Sample-Optimal PAC Learning of Halfspaces with Malicious Noise

Jie Shen [1]

Abstract

We study efficient PAC learning of homogeneous halfspaces in $\mathbb{R}^d$ in the presence of malicious noise of Valiant (1985). This is a challenging noise model and only until recently has near-optimal noise tolerance bound been established under the mild condition that the unlabeled data distribution is isotropic log-concave. However, it remains unsettled how to obtain the optimal sample complexity simultaneously. In this work, we present a new analysis for the algorithm of Awasthi et al. (2017) and show that it essentially achieves the near-optimal sample complexity bound of $\tilde{O}(d)$, improving the best known result of $\tilde{O}(d^2)$. Our main ingredient is a novel incorporation of a matrix Chernoff-type inequality to bound the spectrum of an empirical covariance matrix for well-behaved distributions, in conjunction with a careful exploration of the localization schemes of Awasthi et al. (2017). We further extend the algorithm and analysis to the more general and stronger nasty noise model of Bshouty et al. (2002), showing that it is still possible to achieve near-optimal noise tolerance and sample complexity in polynomial time.

1. Introduction

In this paper, we study computationally efficient PAC learning of homogeneous halfspaces – arguably one of the most important problems in machine learning (Valiant, 1984). In the absence of noise, the problem is well understood and can be efficiently solved by linear programming (Maass & Turán, 1994) or the Perceptron (Rosenblatt, 1958). However, when the unlabeled data[1] or the labels are corrupted, it becomes subtle to develop polynomial-time algorithms that are resilient to the noise (Valiant, 1985; Angluin & Laird, 1988; Kearns & Li, 1988; Kearns et al., 1992).

Generally speaking, a large body of existing works study the problem of learning halfspaces under *label* noise. This includes early works on random classification noise where the label of each instance is independently flipped with a fixed probability (Blum et al., 1996), a more general model termed Massart noise where the probability of flipping a given label may vary from instance to instance but is bounded away from $\frac{1}{2}$ (Sloan, 1988; Massart & Nédélec, 2006), the Tybakov noise where the flipping probability can be arbitrarily close to $\frac{1}{2}$ for a fraction of samples (Tsybakov, 2004), and the much stronger adversarial (i.e. agnostic) noise where the adversary may choose an arbitrary joint distribution over the instance and label spaces (Haussler, 1992; Kearns et al., 1992; Kalai et al., 2005; Daniely, 2015). When only the labels are corrupted, significant progress towards establishing near-optimal performance guarantees has been witnessed in recent years; see, e.g. Awasthi et al. (2017); Diakonikolas et al. (2019; 2020a;b;c); Zhang et al. (2020); Shen (2020).

Compared to the fruitful set of positive results on efficient learning of halfspaces under label noise, less is known for the significantly more challenging regime where *both* instances and labels are corrupted. Specifically, one of such strong noise models that has played a crucial role in learning theory is the malicious noise model of Valiant (1985); Kearns & Li (1988), defined as follows:

Definition 1 (Malicious noise)**.** Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$ be the instance and label space, respectively. Let D be an unknown distribution over $\mathcal{X}$, and $w^* \in \mathbb{R}^d$ be an unknown halfspace. Each time the learner requests a sample, with probability $1 - \eta$, the adversary draws x from D and returns the clean sample $(x, \mathrm{sign}\,(w^* \cdot x))$; with probability η, it may return an arbitrary pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ called dirty sample. The parameter $\eta \in [0, \frac{1}{2})$ is termed noise rate.

Notably, when the adversary is allowed to search for dirty samples, it is assumed to have unlimited computational power and can construct the sample based on the state of the learning algorithm and the history of its outputs. Since this is a much more demanding noise model (compared to label-only noise), even the achievability of optimal noise tolerance by efficient algorithms remained unsettled for decades. For example, the early work of Kearns & Li (1988) presented a general analysis showing that even without any distributional assumptions, it is possible to tolerate the malicious noise at a rate of $\Omega(\epsilon/d)$, but a noise rate greater

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

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

[1] We will also refer to unlabeled data as instances in this paper, and refer to labeled data as samples.

than $\frac{\epsilon}{1+\epsilon}$ cannot be tolerated, where $\epsilon \in (0,1)$ is the target error rate given to the learner. The noise model was then broadly studied in the literature, though the learning algorithms may be inefficient; see e.g. Schapire (1992); Bshouty (1998); Cesa-Bianchi et al. (1999). Under different distributional assumptions, there are more positive results for efficient learning with malicious noise. In particular, when the distribution D is uniform over the unit sphere, Kalai et al. (2005) developed an efficient learning algorithm and obtained a noise tolerance $\Omega(\epsilon/d^{1/4})$, which was later improved to $\Omega(\epsilon^2/\log(d/\epsilon))$ in terms of the dependence on the dimension by Klivans et al. (2009). It is, however, well recognized that the uniform distribution is often restrictive in practice. As a remedy, Klivans et al. (2009) also investigated the remarkably more general isotropic log-concave distributions (Lovász & Vempala, 2007), and showed for the first time a noise tolerance of $\Omega(\epsilon^3/\log^2(d/\epsilon))$ under such mild condition. Unfortunately, owing to the strong power of the adversary, the barrier of achieving the information-theoretic limit of $\frac{\epsilon}{1+\epsilon}$ was not broken for many years (even under the uniform distribution). Very recently, a near-optimal noise tolerance of the form $\Omega(\epsilon)$ was established by Awasthi et al. (2017) for isotropic log-concave distributions through a dedicated iterative localization technique, which stands for the state of the art.

In addition to the degree of noise tolerance, another yet important quantity that characterizes the performance of a learning algorithm is sample complexity. Unfortunately, it turns out that none of the prior works obtained near-optimal sample complexity and noise tolerance simultaneously under the mild condition that the (clean) instances are drawn from an isotropic log-concave distribution. In particular, Awasthi et al. (2017); Shen & Zhang (2021) obtained state-of-the-art noise tolerance but the sample complexity of Awasthi et al. (2017) reads as $\tilde{O}(d^3)$. Shen & Zhang (2021) considered learning of s-sparse halfspaces with malicious noise and showed through a refined analysis that a sample size of $\tilde{O}(s^2 \cdot \operatorname{polylog}(d))$ suffices; when specified to the non-sparse setting (which is the focus of this paper), it still leads to a suboptimal bound of $\tilde{O}(d^2)$. Prior to these two recent works, even a noise tolerance of the form $\Omega(\epsilon)$ was not established, nor an optimal sample complexity bound. On the other hand, it is worth mentioning that under the fairly restrictive uniform distribution over the unit ball, the analysis of Awasthi et al. (2017) does imply a near-optimal sample complexity bound of $\tilde{O}(d)$. In this regard, a natural question is: *can we design a computationally efficient algorithm that is able to tolerate $\Omega(\epsilon)$ malicious noise while enjoying the optimal sample complexity bound of $O(d)$ under isotropic log-concave distributions?*

In this paper, we answer the question in the affirmative. First, we formally describe our assumption on clean instances.

Assumption 1. The distribution D is isotropic log-concave over $\mathbb{R}^d$; namely, it has zero mean and unit covariance matrix, and the logarithm of its density function is concave.

Observe that the family of isotropic log-concave distributions covers prominent distributions such as Gaussian, exponential, and logistic distributions (Lovász & Vempala, 2007; Vempala, 2010). In particular, general isotropic log-concave distributions are asymmetric in nature and the magnitude of the instances drawn from them is often dimension-dependent, making it nontrivial to extend results developed for the uniform distribution over the unit ball.

1.1. Main results

Recall that D and w^* are the underlying distribution and the correct halfspace as stated in Definition 1, respectively. For any homogeneous halfspace $h_w : x \mapsto \operatorname{sign}(w \cdot x)$, let $\operatorname{err}_D(w) := \operatorname{Pr}_{x\sim D}(\operatorname{sign}(w \cdot x) \neq \operatorname{sign}(w^* \cdot x))$ be the error rate of w with respect to D and w^*. The following is our main result.

Theorem 1. *Consider the malicious noise model under Assumption 1. There is an algorithm such that for any target error rate $\epsilon \in (0,1)$ and any failure probability $\delta \in (0,1)$, if $\eta \leq O(\epsilon)$, it outputs a halfspace $\tilde{w}$ satisfying $\operatorname{err}_D(\tilde{w}) \leq \epsilon$ with probability $1-\delta$. The running time is $\operatorname{poly}\left(d, \frac{1}{\epsilon}, \frac{1}{\delta}\right)$ and the sample complexity is $\frac{d}{\epsilon} \cdot \operatorname{polylog}\left(d, \frac{1}{\epsilon}, \frac{1}{\delta}\right)$.*

We highlight that this is the first result for efficient PAC learning of homogeneous halfspaces with both near-optimal malicious noise tolerance and sample complexity under isotropic log-concave distributions. On the algorithmic spectrum, we in fact show that the active learning algorithm proposed by Awasthi et al. (2017) inherently enjoys the announced properties and the noise tolerance bound in Theorem 1 directly inherits from their results. Regarding sample complexity, their original analysis made use of the pseudo-dimension from VC theory (Anthony & Bartlett, 1999) to give an $\tilde{O}(d^3)$ sample complexity bound which is suboptimal. Even using a careful Rademacher complexity bound, we would only obtain an $\tilde{O}(d^2)$ bound. Our improvement comes from a reformulation of the objective function used by Awasthi et al. (2017) and a novel utilization of a matrix Chernoff-type inequality due to Tropp (2012), together with a careful exploration of the localization schemes of Awasthi et al. (2017); see Section 2 for more details.

1.2. Extension to the nasty noise model

We also consider learning of homogeneous halfspaces with nasty noise of Bshouty et al. (2002), which is a strict generalization and is stronger than the malicious noise model.

Definition 2 (Nasty noise)**.** The learner specifies the total number of needed samples N. The adversary takes as input N, draws such many independent instances from D, and labels them correctly according to w^*. Then it may replace an arbitrary η fraction of them with arbitrary samples in $\mathcal{X} \times \mathcal{Y}$. The corrupted sample set is returned to the learner.

Observe that when $N = 1$, it reduces to the malicious noise model. The additional power of the nasty adversary is that when $N > 1$, it may inspect *all* the clean samples, and then decides which of them will be replaced, while in the malicious noise model it can only inject dirty samples (when it is permitted). Note that such extra power of erasing clean instances may modify the marginal distribution of the clean instances returned to the learner, which is one of the technical barriers that we have to carefully address.

For the problem of learning homogeneous halfspaces with nasty noise, we show that the algorithm of Awasthi et al. (2017) still works well (hence, our contribution is a new analysis). We have the following performance guarantee.

Theorem 2. *Consider the nasty noise model under Assumption 1. There is an algorithm such that for any target error rate $\epsilon \in (0, 1)$ and any failure probability $\delta \in (0, 1)$, if $\eta \leq O(\epsilon)$, it outputs a halfspace $\tilde{w}$ satisfying $\mathrm{err}_D(\tilde{w}) \leq \epsilon$ with probability $1 - \delta$. The running time is $\mathrm{poly}\left(d, \frac{1}{\epsilon}, \frac{1}{\delta}\right)$ and the sample complexity is $\frac{d}{\epsilon} \cdot \mathrm{polylog}\left(d, \frac{1}{\epsilon}, \frac{1}{\delta}\right)$.*

Since the malicious noise is a special case of the nasty noise, the information-theoretic limit of the noise tolerance established in Kearns & Li (1988), i.e. $\frac{\epsilon}{1+\epsilon}$, also applies to the nasty noise. In other words, the noise tolerance in the above theorem is near-optimal as well.

Another salient feature coming with the learning algorithm we consider, i.e. the one developed in Awasthi et al. (2017), is label efficiency; that is, the label complexity of the algorithm is $d \cdot \mathrm{polylog}\left(d, \frac{1}{\epsilon}, \frac{1}{\delta}\right)$ which has an exponential improvement on the dependence of $\frac{1}{\epsilon}$; see Appendix C.7 for the proof. To the best of our knowledge, this is also the first label-efficient algorithm that tolerates the nasty noise.

We remark, however, that in many prior works, the learner typically makes a one-time call throughout the learning process to gather all the labeled instances (Bshouty et al., 2002; Diakonikolas et al., 2018). Since we will study an algorithm that proceeds in multiple phases and draws samples adaptively, we consider a natural relaxation which allows the learner to make a one-time call per phase, with a total number of calls being $O(\log \frac{1}{\epsilon})$.[2] Therefore, our results under the nasty noise model are *not* strictly comparable to prior results such as Diakonikolas et al. (2018). It remains open of how to design an efficient algorithm which gathers all samples in one batch while still enjoying near-optimal nasty noise tolerance and sample complexity simultaneously.

1.3. Related works

The malicious and nasty noise models are strong contamination models for the problem of robustly learning Boolean functions. It turns out that most prior works on learning of halfspaces concentrated on obtaining optimal noise tolerance while not pursuing the $O(d)$ sample complexity, in that the former problem alone is already quite challenging (Kearns & Li, 1988; Awasthi et al., 2017). Diakonikolas et al. (2018) considered learning of more general concept classes, e.g. low-degree polynomial threshold functions and intersections of halfspaces, and showed that the underlying concept can be efficiently learned with $\tilde{O}(d^{\gamma})$ sample complexity for some unspecified constant $\gamma > 1$. When adapted to our setting (i.e. learning homogeneous halfspaces under isotropic log-concave marginal distributions), Theorem 1.5 of Diakonikolas et al. (2018) only gives noise tolerance $\eta \leq O(\epsilon^{\gamma'})$ for some constant $\gamma' > 1$ which is suboptimal.

Recent works such as Diakonikolas et al. (2016); Lai et al. (2016) studied mean estimation under a nasty-type model where in addition to returning dirty instances, the adversary has also the power of eliminating a few clean instances. The key technique of robust mean estimation is to use the spectral norm of the empirical covariance matrix to detect dirty instances, and a sample complexity bound of $\tilde{O}(d)$ was obtained, typically under Gaussian distributions rather than the more general isotropic log-concave distributions. More recently, such technique was extensively investigated for a variety of problems such as clustering and linear regression; we refer the readers to the comprehensive survey of Diakonikolas & Kane (2019). From a high level, the idea of certifying clean instances with a small second order moment roots in a much earlier work by Blum et al. (1996), and was then serving as a crucial component in learning halfspaces with malicious noise (Klivans et al., 2009). We note, however, that near-optimal sample complexity for learning halfspaces under isotropic log-concave marginal distributions is not implied by these results.

Notations. For a unit vector $u \in \mathbb{R}^d$ and a positive scalar b, we will frequently use $X_{u,b}$ to denote the band $\{x \in \mathbb{R}^d : |u \cdot x| \leq b\}$. Let T be a set of unlabeled data. We will use $\hat{T}$ to denote its labeled set, i.e. $\{(x, y_x) : x \in T\}$ where y_x is the label that the adversary is committed to. We write $\tilde{O}(f) := O(f \cdot \mathrm{polylog}(f))$. The letters c and C, and their subscript variants such as c_1, C_1, are reserved for specific absolute constants; see Appendix A.

Roadmap. In Section 2, we briefly describe the algorithm of Awasthi et al. (2017), followed by a refined theoretical analysis on the sample complexity. In Section 3, we extend the algorithm and analysis to the nasty noise model. We conclude this paper in Section 4, and defer all the proof details to the appendix.

[2] The crucial difference between the algorithm we consider and prior passive learning algorithms lies in the number of rounds that the learner communicates with the adversary.

2. Learning with Malicious Noise

We elaborate on our analytic tools used to obtain the near-optimal sample complexity bound in this section. Since we will give a new analysis for the algorithm developed by Awasthi et al. (2017), we first briefly introduce their main mechanisms; readers are referred to their original work for more detailed technical descriptions.

To improve readability, throughout this section, we will always implicitly assume that Assumption 1 is satisfied.

2.1. The approach of Awasthi et al. (2017)

The malicious-noise-tolerant algorithm, i.e. Algorithm 2 of Awasthi et al. (2017), is built upon the celebrated margin-based active learning framework of Balcan et al. (2007). For convenience, we record it in Algorithm 1 with a minor simplification (to be clarified). At a high level, it proceeds in $K = O(\log \frac{1}{\epsilon})$ phases, where the key idea is to find in each phase an empirical minimizer of a certain hinge loss that is a good proxy of the loss on clean samples drawn from a localized instance space. The margin-based framework will then assert that this suffices for PAC learnability. To this end, in each phase it has three major steps: rejection sampling, soft outlier removal, and hinge loss minimization.

Let $X_{u,b} := \{x \in \mathbb{R}^d : |u \cdot x| \leq b\}$ for some given unit vector u and certain scalar $b \in [\epsilon, O(1)]$ where $\epsilon \in (0, 1)$ is the given target error rate; in the notation of Algorithm 1, u should be thought of as w_{k-1} and $b = b_k$. Let $D_{u,b}$ be the distribution D conditioned on the event that $x \in X_{u,b}$. During rejection sampling, the learner calls the adversary $\mathrm{EX}_\eta^x(D, w^*)$ to collect a set T of unlabeled data lying in the band $X_{u,b}$.[3] Since the set T is corrupted by the adversary, the goal of soft outlier removal is to find proper weights for all instances in T such that the reweighted hinge loss over T is almost equal to the one evaluated on clean samples. Based on the detection results, during hinge loss minimization, the learner makes an additional call to the oracle EX^y to reveal the labels and finds an empirical minimizer of the reweighted hinge loss:

$$\ell_\tau(w; p \circ \hat{T}) := \frac{1}{|T|} \sum_{(x,y) \in \hat{T}} p(x) \cdot \max\Big\{0, 1 - \frac{1}{\tau} y w \cdot x\Big\}.$$

We remark that the sample complexity at phase k refers to the number of calls to $\mathrm{EX}_\eta^x(D, w^*)$, and the label complexity refers to that of EX^y.

It is known from standard margin-based active learning results that if the soft outlier removal step finds good weights in all the phases, then the final output of Algorithm 1 will have small error rate (Balcan et al., 2007; Awasthi et al., 2017). Therefore, most of our discussions will be dedicated to this crucial step. Note that the sample complexity refers to the total number of calls to $\mathrm{EX}_\eta^x(D, w^*)$ (which happens during rejection sampling) and the label complexity refers to that of EX^y (which happens during loss minimization).

Algorithm 1 Efficient and Sample-Optimal Algorithm Tolerating Malicious Noise

```
Require: Error rate $\epsilon$, failure probability $\delta$, instance generation oracle $\mathrm{EX}_\eta^x(D, w^*)$, label revealing oracle $\mathrm{EX}^y$.
Ensure: Halfspace $\tilde{w}$ with $\mathrm{err}_D(\tilde{w}) \leq \epsilon$ with probability $1 - \delta$.
Initialize $w_0$ as the zero vector in $\mathbb{R}^d$.
$K \leftarrow O(\log \frac{1}{\epsilon})$.
for phases $k = 1, 2, \dots, K$ do
    Clear the working set $T$.
    $b_k \leftarrow \Theta(2^{-k})$, $r_k \leftarrow \Theta(2^{-k})$, $\tau_k \leftarrow \Theta(2^{-k})$.
    Call $\mathrm{EX}_\eta^x(D, w^*)$ for $N_k$ times to form instance set $A$. If $k = 1$, $T \leftarrow A$; otherwise, $T \leftarrow \{x \in A : |w_{k-1} \cdot x| \leq b_k\}$.
    Apply Algorithm 2 to $T$ with $u \leftarrow w_{k-1}$, $b \leftarrow b_k$, $r \leftarrow r_k$, $\xi \leftarrow \frac{1}{2} - \Theta(1)$, $c \leftarrow 2C_2$, and let $q = \{q(x)\}_{x \in T}$ be the returned function. Normalize $q$ to form a probability distribution $p$ over $T$.
    $W_k \leftarrow \{w : \|w\|_2 \leq 1, \|w - w_{k-1}\|_2 \leq r_k\}$, $\hat{T} \leftarrow$ call $\mathrm{EX}^y$ to reveal the labels of $T$. Find $v_k \in W_k$ with
        $$\ell_{\tau_k}(v_k; p \circ \hat{T}) \leq \min_{w \in W_k} \ell_{\tau_k}(w; p \circ \hat{T}) + O(1).$$
    $w_k \leftarrow \frac{v_k}{\|v_k\|_2}$.
end for
return $\tilde{w} \leftarrow w_K$.
```

Decompose $T = T_{\mathrm{C}} \cup T_{\mathrm{D}}$ where T_{C} denotes the set of clean instances in T and T_{D} for the dirty instances. The key algorithmic insight of Awasthi et al. (2017) is that in order to guarantee the success of soft outlier removal, i.e. Algorithm 2 finds a feasible function $q : T \to [0, 1]$ in polynomial time, it is equivalent for the following to hold for some absolute constant $c > 0$:

$$\sup_{w \in W} \frac{1}{|T_{\mathrm{C}}|} \sum_{x \in T_{\mathrm{C}}} (w \cdot x)^2 \leq c(b^2 + r^2), \tag{1}$$

where

$$W := \big\{w \in \mathbb{R}^d : \|w\|_2 \leq 1,\ \|w - u\|_2 \leq r\big\}. \tag{2}$$

In order to prove (1), Awasthi et al. (2017) showed the following useful result.

Lemma 3. *There is an absolute constant $C_2 \geq 1$ such that*

$$\sup_{w : \|w - u\|_2 \leq r} \mathbb{E}_{x \sim D_{u,b}} \big[(w \cdot x)^2\big] \leq C_2(b^2 + r^2).$$

[3] The algorithm of Awasthi et al. (2017) is active in nature. Thus, the adversary initially hides the label and only returns the instance; the learner must make a separate call to reveal the label.

```
Algorithm 2 Localized Soft Outlier Removal
Require: Reference unit vector $u$, band width $b > 0$, radius
    $r = \Theta(b)$, empirical noise rate $\xi \in [0, 1/2)$, absolute
    constant $c > 0$, a set $T$ of instances drawn from $D_{u,b}$.
Ensure: A function $q : T \to [0, 1]$.
Let $W = \{w \in \mathbb{R}^d : \|w\|_2 \le 1,\ \|w - u\|_2 \le r\}$.
Find a function $q : T \to [0, 1]$ satisfying the following:
    1. for all $x \in T, 0 \le q(x) \le 1$;
    2. $\sum_{x \in T} q(x) \ge (1 - \xi)|T|$;
    3. $\sup_{w \in W} \frac{1}{|T|} \sum_{x \in T} q(x)(w \cdot x)^2 \le c(b^2 + r^2)$.
return $q$.
```

On the other hand, Anthony & Bartlett (1999) proved that with high probability,

$$\sup_{w \in W} \left| \frac{1}{|T_{\mathrm{C}}|} \sum_{x \in T_{\mathrm{C}}} (w \cdot x)^2 - \mathbb{E}_{x \sim D_{u,b}} \left[(w \cdot x)^2 \right] \right| \le \alpha \quad (3)$$

provided $|T_{\mathrm{C}}| = O\left(\frac{(\rho^+ - \rho^-)^2}{\alpha^2} d\right)$ where $\rho^- := \inf_{w \in W} (w \cdot x)^2$ and $\rho^+ := \sup_{w \in W} (w \cdot x)^2$.

Hence, Awasthi et al. (2017) combined Lemma 3 and Eq. (3) with $\alpha = C_2(b^2 + r^2)$ and showed that (1) holds with high probability if $|T_{\mathrm{C}}| = O\left(\frac{(\rho^+ - \rho^-)^2}{\alpha^2} d\right)$. If the unlabeled data distribution D were uniform over the unit sphere, then this bound would read as $O(d/\alpha^2)$ which has optimal dependence on d. However, since we are considering the significantly more general family of log-concave distributions, this bound becomes suboptimal.

Lemma 4. *Consider $x \sim D_{u,b}$. Then with probability $1 - \delta$, $(\rho^+ - \rho^-)^2 \le O(d^2 \cdot b^4 \log^4 \frac{1}{b\delta})$.*

Therefore, the VC theory only leads to a suboptimal sample size $|T_{\mathrm{C}}| = O(d^3)$, which implies that the number of calls to the instance generation oracle must be $O(d^3)$.

We note that while Rademacher complexity may sometimes offer improved sample complexity as illustrated by Zhang (2018); Shen & Zhang (2021), for our problem it only gives suboptimal guarantee of $|T_{\mathrm{C}}| = \tilde{O}(d^2)$; see Appendix B. Since the trouble roots in the suboptimal concentration bound of (3) which only involves quadratic functions, one may also wants to apply the well-known Hanson-Wright inequality (Rudelson & Vershynin, 2013) for better bound. The main barrier to apply it is that this inequality requires a sub-gaussian tail for the random vectors while that of log-concave distributions behaves as sub-exponential (see Part 5 of Lemma 24).

2.2. Our results and techniques

In contrast to the quadratic dependence on the dimension d, we show that (1) holds as soon as $|T_{\mathrm{C}}| = \tilde{O}(d)$.

Theorem 5. *With probability $1 - \delta$, Eq. (1) holds if $|T_{\mathrm{C}}| \ge d \cdot \mathrm{polylog}\left(d, \frac{1}{b}, \frac{1}{\delta}\right)$.*

Our technical novelty to show Theorem 5 is to move away from uniform concentration inequalities used by prior works. Rather, we reformulate the objective function of (1) which naturally leads to bounding the spectrum of a sum of random matrices. We then crucially explore the power of the localization scheme for the instance and concept spaces as used in Algorithm 1, and show that such spectrum norm acts as a constant over the phases, leading to the announced sample complexity.

First of all, we use the basic fact that for any $a_1, a_2 \in \mathbb{R}$, $(a_1 + a_2)^2 \le 2(a_1^2 + a_2^2)$, and obtain that

$$\sup_{w \in W} \sum_{x \in T_{\mathrm{C}}} (w \cdot x)^2 \le \sup_{w \in W} \sum_{x \in T_{\mathrm{C}}} \big((w - u) \cdot x\big)^2 + \sum_{x \in T_{\mathrm{C}}} (u \cdot x)^2.$$

Recall that in view of rejection sampling (namely localization in the instance space), for all $x \in T_{\mathrm{C}}$, it was drawn from D conditioned on the event $|u \cdot x| \le b$, implying $(u \cdot x)^2 \le b^2$ with certainty. Hence, it remains to upper bound $\sup_{w \in W} \big((w - u) \cdot x\big)^2$. By the definition of W in (2), we know that $w - u \in r \cdot V$ where $V := \{v : \|v\|_2 \le 1\}$. It thus follows that

$$\sup_{w \in W} \big((w - u) \cdot x\big)^2 \le r^2 \sup_{v \in V} (v \cdot x)^2 = r^2 \sup_{v \in V} v^\top (x x^\top) v.$$

Putting all pieces together, we have that the left-hand side of (1) can be upper bounded as follows:

$$\sup_{w \in W} \frac{1}{|T_{\mathrm{C}}|} \sum_{x \in T_{\mathrm{C}}} (w \cdot x)^2 \le r^2 \sup_{v \in V} v^\top M v + b^2, \qquad (4)$$

where $M = \left(\frac{1}{|T_{\mathrm{C}}|} \sum_{x \in T_{\mathrm{C}}} x x^\top \right)$. Observe that $v^\top M v$ corresponds to an eigenvalue of the matrix M. This motivates the consideration of the spectrum norm of the random matrix M, and is exactly where we need the matrix Chernoff bound of Tropp (2012).

Lemma 6 (Matrix Chernoff inequality)**.** *Consider a finite sequence $\{M_i\}_{i=1}^n$ of independent, random, self-adjoint matrices with dimension d. Assume that each random matrix satisfies $M_i \succeq 0$ and $\lambda_{\max}(M_i) \le \Lambda$ almost surely where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue. Define $\mu_{\max} := \lambda_{\max}(\sum_{i=1}^n \mathbb{E}[M_i])$. Then for all $\alpha \ge 0$, with probability at least $1 - d \cdot \left[\frac{e^\alpha}{(1+\alpha)^{1+\alpha}} \right]^{\frac{\mu_{\max}}{\Lambda}}$,*

$$\lambda_{\max}\left(\sum_{i=1}^n M_i \right) \le (1 + \alpha)\mu_{\max}.$$

To apply the above lemma, we will set $M_i = x_i x_i^\top$ for each $x_i \in T_\mathrm{C}$. We also establish the following result to estimate the two important quantities Λ and $\mu_{\max}$, where the proof crucially explores the localization scheme in the concept and instance spaces.

Lemma 7. *Suppose x is randomly drawn from $D_{u,b}$. Then*

$$\lambda_{\max}\left(\mathbb{E}\left[xx^\top\right]\right) \leq \frac{4C_2(b^2+r^2)}{r^2}.$$

In addition, with probability $1-\delta$,

$$\lambda_{\max}\left(xx^\top\right) \leq K_1 \cdot d \log^2 \frac{1}{b\delta}$$

for some constant $K_1 > 0$.

By setting $\alpha = 1$ in Lemma 6 and incorporating the results in Lemma 7, we have the following:

Proposition 8. *Let T_C be a set of i.i.d. instances drawn from $D_{u,b}$. If $|T_\mathrm{C}| \geq d \cdot \mathrm{polylog}\left(d, \frac{1}{b}, \frac{1}{\delta}\right)$, then with probability $1-\delta$, $\lambda_{\max}(M) \leq O(1)$.*

Now we are in the position to prove Theorem 5.

Proof of Theorem 5. In fact, by (4) and Proposition 8 we immediately have

$$\sup_{w\in W} \frac{1}{|T_\mathrm{C}|} \sum_{x\in T_\mathrm{C}} (w\cdot x)^2 \leq O(r^2) + b^2 \leq O(r^2+b^2),$$

which is the desired result. □

Next, we need to translate the bound of $|T_\mathrm{C}|$ to that of the number of calls to $\mathrm{EX}_\eta^x(D, w^*)$. To do so, we give a sufficient condition on the number of calls to $\mathrm{EX}_\eta^x(D, w^*)$ under which, there are as many instances in T_C as required in Theorem 5. This has been set out in Awasthi et al. (2017) where the primary observation is that under Assumption 1, the probability mass of the band $X_{u,b}$ is $\Theta(b)$. Hence, by calling $\mathrm{EX}_\eta^x(D, w^*)$ for $O(n/b)$ times it is guaranteed to gather n instances to form T. Also, it is possible to show that the empirical noise rate within T will be $O(\xi)$ where $\xi \in [0, 1/2)$ is a small constant, implying that $|T_\mathrm{C}| \geq \frac{1}{2}n$. By backward induction, the sample complexity in each phase is $\tilde{O}(d/b)$.

Formally, we have the following two lemmas.

Lemma 9. *Assume $\eta < \frac{1}{2}$. By making a number of $N = O\left(\frac{1}{b}\left(n + \log\frac{1}{\delta}\right)\right)$ calls to $\mathrm{EX}_\eta^x(D, w^*)$, we will obtain n instances to form T with probability $1-\delta$.*

Lemma 10. *Assume $\eta \leq c_5\epsilon$ for small constant $c_5 > 0$. If $|T| \geq 24 \ln\frac{1}{\delta}$, then with probability $1-\delta$, $|T_\mathrm{C}| \geq \frac{3}{4}|T|$.*

Algorithmic simplification. The last ingredient of Algorithm 1 is hinge loss minimization. A slight improvement in light of our new sample complexity bound is that there is no need to perform random sampling of the instances in T as in Awasthi et al. (2017). This is because in their original analysis, $|T| = \tilde{O}(d^3)$ and for the sake of optimizing label complexity without sacrificing the error rate, random sampling of a subset of size $\tilde{O}(d)$ is an elegant approach. In contrast, we already have shown that the size of T itself is $\tilde{O}(d)$, which can be labeled entirely by querying EX^y.

We are now in the position to prove Theorem 1. We note that the key difference from the analysis in Awasthi et al. (2017) is how we show that the $\tilde{O}(d)$ sample size suffices for soft outlier removal. We will therefore highlight this distinction in our proof. For the full and detailed proof, readers can either refer to their original paper or to our full proof of Theorem 2 in Section 3 since the malicious noise is a special case of the nasty noise.

Proof Sketch of Theorem 1. Consider phase $k \leq K$. Combining Lemma 9 and Lemma 10, we know that it suffices to call $\mathrm{EX}_\eta^x(D, w^*)$ for $\tilde{O}(d/b_k)$ times to ensure the carnality of T_C satisfies the condition in Theorem 5; hence the soft outlier removal step succeeds. On the other side, as shown in Proposition 11 of Shen & Zhang (2021), such sample complexity bound also suffices to guarantee the uniform concentration of hinge loss. Since $b_k \geq \epsilon$ for all $k \leq K$, the per-phase sample complexity is $\tilde{O}(d/b_k) \leq \tilde{O}(d/\epsilon)$. Recall that there are a total of $O(\log\frac{1}{\epsilon})$ phases. Hence, the overall sample complexity is $\tilde{O}(d/\epsilon) \cdot \log\frac{1}{\epsilon} = \tilde{O}(d/\epsilon)$.

The analyses of failure probability, error rate, and computational complexity are standard; see, e.g. Awasthi et al. (2017); Shen & Zhang (2021). □

3. Learning with Nasty Noise

In this section we show that the algorithm and analysis of Awasthi et al. (2017) can be modified to tolerate the nasty noise of Bshouty et al. (2002), with near-optimal noise tolerance and sample complexity.

Although under the nasty noise, we can still decompose $T = T_\mathrm{C} \cup T_\mathrm{D}$ where T_C is the set of clean instances in $X_{u,b}$, the main technical challenge to generalize the results of the preceding section to the nasty noise model is that the instances in T_C may no longer be i.i.d. draws from $D_{u,b}$ due to the extra operation of erasing clean instances. Therefore, many crucial results such as Proposition 8 do not hold, making it subtle to characterize the performance of soft outlier removal. Denote by T_E the clean instances residing in $X_{u,b}$ but were replaced with dirty instances by the adversary. What we can say is that $T_\mathrm{C} \cup T_\mathrm{E}$ are i.i.d. draws from $D_{u,b}$, though T_E is not accessible to the learner. Our main technical insight to handle the nasty noise is that if the nasty noise rate η is small, and we perform instance localization (i.e. rejection sampling) as in Algorithm 1, then it is still possible to construct an *extended* empirical distribution over $T \cup T_{\mathrm{E}}$ through soft outlier removal, under which the reweighted hinge loss on T is almost a good proxy to the hinge loss on the *original* clean instances $T_{\mathrm{C}} \cup T_{\mathrm{E}}$. We show that to do so, it suffices to have $|T| = \tilde{O}(d)$. Then we can reuse the analysis in Section 2 to show that the margin-based active learning algorithm PAC learns the underlying halfspace in polynomial time using $\tilde{O}(d/\epsilon)$ samples.

First of all, we formally introduce the problem setup with a few useful notations that will frequently be used in our subsequent analysis.

Passive learning under nasty noise. In the passive learning model, the sample generation oracle $\mathrm{EX}_\eta(D, w^*; N)$ takes as input a sample size N requested by the learner, draws N i.i.d. instances $x_1, \dots, x_N$ from D and labels them correctly, i.e. each $y_i = \operatorname{sign}(w^* \cdot x_i)$, which forms the labeled clean instance set $\hat{A}' = \{(x_1, y_1), \dots, (x_N, y_N)\}$. It then chooses any $N_{\mathrm{D}} = \eta N$ samples in $\hat{A}'$, and replaces them with arbitrary pairs in $\mathcal{X} \times \mathcal{Y}$. This corrupted sample set, denoted by $\hat{A}$, is returned to the learner. If we pass the parameter $N = 1$ to $\mathrm{EX}_\eta(D, w^*; N)$ and repeatedly call it, then the problem reduces to learning with malicious noise. However, if we must pass a large parameter N, the adversary under nasty noise is more powerful than that under malicious noise: it can inspect all the clean samples and decide which of them to corrupt. In passive learning, $\mathrm{EX}_\eta(D, w^*; N)$ is often called only once with N being the total number of samples needed by the learner (Bshouty et al., 2002; Diakonikolas et al., 2018).

Active learning under nasty noise. In the active learning setting, the sample generation process remains unchanged. However, instead of having direct access to $\mathrm{EX}_\eta(D, w^*; N)$ which returns the labeled set, the learner calls $\mathrm{EX}_\eta^x(D, w^*; N)$ to obtain the unlabeled corrupted instance set A (i.e. A is obtained by removing all the labels in $\hat{A}$). It can then decide to reveal the labels for some of the instances in A by calling EX^y. The sample complexity refers to the size of A, and the label complexity refers to the number of calls to EX^y.

3.1. Algorithm

The nasty-noise-tolerant algorithm is given in Algorithm 3, which is almost the same as Algorithm 1, with the major difference that the learner passes a parameter N to $\mathrm{EX}_\eta^x(D, w^*; N)$ at the beginning of each phase and only keeps those lying in a band to form the actually used instance set T. Then, during hinge loss minimization, the label revealing oracle EX^y is called to reveal the labels of all instances in T. Since the size of T is no greater than N, the number of unlabeled samples, such active learning scheme reduces the labeling cost.

Algorithm 3 Efficient and Sample-Optimal Algorithm Tolerating Nasty Noise

```
Require: Error rate $\epsilon$, failure probability $\delta$, instance generation oracle $\mathrm{EX}_\eta^x(D, w^*; N)$, label revealing oracle $\mathrm{EX}^y$.
Ensure: Halfspace $\tilde{w}$ with $\mathrm{err}_D(\tilde{w}) \le \epsilon$ with probability $1 - \delta$.
Initialize $w_0$ as the zero vector in $\mathbb{R}^d$.
$K \leftarrow O(\log \frac{1}{\epsilon})$.
for phases $k = 1, 2, \dots, K$ do
    Clear the working set $T$.
    $b_k \leftarrow \Theta(2^{-k})$, $r_k \leftarrow \Theta(2^{-k})$, $\tau_k \leftarrow \Theta(2^{-k})$.
    Call $\mathrm{EX}_\eta^x(D, w^*; N)$ with $N = N_k$ to form instance set $A$. If $k = 1$, $T \leftarrow A$; otherwise, $T \leftarrow \{x \in A : |w_{k-1} \cdot x| \le b_k\}$.
    Apply Algorithm 2 to $T$ with $u \leftarrow w_{k-1}$, $b \leftarrow b_k$, $r \leftarrow r_k$, $\xi \leftarrow \xi_k$, $c \leftarrow 4C_2$, and let $q = \{q(x)\}_{x \in T}$ be the returned function. Normalize $q$ to form a probability distribution $p$ over $T$.
    $W_k \leftarrow \{w : \|w\|_2 \le 1, \|w - w_{k-1}\|_2 \le r_k\}$, $\hat{T} \leftarrow$ call $\mathrm{EX}^y$ to reveal the labels of $T$. Find $v_k \in W_k$ with
        $$\ell_{\tau_k}(v_k; p \circ \hat{T}) \le \min_{w \in W_k} \ell_{\tau_k}(w; p \circ \hat{T}) + \kappa.$$
    $w_k \leftarrow \frac{v_k}{\|v_k\|_2}$.
end for
return $\tilde{w} \leftarrow w_K$.
```

Observe that we make a total of K calls to $\mathrm{EX}_\eta^x(D, w^*; N)$, which is different from the passive learning algorithms of Bshouty et al. (2002); Diakonikolas et al. (2018) in that they call the oracle only once throughout learning; thus the results here are not strictly comparable to theirs but are still more general than what we have presented in Section 2.

3.1.1. Hyper-parameter setting

We elaborate on our hyper-parameter setting that is used in Algorithm 3 and in our analysis; note that such concrete setting also applies to Algorithm 1. Let $g(t) = c_2 \left(2t \exp(-t) + \frac{c_3 \pi}{4} \exp\left(-\frac{c_4 t}{4\pi}\right) + 16 \exp(-t)\right)$, where the constants are specified in Appendix A. Observe that there exists an absolute constant $\bar{c} \ge 8\pi/c_4$ satisfying $g(\bar{c}) \le 2^{-8}\pi$, since the continuous function $g(t) \to 0$ as $t \to +\infty$ and all the involved quantities in $g(t)$ are absolute constants. Given such constant $\bar{c}$, we set the constant $\kappa = \exp(-\bar{c})$, $r_1 = 1$ and $r_k = 2^{-k-6}$ for $k \ge 2$, $b_k = \bar{c} \cdot r_k$, $\tau_k = c_0 \kappa \cdot \min\{b_k, 1/9\}$, $\delta_k = \frac{\delta}{(k+1)(k+2)}$, and choose $\xi_k = \min\left\{\frac{1}{2}, \frac{\kappa^2}{16}\left(1 + 4\sqrt{C_2} z_k / \tau_k\right)^{-2}\right\}$ where $z_k = \sqrt{b_k^2 + r_k^2}$. It is easy to see that all ξ_k's are lower bounded by a constant $c_6 := \min\left\{\frac{1}{2}, \frac{\kappa^2}{16}\left(1 + \frac{4}{c_0 \kappa \bar{c}}\sqrt{C_2 \bar{c}^2 + C_2}\right)^{-2}\right\}$

and are upper bounded by $\frac{1}{2}$, thus they behave as $\frac{1}{2} - \Theta(1)$. Our theoretical guarantee holds for any noise rate $\eta \leq c_5 \epsilon$, where the constant $c_5 := \frac{c_8}{2\pi} \bar{c} c_1 c_6$.

We set the total number of phases $K = \log\left(\frac{\pi}{32 c_1 \epsilon}\right)$. For any phase $k \geq 1$, we set $N_k = \frac{d}{b_k} \cdot \operatorname{polylog}\left(d, \frac{1}{b_k}, \frac{1}{\delta_k}\right)$ which is the number of instances requested by the learner.

3.2. Analysis

Throughout the section, we always presume that we are addressing the nasty noise model under Assumption 1.

We decompose $A = A_{\mathrm{C}} \cup A_{\mathrm{D}}$, where A_{C} is the set of clean instances in A and A_{D} consists of the dirty instances in A. Let A' be the unlabeled clean instance set obtained by removing all labels in $\hat{A}'$. We introduce the instance set $A_{\mathrm{E}} = A' \backslash A_{\mathrm{C}}$, which was erased from A' by the adversary.

The following lemma follows directly from the noise model and the Chernoff bound, which states that there are not too many dirty instances in A.

Lemma 11. *Consider the nasty noise model with noise rate $\eta \leq c_5 \epsilon$. Then $|A_{\mathrm{D}}| \leq \frac{1}{2} c_8 \xi b N$ and $|A_{\mathrm{C}}| \geq (1 - \frac{1}{2} c_8 \xi b) N$.*

Next, we have an important consequence showing that when localizing the instances in the band $X_{u,b}$, the nasty noise rate stays as a small constant and there are sufficient clean instances in A that are retained.

Lemma 12. *Let $\eta \leq c_5 \epsilon$. By calling $\mathrm{EX}_\eta^x(D, w^*; N)$, the following hold simultaneously with probability $1 - \delta$:*

1. $\frac{|T_{\mathrm{D}}|}{|T|} \leq \xi$;
2. $|T_{\mathrm{C}}| \geq \frac{1}{2} c_8 (1 - \xi) b N$ *and* $|T_{\mathrm{E}}| \leq \frac{1}{2} c_8 \xi b N$.

By Part 2 of the above lemma, we know that $|T_{\mathrm{C}} \cup T_{\mathrm{E}}| \geq \Omega(bN)$. Hence results similar to Proposition 8 immediately hold on the i.i.d. instance set $T_{\mathrm{C}} \cup T_{\mathrm{E}}$ provided that N is large enough.

Proposition 13. *Let $M = \frac{1}{|T_{\mathrm{C}} \cup T_{\mathrm{E}}|} \sum_{x \in T_{\mathrm{C}} \cup T_{\mathrm{E}}} x x^\top$. If $N \geq \frac{d}{b} \cdot \operatorname{polylog}\left(d, \frac{1}{\delta}\right)$, then with probability $1 - \delta$, $\lambda_{\max}(M) \leq O(1)$.*

The above proposition suffices to show that results similar to Theorem 5 hold on $T_{\mathrm{C}} \cup T_{\mathrm{E}}$. Thus, if the learner were given $T \cup T_{\mathrm{E}}$, then Proposition 13 would imply the success of soft outlier removal under the nasty noise. Nevertheless, T_{E} is in reality inaccessible to the learner; we will hence need a more careful analysis to establish the performance guarantee, which is the theme of the next theorem.

Theorem 14. *Let $\eta \leq c_5 \epsilon$ and $N \geq \frac{d}{b} \cdot \operatorname{polylog}\left(d, \frac{1}{\delta}\right)$. With probability $1 - \delta$, Algorithm 2 outputs a function $q : T \rightarrow [0, 1]$ in polynomial time with the following properties:*

1. *for all* $x \in T$, $q(x) \in [0, 1]$;
2. $\frac{1}{|T|} \sum_{x \in T} q(x) \geq 1 - \xi$;
3. $\sup_{w \in W} \frac{1}{|T|} \sum_{x \in T} q(x) (w \cdot x)^2 \leq c\left(b^2 + r^2\right)$.

Observe that the above theorem already guarantees an $\tilde{O}(d/b)$ sample complexity bound for the success of soft outlier removal. It remains to show that the output of Algorithm 3 has small error rate with respect to D and w^*. To this end, we need to characterize the performance of hinge loss minimization. Our approach is to link the reweighted hinge loss over T to the hinge loss over $T_{\mathrm{C}} \cup T_{\mathrm{E}}$. This is because the latter is a good approximation to the expected hinge loss on clean samples in light of uniform concentration, which itself acts as a surrogate of a localized error rate (that is of our interest).

Let $\hat{T}_{\mathrm{C}} = \{(x, \operatorname{sign}(w^* \cdot x)) : x \in T_{\mathrm{C}}\}$ be the (unrevealed) labeled set of T_{C} (note that T_{C} only contains clean instances, hence they are labeled correctly by the adversary); likewise we denote by $\hat{T}_{\mathrm{E}}$ the (unrevealed) labeled set of T_{E}. Define $\ell_\tau(w; p \circ \hat{T}) = \frac{1}{|T|} \sum_{x \in T} p(x) \cdot \max\left\{0, 1 - \frac{1}{\tau} y_x w \cdot x\right\}$ be the reweighted hinge loss over T where y_x denotes the label of x that the adversary is committed to and $p(x)$ was calculated in Step 7 of Algorithm 3.

Proposition 15. *Let $\eta \leq c_5 \epsilon$. If $N \geq \frac{d}{b} \cdot \operatorname{polylog}\left(d, \frac{1}{\delta}\right)$, then with probability $1 - \delta$,*

$$\sup_{w \in W} \left| \ell_\tau(w; \hat{T}_{\mathrm{C}} \cup \hat{T}_{\mathrm{E}}) - \ell_\tau(w; p \circ \hat{T}) \right| \leq \kappa,$$

where κ was defined in Section 3.1.1.

The above robust approximation of hinge loss combined with standard uniform concentration bounds (which require sample complexity $\tilde{O}(d)$) suffices to establish the following key lemma: the error rate within the band is a constant.

Lemma 16. *Let $\eta \leq c_5 \epsilon$. Consider phase k of Algorithm 3 with hyper-parameter settings in Section 3.1.1. If $w^* \in W_k$, then with probability $1 - \frac{\delta}{(k+1)(k+2)}$,*

$$\operatorname{err}_{D_{w_{k-1}, b_k}}(v_k) \leq 6\kappa.$$

It is important to note a small constant error rate within the band implies an $O(\epsilon)$ error rate of the final output $\tilde{w}$ over the distribution D – a well-known fact in margin-based active learning framework (Balcan et al., 2007; Awasthi et al., 2017). Therefore, Lemma 16 has an immediate implication of the correctness of Theorem 2; the full proof can be found in Appendix C.7.

4. Conclusion

This paper provides an improved analysis on the sample complexity of a well-established algorithm for learning of homogeneous halfspaces under the malicious noise. It is shown that by leveraging a matrix Chernoff-type inequality with localization, the obtained sample complexity is optimal up to logarithmic factors. We also extend our analysis to the stronger nasty noise model, and show the achievability of near-optimal noise tolerance and sample complexity by an efficient algorithm when the learner is permitted to communicate with the adversary for multiple rounds.

Acknowledgements

We thank Chicheng Zhang for bringing the paper of Tropp (2012) into our attention, and for insightful discussions on proof ideas for learning with nasty noise. We also thank Jing Wang for valuable feedback on the merit of the work, and thank the anonymous reviewers for helpful suggestions on improving the presentation. This work is supported by NSF-IIS-1948133 and the startup funding of Stevens Institute of Technology.

References

Angluin, D. and Laird, P. D. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Anthony, M. and Bartlett, P. L. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

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

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

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

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local Rademacher complexities. *The Annals of Statistics*, 33 (4):1497 – 1537, 2005.

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

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

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

Cesa-Bianchi, N., Dichterman, E., Fischer, P., Shamir, E., and Simon, H. U. Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM*, 46(5): 684–719, 1999.

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

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

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Robust estimators in high dimensions without the computational intractability. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, pp. 655–664, 2016.

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

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

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

Diakonikolas, I., Kane, D. M., Kontonis, V., Tzamos, C., and Zarifis, N. A polynomial time algorithm for learning halfspaces with Tsybakov noise. *CoRR*, abs/2010.01705, 2020b.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Learning halfspaces with Massart noise under structured distributions. In *Proceedings of the 33rd Annual Conference on Learning Theory*, pp. 1486–1513, 2020c.

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

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

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

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

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

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

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

Maass, W. and Turán, G. How fast can a threshold gate learn? In *Proceedings of a workshop on computational learning theory and natural learning systems (vol. 1): constraints and prospects*, pp. 381–414, 1994.

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

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

Rudelson, M. and Vershynin, R. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

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

Shen, J. On the power of localized Perceptron for label-optimal learning of halfspaces with adversarial noise. *CoRR*, abs/2012.10793, 2020.

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

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

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Tsybakov, A. B. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

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

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

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

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

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