/|T|$. On the other hand, we have the following result which will be proved later on.

Claim 34 $\sum_{x\in T} p(x)\ell(w;x,y_x) \leq 1 + \sqrt{10C_2}\cdot\frac{z_k}{\tau_k}$.

Therefore, continuing Eq. (D.10) we have

$$\ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}}) \leq \ell_{\tau_k}(w;p) + 2\xi_k\left(1 + \sqrt{10C_2}\cdot\frac{z_k}{\tau_k}\right) + \sqrt{10C_2\xi_k}\cdot\frac{z_k}{\tau_k}.$$

which proves the first inequality of the proposition.

Step 2. We move on to prove the second inequality of the theorem, i.e. using $\ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}})$ to upper bound $\ell_{\tau_k}(w;p)$. Let us denote by $p_{\mathrm{D}} = \sum_{x\in T_{\mathrm{D}}} p(x)$ the probability mass on dirty instances. Then

$$p_{\mathrm{D}} = \frac{\sum_{x\in T_{\mathrm{D}}} q(x)}{\sum_{x\in T} q(x)} \leq \frac{|T_{\mathrm{D}}|}{(1-\xi_k)|T|} \leq \frac{\xi_k}{1-\xi_k} \leq 2\xi_k, \qquad \text{(D.11)}$$

where the first inequality follows from $q(x) \leq 1$ and Part 2 of Proposition 32, the second inequality follows from (D.7), and the last inequality is by our choice $\xi_k \leq 1/2$.

Note that by Part 2 of Proposition 32 and the choice $\xi_k \leq 1/2$, we have $\sum_{x\in T} q(x) \geq (1-\xi_k)|T| \geq |T|/2$. Hence

$$\sum_{x\in T} p(x)(w\cdot x)^2 = \frac{1}{\sum_{x\in T} q(x)}\sum_{x\in T} q(x)(w\cdot x)^2 \leq \frac{2}{|T|}\sum_{x\in T} q(x)(w\cdot x)^2 \leq 10C_2 z_k^2 \qquad \text{(D.12)}$$

where the last inequality holds because of (D.5). Thus,

$$\begin{aligned}
\sum_{x\in T_{\mathrm{D}}} p(x)\ell(w;x,y_x) &\le \sum_{x\in T_{\mathrm{D}}} p(x)\left(1+\frac{|w\cdot x|}{\tau_k}\right)\\
&= p_{\mathrm{D}} + \frac{1}{\tau_k}\sum_{x\in T_{\mathrm{D}}} p(x)\,|w\cdot x|\\
&= p_{\mathrm{D}} + \frac{1}{\tau_k}\sum_{x\in T}\left(\mathbf{1}_{\{x\in T_{\mathrm{D}}\}}\sqrt{p(x)}\right)\cdot\left(\sqrt{p(x)}\,|w\cdot x|\right)\\
&\le p_{\mathrm{D}} + \frac{1}{\tau_k}\sqrt{\sum_{x\in T}\mathbf{1}_{\{x\in T_{\mathrm{D}}\}}p(x)}\cdot\sqrt{\sum_{x\in T}p(x)(w\cdot x)^2}\\
&\overset{\text{(D.12)}}{\le} p_{\mathrm{D}} + \sqrt{p_{\mathrm{D}}}\cdot\sqrt{10C_2}\cdot\frac{z_k}{\tau_k}.
\end{aligned}$$

With the result on hand, we bound $\ell_{\tau_k}(w;p)$ as follows:

$$\begin{aligned}
\ell_{\tau_k}(w;p) &= \sum_{x\in T_{\mathrm{C}}} p(x)\ell(w;x,y_x) + \sum_{x\in T_{\mathrm{D}}} p(x)\ell(w;x,y_x)\\
&\le \sum_{x\in T_{\mathrm{C}}} \ell(w;x,y_x) + \sum_{x\in T_{\mathrm{D}}} p(x)\ell(w;x,y_x)\\
&= \ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}}) + \sum_{x\in T_{\mathrm{D}}} p(x)\ell(w;x,y_x)\\
&\le \ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}}) + p_{\mathrm{D}} + \sqrt{p_{\mathrm{D}}}\cdot\sqrt{10C_2}\cdot\frac{z_k}{\tau_k}\\
&\overset{\text{(D.11)}}{\le} \ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}}) + 2\xi_k + \sqrt{20C_2\xi_k}\cdot\frac{z_k}{\tau_k},
\end{aligned}$$

which proves the second inequality of the proposition.

Putting together. We would like to show $\left|\ell_{\tau_k}(w;p) - \ell_{\tau_k}(w;\tilde{T}_{\mathrm{C}})\right| \le \kappa$. Indeed, this is guaranteed by our setting of ξ_k in Section 3.2 which ensures that ξ_k simultaneously fulfills the following three constraints:

$$\begin{aligned}
2\xi_k\left(1+\sqrt{10C_2}\cdot\frac{z_k}{\tau_k}\right) + \sqrt{10C_2\xi_k}\cdot\frac{z_k}{\tau_k} &\le \kappa,\\
2\xi_k + \sqrt{20C_2\xi_k}\cdot\frac{z_k}{\tau_k} \le \kappa, \quad \text{and} \quad \xi_k &\le \frac{1}{2}.
\end{aligned}$$

This completes the proof. ■

Proof [Proof of Claim 34] Since $\ell(w; x, y_x) \leq 1 + \frac{|w \cdot x|}{\tau_k}$, it follows that

$$\begin{aligned}\sum_{x \in T} p(x)\ell(w; x, y_x) &\leq \sum_{x \in T} p(x)\left(1 + \frac{|w \cdot x|}{\tau_k}\right) \\ &= 1 + \frac{1}{\tau_k}\sum_{x \in T} p(x)\,|w \cdot x| \\ &\leq 1 + \frac{1}{\tau_k}\sqrt{\sum_{x \in T} p(x)(w \cdot x)^2} \\ &\overset{(D.12)}{\leq} 1 + \sqrt{10C_2} \cdot \frac{z_k}{\tau_k},\end{aligned}$$

which completes the proof of Claim 34. ■

The following result is a simple application of Proposition 26. It shows that the loss evaluated on clean instances concentrates around the expected loss.

Proposition 35 (Restatement of Proposition 11) *Consider phase k of Algorithm 1. Suppose that Assumption 1 and 2 are satisfied, and assume $\eta \leq c_5\epsilon$. Then with probability $1 - \frac{\delta_k}{4}$ over the draw of $\bar{T}$, for all $w \in W_k$ we have*

$$\left|L_{\tau_k}(w) - \ell_{\tau_k}(w; \tilde{T}_{\mathrm{C}})\right| \leq \kappa.$$

where $L_{\tau_k}(w) := \mathbb{E}_{x \sim D_{w_{k-1}, b_k}}\left[\ell_{\tau_k}(w; x, \operatorname{sign}(w^* \cdot x))\right]$.

Proof The choice of n_k, i.e. the size of $\left|\bar{T}\right|$, ensures that with probability $1 - \frac{\delta_k}{8}$, $|T_{\mathrm{C}}|$ is at least $\zeta \log \zeta$ where $\zeta = K \cdot s \log^2 \frac{d}{b_k} \cdot \log \frac{d}{\delta_k}$ for some constant $K > 0$ in view of Lemma 30. This observation in allusion to Proposition 26 and union bound, immediately gives the desired result. ■

D.4. Analysis of random sampling

Proposition 36 (Restatement of Proposition 12) *Consider phase k Algorithm 1. Suppose that Assumption 1 and 2 are satisfied, and assume $\eta \leq c_5\epsilon$. Set n_k and m_k as in Section 3.2. Then with probability $1 - \frac{\delta_k}{4}$ over the draw of S_k, for all $w \in W_k$ we have*

$$\left|\ell_{\tau_k}(w; p) - \ell_{\tau_k}(w; S_k)\right| \leq \kappa.$$

Proof

Since we applied pruning to remove all instances with large ℓ_∞-norm, this proposition can be proved by a standard concentration argument for uniform convergence of linear classes under distributions with ℓ_∞ bounded support. We include the proof for completeness.

Note that the randomness is taken over the i.i.d. draw of m_k samples from T according to the distribution p over T. Thus, for any $(x, y) \in S_k$, $\mathbb{E}[\ell_{\tau_k}(w; x, y)] = \ell_{\tau_k}(w; p)$. Moreover, let $R_k = \max_{x \in T} \|x\|_\infty$. Any instance x drawn from T satisfies $\|x\|_\infty \leq R_k$ with probability 1. It is also easy to verify that

$$\ell_{\tau_k}(w; x, y) \leq 1 + \frac{|w \cdot x|}{\tau_k} \leq 1 + \frac{(w - w_{k-1}) \cdot x}{\tau_k} + \frac{|w_{k-1} \cdot x|}{\tau_k} \leq 1 + \frac{\rho_k R_k}{\tau_k} + \frac{b_k}{\tau_k}.$$

By Theorem 8 of Bartlett and Mendelson (2002) along with standard symmetrization arguments, we have that with probability at least $1 - \frac{\delta_k}{4}$,

$$\left|\ell_{\tau_k}(w;p) - \ell_{\tau_k}(w;S_k)\right| \le \left(1 + \frac{\rho_k R_k}{\tau_k} + \frac{b_k}{\tau_k}\right)\sqrt{\frac{\ln(4/\delta_k)}{2m_k}} + \mathcal{R}(\mathcal{F};S_k) \tag{D.13}$$

where $\mathcal{R}(\mathcal{F};S_k)$ denotes the Rademacher complexity of function class $\mathcal{F}$ on the labeled set S_k, and $\mathcal{F} := \{\ell_{\tau_k}(w;x,y) : w \in W_k\}$. In order to calculate $\mathcal{R}(\mathcal{F};S_k)$, we observe that each function $\ell_{\tau_k}(w;x,y)$ is a composition of $\phi(a) = \max\{0, 1 - \frac{1}{\tau_k}ya\}$ and function class $\mathcal{G} := \{x \mapsto w \cdot x : w \in W_k\}$. Since $\phi(a)$ is $\frac{1}{\tau_k}$-Lipschitz, by contraction property of Rademacher complexity, we have

$$\mathcal{R}(\mathcal{F};S_k) \le \frac{1}{\tau_k}\mathcal{R}(\mathcal{G};S_k). \tag{D.14}$$

Let $\sigma = \{\sigma_1, \ldots, \sigma_{m_k}\}$ where the σ_i's are i.i.d. draw from the Rademacher distribution, and let $V_k = B_2(0,r_k) \cap B_1(0,\rho_k)$. We compute $\mathcal{R}(\mathcal{G};S_k)$ as follows:

$$\begin{aligned}
\mathcal{R}(\mathcal{G};S_k) &= \frac{1}{m_k}\mathbb{E}_\sigma\left[\sup_{w\in W_k} w \cdot \left(\sum_{i=1}^{m_k}\sigma_i x_i\right)\right] \\
&= \frac{1}{m_k}\mathbb{E}_\sigma\left[w_{k-1}\cdot\left(\sum_{i=1}^{m_k}\sigma_i x_i\right)\right] + \frac{1}{m_k}\mathbb{E}_\sigma\left[\sup_{w\in W_k}(w - w_{k-1})\cdot\left(\sum_{i=1}^{m_k}\sigma_i x_i\right)\right] \\
&= \frac{1}{m_k}\mathbb{E}_\sigma\left[\sup_{v\in V_k} v\cdot\left(\sum_{i=1}^{m_k}\sigma_i x_i\right)\right] \\
&\le \rho_k R_k\sqrt{\frac{2\log(2d)}{m_k}},
\end{aligned}$$

where the first equality is by the definition of Rademacher complexity, the second equality simply decompose w as a sum of w_{k-1} and $w - w_{k-1}$, the third equality is by the fact that every σ_i has zero mean, and the inequality applies Lemma 43. We combine the above result with (D.13) and (D.14), and obtain that with probability $1 - \frac{\delta_k}{4}$,

$$\left|\ell_{\tau_k}(w;p) - \ell_{\tau_k}(w;S_k)\right| \le \left(1 + \frac{\rho_k R_k}{\tau_k} + \frac{b_k}{\tau_k}\right)\sqrt{\frac{\ln(4/\delta_k)}{m_k}} + \frac{\rho_k R_k}{\tau_k}\sqrt{\frac{2\log(2d)}{m_k}}. \tag{D.15}$$

Recall that we remove all instances with large ℓ_∞-norm in the pruning step of Algorithm 1. In particular, we have

$$R_k \le c_9 \log\frac{48 n_k d}{b_k \delta_k}.$$

Plugging this upper bound into (D.15) and using our hyper-parameter setting gives

$$\left|\ell_{\tau_k}(w;p) - \ell_{\tau_k}(w;S_k)\right| \le K_1 \cdot \sqrt{s}\log\frac{n_k d}{b_k\delta_k}\left(\sqrt{\frac{\log(1/\delta_k)}{m_k}} + \sqrt{\frac{\log d}{m_k}}\right)$$

for some constant $K_1 > 0$. Hence,

$$m_k = O\left(s \log^2 \frac{n_k d}{b_k \delta_k} \cdot \log \frac{d}{\delta_k}\right) = \tilde{O}\left(s \log^2 \frac{d}{b_k \delta_k} \cdot \log \frac{d}{\delta_k}\right)$$

suffices to ensure $\left|\ell_{\tau_k}(w; p) - \ell_{\tau_k}(w; S_k)\right| \leq \kappa$ with probability $1 - \frac{\delta_k}{4}$. ■

D.5. Analysis of Per-Phase Progress

Let $L_{\tau_k}(w) = \mathbb{E}_{x \sim D_{w_{k-1}, b_k}}\left[\ell_{\tau_k}(w; x, \text{sign}\,(w^* \cdot x))\right]$.

Lemma 37 (Lemma 3.7 of Awasthi et al. (2017)) *Suppose Assumption 1 is satisfied. Then*

$$L_{\tau_k}(w^*) \leq \frac{\tau_k}{c_0 \min\{b_k, 1/9\}}.$$

In particular, by our choice of τ_k

$$L_{\tau_k}(w^*) \leq \kappa.$$

Lemma 38 *For any* $1 \leq k \leq k_0$, *if* $w^* \in W_k$, *then with probability* $1 - \delta_k$, $\text{err}_{D_{w_{k-1}, b_k}}(v_k) \leq 8\kappa$.

Proof Observe that with the setting of N_k, we have with probability $1 - \delta_k$ over all the randomness in phase k, Lemma 28, Proposition 33, Proposition 35 and Proposition 36 hold simultaneously. Now we condition on the event that all of these properties are satisfied, which implies for all $w \in W_k$,

$$\left|L_{\tau_k}(w) - \ell_{\tau_k}(w; S_k)\right| \leq 3\kappa. \tag{D.16}$$

We have

$$\begin{aligned}\text{err}_{D_{w_{k-1}, b_k}}(v_k) \leq L_{\tau_k}(v_k) \overset{\zeta_1}{\leq} \ell_{\tau_k}(v_k; S_k) + 3\kappa \overset{\zeta_2}{\leq} \min_{w \in W_k} \ell_{\tau_k}(w; S_k) + 4\kappa \overset{\zeta_3}{\leq} \ell_{\tau_k}(w^*; S_k) + 4\kappa \\ \leq L_{\tau_k}(w^*) + 7\kappa.\end{aligned}$$

In the above, the first inequality follows from the fact that hinge loss upper bounds the 0/1 loss, ζ_1 and the last inequality applies (C.1), ζ_2 is by the definition of v_k (see Algorithm 1), and ζ_3 is by our assumption that w^* is feasible. The proof is complete in view of Lemma 37. ■

Lemma 39 *For any* $1 \leq k \leq k_0$, *if* $w^* \in W_k$, *then with probability* $1 - \delta_k$, $\theta(v_k, w^*) \leq 2^{-k-8}\pi$.

Proof

For $k = 1$, by Lemma 38 and that we actually sample from D, we have

$$\Pr_{x \sim D}\left(\text{sign}\,(v_1 \cdot x) \neq \text{sign}\,(w^* \cdot x)\right) \leq 8\kappa.$$

Hence Part 4 of Lemma 14 indicates that

$$\theta(v_1, w^*) \leq 8c_2\kappa = 16c_2\kappa \cdot 2^{-1}. \tag{D.17}$$

Now we consider $2 \leq k \leq k_0$. Denote $X_k = \{x : |w_{k-1} \cdot x| \leq b_k\}$, and $\bar{X}_k = \{x : |w_{k-1} \cdot x| > b_k\}$. We will show that the error of v_k on both X_k and $\bar{X}_k$ is small, hence v_k is a good approximation to w^*.

First, we consider the error on X_k, which is given by

$$\begin{aligned}
&\Pr_{x\sim D}\left(\operatorname{sign}\left(v_k \cdot x\right) \neq \operatorname{sign}\left(w^* \cdot x\right), x \in X_k\right) \\
&= \Pr_{x\sim D}\left(\operatorname{sign}\left(v_k \cdot x\right) \neq \operatorname{sign}\left(w^* \cdot x\right) \mid x \in X_k\right) \cdot \Pr_{x\sim D}(x \in X_k) \\
&= \operatorname{err}_{D_{w_{k-1},b_k}}(v_k) \cdot \Pr_{x\sim D}(x \in X_k) \\
&\leq 8\kappa \cdot 2b_k \\
&= 16\kappa b_k, \qquad \text{(D.18)}
\end{aligned}$$

where the inequality is due to Lemma 38 and Lemma 19. Note that the inequality holds with probability $1 - \delta_k$.

Next we derive the error on $\bar{X}_k$. Note that Lemma 10 of Zhang (2018) states for any unit vector u, and any general vector v, $\theta(v, u) \leq \pi \|v - u\|_2$. Hence,

$$\theta(v_k, w^*) \leq \pi \left\|v_k - w^*\right\|_2 \leq \pi(\|v_k - w_{k-1}\|_2 + \left\|w^* - w_{k-1}\right\|_2) \leq 2\pi r_k.$$

Recall that we set $r_k = 2^{-k-3} < 1/4$ in our algorithm and choose $b_k = \bar{c} \cdot r_k$ where $\bar{c} \geq 8\pi/c_4$, which allows us to apply Lemma 15 and obtain

$$\begin{aligned}
\Pr_{x\sim D}\left(\operatorname{sign}\left(v_k \cdot x\right) \neq \operatorname{sign}\left(w^* \cdot x\right), x \notin X_k\right) &\leq c_3 \cdot 2\pi r_k \cdot \exp\left(-\frac{c_4 \bar{c} \cdot r_k}{2 \cdot 2\pi r_k}\right) \\
&= 2^{-k} \cdot \frac{c_3 \pi}{4} \exp\left(-\frac{c_4 \bar{c}}{4\pi}\right).
\end{aligned}$$

This in allusion to (D.18) gives

$$\operatorname{err}_D(v_k) \leq 16\kappa \cdot \bar{c} \cdot r_k + 2^{-k} \cdot \frac{c_3 \pi}{4} \exp\left(-\frac{c_4 \bar{c}}{4\pi}\right) = \left(2\kappa \bar{c} + \frac{c_3 \pi}{4} \exp\left(-\frac{c_4 \bar{c}}{4\pi}\right)\right) \cdot 2^{-k}.$$

Recall that we set $\kappa = \exp(-\bar{c})$ and denote by $f(\bar{c})$ the coefficient of 2^{-k} in the above expression. By Part 4 of Lemma 14

$$\theta(v_k, w^*) \leq c_2 \operatorname{err}_D(v_k) \leq c_2 f(\bar{c}) \cdot 2^{-k}. \qquad \text{(D.19)}$$

Now let $g(\bar{c}) = c_2 f(\bar{c}) + 16 c_2 \exp(-\bar{c})$. By our choice of $\bar{c}$, $g(\bar{c}) \leq 2^{-8}\pi$. This ensures that for both (D.17) and (D.19), $\theta(v_k, w^*) \leq 2^{-k-8}\pi$ for any $k \geq 1$. ■

Lemma 40 *For any $1 \leq k \leq k_0$, if $\theta(v_k, w^*) \leq 2^{-k-8}\pi$, then $w^* \in W_{k+1}$.*

Proof We first show that $\|w_k - w^*\|_2 \le r_{k+1}$. Let $\hat{v}_k = v_k / \|v_k\|_2$. By algebra $\|\hat{v}_k - w^*\|_2 = 2\sin\frac{\theta(v_k,w^*)}{2} \le \theta(v_k, w^*) \le 2^{-k-8}\pi \le 2^{-k-6}$. Now we have

$$\begin{aligned}
\left\|w_k - w^*\right\|_2 &= \left\|\mathcal{H}_s(v_k)/\left\|\mathcal{H}_s(v_k)\right\|_2 - w^*\right\|_2 \\
&= \left\|\mathcal{H}_s(\hat{v}_k)/\left\|\mathcal{H}_s(\hat{v}_k)\right\|_2 - w^*\right\|_2 \\
&\le 2\left\|\mathcal{H}_s(\hat{v}_k) - w^*\right\|_2 \\
&\le 4\left\|\hat{v}_k - w^*\right\|_2 \\
&\le 2^{-k-4} \\
&= r_{k+1}.
\end{aligned}$$

By the sparsity of w_k and w^*, and our choice $\rho_{k+1} = \sqrt{2s}r_{k+1}$, we always have

$$\left\|w_k - w^*\right\|_1 \le \sqrt{2s}\left\|w_k - w^*\right\|_2 \le \sqrt{2s}r_{k+1} = \rho_{k+1}.$$

The proof is complete. ■

D.6. Proof of Theorem 4

Proof We will prove the theorem with the following claim.

Claim 41 *For any $1 \le k \le k_0$, with probability at least $1 - \sum_{i=1}^{k}\delta_i$, w^* is in W_{k+1}.*

Based on the claim, we immediately have that with probability at least $1 - \sum_{k=1}^{k_0}\delta_k \ge 1 - \delta$, w^* is in W_{k_0+1}. By our construction of W_{k_0+1}, we have

$$\left\|w^* - w_{k_0}\right\|_2 \le 2^{-k_0-4}.$$

This, together with Part 4 of Lemma 14 and the fact that $\theta(w^*, w_{k_0}) \le \pi\left\|w^* - w_{k_0}\right\|_2$ (see Lemma 10 of Zhang (2018)), implies

$$\mathrm{err}_D(w_{k_0}) \le \frac{\pi}{c_1}\cdot 2^{-k_0-4} = \epsilon.$$

Finally, we derive the sample complexity and label complexity. Recall that n_k was involved in Proposition 32, i.e. the quantity $|T|$, where we required

$$n_k = \tilde{O}\left(s^2\log^4\frac{d}{b_k}\cdot\left(\log d + \log^2\frac{1}{\delta_k}\right) + \log\frac{1}{\delta_k}\right) = \tilde{O}\left(s^2\log^4\frac{d}{b_k}\cdot\left(\log d + \log^2\frac{1}{\delta_k}\right)\right).$$

It is also involved in Proposition 36, where we need

$$m_k = O\left(s\log^2\frac{n_k d}{b_k\delta_k}\cdot\log\frac{d}{\delta_k}\right)$$

and $n_k \ge m_k$ since S_k is a labeled subset of T. As m_k has a cubic dependence on $\log\frac{1}{\delta_k}$, our final choice of n_k is given by

$$n_k = \tilde{O}\left(s^2\log^4\frac{d}{b_k}\cdot\left(\log d + \log^3\frac{1}{\delta_k}\right)\right). \tag{D.20}$$

This in turn gives

$$m_k = \tilde{O}\left(s \log^2 \frac{d}{b_k \delta_k} \cdot \log \frac{d}{\delta_k}\right). \tag{D.21}$$

Therefore, by Lemma 28 we obtain an upper bound of the sample size N_k at phase k as follows:

$$N_k = \tilde{O}\left(\frac{s^2}{b_k} \log^4 \frac{d}{b_k} \cdot \left(\log d + \log^3 \frac{1}{\delta_k}\right)\right) \leq \tilde{O}\left(\frac{s^2}{\epsilon} \log^4 d \left(\log d + \log^3 \frac{1}{\delta}\right)\right),$$

where the last inequality follows from $b_k = \Omega(\epsilon)$ for all $k \leq k_0$ and our choice of δ_k. Consequently, the total sample complexity

$$N = \sum_{k=1}^{k_0} N_k \leq k_0 \cdot \tilde{O}\left(\frac{s^2}{\epsilon} \log^4 d \left(\log d + \log^3 \frac{1}{\delta}\right)\right) = \tilde{O}\left(\frac{s^2}{\epsilon} \log^4 d \left(\log d + \log^3 \frac{1}{\delta}\right)\right).$$

Likewise, we can show that the total label complexity

$$m = \sum_{k=1}^{k_0} m_k \leq k_0 \cdot \tilde{O}\left(s \log^2 \frac{d}{\epsilon\delta} \cdot \log \frac{d}{\delta}\right) = \tilde{O}\left(s \log^2 \frac{d}{\epsilon\delta} \cdot \log \frac{d}{\delta} \cdot \log \frac{1}{\epsilon}\right).$$

It remains to prove Claim 41 by induction. First, for $k = 1$, $W_1 = B_2(0,1) \cap B_1(0, \sqrt{s})$. Therefore, $w^* \in W_1$ with probability 1. Now suppose that Claim 41 holds for some $k \geq 2$, that is, there is an event E_{k-1} that happens with probability $1 - \sum_i^{k-1} \delta_i$, and on this event $w^* \in W_k$. By Lemma 39 we know that there is an event F_k that happens with probability $1 - \delta_k$, on which $\theta(v_k, w^*) \leq 2^{-k-8}\pi$. This further implies that $w^* \in W_{k+1}$ in view of Lemma 40. Therefore, consider the event $E_{k-1} \cap F_k$, on which $w^* \in W_{k+1}$ with probability $\Pr(E_{k-1}) \cdot \Pr(F_k \mid E_{k-1}) = (1 - \sum_i^{k-1} \delta_i)(1 - \delta_k) \geq 1 - \sum_{i=1}^{k} \delta_i$. ■

Appendix E. Miscellaneous Lemmas

Lemma 42 (Chernoff bound) *Let $Z_1, Z_2, \ldots, Z_n$ be n independent random variables that take value in $\{0,1\}$. Let $Z = \sum_{i=1}^n Z_i$. For each Z_i, suppose that $\Pr(Z_i = 1) \leq \eta$. Then for any $\alpha \in [0,1]$*

$$\Pr\left(Z \geq (1+\alpha)\eta n\right) \leq e^{-\frac{\alpha^2 \eta n}{3}}.$$

When $\Pr(Z_i = 1) \geq \eta$, for any $\alpha \in [0,1]$

$$\Pr\left(Z \leq (1-\alpha)\eta n\right) \leq e^{-\frac{\alpha^2 \eta n}{2}}.$$

Lemma 43 (Theorem 1 of Kakade et al. (2008)) *Let $\sigma = (\sigma_1, \ldots, \sigma_n)$ where σ_i's are independent draws from the Rademacher distribution and let $x_1, \ldots, x_n$ be given instances in $\mathbb{R}^d$. Then*

$$\mathbb{E}_\sigma\left[\sup_{w \in B_1(0,\rho)} \sum_{i=1}^n \sigma_i w \cdot x_i\right] \leq \rho\sqrt{2n \log(2d)} \max_{1 \leq i \leq n} \|x_i\|_\infty.$$