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

Jie Shen

Stevens Institute of Technology

`jie.shen@stevens.edu`

January 21, 2021

Abstract

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

1 Introduction

In many practical applications, there are massive amounts of unlabeled data but labeling is expensive. This distinction has driven the study of active learning [CAL94, BBZ07, Das11, Han14, ABL17], where labels are initially hidden and the learner must pay for each label it wishes to be revealed. The goal is to design querying strategies to avoid less informative labeling requests, e.g. the labels that can be inferred from previously seen samples. Parallel to active learning, online learning concerns the scenario where the learner observes a stream of samples and makes real-time model updating in order to compete with the best model obtained by seeing all the history data in a batch [Ros58, LW89, CBLW96, Zin03, Sha12, Haz19]. Online learning algorithms have also been broadly investigated in machine learning, and have found various successful applications owing to its potential of savings in memory cost, low computational cost per sample, and its generalization ability [CCG04, KT08].

In this paper, we study the important problem of active learning of homogeneous halfspaces in the online setting, where the learner observes a stream of unlabeled data and makes spot decision of whether or not to query the labels. The goal is to achieve the best of the two worlds: label efficiency from active learning and computational efficiency from online learning. In this spectrum, there are a

---

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

number of early works that share the same merit with this paper. For example, [FSST97] proposed a query-by-committee learning algorithm and [DKM05] developed a Perceptron-like algorithm, both of which are implemented in an online fashion and enjoy a near-optimal label complexity bound of $\tilde{O}(d \log \frac{1}{\epsilon})$ where d is the dimension of the instance[2] and ϵ is the target error rate. However, there are two crucial assumptions made by these works that seem too stringent: 1) the marginal distribution over the unlabeled data is uniform on the unit sphere in $\mathbb{R}^d$; and 2) there exists a perfect halfspace that incurs zero error rate with respect to the underlying distribution. In this regard, a natural question is: *can we design an online active learner, that provably works under a significantly more general family of marginal distributions while achieving arbitrarily small error rate without the realizability condition?*

To be more concrete, we are interested in designing an online active learning algorithm that PAC learns some underlying halfspace [Val84] when the instances are drawn from an isotropic log-concave distribution [LV07] and the labels are corrupted by the adversarial noise [Hau92, KSS92]. It is worth mentioning that the family of isotropic log-concave distributions is a significant generalization of the uniform distribution since it includes a variety of prevalent distributions such as Gaussian, exponential, logistic. It is also known that establishing performance guarantees under this family is often technically subtle compared to that of uniform distribution due to the asymmetricity nature of log-concave distributions [Vem10, BL13, DKS18]. Returning to the noise model, we note that the adversarial noise, where the adversary may choose an *arbitrary* joint distribution such that the unlabeled data distribution is isotropic log-concave and the overall probability of a noisy label is constrained to be at most ν, is a realistic yet remarkably challenging regime, as suggested by many hardness results [FGKP06, GR09, Dan16, DKZ20, BH20].

Under different assumptions on the adversarial noise rate (which might be suboptimal), a large body of works have established PAC guarantees for online learning with adversarial noise. Unfortunately, none of them resolves the aforementioned question in full. For example, under uniform marginal distributions, Theorem 3 of [KKMS05] showed that a surprisingly simple averaging scheme already tolerates noise rate $\nu = \tilde{\Omega}(\epsilon)$, though the label and sample complexity are both $\tilde{O}(d^2/\epsilon^2)$.[3] The same algorithm was then revisited by [KLS09] under isotropic log-concave distributions, with a worse noise tolerance of $\nu = \tilde{\Omega}(\epsilon^3)$. Very recently, [DKTZ20] proposed a novel objective function by optimizing which with projected online gradient descent, one is guaranteed to tolerate the adversarial noise of $\nu = \tilde{\Omega}(\epsilon)$. Notably, their analysis applies to marginal distributions that are more general than isotropic log-concave. Compared to these passive learning algorithms which have label and sample complexity polynomial in $\frac{1}{\epsilon}$, the work of [YZ17] is more in line with this paper in the sense that they considered the active learning setting. Hence, their label complexity bound has an exponentially better dependence on $\frac{1}{\epsilon}$. However, the noise tolerance of [YZ17] reads as $\nu = \tilde{\Omega}(\epsilon/\log d)$ and their analysis is applicable only to uniform distributions due to the crucial need of the symmetricity of marginal distributions. As we can see in Table 1, none of the prior works subsumes others.

1.1 Main results

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

---

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

[3][KKMS05] analyzed two algorithms: a (batch) polynomial regression and an online averaging. Here we are referring to the online averaging approach. We defer the comparison with the former to Section 4.1 when we are in the position to discuss the connection between adversarial noise and agnostic learning.

Table 1: Comparison to state-of-the-art online algorithms that are robust to adversarial noise. Prior online algorithms cannot even incorporate attribute efficiency. Even under the non-sparse case (i.e. $s = d$), our algorithm (Theorem 1) has better noise tolerance and works under more general distributions than [YZ17], and has improved label and sample complexity compared to [KKMS05] and [DKTZ20].

Work	Log-concave?	Label complexity	Sample complexity	Noise tolerance
Theorem 3 of [KKMS05]	✗	$\tilde{O}(d^2/\epsilon^2)$	$\tilde{O}(d^2/\epsilon^2)$	$\nu = \tilde{\Omega}(\epsilon)$
[YZ17]	✗	$\tilde{O}(d\log\frac{1}{\epsilon})$	$\tilde{O}(d/\epsilon)$	$\nu = \tilde{\Omega}(\epsilon/\log d)$
[DKTZ20]	✓	$\tilde{O}(d/\epsilon^4)$	$\tilde{O}(d/\epsilon^4)$	$\nu = \tilde{\Omega}(\epsilon)$
This work (Theorem 1)	✓	$\tilde{O}(d \cdot \mathrm{polylog}(\frac{1}{\epsilon}))$	$\tilde{O}(d/\epsilon)$	$\nu = \Omega(\epsilon)$
This work (Theorem 2)	✓	$\tilde{O}(s \cdot \mathrm{polylog}(d, \frac{1}{\epsilon}))$	$\tilde{O}(\frac{s}{\epsilon} \cdot \mathrm{polylog}(d))$	$\nu = \Omega(\epsilon)$

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

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

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

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

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

We compare with state-of-the-art online algorithms of [KKMS05, YZ17, DKTZ20] that are tolerant to adversarial noise. [YZ17] presented an active learner and obtained analogous label and sample complexity to this work. However, their noise tolerance reads as $\nu = \tilde{\Omega}(\epsilon/\log d)$ while our algorithm is able to tolerate $\nu = \Omega(\epsilon)$; in addition, our results apply to significantly broader marginal distributions. Since [KKMS05, DKTZ20] considered passive learning, our active learning algorithm naturally enjoys label complexity that has exponentially better dependence on ϵ. Even for the sample complexity, we obtain improved dependence on d and ϵ. On the other side, it is worth mentioning that all these three algorithms run faster than our algorithm, and the analysis of [DKTZ20] works under more general marginal distributions.

In addition to the properties aforementioned, we show that our algorithm can essentially incorporate attribute efficiency [Lit87]. That is, when the concept class consists of s-sparse halfspaces, the obtained label and sample complexity scale as $\tilde{O}(s \cdot \mathrm{polylog}(d))$. This characteristic is especially useful when there is limited availability of samples, a problem that has been studied for decades in machine learning and statistics; see, e.g. [CDS98, Tib96, KS04, CT05, Fel07, PVY17].

We have the following result for learning sparse halfspaces.

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

Observe that Theorem 1 is a special case of the above by setting $s = d$. We note that the state-of-the-art online algorithms of [KKMS05, YZ17, DKTZ20] do not enjoy attribute efficiency – it is yet nontrivial for them to encompass this property. Hence we have exponentially better dependence on the dimension d in label and sample complexity. See Table 1 for a summary of the comparison.

Finally, through an interesting observation made in [ABL17], it is possible to translate our main results for the adversarial noise model to the agnostic model of [Hau92, KSS92] where no assumption is made on the noise rate ν. Let $\text{OPT} = \min_{w \in \mathcal{C}} \text{err}_D(w)$.

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

1.2 Overview of our techniques

Our algorithm is inspired in part by [ZSA20]. We present an overview of the techniques below, and highlight the algorithmic connection to them as well as our novelty.

1) Active learning via stagewise online mirror descent. The first ingredient in our algorithm is a novel perspective of approaching active learning of halfspaces via stagewise online learning, recently utilized by [ZSA20] for learning halfspaces with benign noise. In each phase, given an initial halfspace $w_0 \in \mathbb{R}^d$, regardless of how the samples are generated, standard regret bound established for online mirror descent with linear loss $\langle w, \alpha g_t \rangle$ and ℓ_p-norm regularizer $\Phi(w)$ implies that the produced sequence of iterates $\{w_{t-1}\}_{t=1}^T$ must satisfy the following with certainty:

$$\frac{1}{T} \sum_{t=1}^T \langle u, -g_t \rangle \leq \frac{1}{T} \sum_{t=1}^T \langle w_{t-1}, -g_t \rangle + \frac{\mathcal{B}_\Phi(u, w_0)}{\alpha T} + \frac{\alpha}{T} \sum_{t=1}^T \|g_t\|_q^2, \tag{1.1}$$

where $u \in \mathbb{R}^d$ is the underlying halfspace we aim to approximate, $\mathcal{B}_\Phi(\cdot, \cdot)$ denotes Bregman divergence induced by Φ, and $q \in (0, 1)$ is such that $\frac{1}{p} + \frac{1}{q} = 1$. Our goal is threefold: attribute efficiency, label efficiency, and small error rate. We will first specify $p \approx 1$ to achieve attribute efficiency which is a well-known technique in online learning [GLS01, Gen03]. In order to reduce the error rate of the initial halfspace w_0 with a few label queries, we need to design suitable gradients g_t and choose proper step size α and iteration number T such that a) the right-hand side of (1.1) is as small as $O(\theta(w_0, u))$; and b) the left-hand side is bounded from below by $\theta(\bar{w}, u)$ for certain halfspace $\bar{w}$ that depends on the sequence $\{w_{t-1}\}_{t=1}^T$. It is then possible to show that $\theta(\bar{w}, u) \leq \frac{1}{2} \cdot \theta(w_0, u)$, and we can use $\bar{w}$ as the initial iterate for the next phase of online mirror descent to reduce the distance to u with geometric rate. Therefore, the crucial challenges lie in the design of g_t to accommodate specific noise model and an associated sampling scheme (since g_t depends on the sample). These are also the key technical differences between our work and [ZSA20], which we elaborate on below.

2) A semi-random gradient update for tolerating adversarial noise. In [ZSA20], the gradient g_t is heavily tailored to the bounded noise condition [MN06]. We find it technically hard to reuse it for the adversarial noise since it is well known that the latter is a more involved noise model that will always violate the conditions assumed in the former.[4] Therefore, we consider an alternative yet fairly natural candidate: we choose g_t as the original gradient used in the Perceptron algorithm. That is, given the currently learned halfspace w_{t-1} and a new labeled sample (x_t, y_t), we set $g_t = -y_t x_t \cdot \mathbf{1}_{\{y_t \neq \text{sign}(w_{t-1} \cdot x_t)\}}$. It remains to develop a plausible sampling scheme so that a)

---

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

$-g_t$ will have nontrivial correlation with the underlying halfspace u; and b) only most informative instances are sampled for labeling. To this end, we propose a new sampling region $X_{\hat{w}_{t-1},b} := \{x \in \mathbb{R}^d : 0 < \hat{w}_{t-1} \cdot x \leq b\}$, where $\hat{w}_{t-1} = \frac{w_{t-1}}{\|w_{t-1}\|}$. Such time-varying region is different from active learning using empirical risk minimization [BBZ07, ABL17, Zha18] as in these works x is sampled from the full band $|\hat{w}_0 \cdot x| \leq b$. On one hand, using our sampling region leads to a linear loss $\langle w_{t-1}, g_t \rangle$ at most $O(b)$, while the band used by ERM would result in a loss of $O(b\sqrt{s}\log d)$. Observe that a tighter control on the loss implies tighter upper bound in (1.1). On the other hand, we discover that by restricting on querying the label of instances in $X_{\hat{w}_{t-1},b}$, we are reducing the randomness of model updating because now $g_t = x_t \cdot \mathbf{1}_{\{y_t \neq 1\}}$, i.e. we update the model only when the returned label $y_t = -1$. It turns out that such *semi-randomness* facilitates our control of the correlation between each $-g_t$ and the underlying halfspace u. We note that the semi-random updating rule is inspired by [YZ17], where they used a much narrower sampling region $\{x : \frac{b}{2\sqrt{d}} \leq \hat{w}_{t-1} \cdot x \leq \frac{b}{\sqrt{d}}\}$ and a carefully rescaled Perceptron gradient $g_t = (\hat{w}_{t-1} \cdot x_t) \cdot x_t \cdot \mathbf{1}_{\{y_t \neq 1\}}$ to accommodate their projection-free algorithm for learning under the uniform marginal distribution. In contrast, we incorporate different sampling strategy and gradients into (projected) mirror descent for PAC learning under isotropic log-concave marginal distributions.

3) A new characterization of the correlation between gradient and the underlying halfspace. Our last ingredient is applying localization in the concept space [ABL17]. Roughly speaking, before running online mirror descent, it is possible to construct an ℓ_2-ball where the underlying halfspace u resides in. Such trust region will be serving as the convex constraint set for online minimization. Using a novel analysis, we show that this interesting observation in allusion to the dedicated design of gradients implies that under the adversarial noise model,

$$\mathbb{E}_{(x_t,y_t)\sim D_{\hat{w}_{t-1},b}}[\langle u, -g_t \rangle] \geq f_{u,b}(w_{t-1}) - \beta \cdot \theta(w_0, u), \tag{1.2}$$

where we have the potential function

$$f_{u,b}(w_{t-1}) := \mathbb{E}_{(x_t,y_t)\sim D_{\hat{w}_{t-1},b}}\left[|u \cdot x_t| \cdot \mathbf{1}_{\{u \cdot x_t < 0\}}\right], \tag{1.3}$$

and $\beta > 0$ is some quantity to be controlled. We argue that the function $f_{u,b}(w_{t-1})$ serves almost as a measure of $\theta(w_{t-1}, u)$; hence combining it with (1.1) we have that the average of $\theta(w_{t-1}, u)$ is upper bounded by $\frac{1}{2} \cdot \theta(w_0, u)$. This observation, in conjunction with a non-standard online-to-batch conversion, results in the desired halfspace $\bar{w}$. We note two different aspects compared to [ZSA20]. First, our potential function $f_{u,b}$ is slightly distinct since we are considering a smaller sampling region. Second and more importantly, when deriving the lower bound for the correlation of u and $-g_t$, we carry out a more involved analysis but still incur a negative penalty $-\beta \cdot \theta(w_0, u)$ resulted from the adversarial noise model. In contrast, this term does not appear in [ZSA20] (since the bounded noise model they studied is more benign). Manipulating the penalty turns out to be subtle since if the factor β is large, there is no hope to upper bound the average of $f_{u,b}(w_{t-1})$ by $\frac{1}{2} \cdot \theta(w_0, u)$. We circumvent the technical issue by showing that β is dominated by ν/b which is small as soon as $\nu \leq c_0 \epsilon$ for sufficiently small constant c_0 and b is carefully chosen to be greater than ϵ; see Lemma 13 in the appendix.

1.3 Related Works

Label-efficient learning has also been broadly studied since gathering high quality labels is often expensive [CAL94, Das05, Das11]. The prominent approaches include disagreement-based active learning [Han11, Han14], margin-based active learning [BBZ07, BL13, ABHU15], selective sampling [CCG11, DGS12], and adaptive one-bit compressed sensing [ZYJ14, BFN+17]. There are also

a number of interesting works that appeal to extra information to mitigate the labeling cost, such as comparison [XZS⁺17, KLMZ17, HKLM20, ZS20] and search [BHLZ16].

Adversarial noise is closely related to the agnostic model, which was studied in [Hau92] and then coined out by [KSS92]. Under the uniform marginal distributions, [KKMS05] obtained the best error rate (see Section 4.1 for a precise statement), though the running time and sample complexity is $O(d^{1/\epsilon^4})$. This bound has been proved almost best possible in very recent works under the statistical query model [DKZ20, GGK20]. Interestingly, [Dan15] characterized the tradeoff between the error rate and running time under the uniform marginal distribution by combining the techniques of polynomial regression [KKMS05] and localization [ABL17]. By comprising on the error rate, [KLS09] presented an averaging-based algorithm and showed how to boost it to tolerate an adversarial noise rate $\nu = \tilde{\Omega}(\epsilon^3)$ in polynomial time when the marginal distribution is isotropic log-concave. Such noise tolerance has been improved by a series of recent ERM-based works [BBL09, BHLZ10, ZC14, ABHZ16, ABL17, Zha18, DKS18], among which $\nu = \Omega(\epsilon)$ is the best known noise tolerance. However, solving an ERM often requires more memory storage and computational cost per sample than online methods [SS07].

Achieving attribute efficiency has been a long-standing goal in machine learning and statistics [Blu90, BHL95], and has been pursued in online classification [Lit87], learning decision lists [Ser99, KS04, LS06], compressed sensing [Don06, CW08, TW10, SL18], one-bit compressed sensing [BB08, PV16], and variable selection [FL01, FF08, Zha10, SL17a, SL17b]. It is worth mentioning that [ABHZ16] gave a label-inefficient algorithm for uniformly learning sparse halfspaces with adversarial noise while this work and the closely related works consider non-uniform learning.

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

2 Preliminaries

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

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

In our active learning algorithm, we will often want to draw instances from D_X conditioned on

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

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

We reserve p and q for specific values: $p = \frac{\ln(8d)}{\ln(8d)-1}$ and $q = \ln(8d)$ (note that $\frac{1}{p} + \frac{1}{q} = 1$). We will use the ℓ_p-norm regularizer, that is, $\Phi(w) = \frac{1}{2(p-1)}\|w - v\|_p^2$ for some given vector v. It is known that $\Phi(w)$ is 1-strongly convex with respect to the ℓ_p-norm [SS07, Lemma 17]. We denote the Bregman divergence induced by $\Phi(w)$ by $\mathcal{B}_\Phi(w; w') = \Phi(w) - \Phi(w') - \langle \nabla\Phi(w'), w - w' \rangle$. Observe that $\mathcal{B}_\Phi(w; v) = \Phi(w)$ where v is the reference vector appearing in $\Phi(w)$.

We will sometimes phrase our theoretical guarantee in terms of angles between two halfspaces. The following lemma, due to [BL13], is useful to convert the guarantee of angles to that of error rate.

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

We remark that a more difficult noise model is the agnostic model [Hau92, KSS92, KKMS05], where only Assumption 1 is satisfied and the goal is to output a halfspace that approximates the best halfspace in $\mathcal{C}$. The results from our theorems under the adversarial noise model translate immediately into PAC guarantees for the agnostic model; see Section 4.1 for more details.

3   Main Algorithm

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

Since INITIALIZE invokes REFINE as well, we will introduce the latter first. Generally speaking, the REFINE algorithm, i.e. Algorithm 3, belongs to the family of online mirror descent algorithms with Perceptron gradient and ℓ_p-norm regularization. The crucial ingredients that make it attribute and label efficient are a carefully crafted constraint set, a time-varying sampling region, and semi-random gradients, as we described in Section 1.2. In particular, the constraint set $\mathcal{K}$ is constructed in such a way that the underlying halfspace u is guaranteed to stay in it (with overwhelming probability). Thus, it serves as a trust region into which all the iterates w_t are projected back. It is worth mentioning that we did not put an ℓ_1-norm constraint in $\mathcal{K}$; this is because we already have utilized the ℓ_p-norm regularization to simultaneously guarantee attribute efficiency [GLS01, Gen03] and the stability of online minimization [Sha12, Ora19]. The second component in REFINE is the time-varying sampling region $X_{\hat{w}_{t-1},b} = \{x \in \mathbb{R}^d : 0 < \hat{w}_{t-1} \cdot x \leq b\}$, which results in a linear loss as small as $O(b)$ in each iteration. In contrast, a naive online approach to simulate the ERM algorithm of [Zha18] would lead to a loss as large as $O(\sqrt{s}\log d)$. The idea of using time-varying sampling

---

Algorithm 1 Main Algorithm

Require: Target error rate $\epsilon \in (0,1)$, failure probability $\delta \in (0,1)$, sparsity s.
Ensure: Halfspace $\tilde{u} \in \mathbb{R}^d$ such that $\Pr_{(x,y)\sim D}(\mathrm{sign}(\tilde{u} \cdot x) \neq \mathrm{sign}(u \cdot x)) \leq \epsilon$.

1: $v_0 \leftarrow \text{INITIALIZE}(\frac{\delta}{2}, s)$.
2: $K \leftarrow \lceil \log \frac{\bar{c}\pi}{8\epsilon} \rceil$ where $\bar{c}$ is defined in Lemma 4.
3: **for** $k = 1, 2, \ldots, K$ **do**
4:   $v_k \quad \leftarrow \quad \text{REFINE}(v_{k-1}, \frac{\delta}{2k(k+1)}, s, \alpha_k, b_k, \mathcal{K}_k, \Phi_k, T_k)$, where step size $\alpha_k =$
   $\tilde{\Theta}\left(2^{-k} \cdot \left(\log \frac{d \cdot k^2 \cdot 2^k}{\delta}\right)^{-2}\right)$, band width $b_k = \Theta\left(2^{-k}\right)$, constraint set

$$\mathcal{K}_k = \left\{ w : \|w - v_{k-1}\| \leq \pi \cdot 2^{-k-2}, \ \|w\| \leq 1 \right\},$$

   regularizer $\Phi_k(w) = \frac{1}{2(p-1)}\|w - v_{k-1}\|_p^2$, number of iterations $T_k = \tilde{O}\left(s \log d \cdot \left(\log \frac{d \cdot k^2 \cdot 2^k}{\delta}\right)^2\right)$.

5: **end for**
6: **return** $\tilde{u} \leftarrow v_K$.

Algorithm 2 INITIALIZE

Require: Failure probability δ', sparsity s.
Ensure: An s-sparse halfspace v_0 such that $\theta(v_0, u) \leq \frac{\pi}{8}$.

1: $(x_1, y_1), \ldots, (x_m, y_m) \leftarrow$ call EX to draw m instances, and query EX_y for their labels, where $m = O(s \log \frac{d}{\delta'})$.
2: Compute $w_{\mathrm{avg}} = \frac{1}{m} \sum_{i=1}^{m} y_i x_i$.
3: Let $w^\sharp = \frac{\mathcal{H}_{\tilde{s}}(w_{\mathrm{avg}})}{\|\mathcal{H}_{\tilde{s}}(w_{\mathrm{avg}})\|}$, where $\tilde{s} = 81 \cdot 2^{40} s$.
4: Let $\mathcal{K} = \left\{ w \in \mathbb{R}^d : \|w\| \leq 1, \ w \cdot w^\sharp \geq \frac{1}{9 \cdot 2^{20}} \right\}$ and find a point $w_0 \in \mathcal{K} \cap \{w \in \mathbb{R}^d : \|w\|_1 \leq \sqrt{s}\}$.

5: **return** $v_0 \leftarrow \text{REFINE}(w_0, \frac{\delta'}{2}, s, \alpha, b, \mathcal{K}, \Phi, T)$, where step size $\alpha = \tilde{\Theta}\left(\log^{-2} \frac{d}{\delta'}\right)$, band width $b = \frac{1}{81 \cdot 2^{22}}$, regularizer $\Phi(w) = \frac{1}{2(p-1)}\|w - w_0\|_p^2$, and number of iterations $T = \tilde{O}\left(s \log d \cdot \left(\log \frac{d}{\delta'}\right)^2\right)$.

---

paradigm has appeared in a few online active learning algorithms [YZ17, ZSA20]. Ours is less restrictive than the one in [YZ17], and is more dedicated to the much more challenging adversarial noise compared to the bounded noise model considered in [ZSA20]. Along with the new sampling region is a semi-random Perceptron gradient. Recall that the original gradient used in Perceptron is given by $g_t = -y_t x_t \cdot \mathbf{1}_{\{y_t \neq \mathrm{sign}(w_{t-1} \cdot x_t)\}}$. Since $x_t \in X_{\hat{w}_{t-1}, b}$, we update the model only when the label y_t returned by the adversary equals -1, thus inducing the gradient displayed in Algorithm 3. Finally, after running mirror descent for T iterations, we perform an averaging scheme followed by hard thresholding, to ensure that the output $\tilde{w}$ belongs the to concept class. We remark that the running time of REFINE is polynomial in d, since in each iteration t, updating the model requires solving a convex program. Regarding the label complexity, it is exactly equal to the total iteration number T. We also remark that obtaining x_t can be done by calling EX for $O(b^{-1})$ times since the probability mass of $X_{\hat{w}_{t-1}, b}$ on D_X is $\Theta(b)$; see Lemma 29.

Now we elaborate on the INITIALIZE algorithm, namely Algorithm 2. Technically speaking, one important condition for the success of our analysis is that an overwhelming portion of the iterates must have acute angles with the underlying halfspace u. Therefore, the hypothesis testing approach proposed in [ABL17] does not work out in our case since we will lose control of the intermediate iterates. To circumvent the technical challenge, we tailor the averaging-based initialization scheme

Algorithm 3 REFINE

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

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

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

---

of [ZSA20] to the adversarial noise model. It is possible to show that as far as the noise rate ν is low, w_{avg} has a positive correlation with u, and performing hard thresholding almost preserves it, i.e. $u \cdot w^\sharp = \Omega(1)$. Therefore, we obtain a good reference vector $w^\sharp$ which is guaranteed to have an acute angle with u. Based on an enhanced constraint set that takes the correlation into consideration, we are able to show that most of the iterates are admissible, and hence our analysis of the REFINE algorithm can be reused to show that the output v_0 will have a small acute angle with u. It is also important to note that we are not making efforts to optimize the constants that appear in the INITIALIZE algorithm.

4 Performance Guarantee

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

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

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

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

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

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

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

4.1 Implications to agnostic learning

In the agnostic model [Hau92, KSS92], the adversary chooses a joint distribution D over $\mathbb{R}^d \times \{-1, 1\}$ and fixes it throughout the learning process. Let $\text{OPT} = \min_{w \in \mathcal{C}} \text{err}_D(w)$. The goal of the learner is to output a hypothesis $\tilde{u}$ such that $\text{err}_D(\tilde{u}) \leq c \cdot \text{OPT} + \epsilon$ for some approximation factor $c \geq 1$. The crucial difference from the adversarial noise model is that now Assumption 2 may not be satisfied (in other words, OPT can be very large compared to the target error rate ϵ).

[KKMS05] developed a polynomial regression algorithm that achieves approximation guarantee with $c = 1$, where the computational and sample complexity are both $O(d^{2^{\text{poly}(1/\epsilon)}})$ for learning under isotropic log-concave distributions and are $O(d^{1/\epsilon^4})$ for uniform distributions.

On the other side, [KKMS05] and a number of recent works also obtained weaker (yet still quite nontrivial) approximation guarantee of $O(\text{OPT}) + \epsilon$ with running time and sample complexity polynomial in d and $\frac{1}{\epsilon}$; see, for example, [ABL17, Zha18, DKS18, DKTZ20]. We follow this line, and remark that our main result, Theorem 5, can be immediately translated into the constant approximation guarantee under the agnostic model. In fact, Lemma C.1 of [ABL17] made an interesting observation that if an algorithm can tolerate adversarial noise of $\nu = \Omega(\epsilon)$, then it essentially achieves error rate of $O(\text{OPT}) + \epsilon$ in the agnostic model, with the same running time, label complexity, and sample complexity (up to a constant multiplicative factor). We therefore have the following result, which is a formal statement of Theorem 3, by combining Theorem 5 we established and Lemma C.1 of [ABL17]; we omit the proof since it is fairly straightforward.

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

4.2 Proof of Theorem 5

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

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

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

We first explain the results in Theorem 12. The label complexity of REFINE is straightforward since we request one label per iteration, and the total number of iterations is T. We start with the analysis of the sample complexity of REFINE, i.e. the number of calls to EX in Step 2 therein. Since the marginal distribution is assumed to be isotropic log-concave, Lemma 29 shows that $\Pr_{x_t \sim D_X}(x_t \in$

$X_{\hat{w}_{t-1},b}) \geq c_2 b$ for some absolute constant $c_2 > 0$. Thus, by Chernoff bound, we need to call EX for $O(b^{-1} + \log \frac{T}{\delta'})$ times in order to obtain one x_t with probability $1 - \frac{\delta'}{2T}$. Thus, by union bound over the T iterations in REFINE, with probability $1 - \frac{\delta'}{2}$, the total number of calls to EX is $O(T/b + T \log \frac{T}{\delta'})$. Note that this is also the computational cost for sampling. On the other side, updating the iterates, i.e. Step 3, involves solving a convex program in $\mathbb{R}^d$, which has a running time polynomial in d per iteration. Thus, the overall computational cost of REFINE is $T \cdot \text{poly}(d, \frac{1}{b}, \log \frac{1}{\delta'})$.

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

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

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

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

5 Conclusion and Future Works

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

We discuss a few important directions for future investigation. First, it is interesting to develop online PAC algorithms with $\text{OPT} + \epsilon$ approximation error under the agnostic model, by leveraging, for example, the polynomial regression technique [KKMS05] into the online mirror descent framework. Second, it is useful to extend the analysis to more general concept classes such as intersections of halfspaces [KOS02, DKS18]. It will also be important to design PAC algorithms that leverage additional types of queries such as pairwise comparison [KLMZ17, XZS+17] in the scenario where labels are extremely demanding (e.g. medical data), or to develop new projection-free algorithms for even faster computation.

References

[ABHU15] 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.

[ABHZ16] 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.

[ABL17] 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.

[BB08] 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.

[BBL09] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

[BBZ07] 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.

[BFN+17] 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.

[BH20] Maria-Florina Balcan and Nika Haghtalab. Noise in classification. In Tim Roughgarden, editor, *Beyond the Worst-Case Analysis of Algorithms*, pages 361–381. Cambridge University Press, 2020.

[BHL95] 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.

[BHLZ10] Alina Beygelzimer, Daniel J. Hsu, John Langford, and Tong Zhang. Agnostic active learning without constraints. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, pages 199–207, 2010.

[BHLZ16] 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.

[BL13] 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.

[Blu90] 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.

[CAL94] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[CBLW96] Nicolò Cesa-Bianchi, Philip M. Long, and Manfred K. Warmuth. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7:604–619, 1996.

[CCG04] Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

[CCG11] Giovanni Cavallanti, Nicolò Cesa-Bianchi, and Claudio Gentile. Learning noisy linear classifiers via adaptive and selective sampling. *Machine Learning*, 83(1):71–102, 2011.

[CDS98] 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.

[CT05] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[CW08] Emmanuel J. Candès and Michael B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[Dan15] Amit Daniely. A PTAS for agnostically learning halfspaces. In *Proceedings of the 28th Annual Conference on Learning Theory*, volume 40, pages 484–502, 2015.

[Dan16] Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pages 105–117, 2016.

[Das05] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, pages 235–242, 2005.

[Das11] Sanjoy Dasgupta. Active learning. *Encyclopedia of Machine Learning*, 2011.

[DGS12] 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.

[DKM05] 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.

[DKS18] 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.

[DKTZ20] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Non-convex SGD learns halfspaces with adversarial label noise. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020.

[DKZ20] Ilias Diakonikolas, Daniel Kane, and Nikos Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under gaussian marginals. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020.

[Don06] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[Fel07]   Vitaly Feldman. Attribute-efficient and non-adaptive learning of parities and DNF expressions. *Journal of Machine Learning Research*, 8:1431–1460, 2007.

[FF08]   Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of Statistics*, 36(6):2605–2637, 2008.

[FGKP06]   Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 563–574, 2006.

[FL01]   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.

[FSST97]   Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[Gen03]   Claudio Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

[GGK20]   Surbhi Goel, Aravind Gollakota, and Adam R. Klivans. Statistical-query lower bounds via functional gradients. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020.

[GLS01]   Adam J. Grove, Nick Littlestone, and Dale Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43(3):173–210, 2001.

[GR09]   Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[Han11]   Steve Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

[Han14]   Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2-3):131–309, 2014.

[Hau92]   David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[Haz19]   Elad Hazan. Introduction to online convex optimization. *CoRR*, abs/1909.05207, 2019.

[HKLM20]   Max Hopkins, Daniel Kane, Shachar Lovett, and Gaurav Mahajan. Noise-tolerant, reliable active classification with comparison queries. In *Proceedings of the 33rd Annual Conference on Learning Theory*, pages 1957–2006, 2020.

[KKMS05]   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.

[KLMZ17]   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.

[KLS09]   Adam R. Klivans, Philip M. Long, and Rocco A. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10:2715–2740, 2009.

[KOS02]   Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning intersections and thresholds of halfspaces. In *Proceedings fo the 43rd Annual IEEE Symposium on Foundations of Computer Science*, pages 177–186, 2002.

[KS04]   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.

[KSS92]   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.

[KT08]   Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, pages 801–808, 2008.

[Lit87]   Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. In *Proceedings of the 28th Annual IEEE Symposium on Foundations of Computer Science*, pages 68–77, 1987.

[LS06]   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.

[LV07]   László Lovász and Santosh S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3):307–358, 2007.

[LW89]   Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pages 256–261, 1989.

[MN06]   Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, pages 2326–2366, 2006.

[Ora19]   Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019.

[PV16]   Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, 2016.

[PVY17]   Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

[Ros58]   Frank Rosenblatt. The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958.

[Ser99]   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.

[Sha12]   Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

[SL17a]   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, 2017.

[SL17b]   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, 2017.

[SL18]   Jie Shen and Ping Li. A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18(208):1–42, 2018.

[SS07]   Shai Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. PhD thesis, Hebrew University, 2007.

[SZ20]   Jie Shen and Chicheng Zhang. Attribute-efficient learning of halfspaces with malicious noise: Near-optimal label complexity and noise tolerance. *CoRR*, abs/2006.03781, 2020.

[Tib96]   Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[TW10]   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.

[Val84]   Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vem10]   Santosh S. Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *Journal of the ACM*, 57(6):32:1–32:14, 2010.

[XZS$^+$17] 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.

[YZ17]   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.

[ZC14]   Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, pages 442–450, 2014.

[Zha10]   Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010.

[Zha18]   Chicheng Zhang. Efficient active learning of sparse halfspaces. In *Proceedings of the 31st Annual Conference On Learning Theory*, pages 1856–1880, 2018.

[Zin03]   Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

[ZS20]   Shiwei Zeng and Jie Shen. Learning halfspaces with pairwise comparisons: Breaking the barriers of query complexity via crowd wisdom. *CoRR*, abs/2011.01104, 2020.

[ZSA20]   Chicheng Zhang, Jie Shen, and Pranjal Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020.

[ZYJ14]   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.

A Proof Details

From now on, we always implicitly require that Assumption 1 and Assumption 2 hold if not specified. Since the INITIALIZE algorithm also depends on REFINE, we will start our analysis with the latter.

First of all, we recall a few important notations. Given $w \in \mathbb{R}^d$, the sampling region for the unlabeled example is a band, which is given by

$$X_{\hat{w},b} := \{x \in \mathbb{R}^d : 0 < \hat{w} \cdot x \le b\}, \text{ where } \hat{w} := \frac{w}{\|w\|}. \tag{A.1}$$

We denoted by $D_{X|\hat{w},b}$ the distribution of D_X conditioned on $x \in X_{\hat{w},b}$, and by $D_{\hat{w},b}$ the distribution of D conditioned on $x \in X_{\hat{w},b}$.

Recall that given $w \in \mathbb{R}^d$ and $x \sim D_{X|\hat{w},b}$, our prediction $\hat{y} = \text{sign}(w \cdot x) = 1$ and we set the gradient in REFINE as

$$g = x \cdot \mathbf{1}_{\{y=-1\}},$$

where y is the label returned by the adversary. Also recall that we defined the potential function

$$f_{u,b}(w) = \mathbb{E}_{x \sim D_{X|\hat{w},b}} \left[|u \cdot x| \cdot \mathbf{1}_{\{u \cdot x < 0\}} \right].$$

Finally, we note that the capital letters C and K, and their subscript variants such as C_1 and K_1, are used to denote absolute constants whose values may differ from appearance to appearance. However, we reserve c_0, c_1, and $c_2 > 0$ for specific absolute constants: c_0 is a sufficiently small constant such that the noise rate $\nu \le c_0 \epsilon$, c_1 and c_2 are specified in Lemma 28 and Lemma 29 respectively.

A.1 Analysis of REFINE

Intuitively, since we are performing gradient descent, we would hope that the negative gradient has a nontrivial correlation with the underlying halfspace u. The following lemma formalizes the intuition.

Lemma 13. *Given $w \in \mathbb{R}^d$ with $\|w - u\| \le r$ and $b > 0$, let $g = x \cdot \mathbf{1}_{\{y=-1\}}$ where $(x, y) \sim D_{\hat{w},b}$. If the noise rate $\nu \le c_0 b$ for some absolute constant $c_0 > 0$, then*

$$\mathbb{E}\left[u \cdot (-g) \right] \ge f_{u,b}(w) - \sqrt{\frac{c_0 c_1}{c_2}} \cdot (b + r),$$

where the expectation is taken over the random draw of (x, y).

Proof. By the definition of g, it follows that

$$\mathbb{E}\left[u \cdot (-g) \right] = \mathbb{E}\left[-(u \cdot x) \cdot \mathbf{1}_{\{y=-1\}} \right]. \tag{A.2}$$

As we can rewrite the indicator function $\mathbf{1}_{\{y=-1\}}$ in an equivalent form as follows:

$$\mathbf{1}_{\{y=-1\}} = \mathbf{1}_{\{u \cdot x < 0\}} + \mathbf{1}_{\{u \cdot x > 0, y=-1\}} - \mathbf{1}_{\{y=1, u \cdot x < 0\}}, \tag{A.3}$$

(A.2) can be written as

$$\mathbb{E}\left[u \cdot (-g) \right] = \underbrace{\mathbb{E}\left[-(u \cdot x) \cdot \mathbf{1}_{\{u \cdot x < 0\}} \right]}_{E_1} + \underbrace{\mathbb{E}\left[-(u \cdot x) \cdot \left(\mathbf{1}_{\{u \cdot x > 0, y=-1\}} - \mathbf{1}_{\{y=1, u \cdot x < 0\}} \right) \right]}_{E_2}.$$

First, we argue that $E_1 = f_{u,b}(w)$. In fact, when $u \cdot x \geq 0$, $E_1 = 0 = f_{u,b}(w)$; when $u \cdot x < 0$, $E_1 = \mathbb{E}[|u \cdot x|] = f_{u,b}(w)$.

Let us now consider the term E_2, whose absolute value can be bounded by

$$|E_2| \leq \mathbb{E}\left[|u \cdot x| \cdot \mathbf{1}_{\{\text{sign}(u \cdot x) \neq y\}}\right] \leq \sqrt{\mathbb{E}\left[(u \cdot x)^2\right] \cdot \mathbb{E}\left[\mathbf{1}_{\{\text{sign}(u \cdot x) \neq y\}}\right]}.$$

By Lemma 28, $\mathbb{E}\left[(u \cdot x)^2\right] \leq c_1(b^2 + r^2)$. On the other side, from the definition of ν-adversarial noise, we have

$$\mathbb{E}\left[\mathbf{1}_{\{\text{sign}(u \cdot x) \neq y\}}\right] = \Pr_{(x,y)\sim D_{\hat{w},b}}(y \neq \text{sign}(u \cdot x)) \leq \frac{\Pr_{(x,y)\sim D}(y \neq \text{sign}(u \cdot x))}{\Pr_{x \sim D_X}(x \in X_{\hat{w},b})} \leq \frac{\nu}{c_2 \cdot b}.$$

In the first inequality of the above expression, we use the fact that for an event A, $\Pr_{(x,y)\sim D_{\hat{w},b}}(A) = \Pr_{(x,y)\sim D}(A \mid x \in X_{\hat{w},b}) \leq \frac{\Pr_{(x,y)\sim D}(A)}{\Pr_{x\sim D_X}(x \in X_{\hat{w},b})}$. In the second inequality, we use Lemma 29 to bound the denominator from below.

Therefore,

$$|E_2| \leq \sqrt{c_1(b^2 + r^2) \cdot \frac{\nu}{c_2 \cdot b}} \leq \sqrt{\frac{c_0 c_1}{c_2}(b^2 + r^2)} \leq \sqrt{\frac{c_0 c_1}{c_2}} \cdot (b + r).$$

Now combining the above estimate and that of E_1, we prove the lemma. $\qquad\square$

Lemma 14. *There exists an absolute constant $C > 0$ such that the following holds. Suppose the algorithm* REFINE *is run with initialization w_0, step size $\alpha > 0$, bandwidth $b > 0$, convex constraint set $\mathcal{K}$, regularizer $\Phi(w) = \frac{1}{2(p-1)}\|w - w_0\|_p^2$, number of iterations T, where the following are satisfied:*

1. $\|w_0 - u\|_1 \leq \rho$;

2. $w_0 \in \mathcal{K}$ and $u \in \mathcal{K}$;

3. *for all $w \in \mathcal{K}$, $\|w - u\| \leq r$.*

Then, with probability $1 - \delta$,

$$C \cdot \frac{1}{T}\sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq (b + r)\left(\frac{\sqrt{\log(1/\delta)}}{\sqrt{T}} + \frac{\log(1/\delta)}{T} + C\sqrt{\frac{c_0 c_1}{c_2}}\right) + \frac{\rho^2 \log d}{\alpha T} + \alpha \cdot \log^2 \frac{Td}{b\delta}.$$

Proof. By standard analysis of online mirror descent (see, e.g. Theorem 6.8 of [Ora19]), we have

$$\alpha \sum_{t=1}^{T} w_{t-1} \cdot g_t - \alpha \sum_{t=1}^{T} u \cdot g_t \leq \mathcal{B}_\Phi(u, w_0) + \sum_{t=1}^{T}\|\alpha g_t\|_q^2.$$

Since $w_{t-1} \cdot g_t = w_{t-1} \cdot x_t \cdot \mathbf{1}_{\{y_t = -1\}}$ and x_t is such that $\hat{w}_{t-1} \cdot x_t > 0$, we have $w_{t-1} \cdot g_t \geq 0$ for all t. Using this observation and dividing both sides by α, we obtain

$$\sum_{t=1}^{T} u \cdot (-g_t) \leq \frac{\mathcal{B}_\Phi(u, w_0)}{\alpha} + \alpha \sum_{t=1}^{T}\|g_t\|_q^2. \tag{A.4}$$

We first present upper bounds for the right-hand side. In particular, note that

$$\mathcal{B}_\Phi(u, w_0) = \frac{1}{2(p-1)}\|u - w_0\|_p^2 \leq \frac{\ln(8d) - 1}{2}\rho^2 \leq \frac{\rho^2 \ln(8d)}{2}, \tag{A.5}$$

where in the first inequality we use the fact that for any $p > 1$, $\|u - w_0\|_p \leq \|u - w_0\|_1$.

For the ℓ_q-norm of g_t, denote $g_t^{(j)}$ the jth coordinate of g_t. We have

$$\|g_t\|_q = \left(\sum_{j=1}^{d} \left|g_t^{(j)}\right|^q\right)^{1/q} \leq (d \|g_t\|_\infty^q)^{1/q} \leq 2 \|g_t\|_\infty \leq 2 \|x_t\|_\infty, \tag{A.6}$$

where the first inequality makes use of the definition of ℓ_∞-norm, the second inequality applies the setting $q = \ln(8d)$, and the last step follows from the setting of g_t. On the other side, it is known that for any $x_t \sim D_{X|\hat{w}_{t-1},b}$, with probability $1 - \frac{\delta}{2T}$, we have $\|x_t\|_\infty \leq K_1 \cdot \log \frac{Td}{b\delta}$ for some absolute constant $K_1 > 0$; see Lemma 30. Hence, the union bound implies that with probability $1 - \frac{\delta}{2}$, $\max_{1 \leq t \leq T} \|x_t\|_\infty \leq K \cdot \log \frac{Td}{b\delta}$, which further gives

$$\max_{1 \leq t \leq T} \|g_t\|_q \leq 2K_1 \cdot \log \frac{Td}{b\delta}. \tag{A.7}$$

Now we consider a lower bound of $\sum_{t=1}^{T} u \cdot (-g_t)$. Define the filtration $\mathcal{F}_t := \sigma(w_0, x_1, y_1, w_1, \ldots, x_{t-1}, y_{t-1}, w_t)$, and denote by $\mathbb{E}_{t-1}[\cdot]$ the expectation over $(x_t, y_t) \sim D_{\hat{w}_{t-1},b}$ conditioning on the past filtration $\mathcal{F}_{t-1}$; likewise for $\text{Pr}_{t-1}(\cdot)$.

By existing tail bound of one-dimensional isotropic log-concave distributions in the band $X_{\hat{w}_{t-1},b}$ (see, e.g. Lemma 3.3 of [ABL17]), and the fact that $\|u - \hat{w}_{t-1}\| \leq 2 \|u - w_{t-1}\| \leq 2r$, we have

$$\text{Pr}_{t-1}\left(|u \cdot x_t| \geq a\right) \leq K \exp\left(-K' \cdot \frac{a}{2r + b}\right),$$

for some constants $K, K' > 0$, implying that

$$\text{Pr}_{t-1}\left(|u \cdot (-g_t)| \geq a\right) \leq K \exp\left(-K' \cdot \frac{a}{2r + b}\right).$$

Now applying Lemma 31 with $Z_t = u \cdot (-g_t)$ therein gives that with probability $1 - \frac{\delta}{2}$,

$$\left|\sum_{t=1}^{T} u \cdot (-g_t) - \mathbb{E}_{t-1}[u \cdot (-g_t)]\right| \leq K_2(b + r)\left(\sqrt{T \log \frac{1}{\delta}} + \log \frac{1}{\delta}\right). \tag{A.8}$$

The above concentration of martingales, in allusion to Lemma 13, gives

$$\sum_{t=1}^{T} u \cdot (-g_t) \geq \sum_{t=1}^{T} f_{u,b}(w_{t-1}) - T \cdot \sqrt{\frac{c_0 c_1}{c_2}} \cdot (b + r) - K_2(b + r)\left(\sqrt{T \log \frac{1}{\delta}} + \log \frac{1}{\delta}\right). \tag{A.9}$$

Combining (A.4), (A.5), (A.7), and (A.9), we obtain

$$C \cdot \frac{1}{T} \sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq (b + r)\left(\frac{\sqrt{\log(1/\delta')}}{\sqrt{T}} + \frac{\log(1/\delta')}{T} + C\sqrt{\frac{c_0 c_1}{c_2}}\right)$$
$$+ \frac{\rho^2 \log d}{\alpha T} + \alpha \cdot \log^2 \frac{Td}{b\delta}$$

for some absolute constant $C > 0$. $\qquad\square$

The following proposition is an immediate result of Lemma 14 by specifying the involved hyperparameters and showing that u stays in the convex constraint set $\mathcal{K}$.

Proposition 15. *Suppose that the adversarial noise rate $\nu \leq c_0\epsilon$ for some sufficiently small absolute constant $c_0 > 0$. Consider running the REFINE algorithm with step size $\alpha = \tilde{\Theta}\left(\theta \cdot \log^{-2} \frac{d}{\delta\theta}\right)$, bandwidth $b = \Theta(\theta)$, convex constraint set $\mathcal{K} = \{w \in \mathbb{R}^d : \|w - w_0\| \leq \theta, \|w\| \leq 1\}$, regularizer $\Phi(w) = \frac{1}{2(p-1)}\|w - w_0\|_p^2$, number of iterations $T = \tilde{O}(s \log d \cdot \log^2 \frac{d}{\delta\theta})$. If the initial iterate w_0 is such that $\|w_0\| = 1$, $\|w_0\|_0 \leq s$, and $\|w_0 - u\| \leq \theta$ for some $\theta \leq \frac{\pi}{16}$, then with probability $1 - \delta$,*

$$\frac{1}{T}\sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq \frac{\theta}{50 \cdot 3^4 \cdot 2^{33}}.$$

Proof. We will first verify that the premises of Lemma 14 are satisfied. First of all, it is easy to see that $\|w_0 - u\|_1 \leq \sqrt{\|w_0 - u\|_0} \cdot \|w_0 - u\| \leq \sqrt{2s} \cdot \theta$. Hence we can choose $\rho = \sqrt{2s} \cdot \theta$ in Lemma 14. Next, we have $\|u - w_0\| \leq \theta$ in view of our condition on w_0, which together with the fact that $\|u\| = 1$ implies $u \in \mathcal{K}$. Last, for all $w \in \mathcal{K}$, by triangle inequality $\|w - u\| \leq \|w - w_0\| + \|w_0 - u\| \leq 2\theta$. Hence we can choose $r = 2\theta$ in Lemma 14.

Now with $\rho = \sqrt{2s} \cdot \theta$ and $r = 2\theta$, Lemma 14 indicates that with probability $1 - \delta$,

$$C \cdot \frac{1}{T}\sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq (b + 2\theta)\left(\frac{\sqrt{\log(1/\delta)}}{\sqrt{T}} + \frac{\log(1/\delta)}{T} + C\sqrt{\frac{c_0 c_1}{c_2}}\right) + \frac{\theta^2 \cdot 2s \log d}{\alpha T} + \alpha \cdot \log^2 \frac{Td}{b\delta}.$$

We need each term on the right-hand side is upper bounded by $\frac{C \cdot \theta}{3 \cdot 50 \cdot 3^4 \cdot 2^{33}}$. First, we choose $b = \Theta(\theta)$. Then for the first term, it suffices to choose $T \geq \log(1/\delta)$ and set c_0 to be a sufficiently small absolute constant (this is possible since C, c_1, and c_2 are all fixed absolute constants). The second and the last terms require $\alpha T \geq \Omega(\theta \cdot s \log d)$ and $\alpha \cdot \log^2 \frac{Td}{b\delta} = O(\theta)$ respectively. The latter implies $\alpha = O(\theta \cdot \log^{-2} \frac{Td}{b\delta})$, thus combining it with the former we need

$$\frac{\theta \cdot s \log d}{T} \leq \theta \cdot \log^{-2} \frac{Td}{b\delta}.$$

This can be satisfied if we choose $T = \tilde{O}(s \log d \cdot \log^2 \frac{d}{\delta\theta})$. Finally, we have $\alpha = \tilde{\Theta}(\theta \cdot \log^{-2} \frac{d}{\delta\theta})$. $\square$

Remark 16. In the above proof, we note that $C\sqrt{\frac{c_0 c_1}{c_2}}$ is an extra term introduced by the adversarial noise model. It is important to observe that c_0 is chosen as a very small *absolute constant*. Thus the noise tolerance still reads as $\nu = \Omega(\epsilon)$. The term $C\sqrt{\frac{c_0 c_1}{c_2}}$ does not appear in the bounded noise analysis though; see Lemma 8 of [ZSA20].

Lemma 17. *Let $\theta \in [0, \frac{\pi}{16}]$ be a given scalar. Let $w_0, \ldots w_{T-1}$ be a sequence of vectors such that for all $1 \leq t \leq T$, $\|w_{t-1} - u\| \leq 2\theta$ and $\|w_{t-1}\| \leq 1$. Further assume that $\frac{1}{T}\sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq \frac{\theta}{50 \cdot 3^4 \cdot 2^{33}}$. Let $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_{t-1}$ and $v = \frac{\mathcal{H}_s(\bar{w})}{\|\mathcal{H}_s(\bar{w})\|}$. Then $\theta(v, u) \leq \frac{\theta}{2}$.*

Proof. Define the index set $S = \left\{t \in [T] : f_{u,b}(w_{t-1}) \geq \frac{\theta}{5 \cdot 3^4 \cdot 2^{21}}\right\}$. It is easy to show that $\frac{|S|}{T} \leq \frac{1}{10 \cdot 2^{12}}$ as otherwise the average of $f_{u,b}(w_{t-1})$ will exceed the assumed upper bound. Therefore, $\frac{|\bar{S}|}{T} \geq 1 - \frac{1}{10 \cdot 2^{12}}$. For all $t \in \bar{S}$ we have $f_{u,b}(w_{t-1}) \leq \frac{\theta}{50 \cdot 3^4 \cdot 2^{21}}$; by Lemma 26, we have $\theta(w_{t-1}, u) \leq \frac{\theta}{5}$ for these t.

Now consider $\theta(w_{t-1}, u)$ for $t \in S$. As we showed in the proof of Proposition 15, we have $\|w_{t-1} - u\| \leq 2\theta$. Since $\|u\| = 1$ and $\|w_{t-1}\| \leq 1$, we use the basic fact that $\theta(w_{t-1}, u) \leq \pi \|w_{t-1} - u\| < 8\theta$.

Now we translate these bounds on the angles to those of the cosine distance, and obtain

$$\frac{1}{T}\sum_{t=1}^{T}\cos\theta(w_{t-1},u) \geq \cos\frac{\theta}{5}\cdot\left(1-\frac{1}{20\cdot 2^{12}}\right)+\cos(8\theta)\cdot\frac{1}{20\cdot 2^{12}}$$

$$\geq\left(1-\frac{\theta^2}{50}\right)\left(1-\frac{1}{20\cdot 2^{12}}\right)+\left(1-\frac{(8\theta)^2}{2}\right)\frac{1}{20\cdot 2^{12}}$$

$$\geq 1-\frac{1}{5}\left(\frac{\theta}{32}\right)^2\geq\cos\frac{\theta}{32}.$$

where in the second inequality we use the fact $\cos\theta\geq 1-\frac{\theta^2}{2}$ for any $\theta\in[0,\pi]$, and in the last inequality we use the fact that $\cos\theta\leq 1-\frac{\theta^2}{5}$.

The above inequality, in combination with Lemma 32 yields the following guarantee for $\bar{w}=\frac{1}{T}\sum_{t=1}^{T}w_{t-1}$:

$$\cos\theta(\bar{w},u)\geq\frac{1}{T}\sum_{t=1}^{T}\cos\theta(w_{t-1},u)\geq\cos\frac{\theta}{32}.$$

Finally, we use Lemma 33 to show that $\theta(v,u)\leq\pi\|v-u\|\leq 4\pi\|\bar{w}-u\|\leq 16\cdot\theta(\bar{w},u)\leq\frac{\theta}{2}$, which concludes the proof. $\qquad\square$

Theorem 18 (Restatement of Theorem 12). *Suppose that the adversarial noise rate $\nu\leq c_0\epsilon$ for some sufficiently small absolute constant $c_0>0$. Consider running the* REFINE *algorithm with step size $\alpha=\tilde{\Theta}\left(\theta\cdot\log^{-2}\frac{d}{\delta'\theta}\right)$, bandwidth $b=\Theta(\theta)$, convex constraint set $\mathcal{K}=\{w\in\mathbb{R}^d:\|w-w_0\|\leq\theta,\|w\|\leq 1\}$, regularizer $\Phi(w)=\frac{1}{2(p-1)}\|w-w_0\|_p^2$, number of iterations $T=\tilde{O}(s\log d\cdot\log^2\frac{d}{\delta'\theta})$. If the initial iterate w_0 is such that $\|w_0\|=1$, $\|w_0\|_0\leq s$, and $\|w_0-u\|\leq\theta$ for some $\theta\leq\frac{\pi}{16}$, then the output of the* REFINE *algorithm, $\tilde{w}$, satisfies $\theta(\tilde{w},u)\leq\frac{\theta}{2}$ with probability $1-\delta'$. In addition, the label complexity of* REFINE *is T, and the sample complexity is $O(T/b+T\log\frac{T}{\delta})$.*

Proof. The result of $\theta(\tilde{w},u)\leq\frac{\theta}{2}$ follows from combining Proposition 15 and Lemma 17. In particular, all the required conditions in Proposition 15 are assumed here. The condition $\|w_{t-1}-u\|$ appearing in Lemma 17 can easily be verified by observing $\|w_{t-1}-w_0\|\leq\theta$ and $\|w_0-u\|\leq\theta$.

The label complexity bound is exactly T since REFINE runs in T iterations and requests one label per iteration. Since the marginal distribution is assumed to be isotropic log-concave, Lemma 29 shows that $\Pr_{x_t\sim D_X}(x_t\in X_{\hat{w}_{t-1},b})\geq c_2 b$ for some absolute constant $c_2>0$. Thus, by Chernoff bound, we need to call EX for $O(b^{-1}+\log\frac{T}{\delta})$ times in order to obtain one x_t with probability $1-\frac{\delta}{2T}$. Thus, by union bound over the T iterations in REFINE, with probability $1-\frac{\delta}{2}$, the total number of calls to EX is $O(T/b+T\log\frac{T}{\delta})$. $\qquad\square$

A.2 Analysis of INITIALIZE

In this subsection, we use $\mathbb{E}[\cdot]$ denote the expectation $\mathbb{E}_{(x,y)\sim D}[\cdot]$ and likewise for $\Pr(\cdot)$.

Lemma 19. *Suppose that $\nu\leq\frac{1}{4}$. Then $\mathbb{E}[y(u\cdot x)]\geq\frac{1}{9\cdot 2^{17}}$.*

Proof. We have

$$\mathbb{E}[y(u\cdot x)]=\mathbb{E}[y(u\cdot x)\mid y=\operatorname{sign}(u\cdot x)]\cdot\Pr(y=\operatorname{sign}(u\cdot x))$$
$$+\mathbb{E}[y(u\cdot x)\mid y\neq\operatorname{sign}(u\cdot x)]\cdot\Pr(y\neq\operatorname{sign}(u\cdot x))$$
$$\geq\mathbb{E}[|u\cdot x|]\cdot(1-\nu)-\mathbb{E}[|u\cdot x|]\cdot\nu$$
$$=(1-2\nu)\mathbb{E}[|u\cdot x|].$$

Since $u \cdot x$ is an isotropic log-concave random variable in $\mathbb{R}$, its density function is lower bounded by 2^{-16} when $|u \cdot x| \leq \frac{1}{9}$ in view of Lemma 29. Thus $\mathbb{E}[|u \cdot x|] \geq \frac{1}{9 \cdot 2^{16}}$. On the other side, we assumed $\nu \leq \frac{1}{4}$. Together, we obtain $\mathbb{E}[y(u \cdot x)] \geq \frac{1}{9 \cdot 2^{17}}$. $\qquad \square$

Lemma 20. *Let $m = O(\log \frac{1}{\delta})$ and let $(x_1, y_1), \ldots, (x_m, y_m)$ be m i.i.d. samples drawn from D. Then with probability $1 - \delta$,*

$$w_{\text{avg}} \cdot u \geq \frac{1}{9 \cdot 2^{18}},$$

where $w_{\text{avg}} := \frac{1}{m} \sum_{i=1}^{m} y_i x_i$.

Proof. First, Lemma 29 shows that $u \cdot x$ is isotropic log-concave, and hence $y(u \cdot x)$ is a $(32, 16)$-subexponential random variable by Lemma 34 of [ZSA20]. Therefore, the standard concentration bound implies that there is an absolute constant $K_1 > 0$, such that if $m = O(\log \frac{1}{\delta})$, with probability $1 - \frac{\delta}{2}$,

$$\left| \frac{1}{m} \sum_{i=1}^{m} y_i (u \cdot x_i) - \mathbb{E}[y(u \cdot x)] \right| \leq K_1 \left(\sqrt{\frac{\log(1/\delta)}{m}} + \frac{\log(1/\delta)}{m} \right) \leq \frac{1}{9 \cdot 2^{18}}.$$

This in allusion to Lemma 19 gives $\frac{1}{m} \sum_{i=1}^{m} y_i (u \cdot x_i) \geq \frac{1}{9 \cdot 2^{18}}$, namely

$$w_{\text{avg}} \cdot u \geq \frac{1}{9 \cdot 2^{18}},$$

which is the desired lower bound. $\qquad \square$

Lemma 21. *Let $\tilde{s} \leq d$ be a positive integer, and set $m = O(\tilde{s} \log \frac{d}{\delta})$. Let $(x_1, y_1), \ldots, (x_m, y_m)$ be m i.i.d. samples drawn from D. Then with probability $1 - \delta$,*

$$w^{\sharp} \cdot u \geq \frac{1}{9 \cdot 2^{20}}.$$

Proof. Denote $w' = \mathcal{H}_{\tilde{s}}(w_{\text{avg}})$. Using Lemma 17 of [ZSA20] we know that with probability $1 - \frac{\delta}{2}$, $\|w'\| \leq 2$. From the choice of m and Lemma 20, we have $w_{\text{avg}} \cdot u \geq \frac{1}{9 \cdot 2^{18}}$ with probability $1 - \frac{\delta}{2}$. We hence condition on both events happening.

Now Lemma 16 of [ZSA20] implies that

$$\left| w' \cdot u - w_{\text{avg}} \cdot u \right| \leq \sqrt{\frac{s}{\tilde{s}}} \|w'\| \leq 2 \sqrt{\frac{s}{\tilde{s}}}.$$

Therefore,

$$w' \cdot u \geq w_{\text{avg}} \cdot u - 2 \sqrt{\frac{s}{\tilde{s}}} \geq \frac{1}{9 \cdot 2^{18}} - 2 \sqrt{\frac{s}{\tilde{s}}}.$$

Now taking $\tilde{s} = 81 \cdot 2^{40} s$ gives us $w' \cdot u \geq \frac{1}{9 \cdot 2^{19}}$. Finally, by algebra

$$w^{\sharp} \cdot u = \frac{1}{\|w'\|} (w' \cdot u) \geq \frac{1}{2} \cdot \frac{1}{9 \cdot 2^{19}} = \frac{1}{9 \cdot 2^{20}}.$$

The proof is complete. $\qquad \square$

Proposition 22. *Suppose that the adversarial noise rate $\nu \leq c_0 \epsilon$ for some sufficiently small absolute constant $c_0 > 0$. Let $\zeta = \frac{1}{9 \cdot 2^{20}}$. Consider running the* INITIALIZE *algorithm with step size $\alpha = \tilde{\Theta}\left(\log^{-2} \frac{d}{\delta}\right)$, bandwidth $b = \Theta(1)$, convex constraint set $\mathcal{K} = \{w \in \mathbb{R}^d : \|w\| \leq 1, \ w^\sharp \cdot w \geq \zeta\}$, regularizer $\Phi(w) = \frac{1}{2(p-1)}\|w - w_0\|_p^2$, number of iterations $T = \tilde{O}(s \log d \cdot \log^2 \frac{d}{\delta})$, where w_0 is an arbitrary point in $\mathcal{K} \cap \{w \in \mathbb{R}^d : \|w\|_1 \leq \sqrt{s}\}$ that can be found in polynomial time. Then with probability $1 - \delta$,*

$$\frac{1}{T} \sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq \frac{\zeta}{20 \cdot 3^4 \cdot 2^{33}}.$$

Proof. We will first verify that the premises of Lemma 14 are satisfied. First of all, it is easy to see that $\|w_0 - u\|_1 \leq 2\sqrt{s}$. Hence we can choose $\rho = 2\sqrt{s}$ in Lemma 14. Next, we have $\|u\| = 1$, which together with Lemma 21 implies $u \in \mathcal{K}$ with probability $1 - \frac{\delta}{2}$. Last, for all $w \in \mathcal{K}$, by triangle inequality $\|w - u\| \leq \|w\| + \|u\| \leq 2$. Hence we can choose $r = 2$ in Lemma 14.

Now with $\rho = \sqrt{s}$ and $r = 2$, Lemma 14 indicates that with probability $1 - \frac{\delta}{2}$,

$$C \cdot \frac{1}{T} \sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq (b+2) \left(\frac{\sqrt{\log(1/\delta)}}{\sqrt{T}} + \frac{\log(1/\delta)}{T} + C\sqrt{\frac{c_0 c_1}{c_2}} \right) + \frac{s \log d}{\alpha T} + \alpha \cdot \log^2 \frac{Td}{b\delta}.$$

We need each term on the right-hand side is upper bounded by $\frac{C \cdot \zeta}{20 \cdot 3^5 \cdot 2^{33}}$. First, we choose $b = \Theta(1)$. Then for the first term, it suffices to choose $T \geq \log(1/\delta)$ and set c_0 to be a sufficiently small absolute constant (this is possible since C, c_1, and c_2 are all fixed absolute constants). The second and the last terms require $\alpha T \geq \Omega(s \log d)$ and $\alpha \cdot \log^2 \frac{Td}{b\delta} = O(1)$ respectively. The latter implies $\alpha = O(\log^{-2} \frac{Td}{b\delta})$, thus combining it with the former we need

$$\frac{s \log d}{T} \leq \cdot \log^{-2} \frac{Td}{b\delta}.$$

This can be satisfied if we choose $T = \tilde{O}(s \log d \cdot \log^2 \frac{d}{\delta})$, which results in $\alpha = \tilde{\Theta}(\log^{-2} \frac{d}{\delta})$. $\square$

Lemma 23. *Set $b = \frac{1}{81 \cdot 2^{22}}$ and let $\zeta = \frac{1}{9 \cdot 2^{20}}$. Let $w_0, \ldots w_{T-1}$ be a sequence of vectors such that for all $1 \leq t \leq T$, $\|w_{t-1}\| \leq 1$. Further assume that $\frac{1}{T} \sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq \frac{\zeta}{20 \cdot 3^4 \cdot 2^{33}}$. Let $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w_{t-1}$ and $v = \frac{\mathcal{H}_s(\bar{w})}{\|\mathcal{H}_s(\bar{w})\|}$. Then $\theta(v, u) \leq \frac{\pi}{8}$.*

Proof. In light of Lemma 21, we know that $u \in \mathcal{K}$. Also, our choices of b and ζ implies that $b \leq \frac{\zeta}{36}$. Define $S = \{1 \leq t \leq T : f_{u,b}(w_{t-1}) \geq \frac{\zeta}{3^4 \cdot 2^{21}}\}$. Then by the second part of Lemma 26, for all $t \in \bar{S}$, we have $\theta(w_{t-1}, u) < \zeta$. For all $t \in S$, we have a trivial estimate of $\theta(w_{t-1}, u) \in [0, \pi]$.

Next we bound the size of S. Using the condition $\frac{1}{T} \sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq \frac{\zeta}{20 \cdot 3^4 \cdot 2^{33}}$, it is possible to show that $\frac{|S|}{T} \leq \frac{1}{20 \cdot 2^{12}}$ and thus $\frac{|\bar{S}|}{T} \geq 1 - \frac{1}{20 \cdot 2^{12}}$.

Now we translate these bounds on the angles to those of the cosine distance, and obtain

$$\frac{1}{T} \sum_{t=1}^{T} \cos \theta(w_{t-1}, u) \geq \cos \zeta \cdot \left(1 - \frac{1}{20 \cdot 2^{12}} \right) + (\cos \pi) \cdot \frac{1}{20 \cdot 2^{12}}$$

$$\geq 1 - \frac{1}{5} \left(\frac{\pi}{128} \right)^2 \geq \cos \frac{\pi}{128}.$$

where in the last inequality we use the fact that $\cos \theta \leq 1 - \frac{\theta^2}{5}$.

The above inequality, in combination with Lemma 32 yields the following guarantee for $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_{t-1}$:

$$\cos\theta(\bar{w}, u) \geq \frac{1}{T}\sum_{t=1}^{T}\cos\theta(w_{t-1}, u) \geq \cos\frac{\pi}{128}.$$

Finally, we use Lemma 33 to show that $\theta(v, u) \leq \pi\|v - u\| \leq 4\pi\|\bar{w} - u\| \leq 16 \cdot \theta(\bar{w}, u) \leq \frac{\pi}{8}$, which concludes the proof. □

Theorem 24 (Restatement of Theorem 11). *Suppose that the adversarial noise rate $\nu \leq c_0\epsilon$ for some sufficiently small absolute constant $c_0 > 0$. Let $\zeta = \frac{1}{9\cdot 2^{20}}$. Consider running the INITIALIZE algorithm with step size $\alpha = \tilde{\Theta}\left(\log^{-2}\frac{d}{\delta}\right)$, bandwidth $b = \Theta(1)$, convex constraint set $\mathcal{K} = \{w \in \mathbb{R}^d : \|w\| \leq 1, \ w^{\sharp} \cdot w \geq \zeta\}$, regularizer $\Phi(w) = \frac{1}{2(p-1)}\|w - w_0\|_p^2$, number of iterations $T = \tilde{O}(s\log d \cdot \log^2\frac{d}{\delta})$, where w_0 is an arbitrary point in $\mathcal{K}\cap\{w \in \mathbb{R}^d : \|w\|_1 \leq \sqrt{s}\}$ that can be found in polynomial time. Then with probability $1 - \delta$, the output of INITIALIZE, v_0, is such that $\theta(v_0, u) \leq \frac{\pi}{8}$.*

Proof. This is an immediate result by combining Proposition 22 and Lemma 23. □

B   The Structure of $f_{u,b}(w)$

Our definition of the potential function $f_{u,b}(w)$ slightly differs from that of [ZSA20]: in this work, the expectation is taken over D conditioned on $\{x \in \mathbb{R}^d : 0 < w \cdot x \leq b\}$ while in [ZSA20] it is conditioned on $\{x \in \mathbb{R}^d : -b \leq w \cdot x \leq b\}$. It can be seen that our function value is always less than that of [ZSA20]. However, we note that the difference in sampling region does not lead to significant difference in the structure of the potential function. In particular, we are still able to show that under certain conditions, $f_{u,b}(w)$ serves as an upper bound of $\theta(w, u)$ – a crucial observation made in [ZSA20].

Lemma 25. *Let w and u be two unit vectors. Suppose $b \in \left[0, \frac{\pi}{72}\right]$. We have*

1. *If $\theta(w, u) \in [36b, \frac{\pi}{2}]$, then $f_{u,b}(w) \geq \frac{\theta(w,u)}{3^4 \cdot 2^{21}}$.*

2. *If $\theta(w, u) \in [\frac{\pi}{2}, \pi - 36b]$, then $f_{u,b}(w) \geq \frac{\pi - \theta(w,u)}{3^4 \cdot 2^{21}}$.*

Proof. The proof of the first part follows closely from [Lemma 22, Part 1] of [ZSA20]. In particular, for the region

$$R_1 := \left\{x \in \mathbb{R}^d : w \cdot x \in [0, b], \ u \cdot x \in \left[-\frac{\sin\theta(w, u)}{18}, -\frac{\sin\theta(w, u)}{36}\right]\right\},$$

their analysis shows that

$$\Pr_{x\sim D_X}(x \in R_1) \geq \frac{b}{9 \cdot 2^{18}}. \tag{B.1}$$

To see why it completes the proof of the first part, observe that

$$\begin{aligned}
\mathbb{E}_{x\sim D_X}\left[|u \cdot x| \cdot \mathbf{1}_{\{0\leq w\cdot x\leq b\}} \cdot \mathbf{1}_{\{u\cdot x<0\}}\right] &\geq \mathbb{E}_{x\sim D_X}\left[|u \cdot x| \cdot \mathbf{1}_{\{x\in R_1\}}\right] \\
&\geq \frac{\sin\theta(w, u)}{36} \cdot \mathbb{E}_{x\sim D_X}\left[\mathbf{1}_{\{x\in R_1\}}\right] \\
&\geq \frac{\theta(w, u)}{72} \cdot \Pr_{x\sim D_X}(x \in R_1) \geq \frac{\theta(w, u) \cdot b}{3^4 \cdot 2^{21}},
\end{aligned}$$

where the first inequality uses the fact that R_1 is a subset of both sets $\left\{x \in \mathbb{R}^d : w \cdot x \in [0, b]\right\}$ and $\left\{x \in \mathbb{R}^d : u \cdot x < 0\right\}$; the second inequality uses the fact that for all x in R_1, $|u \cdot x| \geq \frac{\sin \theta(w,u)}{36}$; the third inequality uses the elementary fact that $\sin \phi \geq \frac{\phi}{2}$ for any angle $\phi \in [0, \frac{\pi}{2}]$.

As $\Pr_{x \sim D_X}\left(w \cdot x \in [0, b]\right) \leq b$ by Lemma 29, we have

$$f_{u,b}(w) = \frac{\mathbb{E}_{x \sim D_X}\left[|u \cdot x| \cdot \mathbf{1}_{\{0 \leq w \cdot x \leq b\}} \cdot \mathbf{1}_{\{u \cdot x < 0\}}\right]}{\Pr_{x \sim D_X}\left(w \cdot x \in [0, b]\right)} \geq \frac{\theta(w, u) \cdot b}{3^4 \cdot 2^{21}} \cdot \frac{1}{b} = \frac{\theta(w, u)}{3^4 \cdot 2^{21}}.$$

This completes the proof of the first part.

For the second part, we will define $\phi = \pi - \theta(w, u)$ and consider

$$R_2 := \left\{x \in \mathbb{R}^d : w \cdot x \in [0, b], \ u \cdot x \in \left[-\frac{\sin \phi}{18}, -\frac{\sin \phi}{36}\right]\right\}.$$

Similar to the region R_1, we have $\Pr_{x \sim D_X}(x \in R_2) \geq \frac{b}{9 \cdot 2^{18}}$. Hence, using the same induction with the proof of first part, we have $f_{u,b}(w) \geq \frac{\phi}{3^4 \cdot 2^{21}} = \frac{\pi - \theta(w,u)}{3^4 \cdot 2^{21}}$. $\qquad \square$

The following lemma connects the potential function $f_{u,b}(w)$ to the angle $\theta(w, u)$.

Lemma 26. *Let w and u be two unit vectors. We have the following:*

1. *Suppose $\theta \in [0, \frac{\pi}{2}]$ is given. Set $b \leq \frac{\theta}{5 \cdot 36}$. If $f_{u,b}(w) \leq \frac{\theta}{5 \cdot 3^4 \cdot 2^{21}}$, then $\theta(w, u) \leq \frac{\theta}{5}$.*

2. *Let $w^\sharp \in \mathbb{R}^d$ be a unit vector with $w^\sharp \cdot u \geq \zeta$ for some $\zeta \in (0, 1)$. Set $b \leq \frac{\zeta}{36}$. If w is in $\mathcal{K} := \{w \in \mathbb{R}^d : \|w\| \leq 1, \ w \cdot w^\sharp \geq \zeta\}$ and $f_{u,b}(w) < \frac{\zeta}{3^4 \cdot 2^{21}}$, then $\theta(w, u) < \zeta$.*

Proof. The first part was already set out in Claim 10 of [ZSA20]. For the second part, first, we show that it is impossible for $\theta(w, u) \geq \frac{\pi}{2}$. Assume for contradiction that this holds. By Lemma 27, for all w in $\mathcal{K}$, we have that $\theta(w, u) \leq \pi - \zeta$. By the choice of b, we know that $36b \leq \zeta$, hence $\theta(w, u) \leq \pi - 36b$. Now using the second part of Lemma 25, we have

$$f_{u,b}(w) \geq \frac{\pi - \theta(w, u)}{3^4 \cdot 2^{21}} \geq \frac{\zeta}{3^4 \cdot 2^{21}}, \tag{B.2}$$

which contradicts with the premise that $f_{u,b}(w) < \frac{\zeta}{3^4 \cdot 2^{21}}$.

Therefore, $\theta(w, u) \in [0, \frac{\pi}{2}]$. We now conduct a case analysis. If $\theta(w, u) \leq 36b$, then by the definition of b, we automatically have $\theta(w, u) < \zeta$. Otherwise, $\theta(w, u) \in [36b, \frac{\pi}{2}]$. In this case, the first part of Lemma 25 implies

$$f_{u,b}(w) \geq \frac{\theta(w, u)}{3^4 \cdot 2^{21}}.$$

This inequality, in conjunction with the condition that $f_{u,b}(w) < \frac{\zeta}{3^4 \cdot 2^{21}}$, implies that $\theta(w, u) \leq \zeta$. In summary, in both cases, we have $\theta(w, u) \leq \zeta$. This completes the proof. $\qquad \square$

Lemma 27 (Lemma 19 of [ZSA20]). *Let $w^\sharp \in \mathbb{R}^d$ be a unit vector, and $\zeta \in (0, 1)$. For any two vectors w and v in the set $\mathcal{K} = \{w \in \mathbb{R}^d : \|w\| \leq 1, \ w \cdot w^\sharp \geq \zeta\}$, it holds that $\theta(w, v) \leq \pi - \zeta$.*

C   Useful Lemmas

We collect a few useful results that are frequently invoked in our analysis.

Lemma 28 (Lemma 3.4 of [ABL17]). *There is an absolute constant $c_1 > 0$ such that the following holds. Let D_X be an isotropic log-concave distribution. Fix $w \in \mathbb{R}^d$. For all v with $\|v - w\| \leq r$, $\mathbb{E}_{x \sim D_{X|\hat{w},b}}[(v \cdot x)^2] \leq c_1(b^2 + r^2)$.*

Lemma 29 ([LV07]). *There exists an absolute constants $c_2, c_3 > 0$ such that the following holds. Let D_X be an isotropic log-concave distribution over $\mathbb{R}^d$.*

1. *Given any unit vector w, $w \cdot x$ is isotropic log-concave if x is drawn from D_X.*

2. *Given any unit vector $w \in \mathbb{R}^d$, $c_2 b \leq \Pr_{x \sim D_X}(w \cdot x \in [0, b]) \leq b$.*

3. *If $d = 1$ or $d = 2$, for all $x \in \mathbb{R}^d$ with $\|x\| \leq \frac{1}{9}$, the density function $p(x) \geq 2^{-16}$.*

Lemma 30 (Lemma 16 of [SZ20]). *There exists an absolute constant $c_3 > 0$ such that the following holds for all isotropic log-concave distributions D_X. Let S be a set of i.i.d. instances drawn from $D_{X|\hat{w},b}$. Then*

$$\Pr_{S \sim D^n_{X|\hat{w},b}} \left(\max_{x \in S} \|x\|_\infty \geq c_3 \log \frac{|S| d}{b\delta} \right) \leq \delta.$$

Lemma 31 (Lemma 36 of [ZSA20]). *Suppose $\{Z_t\}_{t=1}^T$ is sequence of random variables adapted to filtration $\{\mathcal{F}_t\}_{t=1}^T$. Denote by $\Pr_{t-1}(\cdot)$ and $\mathbb{E}_{t-1}[\cdot]$ the probability and expectation conditioned on $\mathcal{F}_{t-1}$, respectively. For every Z_t, suppose that $\Pr_{t-1}(|Z_t| > a) \leq C \exp\left(-\frac{a}{\sigma}\right)$ for some absolute constant $C \geq 1$. Then, with probability $1 - \delta$,*

$$\left| \sum_{t=1}^T Z_t - \mathbb{E}_{t-1}[Z_t] \right| \leq 16\sigma(\ln C + 1) \left(\sqrt{2T \ln \frac{2}{\delta}} + \ln \frac{2}{\delta} \right).$$

Lemma 32 (Lemma 24 of [ZSA20]). *Suppose we have a sequence of unit vectors $w_0, \ldots, w_{T-1}$. Let $\bar{w} = \frac{1}{T} \sum_{t=1}^T w_{t-1}$ be their average. Suppose $\frac{1}{T} \sum_{t=1}^T \cos \theta(w_{t-1}, u) \geq 0$. Then, $\cos \theta(\bar{w}, u) \geq \frac{1}{T} \sum_{t=1}^T \cos \theta(w_{t-1}, u)$.*

Lemma 33. *Let w and v be two vectors in $\mathbb{R}^d$. The following holds:*

1. *If v is a unit vector, then $\|\hat{w} - v\| \leq 2 \|w - v\|$.*

2. *If v is s-sparse, then $\|\mathcal{H}_s(w) - v\| \leq 2 \|w - v\|$.*

3. *If v is a unit vector, then $\theta(w, v) \leq \pi \|w - v\|$; if w is a unit vector as well, then we further have $\|w - v\| \leq \theta(w, v)$.*

Proof. The first two parts are known expansion error of ℓ_2-normalization and hard thresholding, respectively. The proof can be found in, e.g. [SL18] (which also presents a sharp bound for the second part). The last part can be derived by fundamental algebra. $\square$