Attribute-Efficient Learning of Sparse Halfspaces with Constant Malicious Noise Rate

Shiwei Zeng
Augusta University
szeng@augusta.edu

Jie Shen
Stevens Institute of Technology
jie.shen@stevens.edu

Abstract

Attribute-efficient PAC learning of sparse halfspaces has been a fundamental problem in machine learning theory. In recent years, machine learning algorithms are faced with prevalent data corruptions or even malicious attacks. It is of central interest to design computationally and attribute-efficient algorithms that are robust to extreme corruptions. In this paper, we consider that there is a constant amount of malicious noise in the data and show that it is possible to PAC learn an underlying s-sparse halfspace $w^* \in \mathbb{R}^d$ with $O(s^2 \log^5 d)$ samples. Specifically, we follow a recent line of works and assume that the underlying distribution satisfies a concentration condition and a margin condition at the same time. As a complementary result, we provide an information-theoretic sample lower bound under such conditions even for the noiseless case. Evidence shows that our sample complexity could be nearly optimal. To show the robustness of our algorithm, we provide a new gradient analysis that carefully handles the sparsity admitted constraints in hinge loss minimization program, which could be of independent interest.

1 Introduction

In the modern machine learning and artificial intelligence, designing provably robust algorithms that enjoy desirable data-efficiency has been a pressing problem. One main field that has received great attention is *attribute-efficient learning* which leverages underlying model structures into algorithmic design and analysis such that the total sample complexity of the algorithm depends polynomially on the sparsity parameter s of the underlying model and only polylogarithmically on the ambient data dimension d. In this paper, we revisit the fundamental problem of PAC learning of the class of sparse halfspaces, i.e. $\{h_w : x \mapsto \mathrm{sign}(w \cdot x) : w \in \mathbb{R}^d, \|w\|_0 \leq s\}$, and propose an algorithm that enjoys a sample complexity of $O(s^2 \log^5 d)$ under extreme noise conditions.

In the past decade, there has been a rich line of works that address PAC learning problems under different noise conditions. Among them, many have achieved attribute-efficiency in algorithmic design for Massart noise conditions [ABHZ16, ZSA20], adversarial label noise conditions [ABHZ16, She21], and malicious noise conditions [SZ21] under a variety of learning scenarios. However, for challenging noise models such as adversarial lable noise and malicious noise, the best known noise tolerance of algorithms are $\Theta(\epsilon)$, where ϵ denotes the error parameter, until the recent works of [Tal20, She25] that combined two distributional assumptions of the concentration and large margin. Efficient algorithms tolerant up to a constant amount of adversarial noise [Tal20] and malicious noise [She25] are proposed. Here, we formally define the problem of PAC learning of sparse halfspaces under the malicious noise condition.

Definition 1.1 (Learning sparse halfspaces with malicious noise)**.** Let $\mathcal{X} := \mathbb{R}^d$ be the instance space and $\mathcal{Y} := \{-1, +1\}$ be the label space. Let $\mathrm{EX}(\mathcal{D}, w^*, \eta)$ be an adversary oracle with underlying distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, ground truth w^* with $\|w^*\|_0 \leq s$, and noise rate $\eta \in (0, \frac{1}{2}]$ fixed before the learning task begins, such that for all $(x, y) \sim \mathcal{D}, y = \mathrm{sign}(w^* \cdot x)$. Everytime the learner requests

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

a sample from $\mathrm{EX}(\mathcal{D}, w^*, \eta)$, with probability $1-\eta$, the oracle returns a clean sample (x, y) from $\mathcal{D}$; with probability η, the oracle returns an arbitrary sample $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Given any $\epsilon, \delta \in (0, 1)$, any $s < d$, the goal of the learner is to output a halfspace $\hat{w}$ using a number $n \leq \mathrm{poly}(s, \log d)$ of samples such that the error rate $\mathrm{err}_{\mathcal{D}}(\hat{w}) \leq \epsilon$ with probability $1-\delta$ (over the choice of the samples and all internal random bits of the learning algorithm), where $\mathrm{err}_{\mathcal{D}}(w) := \Pr_{(x,y)\sim\mathcal{D}}(y \neq \mathrm{sign}(w \cdot x))$.

In this paper, we propose the first attribute-efficient algorithm for the above learning problem for any η upper bounded by a constant η_0. Specifically, we follow the algorithmic framework of [She25] and refine the optimization program with sparsity admitted constraints that are extensively used in compressed sensing and regression literature [Tib96, CW08, PV13b, PV16]. That is, we incorporate an L_1 norm constraint as a relaxed condition for the s-sparse w, and look for an appropriate vector w within the constraint set $\mathcal{W} := \{w : \|w\|_2 \leq 1, \|w\|_1 \leq \sqrt{s}\}$ that enjoys a small hinge loss on the empirical sample set. We show that the gradient conditions on the optimum $\hat{w}$ ensure its correctness. In more details, we carefully analyze the gradient conditions based on the new constraint set and balance out the influence from both the L_2 and L_1 constraints. As a result, we show that with the modified constraints, when the underlying halfspace is indeed sparse, our algorithm only requires a sample complexity that depends polynomially in $s, \log d$. We summarize our results in Section 1.1.

1.1 Main result

Following the works of [Tal20, She25], our result depends on two distributional assumptions.

Assumption 1.2 (Large-margin)**.** Any finite set S_C of clean samples from $\mathcal{D}$ is γ-margin separable by the target halfspace w^* for some $\gamma > 0$. That is, $\forall (x, y) \in S_C$, $yx \cdot w^* \geq \gamma$.

Assumption 1.3 (Mixture of logconcaves)**.** The marginal distribution is $\mathcal{D}_{\mathcal{X}} = \frac{1}{k}\sum_{j=1}^{k} \mathcal{D}_j$, where each $\mathcal{D}_j$ is logconcave with mean μ_j and covariance matrix Σ_j, satisfying $\|\mu_j\|_2 \leq r$ for some $r > 0$ and $\Sigma_j \preceq \sigma^2 I_d$ for some $\sigma^2 = \frac{1}{d}$.

By utilizing the margin and mixture of logconcaves conditions together, we are able to design an attribute-efficient algorithm that learns the underlying s-sparse halfspace under $\Omega(1)$ malicious noise.

Theorem 1.4 (Main result)**.** *For any $\epsilon, \delta \in (0, 1)$, given that Assumption 1.2 and 1.3 hold for some $\gamma \geq \frac{8(\log\frac{1}{\epsilon}+1)}{\sqrt{d}}$, $r \leq 2\gamma$, $k \leq 64$, and $\eta \leq \eta_0 \leq \frac{1}{2^{28}}$, by drawing a set of $\Omega\left(s^2 \cdot \log^5 d \cdot \log^2 \frac{1}{\epsilon\delta}\right)$ samples from EX$(\mathcal{D}, w^*, \eta)$, Algorithm 1 runs in polynomial time and returns a halfspace $\hat{w}$ such that $\mathrm{err}_{\mathcal{D}}(\hat{w}) \leq \epsilon$ with probability at least $1-\delta$.*

Remark 1.5 (Comparison to existing works). Both two assumptions we made are equivalent to those in the prior works of [Tal20, She25], while their sample complexity is at least $\Omega(d)$. Theorem 1.4 suggests that as long as the underlying w^* has a sparse structure, Algorithm 1 naturally enjoys attribute-efficiency while maintaining its robustness by incorporating sparsity admitted constraints.

As a complementary result, we prove an information-theoretic sample lower bound for learning under the margin and mixture of logconcaves conditions. That is, even under such favorable distributional assumptions, PAC learning is not easier than that under the distribution-free settings.

Lemma 1.6. *Learning of s-sparse halfspaces under Assumption 1.2 for any $\gamma \leq \frac{1}{2}$ and 1.3 in the noiseless setting has an information-theoretic sample complexity lower bound of $\Omega(s \cdot (\log\frac{1}{\epsilon} + \log\frac{d}{s}))$.*

Remark 1.7 (Optimality). The prior work of [DKS17] shows that recovering a $\Theta(s^{-c/4})$-fraction of attribute corruptions of the Gaussian distributions under the Huber contamination model requires at least $\Omega(s^{2-3c})$ samples for any statistical query algorithms running in time $\Omega(d^{\Omega(c\cdot s^c)})$. This implies a quadratic sample lower bound $\Omega(s^2)$ for computationally efficient learning of linear models, i.e. $w \cdot x$, with malicious noise under the statistical query learning setting. Showing information-theoretic lower bound of $\Omega(s^2)$ for computationally efficient algorithms is still an open problem. This evidence shows that our sample complexity could be nearly optimal up to logarithmic factors.

Remark 1.8 (Constant noise rate). Our algorithm is tolerant to an η malicious noise rate as long as $\eta \leq \eta_0$ for some constant η_0. This differs from prior works on attribute-efficient learning with noise rate $O(\epsilon)$ [SZ21], i.e. becoming extremely weak when $\epsilon \to 0$. With carefully optimization over the constants in proofs, the upper bound of η_0 can be further improved to $2^{-9}/k^2$. Yet, this work focus on achieving the attribute-efficiency and we leave these slight improvements to interested readers.

Remark 1.9 (Adversarial label noise). Our main results immediately imply a computationally and attribute-efficient algorithm that can tolerate up to a constant amount of adversarial label noise. Note that under the adversarial label noise model, the adversary is only allowed to corrupt the labels while all instances remain untouched. Hence, with even a simpler algorithm, we are able to learn under a constant adversarial noise rate and keep the sample complexity polynomial in $s, \log d$. In the main paper, we focus on the more challenging malicious noise and defer such discussions to Appendix E.

Remark 1.10 (Non-homogeneous halfspaces). A salient feature of learning under both the mixture of logconcaves and margin conditions is that, the analysis can be adapted to non-homogeneous halfspaces. Though we mainly prove for the homogeneous case, Theorem 15 of [Tal20] shows that the key condition, i.e. the dense pancake condition, remains to hold under Lipschitz linear transformations. This implies that our main results also hold for attribute-efficient learning of non-homogeneous sparse halfspaces under the mixture of logconcaves and margin conditions.

1.2 Technical contributions

Gradient analysis over sparse admitted constraints. The benefit of having both the concentration and margin conditions at the same time is that, when data samples are sufficiently concentrated, the margin will push a high-density region far away from the decision boundary. Both prior works [Tal20, She25] rely on this condition and show that any instance that is surrounded by enough amount of good neighbors is not misclassified by an optimum $\hat{w}$, as the gradients of the instance itself and its neighbors should contribute to a sufficient weight and push the minimization program to move to a candidate that must be correct on it. Hence, a key technical component here is the gradient analysis for any instance that lies in a high-density region. However, this becomes more challenging when both L_2 and L_1 constraints are implemented. It is known from the Karush–Kuhn–Tucker (KKT) condition that when the program outputs an optimum $\hat{w}$, its gradient (or subgradient) condition is controlled by both the objective function and the list of all constraint functions. Observing that our program is a hinge loss minimization with respect to a convex set, we can implement the KKT condition in our gradient analysis. However, it is in question how to balance out both the influence from the L_2 norm constraint $\|w\|_2 \leq 1$ and the L_1 norm constraint $\|w\|_1 \leq \sqrt{s}$.

Our key observation is that, when any one of the constraints is active, i.e. the solution is on the boundary, there exists some subgradient g of the objective function that lies in the span of the (sub)gradients of the active constraints. The crucial step is to find a vector w' that can be seen as a component of $w^* - \hat{w}$. Conditioned on that the weight from clean samples is prevalent, if the vector $-g$ fails to point in the direction of w' (bringing $\hat{w}$ further towards w^*), then it must be because a boundary condition is reached. As a result, we need a w' that is orthogonal to g to establish a set of contradictory conditions and show that $\hat{w}$ is indeed a good enough halfspace. Technically, we will choose $w' = w^* - \hat{w}\langle w^*, \kappa\rangle$, where κ is some vector related to the boundary conditions, and make sure that $g \cdot w' = 0$. Note that we don't need to know the exact vector w', but its existence is enough for us to establish all contradictory conditions. More details can be found in Lemma 4.2 and its proof.

Function class covering instead of geometric coverings. Under the attribute-efficient learning regime, the previous technique of spacial covering of the target set $\mathcal{W}$ no longer works. This is because no matter we use L_1 balls or L_2 balls to cover the set $\mathcal{W}$, the sample complexity will blow up. Instead, we revisit the class setting of function space covering [AB02] and utilize the existing result for $\mathcal{N}_\infty$-covering the inner product function space. As a result, the sample complexity can be well bounded by the L_1 norm bound on w and L_∞ bound on the distance between instances.

Comparison to single isotropic logconcave settings. There is a large line of works for robust learning of halfspaces with or without attribute-efficiency that depend on an assumption that the marginal distribution is a single isotropic logconcave distribution [BL13, ABL17, Zha18, SZ21]. Importantly, the isotropic logconcave itself is a very general class of distributions (including the uniforms and Gaussians). However, it still admits a nice property that the probability density stays the same over all directions, i.e. $w \in \mathbb{R}^d, \|w\|_2 \leq 1$. This allows bounding the L_2 distance between a candidate halfspace $\hat{w}$ and w^* with $O(\epsilon)$ for $\mathrm{err}(\hat{w}) \leq \epsilon$. However, this property does not hold any more for the mixture of logconcaves with distribution means deviating from the origin. Fortunately, we utilize the mixture of logconcave assumption together with the margin to show that under specific conditions, L_2 distance still implies the disagreement probability, resulting in a hardness example. See Section 4.3 and Appendix D for more details.

1.3 Related works

The earliest study of learning with irrelevant attributes could date back to the last century, when [Lit87, Blu90, BHL91] proposed attribute-efficient algorithms for learning concept classes in the online learning settings. Since then, attribute efficiency was studied in different learning problems, such as learning concept classes [Ser99, KS04, LS06, HS07, APVZ14, ZS23], regression [Tib96, LSLC20], compressed sensing [Don06, CW08], basis pursuit [CDS98, Tro04, CT05], variable selection [FL01, FF08, SL17a, SL17b], and statistical estimation [DKK$^+$19, CDK$^+$22, ZS22, DKK$^+$22].

While the applications of attribute-efficient learning appears to be broad, learning of the class of sparse halfspaces remains fundamental and elusive especially in the presense of data corruptions. The problem has been extensively studied by both the compressed sensing and learning theory communities. Early works focus on the noiseless or benign noise conditions [BB08, PV13a, PV13b, Zha18], whereas the recent years have witnessed more sophisticated algorithms for extreme noise conditions [ABHZ16, ZSA20, She21, SZ21]. Even without attribute efficiency, it is known that the best possible noise tolerance for distributionally-independent learning under adversarial label noise or malicious noise is $\Theta(\epsilon)$ ([KL88, KSS94, ABL17]). Under the assumption of concentrated marginals, such as isotropic Gaussian or logconcave types of concentration, the best known noise tolerance was $O(\epsilon)$ ([DKTZ20, She21, SZ21]) until the work of [Tal20] introduced the margin condition together with concentration assumptions, and achieved a constant adversarial label noise tolerance. Later on, [She25] provided an improved algorithm that achieved a constant noise rate for the malicious noise. This is the line of research that we follow. In addition, with the additional margin condition, the algorithmic design is much simpler comparing to previous designs. Specifically, the main algorithm is a surrogate loss minimization with respect to convex constraints, which is reminiscent of [PV13b, PV16]. In other words, under proper yet realistic assumptions, we show that simple algorithms can achieve strong noise tolerance and attribute efficiency at the same time.

2 Preliminaries

Given two vectors u and v, denote by $u \cdot v$ and $\langle u, v\rangle$ their dot product and inner product, respectively.

Dense pancake. For any given sample (x, y), the dense pancake is defined as a set of samples lying around it within a certain distance. The distance is measured with respect to some vector $w \in \mathbb{R}^d$. In this paper, we only consider w with sparse properties. That is, we require the dense pancake condition to hold for any $w \in \mathcal{W}$. Note that this is an essential difference between our dense pancake condition and those were defined in the prior works.

Definition 2.1 (Dense pancake condition)**.** For $(x, y) \in \mathcal{X} \times \mathcal{Y}$, given any vector $w \in \mathcal{W}$ and a thickness parameter $\tau > 0$, the pancake $\mathcal{P}^\tau_w(x, y)$ is defined as

$$\mathcal{P}^\tau_w(x, y) := \{(x', y') \in \mathcal{X} \times \mathcal{Y} : |y'x' \cdot w - yx \cdot w| \le \tau\}.$$

We say that the pancake $\mathcal{P}^\tau_w(x, y)$ is ρ-dense with respect to a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ if $\Pr_{(x',y')\sim\mathcal{D}}((x', y') \in \mathcal{P}^\tau_w(x, y)) \ge \rho$. Let $\mathcal{D}_1, \mathcal{D}_2$ be two (not necessarily different) distributions over $\mathcal{X} \times \mathcal{Y}$. We say that $(\mathcal{D}_1, \mathcal{D}_2)$ satisfies the (τ, ρ, β)-dense pancake condition if for any vector $w \in \mathcal{W}$, it holds that $\Pr_{(x,y)\sim\mathcal{D}_2}(\mathcal{P}^\tau_w(x, y)$ is ρ-dense w.r.t. $\mathcal{D}_1) \ge 1 - \beta$.

Hinge loss. Given any vector w, its hinge loss on any (x, y) is given by $\ell_\gamma(w; (x, y)) := \max\{0, 1 - yx \cdot \frac{w}{\gamma}\}$, for a pre-specified parameter γ. When γ is clear from the context, we omit the subscript. For a sample set S, the hinge loss of w on S is denoted as $\ell(w; S) := \sum_{i\in S} \ell(w; (x_i, y_i))$. If we are further given a weight vector for the sample set S, i.e. $q = (q_1, \ldots, q_{|S|})$, denote the weighted hinge loss of w by $\ell(w; q \circ S) := \sum_{i\in S} q_i \cdot \ell(w; (x_i, y_i))$.

Norms. Formally, we denote by $\|v\|_1$ the L_1 norm of a vector v, by $\|v\|_2$ the L_2 norm, and by $\|v\|_\infty$ the L_∞ norm. Specially, denote by $\|v\|_0 := |\{i : v_i \ne 0\}|$ the cardinality of non-zero elements in v. For a matrix A, we denote by $\|A\|_1$ the entrywise L_1 norm of A, i.e. the sum of absolute values of all its entries; and by $\|A\|_*$ its nuclear norm. For a symmetric A, we use $A \succeq 0$ to denote that it is positive semidefinite.

Our work relies crucially on the gradient (and subgradient) analysis of the hinge loss minimization program. The following definition of the gradient norm over a linear summation of the instances from a set S plays an important role in restricting the influence of their (sub)gradients.

Definition 2.2 (Gradient norm)**.** Given a set $S = \{(x_i, y_i)\}$ and weights $q = (q_1, \ldots, q_{|S|})$, the gradient norm of a weighted S is defined as

$$\mathrm{GradNorm}(q \circ S) := \sup_{a \in [-1,1]^{|S|}} \sup_{w \in \mathcal{W}} \left\langle \sum_{i \in S} a_i q_i x_i, w \right\rangle.$$

Other notations. When $\mathcal{D}, w^*, \eta$ are clear from the context, we write EX instead of $\mathrm{EX}(\mathcal{D}, w^*, \eta)$. We use $[\cdot]$ to denote the set of all positive integer that is smaller or equal to the quantity inside.

3 Algorithm

The main algorithmic framework follows from [She25], however, with carefully designed variations for the underlying sparsity structures and the attribute-efficiency requirements.

There are two major ingradients in Algorithm 1. The first ingredient is a soft ourlier removal scheme that assigns weight $q_i \in [0, 1]$ to all instance $i \in S$ such that their weighted variance on any sparse direction w is properly upper bounded. This is the key step that restricts the influence of data corruptions from the malicious noise. The second ingredient is a hinge loss minimization with respect to a set of L_2 and L_1 norm constraints. Recall that we have assumed that the underlying halfspace w^* is a unit vector in $\mathbb{R}^d$ with $\|w\|_0 \le s$ for some $s < d$. Therefore, we consider a set $\mathcal{W}$ of vectors that is a convex hull of the original hypothesis set, i.e. $\{w : \|w\|_2 = 1, \|w\|_0 \le s\} \subset \mathcal{W}$. This ensures that w^* lies in the feasible set of the minimization program (Algorithm 1, Eq. (1)).

Algorithm 1 Main Algorithm

Require: $\mathrm{EX}(D, w^*, \eta)$, target error rate ϵ, failure probability δ, parameters s, γ.
Ensure: A halfspace $\hat{w}$.

1: Draw $n = C \cdot s^2 \log^5 \frac{d}{\delta\epsilon}$ samples from $\mathrm{EX}(D, w^*, \eta)$ to form a set $S = \{(x_i, y_i)\}_{i=1}^n$ for some large enough constant $C > 0$.
2: Apply Algorithm 2 on S and let q be the returned weight vector.
3: Solve the following weighted hinge loss minimization program:

$$\hat{w} \leftarrow \operatorname*{arg\,min}_{\|w\|_2 \le 1, \|w\|_1 \le \sqrt{s}} \ell_\gamma(w; q \circ S). \tag{1}$$

4: Return $\hat{w}$.

It will be convenient to think of S as a union of a set S_C of clean samples that are indeed drawn from the underlying distribution $\mathcal{D}$ and a set of S_D of samples that are maliciously inserted by the adversary, denoted as $S = S_C \cup S_D$.

3.1 Soft outlier-removal

The robustness of our algorithm to the malicious noise comes mainly from the soft outlier removal scheme, presented in Algorithm 2. Intuitively, it assigns high weights to the clean samples in S_C and low weights to the corrupted samples in S_D, especially to those lying far away from the main group. In more details, let $q = (q_1, \ldots, q_n)$ be a weight vector with each q_i controls the weight of sample $i \in S$, where $n = |S|$. The goal of Algorithm 2 is to return a vector q such that for all $w \in \mathcal{W}$, $\frac{1}{n}\sum_{i=1}^n q_i (w \cdot x_i)^2 \le \bar{\sigma}^2$. This is ensured by Lemma A.8. By controling the weighted variance, the algorithm effectively restricts the influence from the malicious samples, as the adversary can not insert an instance with very large gradient to twist the optimization program or to force it to optimize towards a wrong direction. If there exist some of such malicious instances, they will be automatically down-weighted by the soft outlier removal scheme and assigned a smaller q_i, as a consequence of violating the weighted variance constraint.

The main challenge in implementing such a soft outlier removal scheme is the L_1 constraint in set $\mathcal{W}$. It is known to be NP-hard to find the largest variance of $x_i \in S$ with respect to a $w \in \mathcal{W}$, even without the weight vector q. Hence, we relax this problem to a semidefinite program by constructing a set $\mathcal{M} := \{H : H \succeq 0, \|H\|_1 \le s, \|H\|_* \le 1\}$, and instead look for a largest value of $\frac{1}{n}\sum_{i=1}^n x_i^\top H x$ for $H \in \mathcal{M}$. This is presented in Algorithm 2 with weight vector q. This kind of relaxation was

used in prior attribute-efficiency works (e.g. [SZ21] for localized logconcave denoising), and it is known to be computationally and attribute-efficient at the same time. In addition, by running this relaxed program, the output q is a sufficiently good solution to our original problem. That is, $\frac{1}{n}\sum_{i=1}^{n} q_i(w \cdot x_i)^2$ is well bounded for any $w \in \mathcal{W}$. Our contribution lies in the analysis of the program for a mixture of logconcaves instead of a single localized isotropic logconcave. We reuse part of the proof from [SZ21] but have significantly adapted it to the new conditions (see appendix).

Algorithm 2 Soft Outlier Removal

Require: Sample set $S = \{(x_i, y_i)\}_{i=1}^{n}$, upper bound on empirical noise rate $\xi = 2\eta_0 \geq \frac{|S_D|}{|S|}$, variance parameter $\bar{\sigma} \leftarrow 2 \cdot \sqrt{\left(\frac{1}{d} + r^2\right)}$.

Ensure: A weight vector $q = (q_1, \ldots, q_n)$.

1: Find a vector q feasible to the following semidefinite program:

$$\begin{cases} 0 \leq q_i \leq 1, \ \forall i \in [n], \\ \sum_{i=1}^{n} q_i \geq (1-\xi)n, \\ \sup_{H \succeq 0, \|H\|_1 \leq s, \|H\|_* \leq 1} \frac{1}{n}\sum_{i=1}^{n} q_i x_i^\top H x_i \leq \bar{\sigma}^2. \end{cases}$$

2: Return q.

3.2 Attribute-efficient hinge loss minimization

The last yet the most important ingredient, which is also the main source of our algorithmic robustness towards high noise rate, is the hinge loss minimization over the reweighted samples, i.e. $q \circ S$, with respect to sparse set $\mathcal{W}$. Note that with the L_∞ norm filter and soft outlier removal scheme, we are able to identify and remove those samples whose attributes are abnormal. However, there could be a large amount of label noise that remains unattended. Hence, it is essential to utilize the natural robustness of the hinge loss minimization program to handle these label noise.

Based upon the soft outlier removal scheme, we attain a weight vector q with respect to which the reweighted variance $\frac{1}{|S|}\sum_{i \in S} q_i(x_i \cdot w)^2$ is well bounded for any $w \in \mathcal{W}$. This is sufficient for us to show that within any thick and dense enough pancake, the weights for the clean samples are large enough such that the gradients of the malicious samples may not confuse the minimization program too much. As it will be revealed from our gradient analysis, the gradient norm $\mathrm{GradNorm}(S_D)$ serves as an upper bound on the sum of gradients over any set S_D projecting onto any $w \in \mathcal{W}$. Hence, with guarantees from Algorithm 2, the hinge loss minimization program at Eq. (1) ensures that all label noise is also treated. As a result, we can show that a sufficient amount of samples from the underlying distribution $\mathcal{D}$ will not be misclassified by any returned $\hat{w}$. In other words, the error rate of $\hat{w}$ is bounded. More technically, we consider a KKT condition on the optimum $\hat{w}$ of the hinge loss minimization program at Eq. (1), such that any (x, y) that has a dense pancake with respect to $\hat{w}$ would not be misclassified by $\hat{w}$. The main challenge lies in the analysis of the KKT condition with both L_2 and L_1 constraints. As a result, we utilize the underlying unknown parameters λ_1 and λ_2 to balance the influence of both constraints and show that the gradient from the hinge loss function would still point to the right direction. This helps us establish a sufficient condition for showing the correctness of any (x, y) with a dense pancake.

4 Analysis

We present the gradient analysis in Section 4.1. Statistical results are in Section 4.2. The proof idea of the lower bound is in Section 4.3. We provide a proof sketch of main theorem in Section 4.4.

4.1 Gradient analysis

The prior works on establishing algorithmic robustness through surrogate loss minimization for distributions with both concentration and margin conditions crucially rely on an analysis for some special (x, y) that is stated as follows. Given the concentration property of the underlying distribution, there must be a large portion of instances that are surrounded by each other. Then, for any (x, y) that lies in a high-density region and with its neighbors having the same label, the KKT condition for

any optimum $\hat{w}$ ensures that such an (x, y) is not misclassified by $\hat{w}$. Hence, the following analysis focuses on such a special (x, y) and especially the subgradient conditions around it.

For any given $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we consider the sample set S as the union of three distinct sets with respect to it, i.e. $S = S_P \cup S_{\bar{P}} \cup S_D$, where S_P is the set of clean samples (those in S_C) that lie in the pancake of (x, y), i.e. $S_P = S_C \cap \mathcal{P}^{\tau}_{\hat{w}}(x, y)$, $S_{\bar{P}}$ includes the clean samples lying outside the pancake, i.e. $S_{\bar{P}} = S_C \backslash S_P$, and S_D again denotes the set of malicious samples.

Theorem 4.1 (Correctness of $\hat{w}$ on good (x, y))**.** *Given that Assumption 1.2 hold. For any given (x, y), if its pancake $\mathcal{P}^{\tau}_{\hat{w}}(x, y)$ with $\tau \leq \gamma/2$ is ρ-dense with respect to S_C for some $\rho > 4\eta_0$, and it holds that*

$$\sum_{i \in S_C \cap \mathcal{P}^{\tau}_{\hat{w}}(x,y)} q_i > \frac{4}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D), \tag{2}$$

then (x, y) is not misclassified by $\hat{w}$.

The proof utilizes an optimality condition on the output $\hat{w}$ that involves the analysis on the subgradients of both hinge loss function and constraint functions. Recall that the weighted hinge loss is defined as $\ell(\hat{w}; q \circ S) = \sum_{i \in S} q_i \max\{0, 1 - y_i x_i \cdot \frac{\hat{w}}{\gamma}\}$ (see Section 2). We decompose this loss into three parts, i.e.

$$\ell(\hat{w}; q \circ S) = \ell(\hat{w}; q \circ S_P) + \ell(\hat{w}; q \circ S_{\bar{P}}) + \ell(\hat{w}; q \circ S_D),$$

where $q \circ S'$ for any subset $S' \subseteq S$ denotes a weighted sample set according to the weight vector $q_{S'} := \{q_i, i \in S'\}$ while ignoring the unrelated weights.

We prove the theorem by contradicting the optimality of $\hat{w}$ (Algorithm 1, Eq. (1)) and that a given (x, y) described in Theorem 4.1 is misclassified by $\hat{w}$. While outputing $\hat{w}$, if $\hat{w}$ is the interior of the constraint set $\mathcal{W} := \{w : \|w\|_2 \leq 1, \|w\|_1 \leq \sqrt{s}\}$, the analysis follows directly from the prior works. The analysis for $\hat{w}$ on the boundary is more complicated. Denote by $\hat{z} \in \partial_w \|w\|_1 \mid_{w=\hat{w}}$ a subgradient of L_1 constraint function at $\hat{w}$, and it is known that $\{\hat{w}\} = \partial_w \|w\|_2 \mid_{w=\hat{w}}$ is the gradient set for the L_2 constraint. Specifically, due to the KKT conditon, there exists some $g \in \partial_w \ell(w; q \circ S) \mid_{w=\hat{w}}$ and some $\hat{z}$ such that

$$g + \lambda_1 \cdot \hat{z} + \lambda_2 \cdot \hat{w} = 0,$$

where $\lambda_1, \lambda_2 \geq 0$. More precisely, the complementary slackness implies that $\lambda_1 > 0$ if and only if $\|\hat{w}\|_1 = \sqrt{s}$ and $\lambda_2 > 0$ if and only if $\|\hat{w}\|_2 = 1$. When only one of the constraints is active, the proof is rather intuitive as the span of a subgradient g is clear, i.e. either due to $\hat{z}$ or $\hat{w}$. It is when both of the constraints are in effect a much more challenging case. Hence, we will show the proof by considering that $\lambda_1, \lambda_2 > 0$, and the proof for either $\lambda_1 = 0$ or $\lambda_2 = 0$ follows as special cases.

The analysis relies crucially on the a vector w' that can be considered as the component of $w^* - \hat{w}$ that is orthogonal to some subgradient g. That is, we hope that $g \cdot w' = 0$. To find a vector satisfying this condition, we define

$$w' := w^* - \hat{w} \langle w^*, \kappa \rangle, \tag{3}$$

where $\kappa = \frac{\lambda_1}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \hat{z} + \frac{\lambda_2}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \hat{w}$. The following lemma guarantees that there exists some g such that $g \cdot w' = 0$. The proof is deferred to the appendix.

Lemma 4.2. *Consider the set of subgradients $\partial_w \ell(w; q \circ S)\big|_{w=\hat{w}}$ and some element g in it. Given the optimization program Eq.(1) in Algorithm 1 and $\hat{w}$ returned by it, let w' be the vector defined in Eq. (3), for some $\lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 \neq 0$ and $\hat{z} \in \partial_w \|w\|_1 \big|_{w=\hat{w}}$. Then, there exists $g \in \partial_w \ell(w; q \circ S)\big|_{w=\hat{w}}$ such that $g \cdot w' = 0$.*

The following lemma shows that, if an (x, y) is misclassified by $\hat{w}$, then all clean samples within its pancake will contribute to a good weight $\sum_{i \in S_P} q_i$, together with those from the malicious samples, we construct the following upper bound on the dot product $g \cdot w'$ for any subgradient g.

Lemma 4.3. *Consider Algorithm 1. Suppose Assumption 1.2 holds. For some $\hat{z} \in \frac{\partial \|w\|_1}{\partial w}\big|_{w=\hat{w}}$, define $w' = w^* - \hat{w} \langle w^*, \kappa \rangle$, where $\kappa = \frac{\lambda_1}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \hat{z} + \frac{\lambda_2}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \hat{w}$. Then, we have the following holds. For any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and its pancake $\mathcal{P}^{\tau}_{\hat{w}}(x, y)$ with $\tau \leq \gamma/2$, if (x, y) is misclassified by $\hat{w}$, then for any $g \in \partial_w \ell(w; q \circ S)\big|_{w=\hat{w}}$, we have*

$$g \cdot w' \leq -\frac{1}{2} \sum_{i \in S_P} q_i + \frac{2}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D).$$

Notice that when the weight $\sum_{i\in S_P} q_i$ is large enough, i.e. the condition in Eq.(2) is satisfied, it will push the dot product $g \cdot w'$ towards the negative side, such that $g \cdot w' < 0$ holds for all subgradients $g \in \partial_w \ell(w; q \circ S) \mid_{w=\hat{w}}$. Interestingly, this means that the program will keep optimizing towards w^*, and is contradictory to the condition we establish in Lemma 4.2. Hence, we prove by contradiction that any (x, y) with a large weight $\sum_{i\in S_P} q_i$ is not misclassified by $\hat{w}$. The guarantees are summarized in Theorem 4.1 and the full proof is deferred to Appendix A.1.

4.2 Sample complexity

The sample complexity of our algorithm mainly comes from three sources: 1) the amount of samples required for the empirical dense pancake condition, i.e. $(S_C, \mathcal{D})$ satisfies (τ, ρ, ϵ)-dense pancake condition with respect to any $w \in \mathcal{W}$; 2) the amount of samples that ensures the statistical properties for Algorithm 2. 3) the sample complexity that bounds the empirical noise rate by $\xi = 2\eta_0$.

Theorem 4.4 (Sample complexity for empirical pancake density)**.** *Consider a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ satisfying Assumption 1.3 and 1.2 with $\gamma > 2\tau$, where $\tau = 2\sigma \cdot \left(\log \frac{1}{\beta} + 1\right)$. It holds with probability at least $1 - \delta'$ over the draw of $|S_C| \geq \frac{4k}{1-\beta} \cdot \left(C_0 s \cdot \log^4 d + \log \frac{1}{\delta'\beta}\right)$ samples, $(S_C, \mathcal{D})$ satisfies $\left(2\tau, \frac{1-\beta}{2k}, 2\beta\right)$-dense pancake condition, where $C_0 > 0$ is an absolute constant.*

It is straightforward to show that any distribution $\mathcal{D}$ satisfying both Assumption 1.3 and 1.2 is (τ, ρ, β)-dense with $\tau = 2\sigma \cdot (\log \frac{1}{\beta} + 1)$ and $\rho = \frac{1-\beta}{2k}$ (see Lemma B.1 in Appendix B). The challenge here is to show that by collecting a $O(\mathrm{poly}(s, \log d))$ amount of samples from $\mathcal{D}$, the empirical set S_C is $(2\tau, \rho, 2\beta)$-dense with respect to $\mathcal{D}$ with high probability.

The prior works of [Tal20] proves it using standard geometric covering. Specifically, they first show that the dense pancake condition holds for a fixed vector w except for a small failure probability. Then, it suffices to take a union bound of the failure rate over a τ-cover of the unit ball in $\mathbb{R}^d$, since an L_2 distance of τ between w, w' only increase the pancake width by τ while all other density conditions stay unchanged (see Theorem 20 therein). We remark that this analysis does not work under the sparsity conditions. First, using L_2 balls of radius τ to cover $\mathcal{W}$ yields a covering number of at least $\exp(s/\tau^2)$. For small $\tau = \frac{1}{\sqrt{d}}$, this leads to a blow-up in sample complexity of $\mathrm{poly}(d)$. Moreover, to the best of our knowledge, it remains unknown whether the covering number is still $O(s \log d)$ when using L_1 balls of radius τ to cover the set $\mathcal{W}$.

Hence, instead of geometric covering, we utilize results from the traditional function space covering and construct a τ-cover over the inner product space of $\langle w - w', \bar{x}\rangle$, where $\bar{x}$ represents the distance from any point to the pancake center (x, y). It is known that for any logconcave component, samples are away from each other with L_∞ distance of at most $\sigma \log d$ with high probability. We also know that for any pairs of $w, w' \in \mathcal{W}$, the L_1 distance is at most $\sqrt{s}$. Hence, it suffices to bound the covering number $\mathcal{N}_\infty$ by $\exp\left(\tilde{O}\left(\frac{\sqrt{s^2}(\sigma \log d)^2}{\tau^2}\right)\right)$ using Corollary 5 from [Zha02] based on the norm bounds on both parts for the inner product. Here, $\tilde{O}$ hides logarithmic terms. Notice that $\mathcal{N}_\infty$ differs from L_∞ as it gaurantees the covering of functions to be effective for all instances x in the target sample set. See more details in the full proof of Theorem 4.4 (Appendix B) and in Appendix F,

The following lemma dedicates to the sample complexity for the semidefinite program in Algorithm 2.

Lemma 4.5 (Sample complexity for empirical variance)**.** *For any distribution $\mathcal{D}_\mathcal{X}$ satisfying Assumption 1.3, let $S_C = \{x_1, \ldots, x_{|S_C|}\}$ be a set of i.i.d. unlabeled instances drawn from $\mathcal{D}_\mathcal{X}$. Denote $G(H) := \frac{1}{|S_C|}\sum_{i\in S_C} x_i^\top H x_i - \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[x^\top H x]$. Then with probability $1 - \delta'$, $\sup_{H\in\mathcal{M}} |G(H)| \leq 2\left(\frac{1}{d} + r^2\right)$ given that $|S_C| \geq \Theta\left(s^2 \cdot \log^5 d \log^2 \frac{1}{\delta'}\right)$.*

Together with the following lemma, we obtain the variance parameter in Algorithm 2. That is, if $\forall H \in \mathcal{M}$, $\frac{1}{|S_C|}\sum_{i\in S_C} x_i^\top H x_i \leq 4\left(\frac{1}{d} + r^2\right)$ with high probability, it suffices to set $\bar{\sigma} = 2\sqrt{\frac{1}{d} + r^2}$ for the program to identify the clean samples and downweight all malicious samples.

Lemma 4.6. *Given distribution $\mathcal{D}_\mathcal{X}$ satisfying Assumption 1.3, it holds that $\sup_{H\in\mathcal{M}} \mathbb{E}_{X\sim\mathcal{D}_\mathcal{X}}(x^\top H x) \leq 2\left(\frac{1}{d} + r^2\right)$.*

The following lemma bounds the empirical noise rate $\frac{|S_D|}{|S|}$, which facilitates our analysis.

Lemma 4.7 (Sample complexity for empirical noise rate)**.** *If we draw a sample S from EX with size $|S| \geq \frac{3}{\eta_0} \cdot \log \frac{1}{\delta'}$, then it is guaranteed with probability $1-\delta'$ that $|S_C| \geq (1-2\eta_0)\,|S|$ and $|S_D| \leq 2\eta_0\,|S|$.*

All proofs for the theorem and lemmas in this section can be found in Appendix B.

4.3 Lower bound of sample complexity

Though lower bound of $\Omega(s \log d)$ is already known for general learnability of s-sparse halfspaces in the distribution-free settings. However, for learning from fixed distributions under the Assumption 1.2 and 1.3, it is unknown whether such bound still holds. Interestingly, [SZ21] showed that the same lower bound holds for any single isotropic logconcave marginals. Here, we generalize the result to any mixture of logconcaves even under the margin conditions and derive Lemma 1.6.

It is worth noting that for prior works on robust learning halfspaces depending on isotropic logconcave assumptions, a key property that was significantly used is that bounding the angle between two vectors u, v implies bounding the disagreement probability between halfspace induced by u, v. This is due to that the probability density is uniform over all directions. This results in a key restriction for these algorithms that they only works for homogeneous halfspaces. Now, consider $\mathcal{D}_X, \mathcal{D}$ satisfying conditions in Assumption 1.2 with $\gamma \leq \frac{1}{2}$and 1.3, apparently, the probability density of such underlying distribution is no longer uniform over all directions u, v. As a result, we construct a hardness example, where the distribution density is uniform around w^*, and 0 inside the margin. We show that, the volume of the region A' where the probability density is non-zero is large enough, such that any two u, v with disagreement region falling inside A' will have their L_2 distance, i.e. $\|u-v\|_2 \geq c \cdot \epsilon$, implying a significant disagreement probability, i.e. $\Pr_{x \sim \mathcal{D}_X}(\operatorname{sign}(u \cdot x) \neq \operatorname{sign}(v \cdot x)) \geq \epsilon$. The complete proof is deferred to Appendix D.

4.4 Proof of Theorem 1.4

The proof idea of our main theorem is to utilize the key deterministic result (Theorem A.10) and show that by running Algorithm 1 on a large enough sample set, the returned $\hat{w}$ has low error rate with respect to the underlying distribution $\mathcal{D}$.

Proposition 4.8. *Given that Asssumption 1.3 holds, if we draw a sample set S from EX with size $|S| \geq \Omega(s^2 \cdot \log^5 d \log^2 \frac{1}{\delta'} + \frac{1}{\eta_0} \cdot \log \frac{1}{\delta'})$, with probability at least $1-2\delta'$, the semidefinite program in Algorithm 2 is feasible.*

Proposition 4.8 shows that Program (1) is feasible under a proper selection of the sample size. The proof is in Appendix B. We are now ready to prove our main theorem.

Proof of Theorem 1.4. From Proposition 4.8, we know that the semidefinite program in Algorithm 2 is feasible with probability $1-2\delta'$, given that $|S| \geq \Omega(s^2 \cdot \log^5 d \log^2 \frac{1}{\delta'} + \frac{1}{\eta_0} \cdot \log \frac{1}{\delta'})$. From Theorem 4.4, we obtain that with probability at least $1-\delta'$, $(S_C, \mathcal{D})$ satisfies (τ, ρ, ϵ)-dense pancake condition by setting $|S_C| \geq \Omega\left(\frac{4k}{1-\epsilon} \cdot \left(s \cdot \log^4 d + \log \frac{1}{\delta'\epsilon}\right)\right)$ with $\tau = 2\sigma \cdot \left(\log \frac{1}{\epsilon} + 1\right)$.

Given that Assumption 1.2 holds with $\gamma > 0$, $(S_C, \mathcal{D})$ satisfies (τ, ρ, ϵ)-dense pancake condition with $\tau \leq \frac{\gamma}{2}$, and Program (1) is feasible, it suffices to choose $\gamma \geq 2\tau = \frac{8(\log \frac{1}{\epsilon}+1)}{\sqrt{d}}$ to satisfy the conditions for Theorem A.10. As a result, we have $\rho \geq 16\left(\frac{1}{\gamma\sqrt{d}} + \frac{r}{\gamma} + 1\right)\sqrt{\eta_0}$ when $k \leq 64$ and $\eta_0 \leq \frac{1}{2^{28}}$. By setting $\delta' = \delta/3$, it suffices to choose $\Omega\left(s^2 \log^5 d \cdot \log^2 \frac{1}{\delta\epsilon}\right)$. The proof is complete. □

5 Conclusion

In this paper, we revisit the problem of attribute-efficient PAC learning of sparse halfspaces under the challenging malicious noise model, with a noise rate up to a constant. To the best of our knowledge, this is the first attribute-efficient algorithm that works under a malicious noise rate of $\omega(\epsilon)$. It will be interesting to study if our gradient analysis can be extended to other learning settings, e.g. other surrogate loss functions. It is also interesting to apply our adaptation of single logconcave to mixture of logconcaves analysis in other problems, e.g. active learning under mixture of logconcaves.

Acknowledgments and Disclosure of Funding

This work was supported in part by a startup funding from Augusta University and the National Science Foundation under grant NSF-AF-2239376.

References

[AB02] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2002.

[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 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.

[APVZ14] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning sparse polynomial functions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 500–510. SIAM, 2014.

[BB08] Petros Boufounos and Richard G. Baraniuk. 1-bit compressive sensing. In *42nd Annual Conference on Information Sciences and Systems*, pages 16–21. IEEE, 2008.

[BHL91] Avrim Blum, Lisa Hellerstein, and Nick Littlestone. Learning in the presence of finitely or infinitely many irrelevant attributes. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, pages 157–166. Morgan Kaufmann, 1991.

[BI88] Gyora M. Benedek and Alon Itai. Learnability by fixed distributions. In *Proceedings of the First Annual Workshop on Computational Learning Theory*, pages 80–90. ACM/MIT, 1988.

[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*, volume 30, pages 288–316, 2013.

[Blu90] Avrim Blum. Learning boolean functions in an infinite atribute space (extended abstract). In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing*, pages 64–72. ACM, 1990.

[CDK+22] Yu Cheng, Ilias Diakonikolas, Daniel M. Kane, Rong Ge, Shivam Gupta, and Mahdi Soltanolkotabi. Outlier-robust sparse estimation via non-convex optimization. 2022.

[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.

[DKK+19] Ilias Diakonikolas, Daniel Kane, Sushrut Karmalkar, Eric Price, and Alistair Stewart. Outlier-robust high-dimensional sparse estimation via iterative filtering. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*, 2019.

[DKK+22] Ilias Diakonikolas, Daniel M. Kane, Sushrut Karmalkar, Ankit Pensia, and Thanasis Pittas. Robust sparse mean estimation via sum of squares. In *Proceedings of the 35th Annual Conference on Learning Theory*, pages 4703–4763, 2022.

[DKS17] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *Proceedings of the 58th IEEE Annual Symposium on Foundations of Computer Science*, pages 73–84, 2017.

[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.

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

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

[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.

[HS07] Lisa Hellerstein and Rocco A. Servedio. On PAC learning algorithms for rich boolean function classes. *Theoretical Computer Science*, 384(1):66–76, 2007.

[KL88] Michael J. Kearns and Ming Li. Learning in the presence of malicious errors (extended abstract). In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, pages 267–280. ACM, 1988.

[KMT93] Sanjeev R. Kulkarni, Sanjoy K. Mitter, and John N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11:23–35, 1993.

[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*, volume 3120 of *Lecture Notes in Computer Science*, pages 224–238. Springer, 2004.

[KSS94] Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. *Maching Learning*, 17(2-3):115–141, 1994.

[KST08] Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*, pages 793–800, 2008.

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

[Lon95] Philip M. Long. On the sample complexity of PAC learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

[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 Twentieth Annual Conference on Neural Information Processing Systems*, pages 921–928. MIT Press, 2006.

[LSLC20] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. In *The 23rd International Conference on Artificial Intelligence and Statistics,*, volume 108 of *Proceedings of Machine Learning Research*, pages 411–421. PMLR, 2020.

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

[PV13a] Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

[PV13b] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transaction on Information Theory*, 59(1):482–494, 2013.

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

[RWY11] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

[Ser99] Rocco A. Servedio. Computational sample complexity and attribute-efficient learning. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, pages 701–710. ACM, 1999.

[She21] Jie Shen. On the power of localized Perceptron for label-optimal learning of halfspaces with adversarial noise. In *Proceedings of the 38th International Conference on Machine Learning*, pages 9503–9514, 2021.

[She25] Jie Shen. Efficient PAC learning of halfspaces with constant malicious noise rate. In *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pages 1108–1137. PMLR, 2025.

[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 3124–3134, 2017.

[SZ21] Jie Shen and Chicheng Zhang. Attribute-efficient learning of halfspaces with malicious noise: Near-optimal label complexity and noise tolerance. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pages 1072–1113, 2021.

[Tal20] Kunal Talwar. On the error resistance of hinge-loss minimization. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020.

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

[Tro04] Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[vdVW96] Aad van der Vaart and Jon A Wellner. *Weak convergence and empirical processes.* Springer, 1996.

[Zha02] Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

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

[ZS22] Shiwei Zeng and Jie Shen. List-decodable sparse mean estimation. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*, 2022.

[ZS23] Shiwei Zeng and Jie Shen. Attribute-efficient PAC learning of low-degree polynomial threshold functions with nasty noise. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 40719–40748. PMLR, 2023.

[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.

A Omitted Proofs

We use C with subscripts to denote absolute constants.

A.1 Gradient and subgradient analysis of Algorithm 1 (omitted proofs in Section 4.1)

Note that since we follow the theoretical framework of [She25], some proof structure might be similar. However, we study the problem in the attribute-efficient learning setting and definitions have shifted. We include most of detailed proofs for completeness.

Recall that we have decomposed the hinge loss into different parts to facilitate our analysis. That is, for any given (x, y), we consider that the set S consists of three parts, i.e. $S_P, S_{\bar{P}}, S_D$. Similarly, their contribution to the hinge loss function can be decomposed in the following way:

$$\ell(w; q \circ S) = \ell(w; q \circ S_P) + \ell(w; q \circ S_{\bar{P}}) + \ell(w; q \circ S_D).$$

We summarize some characteristics for the function gradients. Namely, for any data set S and weight vector q, the weighted subgradient is given by $\partial_w \ell(w; q \circ S) = \frac{1}{\gamma} \cdot \sum_{i \in S} \partial f(y_i x_i \cdot \frac{w}{\gamma}) \cdot y_i (q_i x_i)$ where $f(z) := \max\{0, 1 - z\}$ It is worth noting that $\forall z$, $\partial f(z) \subseteq [-1, 0]$.

We consider Algorithm 1 for the rest of this section. The goal is to show that, if some significant (x, y) is misclassified by $\hat{w}$, a large number of good samples around (x, y) will push the program to further optimize over $\hat{w}$. That is, the clean samples in set S_P will have non-zero subgradients that contribute significantly to the updating step. We start with gradients from samples in S_P.

Lemma A.1. *Consider a sample (x, y) and its pancake $\mathcal{P}^{\tau}_{\hat{w}}(x, y)$ with $\tau \leq \gamma/2$. If (x, y) is misclassified by $\hat{w}$, i.e. $yx \cdot \hat{w} \leq 0$, then $\forall i \in S_P$, we have that $y_i x_i \cdot \hat{w} < \gamma$ and $\partial f(y_i x_i \cdot \frac{\hat{w}}{\gamma}) = \{-1\}$.*

Proof. Since $i \in S_P$, according to the dense pancake condition (Definition 2.1), we have that $y_i x_i \cdot \hat{w} \leq yx \cdot \hat{w} + \tau$. Since $yx \cdot \hat{w} \leq 0$ and $\tau \leq \gamma/2 < \gamma$, we have that $y_i x_i \cdot \hat{w} < \gamma$. We conclude that $\partial f(y_i x_i \cdot \frac{\hat{w}}{\gamma}) = \{-1\}$. □

The following lemma connects the gradient norm (Definition 2.2) with the actual subgradients from malicious samples, i.e. S_D. Specifically, the gradient norm serves as an upper bound for their weighted subgradients projecting onto any $w \in \mathcal{W}$.

Lemma A.2. *For any vector $w \in \mathcal{W}$, given any $g \in \partial_w \ell(w; q \circ S_D)$, it holds that $g \cdot w \leq \frac{1}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D)$.*

Proof. For S_D, we have that $\partial_w \ell(w; q \circ S_D) = \frac{1}{\gamma} \cdot \sum_{i \in S_D} \partial f(y_i x_i \cdot \frac{w}{\gamma}) \cdot y_i (q_i x_i)$. Since $\partial f(y_i x_i \cdot \frac{w}{\gamma}) \subseteq [-1, 0], \forall i$, we have that

$$\begin{aligned} g \cdot w &\leq \frac{1}{\gamma} \cdot \sup_{a_i \in [-1,0]} \left\langle \sum_{i \in S_D} a_i y_i (q_i x_i), w \right\rangle \\ &\leq \frac{1}{\gamma} \cdot \sup_{a_i \in [-1,1]} \sup_{w \in \mathcal{W}} \left\langle \sum_{i \in S_D} a_i q_i x_i, w \right\rangle \\ &= \frac{1}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D) \end{aligned}$$

due to Definition 2.2. □

The following lemma contributes to the analysis of subgradients $g \in \partial_w \ell(w; q \circ S)\big|_{w=\hat{w}}$ in the direction of w^*.

Lemma A.3. *Assume that Assumption 1.2 holds. Given any (x, y) and its pancake $\mathcal{P}^{\tau}_{\hat{w}}(x, y)$ with $\tau \leq \gamma/2$, if (x, y) is misclassified by $\hat{w}$, i.e. $yx \cdot \hat{w} \leq 0$, then $\forall g \in \partial_w \ell(w; q \circ S)\big|_{w=\hat{w}}$, we have*

$$g \cdot w^* \leq -\sum_{i \in S_P} q_i + \frac{1}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D)$$

Proof. This lemma can be proved in three steps. Namely, we will show that

1. For any $g_1 \in \partial_w \ell(w; q \circ S_P)\big|_{w=\hat{w}}$, it holds that $g_1 \cdot w^* \leq -\sum_{i \in S_P} q_i$;
2. For any $g_2 \in \partial_w \ell(w; q \circ S_{\bar{P}})\big|_{w=\hat{w}}$, it holds that $g_2 \cdot w^* \leq 0$;
3. For any $g_3 \in \partial_w \ell(w; q \circ S_D)\big|_{w=\hat{w}}$, it holds that $g_3 \cdot w^* \leq \frac{1}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D)$.

To show the first statement, it is known from Lemma A.1 that $\forall i \in S_P$, it holds that $\partial f(y_i x_i \cdot \frac{w}{\gamma}) = \{-1\}$. Hence, $g_1 \cdot w^* = -\frac{1}{\gamma} \cdot \sum_{i \in S_P} q_i y_i x_i \cdot w^*$. From Assumption 1.2, we know that $y_i x_i \cdot w^* \geq \gamma$. In conclusion, $g_1 \cdot w^* \leq -\sum_{i \in S_P} q_i$.

For the second part, note that $\forall i \in S_{\bar{P}}$, we still have $y_i x_i \cdot w^* \geq \gamma$ due to Assumption 1.2. Hence, $g_2 \cdot w^* \leq \frac{1}{\gamma} \cdot \sum_{i \in S_{\bar{P}}} \partial f(y_i x_i \cdot \frac{w}{\gamma}) \cdot q_i (y_i x_i \cdot w^*) \leq 0$ since $\partial f(\cdot)$ is always non-positive.

For the last statement, it follows from Lemma A.2 with $w^* \in \mathcal{W}$. □

The next step is to show that the (negative) weighted subgradients, i.e. $-\partial_w \ell(w; q \circ S)\big|_{w=\hat{w}}$, is always aligned with $w^* - \hat{w}$. This is enabled by some well-defined w', which is concluded in Lemma 4.3, and restated as follows.

Lemma A.4 (Restatment of Lemma 4.3)**.** *Consider Algorithm 1. Suppose Assumption 1.2 holds. For some $\hat{z} \in \frac{\partial \|w\|_1}{\partial w}\big|_{w=\hat{w}}$, define $w' = w^* - \hat{w} \langle w^*, \kappa \rangle$, where $\kappa = \frac{\lambda_1}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \hat{z} + \frac{\lambda_2}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \hat{w}$. Assume that $\langle w^*, \kappa \rangle > 0$. Then, we have the following holds. For any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and its pancake $\mathcal{P}^{\tau}_{\hat{w}}(x, y)$ with $\tau \leq \gamma/2$, if (x, y) is misclassified by $\hat{w}$, then for any $g \in \partial_w \ell(w; q \circ S)\big|_{w=\hat{w}}$, we have*

$$g \cdot w' \leq -\frac{1}{2} \sum_{i \in S_P} q_i + \frac{2}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D).$$

Proof. We will show the inequality in three steps:

1. For any $g_1 \in \partial_w \ell(w; q \circ S_P)\big|_{w=\hat{w}}$, it holds that $g_1 \cdot w' \leq -\frac{1}{2} \sum_{i \in S_P} q_i$.
2. For any $g_2 \in \partial_w \ell(w; q \circ S_{\bar{P}})\big|_{w=\hat{w}}$, it holds that $g_2 \cdot w' \leq 0$.
3. For any $g_3 \in \partial_w \ell(w; q \circ S_D)\big|_{w=\hat{w}}$, it holds that $g_3 \cdot w' \leq \frac{2}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D)$.

Here, we give a general proof for the case when $\lambda_1, \lambda_2 \neq 0$. When either λ_1, λ_2 equals to zero, the proof naturally follows.

To show the first statement, we consider S_P. Given any $\hat{w}$, we have that $\forall i$,

$$y_i x_i \cdot w' = y_i x_i \cdot (w^* - \hat{w} \langle w^*, \kappa \rangle) = y_i x_i \cdot w^* - y_i x_i \cdot \hat{w} \langle w^*, \kappa \rangle$$

The goal is to lower bound the above term with a positive multiplicative factor of γ. By Assumption 1.2, $\forall i \in S_P$, we have that $y_i x_i \cdot w^* \geq \gamma$. In addition, it holds that $y_i x_i \cdot \hat{w} \leq yx \cdot \hat{w} + \tau \leq 0 + \frac{\gamma}{2} = \frac{\gamma}{2}$ due to Definition 2.1 and pancake parameters. Hence, it remains to bound quantity $\langle w^*, \kappa \rangle$. This is easy from the observation that when the program outputs $\hat{w}$, it holds that $\|\hat{w}\|_1 \leq \sqrt{s}$ and $\|\hat{w}\|_2 \leq 1$. Therefore, $\langle w^*, \kappa \rangle = \frac{\lambda_1}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \langle w^*, \hat{z} \rangle + \frac{\lambda_2}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \langle w^*, \hat{w} \rangle \leq \frac{\lambda_1 \sqrt{s}}{\sqrt{s}\lambda_1 + \lambda_2} + \frac{\lambda_2}{\sqrt{s}\lambda_1 + \lambda_2} \leq 1$, where the second transition is due to that $\langle w^*, \hat{z} \rangle \leq \sup_{\|z\|_\infty \leq 1} \langle w^*, z \rangle = \|w^*\|_1 = 1$ and $\langle w^*, \hat{w} \rangle \leq \sup_{\|w\|_2 \leq 1} \langle w^*, w \rangle = \|w^*\|_2 = 1$. Hence, we have that $\forall i, y_i x_i \cdot w' \geq \gamma - \frac{\gamma}{2} \geq \frac{\gamma}{2}$.

On the other hand, Lemma A.1 imlies that $\partial f(y_i x_i \cdot \frac{w}{\gamma}) = \{-1\}$ for $i \in S_P$. Hence,

$$\begin{aligned} \partial_w \ell(w; q \circ S_P)\big|_{w=\hat{w}} \cdot w' &= \frac{1}{\gamma} \cdot \sum_{i \in S_P} q_i \partial f\left(y_i x_i \cdot \frac{w}{\gamma}\right) y_i x_i \cdot w' \\ &\subseteq \left(-\infty, -\frac{1}{2} \sum_{i \in S_P} q_i\right]. \end{aligned}$$

Then, we consider $i \in S_{\bar{P}} \subseteq S_C$. If $y_i x_i \cdot w' \geq 0$, then

$$\begin{aligned}\partial_w \ell(w; q \circ S_{\bar{P}})\big|_{w=\hat{w}} \cdot w' &= \frac{1}{\gamma} \cdot \sum_{i \in S_{\bar{P}}} q_i \partial f\left(y_i x_i \cdot \frac{w}{\gamma}\right) y_i x_i \cdot w' \\ &\subseteq (-\infty, 0], \end{aligned} \tag{4}$$

since $\partial f(\cdot)$ is non-positive. We then consider $y_i x_i \cdot w' < 0$, namely,

$$y_i x_i \cdot w' = y_i x_i \cdot w^* - y_i x_i \cdot \hat{w} \langle w^*, \kappa \rangle < 0.$$

By assumption, $\langle w^*, \kappa \rangle > 0$ (see Lemma A.6 for conditions for it to hold). Then, we have

$$y_i x_i \cdot \hat{w} > \frac{y_i x_i \cdot w^*}{\langle w^*, \kappa \rangle} \geq \gamma,$$

where the last inequality is due to Assumption 1.2. Rescaling by a factor of $\frac{1}{\gamma}$, we have that $y_i x_i \cdot \frac{\hat{w}}{\gamma} > 1$. Thus,

$$\partial_w \ell(w; q \circ S_{\bar{P}})\big|_{w=\hat{w}} \cdot w' = \frac{1}{\gamma} \cdot \sum_{i \in S_{\bar{P}}} q_i \partial f\left(y_i x_i \cdot \frac{\hat{w}}{\gamma}\right) y_i x_i \cdot w' = \{0\}.$$

Together with Eq.(4), we conclude that the second part holds.

The third part follows directly from Lemma A.2 by concluding that $\|w'\|_1 = \|w^* - \hat{w} \langle w^*, \kappa \rangle\|_1 \leq \|w^*\|_1 + \|\hat{w}\|_1 \cdot \langle w^*, \kappa \rangle \leq 2\sqrt{s}$ and $\|w'\|_2 \leq 2$. Namely, $\frac{w'}{2} \in \mathcal{W}$. The proof is complete. □

Lemma A.5 (Restatement of Lemma 4.2)**.** *Consider the set of subgradients $\partial_w \ell(w; q \circ S)\big|_{w=\hat{w}}$ and some element g in it. Given the optimization program Eq.(1) in Algorithm 1 and $\hat{w}$ returned by it, let $w' = w^* - \hat{w} \langle w^*, \kappa \rangle$. By choosing $\kappa = \frac{\lambda_1}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \hat{z} + \frac{\lambda_2}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \hat{w}$ for some $\lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 > 0$ and $\hat{z} \in \partial_w \|w\|_1 \big|_{w=\hat{w}}$, there exists $g \in \partial_w \ell(w; q \circ S)\big|_{w=\hat{w}}$ such that $g \cdot w' = 0$.*

Proof. The design of w' stems from the program Eq.(1), which we restate as follows

$$\begin{aligned} &\arg\min \ell(w; q \circ S) \\ &\text{s.t. } \|w\|_1 \leq \sqrt{s}, \\ &\quad \|w\|_2 \leq 1. \end{aligned}$$

Intuitively, w' is the component of $w^* - \hat{w}$ that is orthogonal to some subgradient of the weighted hinge loss. From the KKT condition, we know that for any output $\hat{w}$ of the program, there exist some g and $\hat{z}$ such that

$$g + \lambda_1 \hat{z} + \lambda_2 \hat{w} = 0$$

for some $\lambda_1, \lambda_2 \geq 0$. Taking this g and $\hat{z}$, consider the design of w', we have that

$$g \cdot w' = -(\lambda_1 \hat{z} + \lambda_2 \hat{w}) \cdot (w^* - \hat{w} \langle w^*, \kappa \rangle)$$

To show $g \cdot w' = 0$, it suffices to show that

$$\begin{aligned} -(\lambda_1 \hat{z} + \lambda_2 \hat{w}) \cdot w^* + (\lambda_1 \hat{z} + \lambda_2 \hat{w}) \cdot \hat{w} \langle w^*, \kappa \rangle = 0, \\ (\lambda_1 \hat{z} + \lambda_2 \hat{w}) \cdot \hat{w} \langle w^*, \kappa \rangle = (\lambda_1 \hat{z} + \lambda_2 \hat{w}) \cdot w^*. \end{aligned}$$

Assuming that both $\lambda_1, \lambda_2 \neq 0$ (either one equals zero would be a special case and follows easily from this proof; both equal zero would lead to an interior), then by the complementary slackness, the boundary conditions are achieved, i.e. $\hat{z} \cdot \hat{w} = \|\hat{w}\|_1 = \sqrt{s}$ and $\hat{w} \cdot \hat{w} = \|\hat{w}\|_2 = 1$. Hence, the equality to be shown becomes

$$(\lambda_1 \sqrt{s} + \lambda_2) \langle w^*, \kappa \rangle = \langle \lambda_1 \hat{z} + \lambda_2 \hat{w}, w^* \rangle .$$

It suffices to choose $\kappa = \frac{\lambda_1 \hat{z} + \lambda_2 \hat{w}}{\lambda_1 \sqrt{s} + \lambda_2}$. The proof is complete. □

Lemma A.6. *Consider that the weight from S_P is prevalent (Lemma A.7). Given program Eq.(1) in Algorithm 1 and $\hat{w}$ returned by it, let $w' = w^* - \hat{w} \langle w^*, \kappa \rangle$ and $\kappa = \frac{\lambda_1}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \hat{z} + \frac{\lambda_2}{\sqrt{s}\lambda_1 + \lambda_2} \cdot \hat{w}$ for some $\lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 > 0$ and $\hat{z} \in \partial_w \|w\|_1 \big|_{w=\hat{w}}$. Then, it holds that $\langle w^*, \kappa \rangle > 0$.*

Proof. According to Lemma A.7, we have that

$$\sum_{i \in S_P} q_i > \frac{1}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D).$$

Together with Lemma A.3, we conclude that $\forall g,\, g \cdot w^* \leq -\sum_{i \in S_P} q_i + \frac{1}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D) < 0$. Hence,

$$(\lambda_1 \sqrt{s} + \lambda_2) \langle w^*, \kappa \rangle = \langle \lambda_1 \hat{z} + \lambda_2 \hat{w}, w^* \rangle = -g \cdot w^* > 0,$$

implying that $\langle w^*, \kappa \rangle > 0$ because $(\lambda_1 \sqrt{s} + \lambda_2) > 0$. The proof is complete. □

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1. We will show by contradition. Let us assume that (x, y) is misclassified by $\hat{w}$.

Case 1. Suppose that $\hat{w}$ is an interior of the constraint set of program (1). From Lemma A.3, we know that $\forall g \in \partial_w \ell(w; q \circ S) \mid_{w=\hat{w}}$, it holds that

$$g \cdot w^* \leq -\sum_{i \in S_P} q_i + \frac{1}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D).$$

From Eq. (2), we conclude that $\forall g, g \cdot w^* < 0$. However, from the optimility condition of program (1), there exists some $g = 0$ such that $g \cdot w^* = 0$, leading to a contradition.

Case 2. Suppose that $\hat{w}$ is on the boundary of the constraint set. From Lemma 4.3, we know that $\forall g$, it holds that

$$g \cdot w' \leq -\frac{1}{2} \sum_{i=1}^{n} q_i + \frac{2}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D).$$

Again from Eq. (2), we conclude that $\forall g, g \cdot w' < 0$. However, by the design of w' and Lemma 4.2, there exists some g such that $g \cdot w' = 0$, leading to contradiction.

Hence, (x, y) is not misclassfied by $\hat{w}$. □

A.2 Analysis of density and noise rate conditions

From Section 4.1, it is evident that the essential condition for an (x, y) to be correctly classified by the optimum $\hat{w}$ is that the weight of the clean samples around it is large enough comparing to the gradient norm from the malicious samples. This condition is given in Eq. (2). The following lemma ensures that the condition is satisfied.

Lemma A.7. *Consider Algorithm 2 and its returned value q. Assume that Program* (1) *is feasible. If (x, y) is a sample such that the pancake $\mathcal{P}^{\tau}_{\hat{w}}(x, y)$ is ρ-dense with respect to S_C with $\rho \geq 16\left(\frac{1}{\gamma\sqrt{d}} + \frac{r}{\gamma} + 1\right)\sqrt{\eta_0}$, then it holds that*

$$\sum_{i \in S_C \cap \mathcal{P}^{\tau}_{\hat{w}}(x,y)} q_i > \frac{4}{\gamma} \cdot \mathrm{GradNorm}(q \circ S_D). \tag{5}$$

Proof. We will prove the lemma by lower bounding the left hand side and upper bounding the right hand side respectively. That is, we will show that $\sum_{i \in S_C \cap \mathcal{P}^{\tau}_{\hat{w}}(x,y)} q_i = \sum_{i \in S_P} q_i > (\rho - 2\xi)\,|S|$; and $\mathrm{GradNorm}(q \circ S_D) \leq \bar{\sigma} \cdot \sqrt{\xi} \cdot |S|$. Then, given these two conditions, Eq. (5) follows from that $\rho \geq 16\Big(\frac{1}{\gamma\sqrt{d}} + \frac{r}{\gamma} + 1\Big)\sqrt{\eta_0}$ and $\xi = 2\eta_0$. In more details, it remains to show that

$$(\rho - 2\xi)\,|S| > \frac{4}{\gamma} \cdot \bar{\sigma} \cdot \sqrt{\xi} \cdot |S|\,,$$

$$\rho - 2\xi > \frac{4\bar{\sigma}}{\gamma} \cdot \sqrt{\xi}.$$

Since $\xi = 2\eta_0$ and $\bar{\sigma} = 2 \cdot \sqrt{\left(\frac{1}{d} + r^2\right)}$, it remains to show

$$\rho > 4\eta_0 + \frac{8\sqrt{2}}{\gamma} \cdot \sqrt{\frac{1}{d} + r^2} \cdot \sqrt{\eta_0}.$$

It suffices to choose any $\rho \geq 16\left(\frac{1}{\gamma\sqrt{d}} + \frac{r}{\gamma} + 1\right)\sqrt{\eta_0}$, due to that $(a+b) \geq \sqrt{a^2+b^2}$ for $a, b \geq 0$ and $\sqrt{\eta_0} \geq \eta_0$.

Lower bound. When Algorithm 2 outputs q, it holds that $\sum_{i \in S} q_i \geq (1-\xi)\,|S|$. In addition, we have that $|S_P| \geq \rho\,|S_C|$ and $|S_{\bar{P}}| \leq (1-\rho)\,|S_C|$, given that the pancake $\mathcal{P}^{\tau}_{\hat{w}}(x, y)$ is ρ-dense w.r.t. S_C. Since $S = S_P \cup S_{\bar{P}} \cup S_D$, it suffices to bound $\sum_{i \in S_{\bar{P}}} q_i \leq (1-\rho)\,|S_C| \leq (1-\rho)\,|S|$ and $\sum_{i \in S_D} q_i \leq |S_D| \leq \xi\,|S|$ where ξ is the upper bound on empirical noise rate. Hence, $\sum_{i \in S_P} q_i = \sum_{i \in S} - \sum_{i \in S_{\bar{P}}} q_i - \sum_{S_D} q_i > (1 - \xi - (1-\rho) - \xi)\,|S| = (\rho - 2\xi)\,|S|$.

Upper bound. Since that q is feasible, we have

$$\sum_{i \in S_D} q_i (w \cdot x_i)^2 \leq \sum_{i \in S} q_i \cdot (w \cdot x_i)^2 \leq \bar{\sigma}^2 \cdot |S|\,.$$

where the first transition is due to that all terms inside the summation is non-negative, and the second transition is due to Lemma A.8. Therefore,

$$\sqrt{\sup_{w \in \mathcal{W}} \sum_{i \in S_D} q_i (w \cdot x_i)^2} \leq \bar{\sigma}\sqrt{|S|},$$

Then,

$$\begin{aligned}
\mathrm{GradNorm}(q \circ S_D) &\leq \sqrt{|S_D|} \cdot \sqrt{\sup_{w \in \mathcal{W}} \sum_{i \in S_D} q_i (w \cdot x_i)^2} \\
&\leq \sqrt{|S_D|} \cdot \bar{\sigma}\sqrt{|S|} \\
&\leq \bar{\sigma} \cdot \sqrt{\xi} \cdot |S|
\end{aligned}$$

where the first transition follows from Lemma A.9. The proof is complete. □

Lemma A.8. *Consider Algorithm 2. If it successfully returns a weight vector q, then for any $w \in \mathcal{W}$, it holds that $\frac{1}{|S|}\sum_{i \in S} q_i \cdot (w \cdot x_i)^2 \leq \bar{\sigma}^2$.*

Proof. By simple calculation, $\sum_{i \in S} q_i \cdot (w \cdot x_i)^2 = \sum_{i \in S} q_i \cdot x_i^\top w w^\top x_i = \sum_{i \in S} q_i \cdot x_i^\top W x_i$ where $W := ww^\top$. Since $\|w\|_1 \leq \sqrt{s}$, we have that $\|W\|_1 = \sum_{k,l \in [d]} |W_{k,l}| = \sum_{k,l \in [d]} |w_k \cdot w_l| \leq \sum_{k \in [d]} |w_k| \cdot \|w\|_1 \leq \|w\|_1 \cdot \|w\|_1 \leq s$. Since $\|w\|_2 \leq 1$, we have that $\|W\|_* \leq 1$. Hence, $W \in \{H : H \preceq 0, \|H\|_1 \leq s, \|H\|_* \leq 1\}$, and $\sum_{i \in S} q_i \cdot (w \cdot x_i)^2 \leq \sup_{H \preceq 0, \|H\|_1 \leq s, \|H\|_* \leq 1} \frac{1}{|S|}\sum_{i \in S} q_i x_i^\top H x_i \leq \bar{\sigma}^2$. The proof is complete. □

Lemma A.9. *Let S' be any sample set and let $q = (q_1, \ldots, q_{|S'|})$ with all $q_i \in [0, 1]$. Then, $\mathrm{GradNorm}(q \circ S') \leq \sqrt{|S'|}\sqrt{\sup_{w \in \mathcal{W}} \sum_{i \in S'} q_i (w \cdot x_i)^2}$.*

Proof. From the definition of $\mathrm{GradNorm}$ (Definition 2.2), we have that

$$\begin{aligned}
\mathrm{GradNorm}(q \circ S') &= \sup_{a \in [-1,1]^{|S'|}} \sup_{w \in \mathcal{W}} \left\langle \sum_{i \in S'} a_i q_i x_i, w \right\rangle \\
&= \sup_{a_i \in [-1,1], i \in S'} \sup_{w \in \mathcal{W}} \left\langle \sum_{i \in S'} a_i q_i x_i, w \right\rangle \\
&= \sup_{a_i \in [-1,1], i \in S'} \sup_{w \in \mathcal{W}} \sum_{i \in S'} a_i q_i \langle x_i, w \rangle \\
&\leq \sup_{a_i \in [-1,1], i \in S'} \sup_{w \in \mathcal{W}} \sqrt{\sum_{i \in S'} a_i^2} \sqrt{\sum_{i \in S'} (q_i \langle x_i, w \rangle)^2} \\
&\leq \sqrt{S'} \sqrt{\sup_{w \in \mathcal{W}} \sum_{i \in S'} q_i (x_i \cdot w)^2}
\end{aligned}$$

where the fourth transition is due to Cauchy-Schwartz inequality and the last transition is due to that $\forall i$, $a_i^2 \leq 1$ and $q_i^2 \leq q_i$. □

A.3 Analysis for deterministic results

We note that this section follows directly from prior work of [She25]. However, we must include it here as we have re-prove every theorems and lemmas in the new problem setting. See Section 2 for more details and we encourage interested readers to compare our problem setting with [She25].

Theorem A.10. *Given that Assumption 1.2 holds with some $\gamma > 0$, $(S_C, \mathcal{D})$ satisfies (τ, ρ, ϵ)-dense pancake condition with $\tau \leq \frac{\gamma}{2}$, the semidefinite program in Algorithm 2 is feasible, and*

$$\rho \geq 16 \left(\frac{1}{\gamma \sqrt{d}} + \frac{r}{\gamma} + 1 \right) \sqrt{\eta_0}$$

then Algorithm 1 returns $\hat{w}$ such that $\mathrm{err}_D(\hat{w}) \leq \epsilon$.

Proof. Since $(S_C, \mathcal{D})$ satisfies (τ, ρ, ϵ)-dense pancake condition with $\tau \leq \frac{\gamma}{2}$ with respect to any w with $\|w\|_1 \leq \sqrt{s}$ and $\|w\|_2 \leq 1$ (Definition 2.1). That is, for any unit vector w such that $\|w\|_1 \leq \sqrt{s}$, $\Pr_{(x,y)\sim\mathcal{D}}(\mathcal{P}_w^\tau(x,y)$ is ρ-dense w.r.t. $S_C) \geq 1 - \epsilon$.

Apparantly, this inequality holds for $\hat{w}$ as well due to the output condition in Eq. (1). Hence, for a $(1-\epsilon)$ probability mass in $\mathcal{D}$, (x, y) has a ρ-dense pancake $\mathcal{P}_{\hat{w}}^\tau(x, y)$. Theorem 4.1 together with Lemma A.7 implies that such (x, y) will not be misclassified by $\hat{w}$. For the remaining ϵ probability mass, they might be misclassified by $\hat{w}$ and hence resulting in an error rate less or equal to ϵ. □

B Analysis on Sample Complexity

Lemma B.1. *Consider a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ satisfying Assumption 1.3 and 1.2 with $\gamma > \tau$ for some $\tau = 2\sigma \cdot \left(\log \frac{1}{\beta} + 1 \right)$. Then, for any $\beta > 0$, $\mathcal{D}$ satisfies $(\tau, \frac{1-\beta}{k}, \beta)$-dense pancake condition.*

Proof. Consider a distribution $\mathcal{D}_1$ that is logconcave with zero-mean and covariance matrix $\Sigma \preceq \sigma^2 I_d$. For any unit vector w, the projection of $\mathcal{D}_1$ onto w is still logconcave. The statistical property of logconcave random variable implies that

$$\Pr_{x \sim \mathcal{D}_1} \left(|w \cdot x| \leq \sigma \cdot \left(\log \frac{1}{\beta} + 1 \right) \right) \geq 1 - \beta.$$

It then immediately follows that, for any logconcave distribution $\mathcal{D}_j$ with mean μ_j and covariance matrix $\Sigma_j \prec \sigma^2 I_d$, we have

$$\Pr_{x \sim \mathcal{D}_j} \left(|w \cdot (x - \mu_j)| \leq \sigma \cdot \left(\log \frac{1}{\beta} + 1 \right) \right) \geq 1 - \beta.$$

Note that this condition also holds for w^*. For any $x \sim \mathcal{D}_j$ such that $|w^* \cdot (x - \mu_j)| \le \sigma \cdot (\log 1/\beta + 1) =: \tau/2 < \gamma$, it holds that $\mathrm{sign}(x \cdot w^*) = \mathrm{sign}(\mu_j \cdot w^*)$.

That is,

$$\Pr_{x \sim \mathcal{D}_j} \left(|(\mathrm{sign}(x \cdot w^*)x - \mathrm{sign}(\mu_j \cdot w^*)\mu_j) \cdot w| \le \tau/2 \right) \ge 1 - \beta.$$

In addition, denoted by $(\mathcal{D}_j, w^*)$ the distribution where the instances is drawn from $\mathcal{D}$ and labeled by w^*, then

$$\Pr_{(x,y) \sim (\mathcal{D}_j, w^*)} \left(\Pr_{(x',y') \sim (\mathcal{D}_j, w^*)} \left(|(y'x' - yx) \cdot w| \le \tau \right) \ge 1 - \beta \right) \ge 1 - \beta.$$

Since $\mathcal{D}_{\mathcal{X}} = \sum_{j=1}^{k} \frac{1}{k} \mathcal{D}_j$, each distribution density of $\mathcal{D}_j$ is attenuated by a multiplicative factor of $\frac{1}{k}$. Hence, we have that

$$\Pr_{(x,y) \sim (\frac{1}{k}\mathcal{D}_j, w^*)} \left(\Pr_{(x',y') \sim (\frac{1}{k}\mathcal{D}_j, w^*)} \left(|(y'x' - yx) \cdot w| \le \tau \right) \ge \frac{1-\beta}{k} \right) \ge \frac{1-\beta}{k}.$$

Summing the probability over all $j \in [k]$, we have that

$$\Pr_{(x,y) \sim \mathcal{D}} \left(\Pr_{(x',y') \sim \mathcal{D}} \left(|(y'x' - yx) \cdot w| \le \tau \right) \ge \frac{1-\beta}{k} \right) \ge 1 - \beta.$$

Note that if this condition holds for all unit vectors, then it also holds for all $w \in \mathcal{W}$. Hence, $\mathcal{D}$ satisfies $(\tau, \frac{1-\beta}{k}, \beta)$-dense pancake condition. □

The following theorem is an extension of Theorem 18 in [Tal20]. Part of the proof follows from theirs and is included for completeness.

Theorem B.2 (Restatement of Theorem 4.4)**.** *Consider a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ satisfying Assumption 1.3 and 1.2 with $\gamma > 2\tau$, where $\tau = 2\sigma \cdot \left(\log \frac{1}{\beta} + 1\right)$. It holds with probability at least $1 - \delta'$ over the draw of $|S_C| \ge \frac{2}{\rho} \cdot \left(C_0 s \cdot \log^4 d + \log \frac{1}{\delta'\beta'}\right)$ samples, $(S_C, \mathcal{D})$ satisfies $(2\tau, \rho, \beta + \beta')$-dense pancake condition, where $C_0 > 0$ is an absolute constant and $\rho = \frac{1-\beta}{2k}$.*

Proof. Recall that $\rho = \frac{1-\beta}{2k}$. From Lemma B.1, we have that $\mathcal{D}$ satisfies $(\tau, 2\rho, \beta)$-dense pancake condition. Follow the first part of the proof for Theorem 18 in [Tal20], we have that

$$\Pr_{S \sim \mathcal{D}^n} \left(\Pr_{(x,y) \sim \mathcal{D}} \left(\frac{1}{n} \sum_{i \in S} \mathbb{1} \left(|(y_i x_i - yx) \cdot w| \le \tau \right) < \rho \right) > \beta' + \beta \right) \le \frac{1}{\beta'} \cdot \exp\left(-\frac{\rho n}{2}\right) \tag{6}$$

holds for any fixed $w \in \mathcal{W} := \{w : \|w\|_1 \le \sqrt{s}, \|w\|_2 \le 1\}$. We say that a sample S is (τ, ρ, β)-good for some w under $\mathcal{D}$ if $\Pr_{(x,y) \sim \mathcal{D}} \left(\frac{1}{n} \sum_{i \in S} \mathbb{1} \left(|(y_i x_i - yx) \cdot w| \le \tau \right) < \rho \right) \le \beta$. Eq. (6) implies that a sample S from $\mathcal{D}^n$ is $(\tau, \rho, \beta' + \beta)$-good with respect to a fixed $w \in \mathcal{W}$, except with probability $\frac{1}{\beta'} \cdot \exp(-\frac{\rho n}{2})$.

Next, we want to show that with a slight relaxation on the pancake thickness, S is not only good for a fixed w but also food for a set of w' close by. Note that for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, if $i \in S \cap \mathcal{P}_w^{\tau}(x, y)$, we have that $|(y_i x_i - yx) \cdot w| \le \tau$. Consider pancake $\mathcal{P}_{w'}^{2\tau}(x, y)$. We know that $S \cap \mathcal{P}_w^{\tau}(x, y) \subseteq S \cap \mathcal{P}_{w'}^{2\tau}(x, y)$ if $\forall i \in S$,

$$|(y_i x_i - yx) \cdot (w' - w)| \le \tau, \tag{7}$$

due to the triangle inequality. It then remains to construct a τ-net of the inner product function space $F := \{f_w(\bar{x}) = w \cdot \bar{x} : \|w\|_1 \le \sqrt{s}, \|\bar{x}\|_\infty \le 2\sigma \cdot \log(d/\delta')\}$ (recall that $\forall i$, $\|y_i x_i - yx\|_\infty \le 2\sigma \cdot \log(d/\delta')$ for (x_i, y_i) and (x, y) from the same component, except for probability δ') and then take a union bound. According to Corollary 5 in [Zha02] (which we put in Lemma F.6 for reference), we have that the ∞-norm covering number of F is upper bounded by $\log \mathcal{N}_\infty(F, \tau, n) \le O\left(\frac{\sqrt{s}^2 \cdot (\sigma \cdot \log(d/\delta'))^2 (2 + \log d)}{\tau^2} \cdot \log(\frac{\sqrt{s} \cdot \sigma \cdot \log(d/\delta')}{\tau} \cdot n)\right) = O(s \log^4 d)$, due to the setting of σ and τ.

Note that we need an ∞-norm covering number because Eq. (7) must hold for all $i \in S$ (see more details in [Zha02]). As a result, we have that

$$\Pr_{S\sim\mathcal{D}^n}\left(\exists w\in\mathcal{W}:\Pr_{(x,y)\sim\mathcal{D}}\left(\frac{1}{n}\sum_{i\in S}\mathbb{1}\left(|(y_ix_i-yx)\cdot w|\le 2\tau\right)<\rho\right)>\beta'+\beta\right)$$
$$\le\frac{1}{\beta'}\cdot\exp\left(C_0\cdot s\log^4 d-\frac{\rho n}{2}\right)$$

where $C_0>0$ is some constant. Bounding the above failure probabililty with δ', it suffices to choose

$$n\ge\frac{2}{\rho}\cdot\left(C_0 s\cdot\log^4 d+\log\frac{1}{\delta'\beta'}\right). \tag{8}$$

□

Lemma B.3 (Restatement of Lemma 4.7)**.** *If we draw a sample S from EX with size $|S|\ge\frac{3}{\eta_0}\cdot\log\frac{1}{\delta'}$, then it is guaranteed with probability $1-\delta'$ that $|S_C|\ge(1-2\eta_0)|S|$ and $|S_D|\le 2\eta_0|S|$.*

Proof. It is known that every time the algorithm request a sample from EX, with probability $\eta\le\eta_0$, it will be a malicious sample $(x,y)\in S_D$. Hence, by Chernoff bound (Lemma F.2) we have that

$$\Pr\left(|S_D|\ge 2\eta_0|S|\right)\le\exp\left(-\frac{\eta_0|S|}{3}\right).$$

Let the probability bounded by δ', we conclude that it suffices to choose $|S|\ge\frac{3}{\eta_0}\cdot\log\frac{1}{\delta'}$. □

The following proof follows the structure from that of Proposition 27 in [SZ21]. However, we note that this proof has been adapted to the mixture of logconcaves with non-zero centers and covariance $\sigma^2 I_d$. In fact, Lemma C.3 and C.4 in Section C are strong evidences for that learning under mixture of logconcaves is significantly different from that under a single isotropic logconcave. For example, most of the previous results on isotropic logconcave only work for learning of homogeneous halfspaces.

Lemma B.4 (Restatement of Lemma 4.5)**.** *For any distribution $\mathcal{D}_\mathcal{X}$ satisfying Assumption 1.3, let $S_C=\{x_1,\dots,x_{|S_C|}\}$ be a set of i.i.d. unlabeled instances drawn from $\mathcal{D}_\mathcal{X}$. Denote $G(H):=\frac{1}{|S_C|}\sum_{i\in S_C}x_i^\top Hx_i-\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[x^\top Hx]$. Then with probability $1-\delta$,*

$$\sup_{H\in\mathcal{M}}|G(H)|\le 2\left(\frac{1}{d}+r^2\right)$$

given that $|S_C|\ge\Theta\left(s^2\cdot\log^5 d\log^2\frac{1}{\delta}\right)$.

Proof. Recall that $\mathcal{M}:=\{H:H\succeq 0,\|H\|_1\le s,\|H\|_*\le 1\}$. We aim to use Adamczak's bound (Lemma F.3) for a function class $\mathcal{F}:=\{x\mapsto x^\top Hx:H\in\mathcal{M}\}$. For the ease of presentation, we further let that $|S_C|=n$ throughout this lemma.

We find a function $F(x)$ that upper bounds $f(x)$ for any $f\in\mathcal{F}$. Given that $|f(x)|=\left|x^\top Hx\right|=\left|\sum_{k,l}H_{k,l}x^{(k)}x^{(l)}\right|\le\left(\sum_{k,l}|H_{k,l}|\right)\cdot\|x\|_\infty^2\le s\|x\|_\infty^2$, we define $F(x)=s\|x\|_\infty^2$.

We first bound the Orlicz norm for $\max_{1\le i\le n}F(x_i)$. According to Lemma C.2, we have that

$$\left\|\max_{1\le i\le n}F(x_i)\right\|_{\psi_{0.5}}=\left\|\sqrt{\max_{1\le i\le n}F(x_i)}\right\|_{\psi_1}^2.$$

Since $F(x_i)=s\|x_i\|_\infty^2$, $\forall 1\le i\le n$, we have that $\max_{1\le i\le n}F(x_i)=s\cdot\max_{1\le i\le n}\|x_i\|_\infty^2$. Moreover, $\sqrt{\max_{1\le i\le n}F(x_i)}=\sqrt{s}\cdot\max_{1\le i\le n}\|x_i\|_\infty$. Hence, Lemma C.2 and C.3 implies that

$$\begin{aligned}\left\|\sqrt{\max_{1\le i\le n}F(x_i)}\right\|_{\psi_1}&=\left\|\sqrt{s}\cdot\max_{1\le i\le n}\|x_i\|_\infty\right\|_{\psi_1}\\&=\sqrt{s}\cdot\left\|\max_{1\le i\le n}\|x_i\|_\infty\right\|_{\psi_1}\\&\le C_2\cdot\sqrt{s}\sigma\left(\frac{r}{\sigma}+1\right)\log(nd).\end{aligned}$$

Therefore, we conclude that

$$\left\| \max_{1\le i\le n} F(x_i) \right\|_{\psi_{0.5}} \le \left(C_2 \cdot \sqrt{s}\sigma \left(\frac{r}{\sigma} + 1 \right) \log(nd) \right)^2. \tag{9}$$

Next, we bound the second term in Adamczak's bound, which includes term $\sup_{f\in\mathcal{F}} \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[(f(x))^2]$. It is equivalent that taking the supremum over $f \in \mathcal{F}$ and taking it over $H \in \mathcal{M}$. Given the definition of $F(x)$, it is easy to see that $(f(x))^2 \le (F(x))^2 = s^2 \|x\|_\infty^4, \forall f \in \mathcal{F}$. Applying Lemma C.2, we have that

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\|x\|_\infty^4] = \| \|x\|_\infty \|_4^4 \le \| \|x\|_\infty \|_{\psi_4}^4 \le \left(4! \, \| \|x\|_\infty \|_{\psi_1} \right)^4,$$

together with the Orlicz norm bound in Lemma C.4, we have that

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[(f(x))^2] &\le \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}} \left[(F(x))^2 \right] \\
&= s^2 \cdot \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}} \left[\|x\|_\infty^4 \right] \\
&\le s^2 \left(4! \, \| \|x\|_\infty \|_{\psi_1} \right)^4 \\
&\le (4! C_2)^4 \cdot s^2 \sigma^4 \left(\frac{r}{\sigma} + 1 \right)^4 \log^4(d).
\end{aligned} \tag{10}$$

The final step is to use Rademacher analysis to bound $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}^n} \left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}(f(x)) \right| \right)$. Let $a = \{a_1, \ldots, a_n\}$ where a_i's are indepedent Rademacher variables. By Lemma 26.2 of [SB14], we have that

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}^n} \left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}(f(x)) \right| \right) \\
&\le \frac{2}{n} \mathbb{E}_{S,a} \left(\sup_{f\in\mathcal{F}} \left| \sum_{i=1}^n a_i f(x_i) \right| \right) \\
&\le \frac{2}{n} \mathbb{E}_{S,a} \left(\sup_{\|H\|_1 \le s} \left| \sum_{i=1}^n a_i x_i^\top H x_i \right| \right)
\end{aligned}$$

Fix a sample set S and consider the randomness in a. Here we will use vectorization for a matrix H, denoted as $\mathrm{vec}(H)$. We have that

$$\begin{aligned}
&\mathbb{E}_a \left(\sup_{\|H\|_1 \le s} \left| \sum_{i=1}^n a_i x_i^\top H x_i \right| \right) \\
&\le \mathbb{E}_a \left(\sup_{H:\|\mathrm{vec}(H)\|_1 \le s} \left| \sum_{i=1}^n a_i \left\langle \mathrm{vec}(H), \mathrm{vec}(x_i x_i^\top) \right\rangle \right| \right) \\
&\le s\sqrt{2n\log(2d^2)} \cdot \max_{1\le i\le n} \left\| \mathrm{vec}(x_i x_i^\top) \right\|_\infty \\
&\le s\sqrt{2n\log(2d^2)} \cdot \max_{1\le i\le n} \|x_i\|_\infty^2,
\end{aligned}$$

where the second inequality is due to Lemma F.4. We then consider expectation over the choice of S

$$\begin{aligned}
&\mathbb{E}_{S,a} \left(\sup_{\|H\|_1 \le s} \left| \sum_{i=1}^n a_i x_i^\top H x_i \right| \right) \\
&\le s\sqrt{2n\log(2d^2)} \cdot \mathbb{E}_S \left(\max_{1\le i\le n} \|x_i\|_\infty^2 \right) \\
&\le s\sqrt{4n\log(2d)} \cdot \left(2! \cdot \left\| \max_{1\le i\le n} \|x_i\|_\infty \right\|_{\psi_1} \right)^2 \\
&\le s\sqrt{4n\log(2d)} \cdot 16 \cdot C_2^2 \cdot \sigma^2 \left(\frac{r}{\sigma} + 1 \right)^2 \log^2(nd),
\end{aligned}$$

where the second inequality is due to Part 2 of Lemma C.2 and the third inequality is due to Lemma C.3.

As a result, we have that

$$\mathbb{E}_{S,a}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}a_i f(x_i)\right|\right) \leq C_3\cdot\sqrt{n\log d}\cdot s\sigma^2\left(\frac{r}{\sigma}+1\right)^2\log^2(nd)$$

for some constant $C_3 > 0$. In addition,

$$\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}^n}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(x_i)-\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}(f(x))\right|\right) \leq \frac{C_3\cdot\sqrt{\log d}}{\sqrt{n}}\cdot s\sigma^2\left(\frac{r}{\sigma}+1\right)^2\log^2(nd). \tag{11}$$

Combining all three steps, i.e. Eq. (11), (10), (9)

$$\sup_{H\in\mathcal{M}}|G(H)| \leq C_4\cdot\left(s\sigma^2\left(\frac{r}{\sigma}+1\right)^2\log^2(nd)\cdot\left(\sqrt{\frac{\log d}{n}}+\sqrt{\frac{\log\frac{1}{\delta}}{n}}+\frac{\log^2\frac{1}{\delta}}{n}\right)\right).$$

Let this quantity be upper bounded by some $t > 0$, then we get that it suffices choose n as follows,

$$n = \Theta\left(\frac{1}{t^2}\cdot s^2\sigma^4\left(\frac{r}{\sigma}+1\right)^4\log^4(nd)\cdot\left(\log d+\log^2\frac{1}{\delta}\right)\right).$$

Let $t = (\sigma + r)^2$, and since $n = O(d)$, it suffices to choose

$$n = \Theta\left(s^2\cdot\log^5(nd)\log^2\frac{1}{\delta}\right) = \Theta\left(s^2\cdot\log^5(d)\log^2\frac{1}{\delta}\right). \tag{12}$$

Recall that $\sigma = \frac{1}{\sqrt{d}}$, and so $t = \left(\frac{1}{\sqrt{d}}+r\right)^2 \leq 2\left(\frac{1}{d}+r^2\right)$.

□

The following lemma shows that for the underlying marginal distribution, which is a uniform mixture of k logconcaves, the variance with respect to matrix $H \in \mathcal{M}$ is upper bounded. This is crucial in setting the value of parameter $\bar{\sigma}$ in Algorithm 2.

Lemma B.5 (Restatement of Lemma 4.6)**.** *Given distribution $\mathcal{D}_{\mathcal{X}}$ satisfying Assumption 1.3, it holds that* $\sup_{H\in\mathcal{M}}\mathbb{E}_{X\sim\mathcal{D}_{\mathcal{X}}}(x^\top Hx) \leq 2\left(\frac{1}{d}+r^2\right)$.

Proof. Recall that $\mathcal{M} := \{H \succeq 0, \|H\|_1 \leq s, \|H\|_* \leq 1\}$, which is the constraint set of program in Algorithm 2. Consider that H is a positive semidefinite matrix with trace norm at most 1, we can eigendecomposite $H = \sum_{i=1}^d \lambda_i v_i v_i^\top$ where $\lambda_i \geq 0$ and $\sum_{i=1}^d \lambda_i \leq 1$. Hence, for any $H \in \mathcal{M}$,

$$x^\top Hx = \sum_{i=1}^{d}\lambda_i x^\top v_i v_i^\top x = \sum_{i=1}^{d}\lambda_i (v_i^\top x)^2 \leq \sum_{i=1}^{d}\lambda_i\left[2(v_i^\top(x-\mu_j))^2+2(v_i^\top\mu_j)^2\right]$$

where μ_j is the mean of the j-th logconcave distribution to which x belongs. Without loss of generatliy, we fix a j and the analysis follows for all $j \in [k]$. Then, for any $H \in \mathcal{M}$,

$$\begin{aligned}
\mathbb{E}_{x\sim\mathcal{D}_j}[x^\top Hx] &\le \mathbb{E}_{x\sim\mathcal{D}_j} \sum_{i=1}^{d} \lambda_i \left[2(v_i^\top (x-\mu_j))^2 + 2(v_i^\top \mu_j)^2\right] \\
&= 2\sum_{i=1}^{d} \lambda_i \left(v_i^\top \mathbb{E}_{x\sim\mathcal{D}_j}[(x-\mu_j)(x-\mu_j)^\top] v_i + (v_i^\top \mu_j)^2\right) \\
&\le 2\sum_{i=1}^{d} \lambda_i \left(\frac{1}{d}\left\|v_i\right\|_2^2 + \left\|\mu_j\right\|_2^2\right) \\
&\le 2\left(\frac{1}{d} + r^2\right).
\end{aligned}$$

By linearity of the expected values, we conclude that $\sup_{H\in\mathcal{M}} \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}(x^\top Hx) \le 2\left(\frac{1}{d} + r^2\right)$. The proof is complete. □

Proposition B.6 (Restatement of Proposition 4.8)**.** *Given that Asssumption 1.3 holds, if we draw a sample set S from EX with size $|S| \ge \Omega(s^2 \cdot \log^5 d \log^2 \frac{1}{\delta'} + \frac{1}{\eta_0} \cdot \log \frac{1}{\delta'})$, with probability at least $1 - 2\delta'$, the semidefinite program in Algorithm 2 is feasible.*

Proof. To show that the program is feasible, we want to show that there exists a q^* that satisfies all three constraints. From Lemma 4.7, we know that $|S_C| \ge (1-2\eta_0)|S|$ with probability $1-\delta'$ if $|S| \ge \frac{3}{\eta_0} \cdot \log \frac{1}{\delta'}$. From Lemma 4.5 and 4.6, we know that if $|S_C| \ge \Omega\left(s^2 \cdot \log^5 d \log^2 \frac{1}{\delta'}\right)$, it holds with probability $1-\delta'$,

$$\sup_{H\in\mathcal{M}} \frac{1}{|S_C|} \sum_{i\in S_C} x_i^\top H x_i \le 4\left(\frac{1}{d} + r^2\right) = \bar{\sigma}^2.$$

Let $\eta_0 \le \frac{1}{4}$, we have that $|S_C| \ge \frac{1}{2}|S|$. Taking the union bound, the above conditions are satisfied with probability at least $1-2\delta'$ if we choose $|S| \ge \Omega\left(s^2 \cdot \log^5(d) \log^2 \frac{1}{\delta'} + \frac{1}{\eta_0} \cdot \log \frac{1}{\delta'}\right)$. Hence, by setting $q_i^* = 1$ for every $i \in S_C$ and setting all other weights to 0, q^* satisfies all three constraints. □

C Orlicz Norm and Concentration Results

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

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

The following follows from Section 1.3 of [vdVW96].

Lemma C.2. *Let Z, Z_1, Z_2 be real-valued random variables. Consider the Orlicz norm with respect to ϕ_α. We have the following:*

1. *$\|\cdot\|_{\psi_\alpha}$ is a norm. For any $\alpha \in \mathbb{R}$, $\|\alpha Z\|_{\psi_\alpha} = |\alpha| \cdot \|Z\|_{\psi_\alpha}$; $\|Z_1 + Z_2\|_{\psi_\alpha} \le \|Z_1\|_{\psi_\alpha} + \|Z_2\|_{\psi_\alpha}$.*
2. *$\|Z\|_p \le \|Z\|_{\psi_p} \le p! \|Z\|_{\psi_1}$ where $\|Z\|_p := (\mathbb{E}[|Z|^p])^{1/p}$.*
3. *For any $p, \alpha > 0$, $\|Z\|_{\psi_p}^\alpha = \|Z^\alpha\|_{\psi_{p/\alpha}}$.*
4. *If $\Pr(|Z| \ge t) \le K_1 \exp(-K_2 t^\alpha)$ for any $t \ge 0$, then $\|Z\|_{\psi_\alpha} \le \left(\frac{2(\log K_1 + 1)}{K_2}\right)^{1/\alpha}$.*
5. *If $\|Z\|_{\psi_\alpha} \le K$, then for all $t \ge 0, \Pr(|Z| \ge t) \le 2\exp\left(-\left(\frac{t}{K}\right)^\alpha\right)$.*

Lemma C.3. *There exist absolute constants $C_1, C_2 > 0$ such that the following holds for any distribution $\mathcal{D}_{\mathcal{X}}$ satisfying Assumption 1.3. Let $S = \{x_1, \ldots, x_n\}$ be a set of n samples from $\mathcal{D}_{\mathcal{X}}$. We have that,*

$$\left\| \max_{x \in S} \|x\|_\infty \right\|_{\psi_1} \leq C_2 \cdot \sigma \left(\frac{r}{\sigma} + 1 \right) \log(nd).$$

In addition,

$$\mathbb{E}_{S \sim \mathcal{D}_{\mathcal{X}}^n} \left(\max_{x \in S} \|x\|_\infty \right) \leq C_1 \cdot \sigma \log(nd) + r.$$

Proof. Let Z be zero-mean σ^2-variance logconcave random variable in $\mathbb{R}$. Lemma F.1 implies that

$$\Pr\left(|Z| > t\right) \leq \exp(-t/\sigma + 1). \tag{13}$$

Since $\mathcal{D}_{\mathcal{X}}$ is a mixture of k logconcave distribution, we denote by $\bar{x}_i$ the i-th instance subtracting its logconcave distribution mean, e.g. $(x_i - \mu_j)$ if x_i is from $\mathcal{D}_j, j \in [k]$. Fix an instance $i \in \{1, \ldots, n\}$ and a coordinate $l \in \{1, \ldots, d\}$. Denote by $\bar{x}_i^{(l)}$ the l-th coordinate of $\bar{x}_i$. Then, we have that Eq. (13) holds for any $\bar{x}_i^{(l)}$. Taking the union bound over all instances and all coordinates gives us that,

$$\Pr_{S \sim D^n} \left(\max_{x \in S} \|\bar{x}\|_\infty > t \right) \leq nd \cdot \exp\left(-\frac{t}{\sigma} + 1 \right).$$

Applying Part 4 of Lemma C.2, we have that

$$\left\| \max_{x \in S} \|\bar{x}\|_\infty \right\|_{\psi_1} \leq C_1 \cdot \sigma \log(nd).$$

Likewise, for each coordinate $\bar{x}_i^{(l)}$, we have that $\Pr\left(\left| \bar{x}_i^{(l)} \right| > t \right) \leq \exp(-t/\sigma + 1)$, which implies that $\Pr\left(\left| x_i^{(l)} \right| > t \right) \leq \exp(-(t-r)/\sigma + 1)$ for all $t > r$, since $\left| \mu_j^{(l)} \right| \leq \|\mu_j\|_2 \leq r, \forall j \in [k]$. For $S \sim D^n$, we have that $\Pr(\max_{x \in S} \|x\|_\infty > t) \leq nd \cdot \exp(r/\sigma + 1) \exp(-t/\sigma)$. We use similar technique to bound the Orlicz norm for $\max_{x \in S} \|x\|_\infty$ as follows,

$$\left\| \max_{x \in S} \|x\|_\infty \right\|_{\psi_1} \leq \frac{2\left((r/\sigma + 1)\log(nd) + 1\right)}{1/\sigma} = C_2 \cdot \sigma \left(\frac{r}{\sigma} + 1 \right) \log(nd).$$

On the other hand, we have that

$$\mathbb{E}_{S \sim D^n} \left(\max_{x \in S} \|x\|_\infty \right) \leq \mathbb{E}_{S \sim D^n} \left(\max_{x \in S} \|\bar{x}\|_\infty + r \right) = \mathbb{E}_{S \sim D^n} \left(\max_{x \in S} \|\bar{x}\|_\infty \right) + r.$$

since $\|\mu\|_\infty \leq \|\mu\|_2 \leq r$.

The lemma then follows from the fact that $\mathbb{E}_{S \sim D^n} \left(\max_{x \in S} \|\bar{x}\|_\infty \right) \leq \left\| \max_{x \in S} \|\bar{x}\|_\infty \right\|_{\psi_1}$. □

Lemma C.4. *There exist absolute constants $C_1, C_2 > 0$ such that the following holds for any logconcave distribution D satisfying Assumption 1.3.*

$$\left\| \|x\|_\infty \right\|_{\psi_1} \leq C_2 \cdot \sigma \left(\frac{r}{\sigma} + 1 \right) \log(d).$$

In addition,

$$\mathbb{E}_{x \sim D} \left(\|x\|_\infty \right) \leq C_1 \cdot \sigma \log(d) + r.$$

Proof. The lemma follows from the analysis in Lemma C.3 with $n = 1$ sample. □

D Lower Bounds

In this section, we show the sample complexity lower bounds for attribute-efficient PAC learning of sparse halfspaces. We use some arguments in prior works of [SZ21] and [DKS17] and show that our sample complexities are indeed nearly optimal. Specifically, Lemma 2 of [SZ21] provided an information-theoretic lower bound of label complexity for attribute-efficiently active learning of s-sparse halfspaces under single isotropic logconcave marginals. We use similar arguments to show a sample complexity lower bound of passive learning under the mixture of logconcaves.

[BI88] proved that a concept class W is (ϵ, δ)-learnable with respect to a fixed distribution $\mathcal{D}_{\mathcal{X}}$ with at least $\log((1-\delta)N(2\epsilon, W, \mathcal{D}_{\mathcal{X}}))$ samples, where $N(\epsilon, W, \mathcal{D}_{\mathcal{X}})$ is the ϵ-covering number of W under the metric of $(W, \mathcal{D}_{\mathcal{X}})$. In more details, this metric measures the distance of two concepts $w_1, w_2 \in W$ using $\Pr_{(x,y)\sim\mathcal{D}_{\mathcal{X}}}(w_1(x) \neq w_2(x))$. In [KMT93], it was shown that the lower bound for the label complexity of active learning is quantitatively the same. [SZ21] then utilizes this result together with known lower bounds of ϵ-packing numbers of s-sparse halfspaces, i.e. $(\frac{1}{\epsilon})^{\Omega(s)}$ [Lon95] and $(\frac{d}{s})^{\Omega(s)}$ [RWY11], to derive the lower bound under isotropic logconcaves. Notice that the result implicitly depends on a fact that the Euclidean distance or the angle between two unit vectors u, v implies the error rate distance under isotropic logconcave distribution, which we include as follows.

Lemma D.1 (Lemma 3 of [BL13])**.** *For any two unit vectors $u, v \in \mathbb{R}^d$, there are absolute constants $c > 0$ such that for any isotropic logconcave distributions D, $c \cdot \theta(u, v) \leq \Pr_{x\sim D}(\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(v \cdot x))$.*

Inspired by Lemma 10 of [Zha18], we also summarize the following lemma.

Lemma D.2. *For any two unit vectors u, v, we have that $\frac{2}{\pi} \cdot \theta(u, v) \leq \|u - v\|_2 \leq \theta(u, v)$.*

Proof. By Jordan's inequality, we have that $\frac{2}{\pi} \cdot \theta \leq \sin\theta \leq \theta$ for any $\theta \in [0, \frac{\pi}{2}]$. Together with the fact that $\frac{\|u-v\|_2}{2} = \sin\frac{\theta(u,v)}{2}$, we obtain the bounds. □

Combining Lemma D.1 and D.2, the result in [SZ21] is obtained. We now show that this bound also holds for mixtures of logconcaves. This requires adapting the proof of Lemma D.1 from isotropic logconcave to mixture of logconcaves. In this paper, we only show the difference and refer interested readers to the proof of Lemma 3 of [BL13]. To that end, we have the following theorem.

Lemma D.3 (Restatement of Lemma 1.6)**.** *Learning of s-sparse halfspaces under Assumption 1.2 for any $\gamma \leq \frac{1}{2}$ and 1.3 in the noiseless setting has an information-theoretic sample complexity lower bound of $\Omega(s \cdot (\log\frac{1}{\epsilon} + \log\frac{d}{s}))$.*

Proof. It suffices to construct a hardness result that there exist a mixture of logconcave distributions D under which (ϵ, δ)-learning of w^* requires at least $\Omega(s \cdot (\log\frac{1}{\epsilon} + \frac{d}{s}))$ samples. Notice that the ϵ-packing number is at least $(\frac{1}{\epsilon})^{\Omega(s)} + (\frac{d}{s})^{\Omega(s)}$. It remains to show that under distribution D, $\exists c > 0$ such that $c \cdot \|u - v\|_2 \leq \Pr_{x\sim D}(\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(v \cdot x))$.

Define $\mathrm{proj}_{u,v}(D)$ to be the distribution given by first picking $x \sim D$ and then orthogonally projecting x onto the plane determined by u and v. [BL13] showed that $\Pr_{x\sim D}(\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(v \cdot x)) = \Pr_{x\sim\mathrm{proj}_{u,v}(D)}(\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(v \cdot x))$. Moreover, let A be the region of disagreement between u and v intersected with the ball of radius $1/9$ in $\mathbb{R}^2$ defined by u and v. Then, $\Pr_{x\sim\mathrm{proj}_{u,v}(D)}(\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(v \cdot x)) \geq \mathrm{vol}(A) \cdot \inf_{x\in A}\mathrm{proj}_{u,v}(D)(x) \geq c \cdot \theta(u, v)$, where $\mathrm{vol}(A)$ is the volume of A and $\mathrm{proj}_{u,v}(D)(x)$ is the probability density of x under $\mathrm{proj}_{u,v}(D)$.

Now, consider D to be a distribution which has 0 probability density inside the region of $B := \{x : |w^* \cdot x| \geq \gamma\}$ and uniform outside it, i.e. $\bar{B}$. Note that this D satisfies both Assumption 1.2 and 1.3. We will show that there is still at least $(\frac{1}{\epsilon})^{\Omega(s)} + (\frac{d}{s})^{\Omega(s)}$ number of s-sparse halfspaces that are ϵ away from each other, i.e. $\Pr_{x\sim D}(\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(v \cdot x)) \geq \epsilon$. To show this, consider any unit vectors u, v such that $\theta(u, w^*) \geq \frac{\pi}{4}$ and $\theta(v, w^*) \geq \frac{\pi}{4}$. Now, $\Pr_{x\sim\mathrm{proj}_{u,v}(D)}(\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(v \cdot x)) \geq \mathrm{vol}(A\backslash B) \cdot \inf_{x\in A\backslash B}\mathrm{proj}_{u,v}(D)(x) \geq c \cdot \theta(u, v) \cdot (1 - (\sqrt{2}\gamma)^2) \geq \frac{c}{2}\theta(u, v)$ for $\gamma \leq \frac{1}{2}$. Here, the second transition is due to that $A \cap B$ is contained in a sector with angle $\theta(u, v)$ and radius $\leq \sqrt{2}\gamma$.

In other words, for any unit vector u such that $\theta(u, w^*) \geq \frac{\pi}{4}$, the Euclidean distance of $c \cdot \epsilon$ implies $\Pr_{x \sim D}(\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x)) \geq \epsilon$. Define $A' := \{u \in \mathbb{R}^d : \|u\|_0 \leq s, \theta(u, w^*) \geq \frac{\pi}{4}\}$. It is not hard to notice that $\text{vol}(A')$ is propotional to that of W. Hence, the packing number stays the same. By taking $\log N(2\epsilon, A', D) = \Omega(s \cdot (\log \frac{1}{\epsilon} + \log \frac{d}{s}))$ we obtain the sample complexity lower bound as presented in the lemma. The proof is complete. □

Last but not least, it was well known that for malicious noise model (or Huber contamination model in unsupervised learning), there exists a statistical-computational gap of $\Omega(s^2)$ for any statistical-query algorithm [DKS17]. This is a strong evidence showing that our sample complexity is almost optimal.

E Attribute-Efficient Learning with Adversarial Label Noise

The algorithm for attribute-efficient learning under a constant adversarial label noise is as follows.

Algorithm 3 Main Algorithm

Require: $\text{EX}(D, w^*, \eta)$, target error rate ϵ, failure probability δ, parameters s, γ.
Ensure: A halfspace $\hat{w}$.

1: Draw $n' = C \cdot s^2 \log^5 \frac{d}{\delta\epsilon}$ samples from $\text{EX}_{\text{adv}}(D, w^*, \eta)$ to form a set $S' = \{(x_i, y_i)\}_{i=1}^{n'}$.
2: Remove all samples (x, y) in S' with $\|x\|_\infty \geq r + \sigma \cdot (\log \frac{n'd}{\delta'} + 1)$ to form a set S.
3: Solve the following hinge loss minimization program:

$$\hat{w} \leftarrow \underset{\|w\|_1 \leq \sqrt{s}, \|w\|_2 \leq 1}{\arg\min} \ell_\gamma(w; S). \tag{14}$$

4: Return $\hat{w}$.

We will use the natural concentration of instances from the mixture of logconcave distributions and show that a bounded variance allows Algorithm 3 to return a good enough halfspaces $\hat{w}$ that has a low error rate with high probability.

Definition E.1 (Learning sparse halfspaces with adversarial label noise)**.** Given any $\epsilon, \delta \in (0, 1)$, and an adversary oracle $\text{EX}_{\text{adv}}(\mathcal{D}, w^*, \eta)$ with underlying distribution $\mathcal{D}$, ground truth w^* with $\|w^*\|_0 \leq s$, and noise rate η fixed before the learning task begins, when the learner requests a set of samples from $\mathcal{D}$, the oracle draws a set $S_0 \sim \mathcal{D}^n$ and then flips an arbitrary η fraction of the labels in S_0, and returns the modified set S to the learner.

Theorem E.2 (Attribute-efficent learning under adversarial label noise)**.** *Given a set S of $n \geq \Omega\left(s \cdot \log^4 \frac{d}{\delta\epsilon}\right)$ samples from $EX_{adv}(\mathcal{D}, w^*, \eta)$ with $\mathcal{D}, w^*$ satisfying Assumption 1.2 with $\gamma \geq \frac{4(\log \frac{1}{\epsilon}+1)}{\sqrt{d}}$ and Assumption 1.3 with $r \leq 2\gamma$, $k \leq 64$, and the adversarial label noise rate $\eta \leq \eta_0 \leq \frac{1}{2^{32}}$, the following holds. For any $\epsilon, \delta \in (0, 1)$, Algorithm 3 returns a $\hat{w}$ such that $\text{err}_{\mathcal{D}}(\hat{w}) \leq \epsilon$ with probability at least $1 - \delta$.*

Proof of Theorem E.2. We again consider that S consists of two sets of samples. i.e. $S_C \cup S_D$ where S_C are the samples not being corrupted, and S_D are those whose labels are flipped. We also consider the set $\bar{S}$ of instances without labels from set S.

We show that similar deterministic guarantees to Theorem A.10 holds for Algorithm 3.

Claim E.3. *If Assumption 1.2 holds, and*

$$c \cdot \rho \geq \left(\frac{1}{\gamma\sqrt{d}} + \frac{r}{\gamma} + 1\right)\sqrt{\eta_0} \tag{15}$$

for some some constant $c \in (0, \frac{1}{32})$, then, the output $\hat{w}$ of Algorithm 3 has error rate $\text{err}_{\mathcal{D}}(\hat{w}) \leq \epsilon$.

Note that $\mathcal{D}$ satisfies some dense pancake condition (Lemma B.1), and S_C is obtained by removing at most η fraction from S. Hence, due to Theorem B.2, it holds with probability at least $1 - \delta'$ that $(S_C, \mathcal{D})$ satisfies the $(\tau, \rho - \eta, \epsilon)$-dense pancake condition with respect to any $w \in \mathcal{W}$ with

$|S| \geq \Omega\left(\frac{1}{\rho} \cdot \left(s \cdot \log^4 d + \log \frac{1}{\delta' \epsilon}\right)\right)$, where $\tau = 4\sigma \cdot \left(\log \frac{1}{\epsilon} + 1\right)$, $\rho = \frac{1-\epsilon}{2k}$. It is easy to see from Eq. (15) that $\eta \leq \eta_0 \leq \frac{\rho}{2}$. Thus, $(S_C, \mathcal{D})$ satisfies $(\tau, \rho/2, \epsilon)$-dense pancake condition. We choose that $\gamma \geq 2\tau = \frac{8(\log \frac{1}{\epsilon}+1)}{\sqrt{d}}$. In the following, we rescale $\rho' = \rho/2$.

Then, we make the following claim.

Claim E.4. *Due to Assumption 1.2, for any given (x, y), if its pancake $\mathcal{P}^{\tau}_{\hat{w}}(x, y)$ with $\tau \leq \gamma/2$ is ρ'-dense with respect to S_C for some $\rho' > 4\eta_0$, and it holds that*

$$|S_C \cap \mathcal{P}^{\tau}_{\hat{w}}(x, y)| > \frac{4}{\gamma} \cdot \mathrm{GradNorm}(S_D), \tag{16}$$

then (x, y) is not misclassified by the $\hat{w}$ returned by Algorithm 3.

The proof follows similarly from that of Theorem 4.1. Notice that we do not require reweighting the samples here, hence, we can reproduce the proof by setting q to an all-one vector. Moreover, we consider the gradient norm on S_D directly. We then show that Eq. (16) holds under the condition of Eq. (15) , which is similar to proving Lemma A.7. We note that $|S_C \cap \mathcal{P}^{\tau}_{\hat{w}}(x, y)| \geq (\rho - \eta)\,|S|$, since $\mathcal{P}^{\tau}_{\hat{w}}(x, y)$ is assumed to be ρ-dense and $|S_D|$ takes up to η-fraction (Definition E.1).

Then, we show that $\mathrm{GradNorm}(S_D) \leq 2\sqrt{\frac{1}{d} + r^2} \cdot \sqrt{\eta}\,|S|$. That is, we apply Lemma A.9 that $\mathrm{GradNorm}(S_D) \leq \sqrt{|S_D|}\sqrt{\sup_{w \in \mathcal{W}} \sum_{i \in S_D} (w \cdot x_i)^2}$ by letting the weight vector q be all ones. In addition, $\sup_{w \in \mathcal{W}} \sum_{i \in S_D} (w \cdot x_i)^2 \leq \sup_{w \in \mathcal{W}} \sum_{i \in S} (w \cdot x_i)^2 \leq 4(\frac{1}{d} + r^2) \cdot |S|$, where the second transition is due to Lemma B.4 and B.5 with the fact that $ww^{\top} \in \mathcal{M}$. Note that this is satisfied with probability $1 - \delta'$ by choosing $|S| \geq \Theta\left(s^2 \cdot \log^5 d \log^2 \frac{1}{\delta'}\right)$.The gradient norm is well bounded. Finally, it is easy to obtain Eq. (16) from Eq. (15).

Therefore, given Claim E.4 holds, Eq. (15) holds with $k \leq 64$ and $\eta_0 \leq \frac{1}{2^{32}}$, and $(S_C, \mathcal{D})$ satisfies $(\tau, \rho/2, \epsilon)$-dense pancake condition, we conclude that all underlying instances but an ϵ fraction is not misclassified by the returned $\hat{w}$. That is, $\mathrm{err}_{\mathcal{D}}(\hat{w}) \leq \epsilon$.

It then remains to conclude the sample complexity for determinimistic result and the confidence bound. Let $\delta' = \delta/2$. Then, $\mathrm{err}_{\mathcal{D}}(\hat{w}) \leq \epsilon$ holds with probability $1 - \delta$. The sample complexity is as presented in the theorem, which is similar to that in Theorem 1.4.

□

F Useful Lemmas

Lemma F.1 (Isotropic logconcave, [LV07])**.** *For any isotropic logconcave distribution D over $\mathbb{R}^d, d \geq 1$, it holds that*

$$\Pr_{x \sim D}(\|x\|_2 \geq t\sqrt{d}) \leq \exp(-t + 1).$$

In addition, any orthogonal projection of D onto subspace of $\mathbb{R}^d$ is still isotropic logconcave.

Lemma F.2 (Chernoff bound)**.** *Let $Z_1, Z_2, \ldots, Z_n$ be n independent random variables that take value in $\{0, 1\}$. Let $Z = \sum_{i=1}^{n} Z_i$. For each Z_i, suppose that $\Pr(Z_i = 1) \leq \eta$. Then for any $\alpha \in [0, 1]$*

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

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

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

The above two probability inequalities hold when η equals exactly $\Pr(Z_i = 1)$.

Lemma F.3 (Adamczak's bound)**.** *For any $\alpha \in (0, 1]$, there exists a constant $\Lambda_\alpha > 0$, such that the following holds. Given any function class $\mathcal{F}$, and a function F such that for any $f \in \mathcal{F}$,*

$|f(x)| \le F(x)$, we have with probability at least $1-\delta$ over the draw of a set $S = \{x_1, \ldots, x_n\}$ of i.i.d. instances from a distribution D,

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

where $\|\cdot\|_{\psi_\alpha}$ denotes the Orlicz norm with parameter α.

Lemma F.4 (Theorem 1 of [KST08])**.** *Let $a = (a_1, \ldots, a_n)$ where a_i's are independent draws from the Rademacher distribution and let $x_1, \ldots, x_n$ be given instances in $\mathbb{R}^d$. Then,*

$$\mathbb{E}_a \left(\sup_{\|w\|_1 \le t} \sum_{i=1}^{n} a_i w \cdot x_i \right) \le t\sqrt{2n\log(2d)} \max_{1\le i\le n} \|x_i\|_\infty .$$

Definition F.5 (p-norm covering number)**.** Let $F := \{f_w : \mathbb{R}^d \to \mathbb{R}, w \in \mathbb{R}^d\}$ be a set of real-valued functions. Consider a set of observations $X^n = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ and a vector $f_w(X^n) = (f_w(x_1), \ldots, f_w(x_n)) \in \mathbb{R}^n$ parameterized by w, the covering number in p-norm, denoted as $\mathcal{N}_p(F, \epsilon, X^n)$, is the minimum number m of a collection of vectors $v_1, \ldots, v_m \in \mathbb{R}^n$ such that $\forall w$, there exists v_j satisfying

$$\|f_w(X) - v_j\|_p = \left(\sum_{i=1}^{n} (f_w(x_i) - v_j^i)^p \right)^{1/p} \le n^{1/p} \cdot \epsilon,$$

where v_j^i denotes the i-th component of vector v_j. We also define $\mathcal{N}_p(F, \epsilon, n) = \sup_{X^n} \mathcal{N}_p(F, \epsilon, X^n)$.

Lemma F.6 (Corollary 5 of [Zha02])**.** *Consider $x, w \in \mathbb{R}^d$. Let F be the set of linear functions on x induced by w, i.e. $F := \{f_w : x \mapsto w \cdot x\}$. Let $\mathcal{N}_\infty(F, \epsilon, n)$ be the ∞-norm covering number of F as defined in Definition F.5. If $\|x\|_\infty \le b$ and $\|w\|_1 \le a$, then $\forall \epsilon > 0$,*

$$\log \mathcal{N}_\infty(F, \epsilon, n) \le \frac{288 a^2 b^2 (2 + \ln d)}{\epsilon^2} \log_2 \left(2 \left\lceil \frac{8ab}{\epsilon} + 2 \right\rceil n + 1 \right).$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarized the problem and our contribution accurately in abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [N/A]

   Justification: This paper studies a fundamental theoretical problem and provides refined guarantees and analysis. There is no obvious liminations under the current setting.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper provides a full set of assumptions and complete proofs for all theoretical results, with some proofs placed in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper is of theoretical nature.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not include experiements.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper is of theoretical nature.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper is of theoretical nature.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper is of theoretical nature.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Research conducted conforms with NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The research is foundational and not tied to any particular applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not have such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All reference are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper only proposes theoretical machine learning algorithms, instead of actual assets that require documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.