Carnegie Mellon University
Research Showcase @ CMU

Machine Learning Department | School of Computer Science

1-3-2014

The Power of Localization for Efficiently Learning Linear Separators with Noise

Pranjal Awasthi
Princeton University

Maria-Florina Balcan
Georgia Institute of Technology, ninamf@cs.cmu.edu

Philip M. Long
Microsoft

Follow this and additional works at: http://repository.cmu.edu/machine_learning

 Part of the Computer Sciences Commons

Published In

Proceedings of the 46th Annual ACM Symposium on Theory of Computing, 449-458.

This Conference Proceeding is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Machine Learning Department by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

The Power of Localization for Efficiently Learning Linear Separators with Noise

Pranjal Awasthi
pawasthi@princeton.edu

Maria Florina Balcan
ninamf@cc.gatech.edu

Philip M. Long
plong@microsoft.com

January 3, 2014

Abstract

We introduce a new approach for designing computationally efficient learning algorithms that are tolerant to noise, one of the most fundamental problems in learning theory. We demonstrate the effectiveness of our approach by designing algorithms with improved noise tolerance guarantees for learning linear separators, the most widely studied and used concept class in machine learning.

We consider two of the most challenging noise models studied in learning theory, the *malicious* noise model of Valiant [Val85, KL88] and the *adversarial* label noise model of Kearns, Schapire, and Sellie [KSS94]. For malicious noise, where the adversary can corrupt an η of fraction both the label part and the feature part, we provide a polynomial-time algorithm for learning linear separators in $\Re^d$ under the uniform distribution with near information-theoretic optimal noise tolerance of $\eta = \Omega(\epsilon)$. This improves significantly over previously best known results of [KKMS05, KLS09]. For the *adversarial label noise* model, where the distribution over the feature vectors is unchanged, and the overall probability of a noisy label is constrained to be at most η, we give a polynomial-time algorithm for learning linear separators in $\Re^d$ under the uniform distribution that can handle a noise rate of $\eta = \Omega(\epsilon)$. This improves significantly over the results of [KKMS05] which either required runtime super-exponential in $1/\epsilon$ (ours is polynomial in $1/\epsilon$) or tolerated less noise.

In the case that the distribution is isotropic log-concave, we present a polynomial-time algorithm for the malicious noise model that tolerates $\Omega\left(\frac{\epsilon}{\log^2(1/\epsilon)}\right)$ noise, and a polynomial-time algorithm for the adversarial label noise model that also handles $\Omega\left(\frac{\epsilon}{\log^2(1/\epsilon)}\right)$ noise. Both of these also improve on results from [KLS09]. In particular, in the case of malicious noise, unlike previous results, our noise tolerance has no dependence on the dimension d of the space.

A particularly nice feature of our algorithms is that they can naturally exploit the power of active learning, a widely studied modern learning paradigm, where the learning algorithm can only receive the classifications of examples when they ask for them. We show that in this model, our algorithms achieve a label complexity whose dependence on the error parameter ϵ is *exponentially better* than that of any passive algorithm. This provides the first polynomial-time active learning algorithm for learning linear separators in the presence of adversarial label noise, as well as the first analysis of active learning under the challenging malicious noise model.

Our algorithms and analysis combine several ingredients including aggressive localization, hinge loss minimization, and a novel localized and soft outlier removal procedure. Our work illustrates an unexpected use of localization techniques (previously used for obtaining better sample complexity results) in order to obtain better noise-tolerant polynomial-time algorithms.

1 Introduction

Overview. Dealing with noisy data is one of the main challenges in machine learning and is a highly active area of research. In this work we study the noisy learnability of linear separators, arguably the most popular class of functions used in practice [CST00]. Linear separators are at the heart of methods ranging from support vector machines (SVMs) to logistic regression to deep networks, and their learnability has been the subject of intense study for over 50 years. Learning linear separators from correctly labeled (non-noisy) examples is a very well understood problem with simple efficient algorithms like Perceptron being effective both in the classic passive learning setting [KV94, Vap98] and in the more modern active learning framework [Das11]. However, for noisy settings, except for the special case of uniform random noise, very few positive algorithmic results exist even for passive learning. In the context of theoretical computer science more broadly, problems of noisy learning are related to seminal results in approximation-hardness [ABSS93, GR06], cryptographic assumptions [BFKL94, Reg05], and are connected to other classic questions in learning theory (e.g., learning DNF formulas [KSS94]), and appear as barriers in differential privacy [GHRU11]. Hence, not surprisingly, designing efficient algorithms for learning linear separators in the presence of adversarial noise (see definitions below) is of great importance.

In this paper we present new techniques for designing efficient algorithms for learning linear separators in the presence of *malicious* and *adversarial* noise. These are two of the most challenging noise models studied in learning theory. The models were originally proposed for a setting in which the algorithm must work for an arbitrarily, unknown distribution. As we will see, bounds on the amount of noise tolerated for this setting, however were very weak, and no significant progress was made for many years. This gave rise to the question of the role that the distribution played in determining the limits of noise tolerance. A breakthrough result of [KKMS05] and subsequent work of [KLS09] showed that indeed better bounds on the level of noise tolerance can be obtained for the uniform and more generally isotropic log-concave distributions. In this paper, we significantly improve these results. For the malicious noise case, where the adversary can corrupt both the label part and the feature part of the observation (and it has unbounded computational power and access to the entire history of the learning algorithm's computation), we design an efficient algorithm that can tolerate near-optimal amount of malicious noise (within constant factor of the statistical limit) for the uniform distribution, and also significantly improves over the previously known results for log-concave distribution. In particular, unlike previous works, our noise tolerance limit has no dependence on the dimension d of the space. We also show similar improvements for adversarial label noise, and furthermore show that our algorithms can naturally exploit the power of active learning. Active learning is a widely studied modern learning paradigm, where the learning algorithm only receives the classifications of examples when it asks for them. We show that in this model, our algorithms achieve a label complexity whose dependence on the error parameter ϵ is *exponentially better* than that of any passive algorithm. This provides the first polynomial-time active learning algorithm for learning linear separators in the presence of adversarial label noise, solving an open problem posed in [BBL06, Mon06]. It also provides as well as the first analysis showing the benefits of active learning over passive learning under the challenging malicious noise model.

Overall, our work illustrates an unexpected use of localization techniques (previously used for obtaining better sample complexity results) in order to obtain better noise-tolerant polynomial-time algorithms. Our work brings a new set of algorithmic and analysis techniques including localization and soft outlier removal, that we believe will have other applications in learning theory and optimization more broadly.

In the following we start by formally defining the learning models we consider, we then present most relevant prior work, and then our main results and techniques.

Passive and Active Learning. Noise Models. In this work we consider the problem of learning linear separators in two important learning paradigms: the classic passive learning setting and the more modern active learning scenario. As typical [KV94, Vap98], we assume that there exists a distribution D over $\Re^d$ and a fixed unknown target function w^*. In the noise-free settings, in the *passive supervised learning* model the algorithm is given access to a distribution oracle $EX(D, w^*)$ from which it can get training samples $(x, \text{sign}(w^* \cdot x))$ where $x \sim D$. The goal of the algorithm is to output a hypothesis w such that $err_D(w) = \Pr_{x\sim D}[\text{sign}(w^* \cdot x) \neq \text{sign}(w \cdot x)] \leq \epsilon$. In the active learning model [CAL94, Das11] the learning algorithm is given as input a pool of unlabeled examples drawn from the distribution oracle. The algorithm can then query for the labels of examples of its choice from the pool. The goal is to produce a hypothesis of low error while also optimizing for the number of label queries (also known as *label complexity*). The hope is that in the active learning setting we can output a classifier of small error by using many fewer label requests than in the passive learning setting by actively directing the queries to informative examples (while keeping the number of unlabeled examples polynomial).

In this work we focus on two important and realistic noise models. The first one is the malicious noise model of [Val85, KL88] where samples are generated as follows: with probability $(1 - \eta)$ a random pair (x, y) is output where $x \sim D$ and $y = \text{sign}(w^* \cdot x)$; with probability η the adversary can output an arbitrary pair $(x, y) \in \Re^d \times \{-1, 1\}$. We will call η the noise rate. Each of the adversary's examples can depend on the state of the learning algorithm and also the previous draws of the adversary. We will denote the malicious oracle as $EX_\eta(D, w^*)$. The goal remains however that of achieving arbitrarily good predictive approximation to the underlying target function with respect to the underlying distribution, that is to output a hypothesis w such that $\Pr_{x\sim D}[\text{sign}(w^* \cdot x) \neq \text{sign}(w \cdot x)] \leq \epsilon$.

In this paper, we consider an extension of the malicious noise model [Val85, KL88] to the the active learning model as follows. There are two oracles, an example generation oracle and a label revealing oracle. The example generation oracle works as usual in the malicious noise model: with probability $(1 - \eta)$ a random pair (x, y) is generated where $x \sim D$ and $y = \text{sign}(w^* \cdot x)$; with probability η the adversary can output an arbitrary pair $(x, y) \in \Re^d \times \{-1, 1\}$. In the active learning setting, unlike the standard malicious noise model, when an example (x, y) is generated, the algorithm only receives x, and must make a separate call to the label revealing oracle to get y. The goal of the algorithm is still to output a hypothesis w such that $\Pr_{x\sim D}[\text{sign}(w^* \cdot x) \neq \text{sign}(w \cdot x)] \leq \epsilon$.

In the adversarial label noise model, before any examples are generated, the adversary may choose a joint distribution P over $\Re^d \times \{-1, 1\}$ whose marginal distribution over $\Re^d$ is D and such that $\Pr_{(x,y)\sim P}(\text{sign}(w^* \cdot x) \neq y) \leq \eta$. In the active learning model, we will have two oracles, and example generation oracle and a label revealing oracle. We note that the results from our theorems in this model translate immediately into similar guarantees for the agnostic model of [KSS94] (used routinely both in passive and active learning (e.g., [KKMS05, BBL06, Han07]) – see Appendix G for details.

We will be interested in algorithms that run in time $poly(d, 1/\epsilon)$ and use $poly(d, 1/\epsilon)$ samples. In addition, for the active learning scenario we want our algorithms to also optimize for the number of label requests. In particular, we want the number of labeled examples to depend only polylogarithmically in $1/\epsilon$. The goal then is to quantify for a given value of ϵ, the tolerable noise rate $\eta(\epsilon)$ which would allow us to design an efficient (passive or active) learning algorithm.

Previous Work. In the context of passive learning, Kearns and Li's analysis [KL88] implies that halfspaces can be efficiently learned with respect to arbitrary distributions in polynomial time while tolerating a malicious noise rate of $\tilde{\Omega}\left(\frac{\epsilon}{d}\right)$. A slight variant of a construction due to Kearns and Li [KL88] shows that malicious noise at a rate greater than $\frac{\epsilon}{1+\epsilon}$, cannot be tolerated by algorithms learning halfspaces when the distribution is uniform over the unit sphere. The $\tilde{\Omega}\left(\frac{\epsilon}{d}\right)$ bound for the distribution-free case was not improved

for many years. Kalai et al. [KKMS05] showed that, when the distribution is uniform, the poly$(d, 1/\epsilon)$-time averaging algorithm tolerates malicious noise at a rate $\Omega(\epsilon/\sqrt{d})$. They also described an improvement to $\tilde{\Omega}(\epsilon/d^{1/4})$ based on the observation that uniform examples will tend to be well-separated, so that pairs of examples that are too close to one another can be removed, and this limits an adversary's ability to coordinate the effects of its noisy examples. [KLS09] analyzed another approach to limiting the coordination of the noisy examples and proposed an outlier removal procedure that used PCA to find any direction u onto which projecting the training data led to suspiciously high variance, and removing examples with the most extreme values after projecting onto any such u. Their algorithm tolerates malicious noise at a rate $\Omega(\epsilon^2/\log(d/\epsilon))$ under the uniform distribution.

Motivated by the fact that many modern learning applications have massive amounts of unannotated or unlabeled data, there has been significant interest in machine learning in designing active learning algorithms that most efficiently utilize the available data, while minimizing the need for human intervention. Over the past decade there has been substantial progress progress on understanding the underlying statistical principles of active learning, and several general characterizations have been developed for describing when active learning could have an advantage over the classic passive supervised learning paradigm both in the noise free settings and in the agnostic case [FSST97, Das05, BBL06, BBZ07, Han07, DHM07, CN07, BHW08, Kol10, BHLZ10, Wan11, Das11, RR11, BH12]. However, despite many efforts, except for very simple noise models (random classification noise [BF13] and linear noise [DGS12]), to date there are no known computationally efficient algorithms with provable guarantees in the presence of noise. In particular, there are no computationally efficient algorithms for the agnostic case, and furthermore no result exists showing the benefits of active learning over passive learning in the malicious noise model, where the feature part of the examples can be corrupted as well. We discuss additional related work in Appendix A.

1.1 Our Results

1. We give a poly$(d, 1/\epsilon)$-time algorithm for learning linear separators in $\Re^d$ under the uniform distribution that can handle a noise rate of $\eta = \Omega(\epsilon)$, where ϵ is the desired error parameter. Our algorithm (outlined in Section 3) is quite different from those in [KKMS05] and [KLS09] and improves significantly on the noise robustness of [KKMS05] by roughly a factor $d^{1/4}$ and on the noise robustness of [KLS09] by a factor $\frac{\log d}{\epsilon}$. Our noise tolerance is near-optimal and is within a constant factor of the statistical lower bound of $\frac{\epsilon}{1+\epsilon}$. In particular we show the following.

   Theorem 1.1. *There is a polynomial-time algorithm A_{um} for learning linear separators with respect to the uniform distribution over the unit ball in $\Re^d$ in the presence of malicious noise such that an $\Omega(\epsilon)$ upper bound on η suffices to imply that for any $\epsilon, \delta > 0$, the output w of A_{um} satisfies $\Pr_{(x,y)\sim D}[\mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w^* \cdot x)] \leq \epsilon$ with probability at least $1 - \delta$.*

2. For the adversarial noise model, we give a poly$(d, 1/\epsilon)$-time algorithm for learning with respect to the uniform distribution that tolerates a noise rate $\Omega(\epsilon)$.

   Theorem 1.2. *There is a polynomial-time algorithm A_{ul} for learning linear separators with respect to the uniform distribution over the unit ball in $\Re^d$ in the presence of adversarial label noise such that an $\Omega(\epsilon)$ upper bound on η suffices to imply that for any $\epsilon, \delta > 0$, the output w of A_{um} satisfies $\Pr_{(x,y)\sim D}[\mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w^* \cdot x)] \leq \epsilon$ with probability at least $1 - \delta$.*

   As a restatement of the above theorem, in the agnostic setting considered in [KKMS05], we can output a halfspace of error at most $O(\eta + \alpha)$ in time poly$(d, 1/\alpha)$. The previous best result of [KKMS05] achieves this by learning a low degree polynomial in time whose dependence on ϵ is exponential.

3. We obtain similar results for the case of isotropic log-concave distributions.

   Theorem 1.3. *There is a polynomial-time algorithm A_{ilcm} for learning linear separators with respect to any isotropic log-concave distribution in $\Re^d$ in the presence of malicious noise such that an $\Omega\left(\frac{\epsilon}{\log^2(\frac{1}{\epsilon})}\right)$ upper bound on η suffices to imply that for any $\epsilon, \delta > 0$, the output w of A_{ilcm} satisfies $\Pr_{(x,y)\sim D}[\mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w^* \cdot x)] \leq \epsilon$ with probability at least $1 - \delta$.*

   This improves on the best previous bound of $\Omega\left(\frac{\epsilon^3}{\log^2(d/\epsilon)}\right)$ on the noise rate [KLS09]. Notice that our noise tolerance bound has no dependence on d.

   Theorem 1.4. *There is a polynomial-time algorithm A_{ilcl} for learning linear separators with respect to isotropic log-concave distribution in $\Re^d$ in the presence of adversarial label noise such that an $\Omega\left(\epsilon/\log^2(1/\epsilon)\right)$ upper bound on η suffices to imply that for any $\epsilon, \delta > 0$, the output w of A_{ilcl} satisfies $\Pr_{(x,y)\sim D}[\mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w^* \cdot x)] \leq \epsilon$ with probability at least $1 - \delta$.*

   This improves on the best previous bound of $\Omega\left(\frac{\epsilon^3}{\log(1/\epsilon)}\right)$ on the noise rate [KLS09].
4. A particularly nice feature of our algorithms is that they can naturally exploit the power of active learning. We show that in this model, the label complexity of both algorithms depends only polylogarithmically in $1/\epsilon$ where ϵ is the desired error rate, while still using only a polynomial number of unlabeled samples (for the uniform distribution, the dependence of the number of labels on ϵ is $O(\log(1/\epsilon))$). Our efficient algorithm that tolerates adversarial label noise solves an open problem posed in [BBL06, Mon06]. Furthermore, our paper provides the first active learning algorithm for learning linear separators in the presence of non-trivial amount of adversarial noise that can affect not only the label part, but also the feature part.

Our work exploits the power of localization for designing noise-tolerant polynomial-time algorithms. Such localization techniques have been used for analyzing sample complexity for passive learning (see [BBM05, BBL05, Zha06, BLL09, BL13]) or for designing active learning algorithms (see [BBZ07, Kol10, Han11, BL13]). In order to make such a localization strategy computationally efficient and tolerate malicious noise we introduce several key ingredients described in Section 1.2.

We note that all our algorithms are proper in that they return a linear separator. (Linear models can be evaluated efficiently, and are otherwise easy to work with.) We summarize our results in Tables 1 and 2.

Table 1: Comparison with previous $\mathrm{poly}(d, 1/\epsilon)$-time algs. for uniform distribution

Passive Learning	Prior work	Our work
malicious	$\eta = \Omega(\frac{\epsilon}{d^{1/4}})$ [KKMS05] $\eta = \Omega(\frac{\epsilon^2}{\log(d/\epsilon)})$ [KLS09]	$\eta = \Omega(\epsilon)$
adversarial	$\eta = \Omega(\epsilon/\sqrt{\log(1/\epsilon)})$ [KKMS05]	$\eta = \Omega(\epsilon)$
Active Learning (malicious and adversarial)	NA	$\eta = \Omega(\epsilon)$

1.2 Techniques

Hinge Loss Minimization As minimizing the 0-1 loss in the presence of noise is NP-hard [JP78, GJ90], a natural approach is to minimize a surrogate convex loss that acts as a proxy for the 0-1 loss. A common

Table 2: Comparison with previous $\mathrm{poly}(d, 1/\epsilon)$-time algorithms isotropic log-concave distributions

Passive Learning	Prior work	Our work
malicious	$\eta = \Omega(\frac{\epsilon^3}{\log^2(d/\epsilon)})$ [KLS09]	$\eta = \Omega(\frac{\epsilon}{\log^2(1/\epsilon)})$
adversarial	$\eta = \Omega(\frac{\epsilon^3}{\log(1/\epsilon)})$ [KLS09]	$\eta = \Omega(\frac{\epsilon}{\log^2(1/\epsilon)})$
Active Learning (malicious and adversarial)	NA	$\Omega(\frac{\epsilon}{\log^2(1/\epsilon)})$

choice in machine learning is to use the hinge loss defined as $\ell_\tau(w, x, y) = \max\left(0, 1 - \frac{y(w\cdot x)}{\tau}\right)$, and, for a set T of examples, we let $\ell_\tau(w, T) = \frac{1}{|T|}\sum_{(x,y)\in T} \ell_\tau(w, x, y)$. Here τ is a parameter that changes during training. It can be shown that minimizing hinge loss with an appropriate normalization factor can tolerate a noise rate of $\Omega(\epsilon^2/\sqrt{d})$ under the uniform distribution over the unit ball in $\Re^d$. This is also the limit for such a strategy since a more powerful malicious adversary with can concentrate all the noise directly opposite to the target vector w^* and make sure that the hinge-loss is no longer a faithful proxy for the 0-1 loss.

Localization in the instance and concept space Our first key insight is that by using an iterative localization technique, we can limit the harm caused by an adversary at each stage and hence can still do hinge-loss minimization despite significantly more noise. In particular, the iterative style algorithm we propose proceeds in stages and at stage k, we have a hypothesis vector w_k of a certain error rate. The goal in stage k is to produce a new vector w_{k+1} of error rate half of w_k. In order to halve the error rate, we focus on a band of size $b_k = \Theta(\frac{2^{-k}}{\sqrt{d}})$ around the boundary of the linear classifier whose normal vector is w_k, i.e. $S_{w_k,b_k} = \{x : |w_k \cdot x| < b_k\}$. For the rest of the paper, we will repeatedly refer to this key region of borderline examples as "the band". The key observation made in [BBZ07] is that outside the band, all the classifiers still under consideration (namely those hypotheses within radius r_k of the previous weight vector w_k) will have very small error. Furthermore, the probability mass of this band under the original distributions is small enough, so that in order to make the desired progress we only need to find a hypothesis of constant error rate over the data distribution conditioned on being within margin b_k of w_k. This insight has been crucially used in the [BBZ07] in order to obtain active learning algorithms with improved label complexity ignoring computational complexity considerations[1].

In this work, we show the surprising fact that this idea can be extended and adapted to produce polynomial time algorithms with improved noise tolerance as well! Not only do we use this localization idea for different purposes, but our analysis significantly departs from [BBZ07]. To obtain our results, we exploit several new ideas: (1) the performance of the rescaled hinge loss minimization in smaller and smaller bands, (2) a careful variance analysis, and (3) another type of localization — we develop and analyze a novel *soft and localized outlier removal* procedure. In particular, we first show that if we minimize a variant of the hinge loss that is rescaled depending on the width of the band, it remains a faithful enough proxy for the 0-1 error even when there is significantly more noise. As a first step towards this goal, consider the setting where we pick τ_k proportionally to b_k, the size of the band, and r_k is proportional to the error rate of w_k, and then minimize a normalized hinge loss function $\ell_{\tau_k}(w, x, y) = \max(0, 1 - \frac{y(w\cdot x)}{\tau_k})$ over vectors $w \in B(w_k, r_k)$. We first show that w^* has small hinge loss within the band. Furthermore, within the band the adversarial examples cannot hurt the hinge loss of w^* by a lot. To see this notice that if the malicious noise rate is η, within S_{w_{k-1},b_k} the effective noise rate is $\Theta(\eta 2^k)$. Also the maximum value of the hinge loss for vectors $w \in B(w_k, 2^{-k})$ is $O(\sqrt{d})$. Hence the maximum amount by which the adversary can affect the hinge loss

[1]We note that the localization considered by [BBZ07] is a more aggressive one than those considered in disagreement based active learning literature [BBL06, Han07, Kol10, Han11, Wan11] and earlier in passive learning [BBM05, BBL05, Zha06].

is $O(\eta 2^k\sqrt{d})$. Using this approach we get a noise tolerance of $\Omega(\epsilon/\sqrt{d})$.

In order to get a much better noise tolerance in the adversarial or agnostic setting, we crucially exploit a careful analysis of the variance of $w \cdot x$ for vectors w close to the current vector w_{k-1}, one can get a much tighter bound on the amount by which an adversary can "hurt" the hinge loss. This then leads to an improved noise tolerance of $\Omega(\epsilon)$.

For the case of malicious noise, in addition we need to deal with the presence of outliers, i.e. points not generated from the uniform distribution. We do this by introducing a *soft localized outlier removal* procedure at each stage (described next). This procedure assigns a weight to each data point indicating how "noisy" the point is. We then minimize the weighted hinge loss. This combined with the variance analysis mentioned above leads to a noise of tolerance of $\Omega(\epsilon)$ in the malicious case as well.

Soft Localized Outlier Removal Outlier removal techniques have been studied before in the context of learning problems [BFKV97, KLS09]. The goal of outlier removal is to limit the ability of the adversary to coordinate the effects of noisy examples – excessive such coordination is detected and removed. Our outlier removal procedure (see Figure 2) is similar in spirit to that of [KLS09] with two key differences. First, as in [KLS09], we will use the variance of the examples in a particular direction to measure their coordination. However, due to the fact that in round k, we are minimizing the hinge loss only with respect to vectors that are close to w_{k-1}, we only need to limit the variance in these directions. This variance is $\Theta(b_k^2)$ which is much smaller than $1/d$. This allows us to limit the harm of the adversary to a greater extent than was possible in the analysis of [KLS09]. The second difference is that, unlike previous outlier removal techniques, we do not remove any examples but instead weigh them appropriately and then minimize the weighted hinge loss. The weights indicate how noisy a given example is. We show that these weights can be computed by solving a linear program with infinitely many constraints. We then show how to design an efficient separation oracle for the linear program using recent general-purpose techniques from the optimization community [SZ03, BM13].

In Section 4 we show that our results hold for a more general class of distributions which we call *admissible* distributions. From Section 4 it also follows that our results can be extended to β-log-concave distributions (for small enough β). Such distributions, for instance, can capture mixtures of log-concave distributions [BL13].

2 Preliminaries

Our algorithms and analysis will use the hinge loss defined as $\ell_\tau(w, x, y) = \max\left(0, 1 - \frac{y(w\cdot x)}{\tau}\right)$, and, for a set T of examples, we let $\ell_\tau(w, T) = \frac{1}{|T|}\sum_{(x,y)\in T} \ell_\tau(w, x, y)$. Here τ is a parameter that changes during training. Similarly, the expected hinge loss w.r.t. D is defined as $L_\tau(w, D) = E_{x\sim D}(\ell_\tau(w, x, \mathrm{sign}(w^* \cdot x)))$. Our analysis will also consider the distribution $D_{w,\gamma}$ obtained by conditioning D on membership in the band, i.e. the set $\{x : \|x\|_2 = 1, |w \cdot x| \leq \gamma\}$.

Since it is very natural, for clarity of exposition, we present our algorithms directly in the active learning model. We will prove that our active algorithm only uses a polynomial number of unlabeled samples, which then immediately implies a guarantee for passive learning setting as well. At a high level, our algorithms are iterative learning algorithms that operate in rounds. In each round k we focus our attention and use points that fall near the current hypothesized decision boundary w_{k-1} and use them in order to obtain a new vector w_k of lower error. In the malicious noise case, in round k we first do a soft outlier removal and then minimize hinge loss normalized appropriately by τ_k. A formal description appears in Figure 1, and a formal description of the outlier removal procedure appears in Figure 2. We will present specific choices of the

Figure 1 COMPUTATIONALLY EFFICIENT ALGORITHM TOLERATING MALICIOUS NOISE

Input: allowed error rate ϵ, probability of failure δ, an oracle that returns x, for (x, y) sampled from $\mathrm{EX}_\eta(f, D)$, and an oracle for getting the label from an example; a sequence of unlabeled sample sizes $n_k > 0$ $k \in Z^+$; a sequence of labeled sample sizes $m_k > 0$; a sequence of cut-off values $b_k > 0$; a sequence of hypothesis space radii $r_k > 0$; a sequence of removal rates ξ_k; a sequence of variance bounds σ_k^2; precision value κ; weight vector w_0.

1. Draw n_1 examples and put them into a working set W.
2. For $k = 1, \ldots, s = \lceil \log_2(1/\epsilon) \rceil$
   (a) Apply the algorithm from Figure 2 to W with parameters $u \leftarrow w_{k-1}$, $\gamma \leftarrow b_{k-1}$, $r \leftarrow r_k$, $\xi \leftarrow \xi_k$, $\sigma^2 \leftarrow \sigma_k^2$ and let q be the output function $q : W \to [0, 1]$. Normalize q to form a probability distribution p over W.
   (b) Choose m_k examples from W according to p and reveal their labels. Call this set T.
   (c) Find $v_k \in B(w_{k-1}, r_k)$ to approximately minimize training hinge loss over T s.t. $\|v_k\|_2 \leq 1$:
   $\ell_{\tau_k}(v_k, T) \leq \min_{w \in B(w_{k-1}, r_k) \cap B(0,1))} \ell_{\tau_k}(w, T) + \kappa/8$
   Normalize v_k to have unit length, yielding $w_k = \frac{v_k}{\|v_k\|_2}$.
   (d) Clear the working set W.
   (e) **Until** n_{k+1} additional data points are put in W, given x for $(x, f(x))$ obtained from $\mathrm{EX}_\eta(f, D)$, **if** $|w_k \cdot x| \geq b_k$, **then** reject x **else** put into W

Output: weight vector w_s of error at most ϵ with probability $1 - \delta$.

parameters of the algorithms in the following sections.

The description of the algorithm and its analysis is simplified if we assume that it starts with a preliminary weight vector w_0 whose angle with the target w^* is acute, i.e. that satisfies $\theta(w_0, w^*) < \pi/2$. We show in Appendix B that this is without loss of generality for the types of problems we consider.

3 Learning with respect to uniform distribution with malicious noise

Let S_{d-1} denote the unit ball in $\mathbf{R}^d$. In this section we focus on the case where the marginal distribution D is the uniform distribution over S_{d-1} and present our results for malicious noise. We present the analysis of our algorithm directly in the active learning model, and present a proof sketch for its correctness in Theorem 3.1 below. The proof of Theorem 1.1 follows immediately as a corollary. Complete proof details are in Appendix C.

Theorem 3.1. *Let w^* be the (unit length) target weight vector. There are absolute positive constants $c_1, ..., c_4$ and a polynomial p such that, an $\Omega(\epsilon)$ upper bound on η suffices to imply that for any $\epsilon, \delta > 0$, using the algorithm from Figure 1 with $\epsilon_0 = 1/8$, cut-off values $b_k = c_1 2^{-k} d^{-1/2}$, radii $r_k = c_2 2^{-k} \pi$, $\kappa = c_3$, $\tau_k = c_4 2^{-k} d^{-1/2}$ for $k \geq 1$, $\xi_k = c\kappa^2$, $\sigma_k = (\frac{r_k^2}{d-1} + b_{k-1}^2)$, a number $n_k = p(d, 2^k, \log(1/\delta))$ of unlabeled examples in round k and a number $m_k = O(d(d + \log(k/\delta)))$ of labeled examples in round k, after $s = \lceil \log_2(1/\epsilon) \rceil$ iterations, we find w_s satisfying $\mathrm{err}(w_s) = \Pr_{(x,y)\sim D}[\mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w^* \cdot x)] \leq \epsilon$ with probability $\geq 1 - \delta$.*

Figure 2 LOCALIZED SOFT OUTLIER REMOVAL PROCEDURE

Input: a set $S = \{(x_1, x_2, \ldots x_n)\}$ samples; the reference unit vector u; desired radius r; a parameter ξ specifying the desired bound on the fraction of clean examples removed; a variance bound σ^2

1. Find $q : S \to [0, 1]$ satisfying the following constraints:
   (a) for all $x \in S$, $0 \leq q(x) \leq 1$
   (b) $\frac{1}{|S|} \sum_{(x,y)\in S} q(x) \geq 1 - \xi$
   (c) for all $w \in B(u, r) \cap B(\mathbf{0}, 1)$, $\frac{1}{|S|} \sum_{x\in S} q(x)(w \cdot x)^2 \leq c\sigma^2$

Output: A function $q : S \to [0, 1]$.

3.1 Proof Sketch of Theorem 3.1

We may assume without loss of generality that all examples, including noisy examples, fall in S_{d-1}. This is because any example that falls outside S_{d-1} can be easily identified by the algorithm as noisy and removed, effectively lowering the noise rate.

A first key insight is that using techniques from [BBZ07], we may reduce our problem to a subproblem concerning learning with respect to a distribution obtained by conditioning on membership in the band. In particular, in Appendix C.1, we prove that, for a sufficiently small absolute constant κ, Theorem 3.2 stated below, together with proofs of its computational, sample and label complexity bounds, suffices to prove Theorem 3.1.

Theorem 3.2. *After round k of the algorithm in Figure 1, with probability at least $1 - \frac{\delta}{k+k^2}$, we have $err_{D_{w_{k-1},b_{k-1}}}(w_k) \leq \kappa$.*

The proof of Theorem 3.2 follows from a series of steps summarized in the lemmas below. First, we bound the hinge loss of the target w^* within the band $S_{w_{k-1},b_{k-1}}$. Since we are analyzing a particular round k, to reduce clutter in the formulas, for the rest of this section, let us refer to ℓ_{τ_k} simply as ℓ and $L_{\tau_k}(\cdot, D_{w_{k-1},b_{k-1}})$ as $L(\cdot)$.

Lemma 3.3. $L(w^*) \leq \kappa/12$.

Proof Sketch: Notice that $y(w^* \cdot x)$ is never negative, so, on any clean example (x, y), we have $\ell(w^*, x, y) = \max\left\{0, 1 - \frac{y(w^*\cdot x)}{\tau_k}\right\} \leq 1$, and, furthermore, w^* will pay a non-zero hinge only inside the region where $|w^* \cdot x| < \tau_k$. Hence, $L(w^*) \leq \Pr_{D_{w_{k-1},b_{k-1}}}(|w^* \cdot x| \leq \tau_k) = \frac{\Pr_{x\sim D}(|w^*\cdot x|\leq\tau_k \,\&\, |w_{k-1}\cdot x|\leq b_{k-1})}{\Pr_{x\sim D}(|w_{k-1}\cdot x|\leq b_{k-1})}$. Using standard tail bounds (see Eq. 1 in Appendix C), we can lower bound the denominator $\Pr_{x\sim D}(|w_{k-1} \cdot x| < b_{k-1}) \geq c_1' b_{k-1}\sqrt{d}$ for a constant c_1'. Also the numerator is at most $\Pr_{x\sim D}(|w^* \cdot x| \leq \tau_k) \leq c_2'\tau_k\sqrt{d}$. For another constant c_2'. Hence, we have $L(w^*) \leq \frac{c_2'\sqrt{d}\tau_k}{c_1'\sqrt{d}b_{k-1}} \leq \kappa/12$, for the appropriate choice of constants c_1' and c_2' and making κ small enough. □

During round k we can decompose the working set W into the set of "clean" examples W_C which are drawn from $D_{w_{k-1},b_{k-1}}$ and the set of "dirty" or malicious examples W_D which are output by the adversary. Next, we will relate the hinge loss of vectors over the weighted set W to the hinge loss over clean examples W_C. In order to do this we will need the following guarantee from the outlier removal subroutine of Figure 2.

Theorem 3.4. *There is a constant c and a polynomial p such that, if $n \geq p(1/\eta, d, 1/\xi, 1/\delta, 1/\gamma)$ examples are drawn from the distribution $D_{u,\gamma}$ (each replaced with an arbitrary unit-length vector with probability $\eta < 1/4$), then by using the algorithm in Figure 1 with $\sigma^2 = \frac{r^2}{d-1} + \gamma^2$, we have that with probability $1-\delta$, the output q of satisfies the following: (a) $\sum_{(x,y)\in S} q(x) \geq (1-\xi)|S|$, and (b) for all unit length w such that $\|w-u\|_2 \leq r$, $\frac{1}{|S|}\sum_{x\in S} q(x)(w\cdot x)^2 \leq c\sigma^2$. Furthermore, the algorithm can be implemented in polynomial time.*

The key points in proving this theorem are the following. We will show that the vector q^* which assigns a weight 1 to examples in W_C and weight 0 to examples in W_D is a feasible solution to the linear program in Figure 2. In order to do this, we first show that the fraction of dirty examples in round k is not too large, i.e., w.h.p., we have $|W_D| = O(\eta|S|)$. Next, we use the improved variance bound from Lemma C.2 regarding $E[(w.x)^2]$ for all w close to u. This bound is $(\frac{r^2}{d-1} + \gamma^2)$. The proof of feasibility follows easily by combining the variance bound with standard VC tools. In the appendix we also show how to solve the linear program in polynomial time. The complete proof of the theorem 3.4 is in Appendix C.

As explained in the introduction, the soft outlier removal procedure allows us to get a much refined bound on the hinge loss over the clean set W_C, i.e., $\ell(w, W_C)$ as compared to the hinge loss over the weighted set W, i.e., $\ell(w,p)$. This is formalized in the following lemma. Here $\ell(w, W_C)$ and $\ell(w,p)$ are defined with respect to the true unrevealed labels that the adversary has committed to.

Lemma 3.5. *There are absolute constants c_1, c_2 and c_3 such that, for large enough d, with probability $1 - \frac{\delta}{2(k+k^2)}$, if we define $z_k = \sqrt{\frac{r_k^2}{d-1} + b_{k-1}^2}$, then for any $w \in B(w_{k-1}, r_k)$, we have $\ell(w, W_C) \leq \ell(w,p) + \frac{c_1\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/32$ and $\ell(w,p) \leq 2\ell(w, W_C) + \kappa/32 + \frac{c_2\eta}{\epsilon} + \frac{c_3\sqrt{\eta/\epsilon}z_k}{\tau_k}$.*

A detailed proof of 3.5 is given in Appendix C. Here were give a few ideas. The loss $\ell(w,x,y)$ on a particular example can be upper bounded by $1 + \frac{|w\cdot x|}{\tau}$. One source of difference between $\ell(w, W_C)$, the loss on the clean examples, and $\ell(w,p)$, the loss minimized by the algorithm, is the loss on the (total fractional) dirty examples that were not deleted by the soft outlier removal. By using the Cauchy-Shwartz inequality, the (weighted) sum of $1 + \frac{|w\cdot x|}{\tau}$ over those surviving noisy examples can be bounded in terms of the variance in the direction w, and the (total fractional) number of surviving dirty examples. Our soft outlier detection allows us to bound the variance of the surviving noisy examples in terms of $\Theta(z_k^2)$. Another way that $\ell(w, W_C)$ can be different from $\ell(w,p)$ is effect of deleting clean examples. We can similarly use the variance on the clean examples to bound this in terms of z. Finally, we can flesh out the detailed bound by exploiting the (soft counterparts of) the facts that most examples are clean and few examples are excluded.

Given, these the proof of Theorem 3.2 can be summarized as follows.

Let $E = \mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k) = \mathrm{err}_{Dw_{k-1},b_{k-1}}(v_k)$ be the probability that we want to bound. Applying VC theory, w.h.p., all sampling estimates of expected loss are accurate to within $\kappa/32$, so we may assume w.l.o.g. that this is the case. Since, for each error, the hinge loss is at least 1, we have $E \leq L(v_k)$. Applying Lemma 3.5 and VC theory, we get, $E \leq \ell(v_k, T) + \frac{c_1\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/8$. Since v_k approximately minimizes the hinge loss, VC theory implies $E \leq \ell(w^*, p) + \frac{c_1\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/3$. Once again applying Lemma 3.5 and VC theory yields $E \leq 2L(w^*) + \frac{c_1\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \frac{c_2\eta}{\epsilon} + \frac{c_3\sqrt{\eta/\epsilon}z_k}{\tau_k} + \kappa/2$. Since $L(w^*) \leq \kappa/12$, we get $E \leq \kappa/6 + \frac{c_1\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \frac{c_2\eta}{\epsilon} + \frac{c_3\sqrt{\eta/\epsilon}z_k}{\tau_k} + \kappa/2$. Now notice that z_k/τ_k is $\Theta(1)$. Hence an $\Omega(\epsilon)$ bound on η suffices to imply, w.h.p., that $\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k) \leq \kappa$.

4 Learning with respect to admissible distributions with malicious noise

One of our main results (Theorem 1.3) concerns isotropic log concave distributions. A probability distribution is *isotropic log-concave* if its density can be written as $\exp(-\psi(x))$ for a convex function ψ, its mean is $\mathbf{0}$, and its covariance matrix is I.

In this section, we extend our analysis from the previous section and show that it works for isotropic log concave distributions, and in fact an even more general class of distributions which we call as *admissible distributions*. In particular this includes the class of isotropic log-concave distributions in $\mathbf{R}^d$ and the uniform distributions over the unit ball in $\mathbf{R}^d$.

Definition 4.1. *A sequence $D_4, D_5, ...$ of probability distributions over $\mathbf{R}^4, \mathbf{R}^5, ...$ respectively is* λ-admissible *if it satisfies the following conditions. (1.) There are $c_1, c_2, c_3 > 0$ such that, for all $d \geq 4$, for x drawn from D_d and any unit length $u \in \mathbf{R}^d$, (a) for all $a, b \in [-c_1, c_1]$ for which $a \leq b$, we have $\Pr(u \cdot x \in [a, b]) \geq c_2|b - a|$ and for all $a, b \in \mathbf{R}$ for which $a \leq b$, $\Pr(u \cdot x \in [a, b]) \leq c_3|b - a|$. (2.) For any $c_4 > 0$, there is a $c_5 > 0$ such that, for all $d \geq 4$, the following holds. Let u and v be two unit vectors in R^d, and assume that $\theta(u, v) = \alpha \leq \pi/2$. Then $\Pr_{x \sim D_d}[\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x)$ and $|v \cdot x| \geq c_5\alpha] \leq c_4\alpha$. (3.) There is an absolute constant c_6 such that, for any $d \geq 4$, for any two unit vectors u and v in R^d we have $c_6\theta(v, u) \leq \Pr_{x \sim D_d}(\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x))$. (4.) There is a constant c_8 such that, for all constant c_7, for all $d \geq 4$, for any a such that, $\|a\|_2 \leq 1$, and $||u - a|| \leq r$, for any $0 < \gamma < c_7$, we have $\mathbf{E}_{x \sim D_{d,u,\gamma}}\left((a \cdot x)^2\right) \leq c_8 \log^\lambda(1 + 1/\gamma)(r^2 + \gamma^2)$. (5.) There is a constant c_9 such that $\Pr_{x \sim D}(||x|| > \alpha) \leq c_9 \exp(-\alpha/\sqrt{d})$.*

For the case of admissible distributions we have the following theorem, which is proved in Appendix D.

Theorem 4.2. *Let a distribution D over R^d be chosen from a λ-admissible sequence of distributions Let w^* be the (unit length) target weight vector. There are settings of the parameters of the algorithm A from Figure 1, such that an $\Omega\left(\frac{\epsilon}{\log^\lambda(\frac{1}{\epsilon})}\right)$ upper bound on the rate η of malicious noise suffices to imply that for any $\epsilon, \delta > 0$, a number $n_k = \text{poly}(d, M^k, \log(1/\delta))$ of unlabeled examples in round k and a number $m_k = O\left(d \log\left(\frac{d}{\epsilon\delta}\right)(d + \log(k/\delta))\right)$ of labeled examples in round $k \geq 1$, and w_0 such that $\theta(w_0, w^*) < \pi/2$, after $s = O(\log(1/\epsilon))$ iterations, finds w_s satisfying $\text{err}(w_s) \leq \epsilon$ with probability $\geq 1 - \delta$.*

If the support of D is bounded in a ball of radius $R(d)$, then, we have that $m_k = O\left(R(d)^2(d + \log(k/\delta))\right)$ label requests suffice.

The above theorem contains Theorem 1.3 as a special case. This is because of the fact that any isotropic log-concave distribution is 2-admissible (see Appendix F.2 for a proof).

5 Adversarial label noise

The intuition in the case of adversarial label noise is the same as for malicious noise, except that, because the adversary cannot change the marginal distribution over the instances, it is not necessary to perform outlier removal. Bounds for learning with adversarial label noise are not corollaries of bounds for learning with malicious noise, however, because, while the marginal distribution over the instances for *all* the examples, clean and noisy, is not affected by the adversary, the marginal distribution over the *clean* examples is changed (because the examples whose labels are flipped are removed from the distribution over clean examples).

Theorem 1.2 and Theorem 1.4, which concern adversarial label noise, can be proved by combining the analysis in Appendix E with the facts that the uniform distribution and i.l.c. distributions are 0-admissible and 2-admissible respectively.

6 Discussion

Localization in this paper refers to the practice of narrowing the focus of a learning algorithm to a restricted range of possibilities (which we know to be safe given the information so far), thereby reducing sensitivity of estimates of the quality of these possibilities based on random data –this in turn leads to better noise tolerance in our work. (Note that, while the examples in the band in round k do not occupy a neighborhood in feature space, they concern differences between hypotheses in a neighborhood around w_{k-1}.) We note that the idea of localization in the concept space is traditionally used in statistical learning theory both in supervised and active learning for getting sharper rates [BBL05, BLL09, Kol10]. Furthermore, the idea of localization in the instance space has been used in margin-based analysis of active learning [BBZ07, BL13]. In this work we used localization in both senses in order to get polynomial-time algorithms with better noise tolerance. It would be interesting to further exploit this idea for other concept spaces.

References

[AB99] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[ABS10] P. Awasthi, A. Blum, and O. Sheffet. Improved guarantees for agnostic learning of disjunctions. *COLT*, 2010.

[ABSS93] S. Arora, L. Babai, J. Stern, and Z. Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. In *Proceedings of the 1993 IEEE 34th Annual Foundations of Computer Science*, 1993.

[Bau90] E. B. Baum. The perceptron algorithm is fast for nonmalicious distributions. *Neural Computation*, 2:248–260, 1990.

[BBL05] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:9:323–375, 2005.

[BBL06] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *ICML*, 2006.

[BBM05] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[BBZ07] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *COLT*, 2007.

[BF13] M.-F. Balcan and V. Feldman. Statistical active learning algorithms. *NIPS*, 2013.

[BFKL94] Avrim Blum, Merrick L. Furst, Michael J. Kearns, and Richard J. Lipton. Cryptographic primitives based on hard learning problems. In *Proceedings of the 13th Annual International Cryptology Conference on Advances in Cryptology*, 1994.

[BFKV97] A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1997.

[BGMN05] F. Barthe, O. Guédon, S. Mendelson, and A. Naor. A probabilistic approach to the geometry of the pn-ball. *The Annals of Probability*, 33(2):480–513, 2005.

[BH12] M.-F. Balcan and S. Hanneke. Robust interactive learning. In *COLT*, 2012.

[BHLZ10] A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *NIPS*, 2010.

[BHW08] M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *COLT*, 2008.

[BL13] M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, 2013.

[BLL09] N. H. Bshouty, Y. Li, and P. M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *JCSS*, 2009.

[BM13] D. Bienstock and A. Michalka. Polynomial solvability of variants of the trust-region subproblem, 2013. Optimization Online.

[BSS12] A. Birnbaum and S. Shalev-Shwartz. Learning halfspaces with the zero-one loss: Time-accuracy tradeoffs. *NIPS*, 2012.

[Byl94] T. Bylander. Learning linear threshold functions in the presence of classification noise. In *Conference on Computational Learning Theory*, 1994.

[CAL94] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2), 1994.

[CGZ10] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Learning noisy linear classifiers via adaptive and selective sampling. *Machine Learning*, 2010.

[CN07] R. Castro and R. Nowak. Minimax bounds for active learning. In *COLT*, 2007.

[CST00] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[Das05] S. Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, volume 18, 2005.

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

[DGS12] O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *JMLR*, 2012.

[DHM07] S. Dasgupta, D.J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. *NIPS*, 20, 2007.

[FGKP06] V. Feldman, P. Gopalan, S. Khot, and A. Ponnuswami. New results for learning noisy parities and halfspaces. In *FOCS*, pages 563–576, 2006.

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

[GHRU11] Anupam Gupta, Moritz Hardt, Aaron Roth, and Jonathan Ullman. Privately releasing conjunctions and the statistical query barrier. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, 2011.

[GJ90] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. 1990.

[GR06] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 2006.

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

[GSSS13] A. Gonen, S. Sabato, and S. Shalev-Shwartz. Efficient pool-based active learning of halfspaces. In *ICML*, 2013.

[Han07] S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

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

[JP78] D. S. Johnson and F. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93 – 107, 1978.

[KKMS05] Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, 2005.

[KL88] Michael Kearns and Ming Li. Learning in the presence of malicious errors. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, 1988.

[KLS09] A. R. Klivans, P. M. Long, and Rocco A. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10, 2009.

[Kol10] V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.

[KSS94] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Mach. Learn.*, 17(2-3), November 1994.

[KV94] M. Kearns and U. Vazirani. *An introduction to computational learning theory*. MIT Press, Cambridge, MA, 1994.

[LS06] P. M. Long and R. A. Servedio. Attribute-efficient learning of decision lists and linear threshold functions under unconcentrated distributions. *NIPS*, 2006.

[LS11] P. M. Long and R. A. Servedio. Learning large-margin halfspaces with more malicious noise. *NIPS*, 2011.

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

[Mon06] Claire Monteleoni. Efficient algorithms for general active learning. In *Proceedings of the 19th annual conference on Learning Theory*, 2006.

[Pol11] D. Pollard. *Convergence of Stochastic Processes*. Springer Series in Statistics. 2011.

[Reg05] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, 2005.

[RR11] M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. In *NIPS*, 2011.

[Ser01] Rocco A. Servedio. Smooth boosting and learning with malicious noise. In *14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, 2001.

[SZ03] J. Sturm and S. Zhang. On cones of nonnegative quadratic functions. *Mathematics of Operations Research*, 28:246–267, 2003.

[Val85] L. G. Valiant. Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial intelligence*, 1985.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[Vem10] S. Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *JACM*, 57(6), 2010.

[Wan11] L. Wang. Smoothness, Disagreement Coefficient, and the Label Complexity of Agnostic Active Learning. *JMLR*, 2011.

[Zha06] T. Zhang. Information theoretical upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

A Additional Related Work

Passive Learning Blum et al. [BFKV97] considered noise-tolerant learning of halfspaces under a more idealized noise model, known as the random noise model, in which the label of each example is flipped with a certain probability, independently of the feature vector. Some other, less closely related, work on efficient noise-tolerant learning of halfspaces includes [Byl94, BFKV97, FGKP06, GR09, Ser01, ABS10, LS11, BSS12].

Active Learning As we have mentioned, most prior theoretical work on active learning focuses on either sample complexity bounds (without regard for efficiency) or on providing polynomial time algorithms in the noiseless case or under simple noise models (random classification [BF13] noise or linear noise [CGZ10, DGS12]).

In [CGZ10, DGS12] online learning algorithms in the selective sampling framework are presented, where labels must be actively queried before they are revealed. Under the assumption that the label conditional distribution is a linear function determined by a fixed target vector, they provide bounds on the regret of the algorithm and on the number of labels it queries when faced with an adaptive adversarial strategy of generating the instances. As pointed out in [DGS12], these results can also be converted to a distributional PAC setting where instances x_t are drawn i.i.d. In this setting they obtain exponential improvement in label complexity over passive learning. These interesting results and techniques are not directly comparable to ours. Our framework is not restricted to halfspaces. Another important difference is that (as pointed out

in [GSSS13]) the exponential improvement they give is not possible in the noiseless version of their setting. In other words, the addition of linear noise defined by the target makes the problem easier for active sampling. By contrast RCN can only make the classification task harder than in the realizable case.

Recently, [BF13] showed the first polynomial time algorithms for actively learning thresholds, balanced rectangles, and homogenous linear separators under log-concave distributions in the presence of random classification noise. Active learning with respect to isotropic log-concave distributions in the absence of noise was studied in [BL13].

B Initializing with vector w_0

Suppose we have an algorithm B as a subroutine that works, given access to such a w_0. Then we can arrive at an algorithm A which works without it as follows. We will describe the procedure below for general admissible distributions. With probability 1, for a random u, either u or $-u$ has an acute angle with w^*. We may then run B with both choices, ϵ set to $\frac{\pi c_6}{4}$ for any admissible distribution. Here c_6 is the constant in Definition 4.1. Then we can use hypothesis testing on $O(\log(1/\delta))$ examples, and, with high probability, find a hypothesis w' with error less than $\frac{\pi c_6}{4}$. Part 3 of Definition 4.1 then implies that A may then set $w_0 = w'$, and call B again.

C Proof of Theorem 3.1

We start by stating state properties of the distribution D which will be useful in our analysis in the next section.

1. [Bau90, BBZ07, KKMS05] For any $C > 0$, there are $c_1, c_2 > 0$ such that, for x drawn from the uniform distribution over S_{d-1} and any unit length $u \in \mathbf{R}^d$,

   - for all $a, b \in [-C/\sqrt{d}, C/\sqrt{d}]$ for which $a \leq b$, we have

   $$c_1|b-a|\sqrt{d} \leq \Pr(u \cdot x \in [a,b]) \leq c_2|b-a|\sqrt{d}, \tag{1}$$

   - and if $b \geq 0$, we have

   $$\Pr(u \cdot x > b) \leq \frac{1}{2}e^{-db^2/2}. \tag{2}$$

2. [BBZ07, BL13] For any $c_6 > 0$, there is a $c_7 > 0$ such that, for all $d \geq 4$, the following holds. Let u and v be two unit vectors in R^d, and assume that $\theta(u,v) = \alpha \leq \pi/2$. Then

   $$\Pr_{x \sim D_d}[\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x) \text{ and } |v \cdot x| \geq c_7 \frac{\alpha}{\sqrt{d}}] \leq c_6 \alpha. \tag{3}$$

C.1 Margin based analysis

The proof of Theorem 3.1 follows the high level structure of the proof of [BBZ07]; the new element is the application of Theorem C.4 which analyzes the performance of the hinge loss minimization algorithm for learning inside the band, which in turn applies Theorem C.1, which analyzes the benefits of our new localized outlier removal procedure.

Proof (of Theorem 1.1): We will prove by induction on k that after $k \leq s$ iterations, we have $\text{err}_D(w_k) \leq 2^{-(k+1)}$ with probability $1 - \delta(1 - 1/(k+1))/2$.

When $k = 0$, all that is required is $\mathrm{err}_D(w_0) \leq 1/2$.

Assume now the claim is true for $k-1$ ($k \geq 1$). Then by induction hypothesis, we know that with probability at least $1 - \delta(1-1/k)/2$, w_{k-1} has error at most 2^{-k}. This implies $\theta(w_{k-1}, w^*) \leq \pi 2^{-k}$.

Let us define $S_{w_{k-1},b_{k-1}} = \{x : |w_{k-1} \cdot x| \leq b_{k-1}\}$ and $\bar{S}_{w_{k-1},b_{k-1}} = \{x : |w_{k-1} \cdot x| > b_{k-1}\}$. Since w_{k-1} has unit length, and $v_k \in B(w_{k-1}, r_k)$, we have $\theta(w_{k-1}, v_k) \leq r_k$ which in turn implies $\theta(w_{k-1}, w_k) \leq r_k$.

Applying Equation 3 to bound the error rate outside the band, we have both:

$$\Pr_x\left[(w_{k-1} \cdot x)(w_k \cdot x) < 0, x \in \bar{S}_{w_{k-1},b_{k-1}}\right] \leq 2^{-(k+4)} \quad \text{and}$$

$$\Pr_x\left[(w_{k-1} \cdot x)(w^* \cdot x) < 0, x \in \bar{S}_{w_{k-1},b_{k-1}}\right] \leq 2^{-(k+4)}.$$

Taking the sum, we obtain $\Pr_x\left[(w_k \cdot x)(w^* \cdot x) < 0, x \in \bar{S}_{w_{k-1},b_{k-1}}\right] \leq 2^{-(k+3)}$. Therefore, we have

$$\mathrm{err}(w_k) \leq (\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k)) \Pr(S_{w_{k-1},b_{k-1}}) + 2^{-(k+3)}.$$

Let c_2' be the constant from Equation 1. We have $\Pr(S_{w_{k-1},b_{k-1}}) \leq 2c_2' b_{k-1}\sqrt{d}$, this implies

$$\mathrm{err}(w_k) \leq (\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k))2c_2' b_{k-1}\sqrt{d} + 2^{-(k+3)} \leq 2^{-(k+1)}\left((\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k))4c_1 c_2' + 1/2\right).$$

Recall that $D_{w_{k-1},b_{k-1}}$ is the distribution obtained by conditioning D on the event that $x \in S_{w_{k-1},b_{k-1}}$. Applying Theorem C.4, with probability $1 - \frac{\delta}{2(k+k^2)}$, w_k has error at most $\kappa = \frac{1}{8c_1 c_2'}$ within $S_{w_{k-1},b_{k-1}}$, implying that $\mathrm{err}(w_k) \leq 2^{-(k+1)}$, completing the proof of the induction, and therefore showing, with probability at least $1-\delta$, $O(\log(1/\epsilon))$ iterations suffice to achieve $\mathrm{err}(w_k) \leq \epsilon$.

A polynomial number of unlabeled samples are required by the algorithm and the number of labeled examples required by the algorithm is $\sum_k m_k = O(d(d + \log\log(1/\epsilon) + \log(1/\delta))\log(1/\epsilon))$. □

C.2 Analysis of the outlier removal subroutine

The analysis of the learning algorithm uses the following theorem (same as Theorem 3.4 in the main body) about the outlier removal subroutine of Figure 2.

Theorem C.1. *There is a polynomial p such that, if $n \geq p(1/\eta, d, 1/\xi, 1/\delta, 1/\gamma)$ examples are drawn from the distribution $D_{u,\gamma}$ (each replaced with an arbitrary unit-length vector with probability $\eta < 1/4$), then, with probability $1-\delta$, the output q of the algorithm in Figure 1 satisfies the following:*

- $\sum_{x \in S} q(x) \geq (1-\xi)|S|$ *(a fraction $1-\xi$ of the weight is retained)*
- *For all unit length w such that $\|w - u\|_2 \leq r$,*

$$\frac{1}{|S|}\sum_{x \in S} q(x)(w \cdot x)^2 \leq 2\left(\frac{r^2}{d-1} + \gamma^2\right). \tag{4}$$

Furthermore, the algorithm can be implemented in polynomial time.

Our proof of Theorem 3.4 proceeds through a series of lemmas. We would like to point out that in the analysis below we will treat each element $x_i \in S$ as distinct (even if $x_i = x_j$ for some j). Obviously, a feasible q satisfies the requirements of the lemma. So all we need to show is

- there is a feasible solution q, and
- we can simulate a separation oracle: given a provisional solution $\hat{q}$, we can find a linear constraint violated by $\hat{q}$ in polynomial time.

We will start by working on proving that there is a feasible q. First of all, a Chernoff bound implies that $n \geq \text{poly}(1/\eta, 1/\delta)$ suffices for it to be the case that, with probability $1-\delta$, at most 2η members of S are noisy. Let us assume from now on that this is the case.

We will show that q^* which sets $q^*(x,y) = 0$ for each noisy point, and $q^*(x,y) = 1$ for each non-noisy point, is feasible. First we get a bound on $E[(a.x)^2]$ for all vectors a close to u. This is formlized in the following lemma

Lemma C.2. *For all a such that $\|u-a\|_2 \leq r$ and $\|a\|_2 \leq 1$*

$$\mathbf{E}_{x\sim U_{u,\gamma}}((a\cdot x)^2) \leq r^2/(d-1) + \gamma^2.$$

Proof. W.l.o.g. we may assume that $u = (1,0,0,...,0)$. We can write $x = (x_1, x_2, \ldots, x_d)$ as $x = (x_1, x')$, so that x' is chosen uniformly over all vectors in $\mathbf{R}^{d-1}$ of length at most $\sqrt{1-x_1^2}$. Let us decompose $\mathbf{E}_{x\sim D}((a\cdot x)^2)$ into parts that we can analyze separately as follows.

$$\mathbf{E}_{x\sim U_{u,\gamma}}((a\cdot x)^2) = a_1^2\mathbf{E}_{x\sim U_{u,\gamma}}(x_1^2) + a_1\sum_{i=2}^{n} a_i\mathbf{E}_{x\sim U_{u,\gamma}}(x_1x_i) + \mathbf{E}_{x\sim U_{u,\gamma}}((x'\cdot a)^2). \tag{5}$$

Thus $\mathbf{E}_{x\sim D}((x'\cdot a)^2)$ is at most the expectation of $(x'\cdot a)^2$ when $x' = (0, x_2, ..., x_d)$ is sampled uniformly from the unit ball in $\mathbf{R}^{d-1}$. Thus

$$\mathbf{E}_{x\sim U_{u,\gamma}}((x'\cdot a)^2) \leq \frac{1}{d-1}\sum_{i=2}^{d} a_i^2 \leq \frac{r^2}{d-1}. \tag{6}$$

Furthermore, since $|x_1| \leq \gamma$ when x is drawn from $U_{u,\gamma}$, we have

$$\mathbf{E}_{x\sim U_{u,\gamma}}(x_1^2) \leq \gamma^2. \tag{7}$$

Finally, by symmetry, $\mathbf{E}_{x\sim U_{u,\gamma}}(x_1x_i) = 0$ for all i. Putting this together with (7), (6) and (5) completes the proof. □

Next, we use VC tools to show the following bound on clean examples.

Lemma C.3. *If we draw ℓ times i.i.d. from D to form C, with probability $1-\delta$, we have that for any unit length a,*

$$\frac{1}{\ell}\sum_{x\in C}(a\cdot x)^2 \leq E[(a\cdot x)^2] + \sqrt{\frac{O(d\log(\ell/\delta)(d+\log(1/\delta)))}{\ell}}.$$

Proof: See Appendix H. □

The above two lemmas imply that $n = \text{poly}\,(d, 1/\eta, 1/\delta, 1/\gamma)$ suffices for it to be the case that, for all $w \in B(u,r)$,

$$\frac{1}{|S|}\sum_{x} q^*(x)(a\cdot x)^2 \leq 2\mathbf{E}[(a\cdot x)^2] \leq 2(\frac{r^2}{d-1} + \gamma^2),$$

so that q^* is feasible.

So what is left is to prove that the convex program has a separation oracle. First, it is easy to check whether, for all $x \in S$, $0 \leq q(x) \leq 1$, and whether $\sum_{x \in S} q(x) \geq (1-\xi)|S|$. An algorithm can first do that. If these pass, then it needs to check whether there is a $w \in B(u,r)$ with $||w||_2 \leq 1$ such that

$$\frac{1}{|S|} \sum_{x \in S} q(x)(w \cdot x)^2 > c(\frac{r^2}{d-1} + \gamma^2).$$

This can be done by finding $w \in B(u,r)$ with $||w||_2 \leq 1$ that maximizes $\sum_{x \in S} q(x)(w \cdot x)^2$, and checking it.

Suppose X is a matrix with a row for each $x \in S$, where the row is $\sqrt{q(x)}x$. Then $\sum_{x \in S} q(x)(w \cdot x)^2 = w^T X^T X w$, and, maximizing this over w is an equivalent problem to minimizing $w^T(-X^T X)w$ subject to $\|w - u\|_2 \leq r$ and $||w|| \leq 1$. Since $-X^T X$ is symmetric, problems of this form are known to be solvable in polynomial time [SZ03] (see [BM13]).

C.3 The error within a band in each iteration

At each iteration, the algorithm of Figure 1 concentrates its attention on examples in the band. Our next theorem (same as Theorem 3.2 in the main body) analyzes its error on these examples.

Theorem C.4. *After round k of the algorithm in Figure 1, with probability $1 - \frac{\delta}{k+k^2}$, we have $err_{D_{w_{k-1},b_{k-1}}}(w_k) \leq \kappa$.*

We will prove Theorem C.4 using a series of lemmas below. First, we bound the hinge loss of the target w^* within the band $S_{w_{k-1},b_{k-1}}$. Since we are analyzing a particular round k, to reduce clutter in the formulas, for the rest of this section, let us refer to

- ℓ_{τ_k} simply as ℓ,
- $L_{\tau_k}(\cdot, D_{w_{k-1},b_{k-1}})$ as $L(\cdot)$.

Lemma C.5. $L(w^*) \leq \kappa/12$.

Proof. Notice that $y(w^* \cdot x)$ is never negative, so, on any clean example (x,y), we have

$$\ell(w^*, x, y) = \max\left\{0, 1 - \frac{y(w^* \cdot x)}{\tau_k}\right\} \leq 1,$$

and, furthermore, w^* will pay a non-zero hinge only inside the region where $|w^* \cdot x| < \tau_k$. Hence,

$$L(w^*) \leq \Pr_{D_{w_{k-1},b_{k-1}}}(|w^* \cdot x| \leq \tau_k) = \frac{\Pr_{x \sim D}(|w^* \cdot x| \leq \tau_k \,\&\, |w_{k-1} \cdot x| \leq b_{k-1})}{\Pr_{x \sim D}(|w_{k-1} \cdot x| \leq b_{k-1})}.$$

Let c_1' and c_2' be the constants in Equation (1) respectively. We can lower bound the denominator $\Pr_{x \sim D}(|w_{k-1} \cdot x| < b_{k-1}) \geq 2c_1' b_{k-1}\sqrt{d}$. Also the numerator is at most $\Pr_{x \sim D}(|w^* \cdot x| \leq \tau_k) \leq 2c_2' \tau_k \sqrt{d}$. Hence, we have $L(w^*) \leq \frac{2c_2'\tau_k}{2c_1' b_{k-1}} = \kappa/12$. (by setting $c_4 = c_1'$ and $c_1 = c_2'/2$.) ☐

During round k we can decompose the working set W into the set of "clean" examples W_C which are drawn from $D_{w_{k-1},b_{k-1}}$ and the set of "dirty" or malicious examples W_D which are output by the adversary. We will next show that the fraction of dirty examples in round k is not too large.

Lemma C.6. *With probability* $1 - \frac{\delta}{6(k+k^2)}$,

$$|W_D| \leq 8c_1c_4\eta n_k 2^k. \tag{8}$$

Proof. From Equation 1 and the setting of our parameters, the probability that an example falls in $S_{w_{k-1},b_{k-1}}$ $2c_1c_4 2^{-k}$. Therefore, with probability $(1 - \frac{\delta}{12(k+k^2)})$, the number of examples we must draw before we encounter n_k examples that fall within $S_{w_{k-1},b_{k-1}}$ is at most $4c_1c_4n_k2^k$. The probability that each unlabeled example we draw is noisy is at most η. Applying a Chernoff bound, with probability at least $1 - \frac{\delta}{12(k+k^2)}$,

$$|W_D| \leq 8c_1c_4\eta n_k 2^k.$$

completing the proof. □

Next, we bound the loss on an example in terms of the norm of x.

Lemma C.7. *For any* $w \in B(w_{k-1}, r_k)$*, and all* x*,*

$$\ell(w, x, y) \leq \frac{4c_2\pi}{c_4}\sqrt{d}.$$

Proof. A simple calculation shows:

$$\begin{aligned}\ell(w, x, y) &\leq 1 + \frac{|w \cdot x|}{\tau_k} \leq 1 + \frac{|w_{k-1} \cdot x| + \|w - w_{k-1}\|_2||x||_2}{\tau_k} \\ &\leq 1 + \frac{b_{k-1} + r_k}{\tau_k} \leq \frac{4c_2\pi}{c_4}\sqrt{d}.\end{aligned}$$

□

Recall that the total variation distance between two probability distributions is the maximum difference between the probabilities that the assign to any event. We can think of q as soft indicator functions for "keeping" examples, and so interpret the inequality $\sum_{x\in W} q(x) \geq (1 - \xi)|W|$ as roughly akin to saying that most examples are kept. This means that distribution p obtained by normalizing q is close to the uniform distribution over W. We make this precise in the following lemma.

Lemma C.8. *The total variation distance between* p *and the uniform distribution over* W *is at most* ξ*.*

Proof. Lemma 1 of [LS06] implies that the total variation distance ρ between q and the uniform distribution over W satisfies

$$\rho = 1 - \sum_{x\in W} \min\left\{q(x), \frac{1}{|W|}\right\}.$$

Since $q(x) \leq 1$ for all x, we have $\sum_{x\in W} q(x) \leq |W|$, so that

$$\rho \leq 1 - \frac{1}{|W|}\sum_{x\in W} \min\{q(x), 1\}.$$

Again, since $q(x) \leq 1$, we have

$$\rho \leq 1 - \frac{(1-\xi)|W|}{|W|} = \xi.$$

□

Next, we will relate the average hinge loss when examples are weighted according to p i.e., $\ell(w, p)$ to the hinge loss averaged over clean examples W_C, i.e., $\ell(w, W_C)$. This is relationship is better than using a uniform bound on the variance since, within the band, projecting the data onto directions close to w_{k-1} will lead to much smaller variance. Specifically, we prove the following lemma (same as Lemma 3.5 in the main body but with precise constants) Here $\ell(w, W_C)$ and $\ell(w, p)$ are defined with respect to the true unrevealed labels that the adversary has committed to.

Lemma C.9. *Define* $z_k = \sqrt{\frac{r_k^2}{d-1} + b_{k-1}^2}$. *For large enough* d, *with probability* $1 - \frac{\delta}{2(k+k^2)}$, *for any* $w \in B(w_{k-1}, r_k)$, *we have*

$$\ell(w, W_C) \leq \ell(w, p) + \frac{32 c_1 c_4 \eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/32 \tag{9}$$

and

$$\ell(w, p) \leq 2\ell(w, W_C) + \kappa/32 + \frac{8 c_1 c_4 \eta}{\epsilon} + \frac{\sqrt{32 c_1 c_4 \eta/\epsilon} z_k}{\tau_k} \tag{10}$$

Proof. As in the analysis of the outlier removal procedure, we will treat each element $(x, y) \in W$ as distinct. Fix an arbitrary $w \in B(w_{k-1}, r_k)$. By the guarantee of Theorem C.1, Lemma C.6, and Lemmas C.2 and C.3 we know that, with probability $1 - \frac{\delta}{2(k+k^2)}$,

$$\frac{1}{|W|}\sum_{x \in W} q(x)(w \cdot x)^2 \leq 4 z_k^2, \tag{11}$$

together with

$$|W_D| \leq 8 c_1 c_4 \eta n_k 2^k \tag{12}$$

and

$$\frac{1}{|W_C|}\sum_{(x,y) \in W_C} (w \cdot x)^2 \leq 2 z_k^2, \tag{13}$$

Assume that (11), (12) and (13) all hold.

Since $\sum_{x \in W} q(x) \geq (1 - \xi_k)|W| \geq |W|/2$, we have that (11) implies

$$\sum_{x \in W} p(x)(w \cdot x)^2 \leq 8 z_k^2. \tag{14}$$

First, let us bound the weighted loss on noisy examples in the training set. In particular, we will show that

$$\sum_{(x,y) \in W_D} p(x)\ell(w, x, y) \leq C_0 \eta 2^k + \xi_k + \frac{\sqrt{2c' C_0 \eta 2^k} z_k}{\tau_k}. \tag{15}$$

To see this, notice that,

$$\sum_{(x,y)\in W_D} p(x)\ell(w,x,y) = \sum_{(x,y)\in W_D} p(x)\max\left\{0, 1-\frac{y(w\cdot x)}{\tau_k}\right\}$$
$$\leq \Pr_p(W_D) + \frac{1}{\tau_k}\sum_{(x,y)\in W_D} p(x)|w\cdot x| = \Pr_p(W_D) + \frac{1}{\tau_k}\sum_{(x,y)\in W} p(x)1_{W_D}(x,y)|w\cdot x|$$
$$\leq \Pr_p(W_D) + \frac{1}{\tau_k}\sqrt{\sum_{(x,y)\in W} p(x)1_{W_D}(x,y)}\sqrt{\sum_{(x,y)\in W} p(x)(w\cdot x)^2} \text{ (by the Cauchy-Shwartz inequality)}$$
$$\leq \Pr_p(W_D) + \frac{\sqrt{8\Pr_p(W_D)}z_k}{\tau_k} \leq 8c_1c_4\eta 2^k + \xi_k + \frac{\sqrt{64c_1c_4\eta 2^k}z_k}{\tau_k}$$

where the second to last inequality follows by by (14) and the last one follows by Lemma C.8 and (8).

Similarly, we will show that

$$\sum_{(x,y)\in W} p(x)\ell(w,x,y) \leq 1 + \frac{4z_k}{\tau_k}. \tag{16}$$

To see this notice that,

$$\sum_{(x,y)\in W} p(x)\ell(w,x,y) = \sum_{(x,y)\in W} p(x)\max\left\{0, 1-\frac{y(w\cdot x)}{\tau_k}\right\}$$
$$\leq 1 + \frac{1}{\tau_k}\sum_{x\in W} p(x)|w\cdot x| \leq 1 + \frac{1}{\tau_k}\sqrt{\sum_{x\in W} p(x)(w\cdot x)^2} \leq 1 + \frac{4z_k}{\tau_k},$$

where the last step follow by (14). Next, we have

$$\ell(w,W_C) = \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + (1_{W_C}(x,y) - q(x))\ell(w,x,y)\right)$$
$$\leq \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + \sum_{(x,y)\in W_C}(1-q(x))\ell(w,x,y)\right)$$
$$\leq \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + \sum_{(x,y)\in W_C}(1-q(x))\left(1+\frac{|w\cdot x|}{\tau_k}\right)\right)$$
$$\leq \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + \xi_k|W| + \frac{1}{\tau_k}\sum_{(x,y)\in W_C}(1-q(x))|w\cdot x|\right)$$
$$\leq \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + \xi_k|W| + \frac{1}{\tau_k}\sqrt{\sum_{(x,y)\in W_C}(1-q(x))^2}\sqrt{\sum_{(x,y)\in W_C}(w\cdot x)^2}\right)$$

by the Cauchy-Shwartz inequality. Recall that $0 \leq q(x) \leq 1$, and $\sum_{x \in W} q(x) \geq (1 - \xi_k)|W|$. Thus,

$$\begin{aligned}
\ell(w, W_C) &\leq \frac{1}{|W_C|} \left(\sum_{(x,y) \in W} q(x) \ell(w, x, y) + \xi_k |W| + \frac{1}{\tau_k} \sqrt{\xi_k |W|} \sqrt{\sum_{x \in W_C} (w \cdot x)^2} \right) \\
&\leq \frac{1}{|W_C|} \left(\sum_{(x,y) \in W} q(x) \ell(w, x, y) + \xi_k |W| + \frac{\sqrt{\xi_k |W| |W_C| 2 z_k^2}}{\tau_k} \right)
\end{aligned}$$

by (25). Since $|W_C| \geq |W|/2$, we have

$$\ell(w, W_C) \leq \frac{1}{|W_C|} \left(\sum_{(x,y) \in W} q(x) \ell(w, x, y) \right) + 2\xi_k + \frac{\sqrt{4 \xi_k z_k^2}}{\tau_k}.$$

We have chosen ξ_k small enough that

$$\begin{aligned}
\ell(w, W_C) &\leq \frac{1}{|W_C|} \left(\sum_{(x,y) \in W} q(x) \ell(w, x, y) \right) + \kappa/32 \\
&= \frac{\sum_{(x,y) \in W} q(x)}{|W_C|} \left(\sum_{(x,y) \in W} p(x) \ell(w, x, y) \right) + \kappa/32 \\
&= \ell(w, p) + \left(\frac{\sum_{(x,y) \in W} q(x)}{|W_C|} - 1 \right) \left(\sum_{(x,y) \in W} p(x) \ell(w, x, y) \right) + \kappa/32 \\
&\leq \ell(w, p) + \left(\frac{|W|}{|W_C|} - 1 \right) \left(\sum_{(x,y) \in W} p(x) \ell(w, x, y) \right) + \kappa/32 \\
&\leq \ell(w, p) + \left(\frac{|W|}{|W_C|} - 1 \right) \left(1 + \frac{4 z_k}{\tau_k} \right) + \kappa/32.
\end{aligned}$$

Applying (12) yields (9).

Also,

$$\begin{aligned}
\ell(w,p) &= \sum_{(x,y)\in W} p(x)\ell(w,x,y) \\
&= \sum_{(x,y)\in W_C} p(x)\ell(w,x,y) + \sum_{(x,y)\in W_D} p(x)\ell(w,x,y) \\
&\leq \sum_{(x,y)\in W_C} p(x)\ell(w,x,y) + 8c_1c_4\eta 2^k + \xi_k + \frac{\sqrt{32c_1c_4\eta 2^k}z_k}{\tau_k} \quad \text{(by (27))}. \\
&= \frac{\sum_{(x,y)\in W_C} q(x)\ell(w,x,y)}{\sum_{(x,y)\in W_C} q(x)} + 8c_1c_4\eta 2^k + \xi_k + \frac{\sqrt{32c_1c_4\eta 2^k}z_k}{\tau_k} \\
&\leq \frac{\sum_{(x,y)\in W_C} \ell(w,x,y)}{\sum_{(x,y)\in W_C} q(x)} + 8c_1c_4\eta 2^k + \xi_k + \frac{\sqrt{32c_1c_4\eta 2^k}z_k}{\tau_k} \quad \text{(since } \forall x, q(x)\leq 1\text{)}. \\
&\leq \frac{\sum_{(x,y)\in W_C} \ell(w,x,y)}{|W_C| - \xi_k|W|} + 8c_1c_4\eta 2^k + \xi_k + \frac{\sqrt{32c_1c_4\eta 2^k}z_k}{\tau_k} \\
&\leq 2\ell(w,W_C) + 8c_1c_4\eta 2^k + \xi_k + \frac{\sqrt{32c_1c_4\eta 2^k}z_k}{\tau_k},
\end{aligned}$$

by (8), which in turn implies (10). □

Finally, we need some bounds about estimates of the hinge loss.

Lemma C.10. *With probability* $1 - \frac{\delta}{2(k+k^2)}$*, for all* $w \in B(w_{k-1}, r_k)$*,*

$$|L(w) - \ell(w, W_C)| \leq \kappa/32 \tag{17}$$

and

$$|\ell(w,p) - \ell(w,T)| \leq \kappa/32. \tag{18}$$

Proof. See Appendix H. □

Proof of Theorem C.4. By Lemma C.10, with probability $1 - \frac{\delta}{2(k+k^2)}$, for all $w \in B(w_{k-1}, r_k)$, (17) and (18) hold. Also with probability $1 - \frac{\delta}{2(k+k^2)}$, both (9) and (10) hold. Let us assume from here on that all of these hold.

Then we have

$$\begin{aligned}
\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k) &= \mathrm{err}_{D_{w_{k-1},b_{k-1}}}(v_k) \\
&\leq L(v_k) \quad \text{(since for each error, the hinge loss is at least 1)} \\
&\leq \ell(v_k, W_C) + \kappa/32 \quad \text{(by (17))} \\
&\leq \ell(v_k,p) + \frac{32c_1c_4\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/16 \quad \text{(by (9))} \\
&\leq \ell(v_k,T) + \frac{32c_1c_4\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/8 \quad \text{(by (18))} \\
&\leq \ell(w^*,T) + \frac{32c_1c_4\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/4 \quad \text{(since } w^* \in B(w_{k-1}, r_k)\text{)} \\
&\leq \ell(w^*,p) + \frac{32c_1c_4\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/3 \quad \text{(by (18))}.
\end{aligned}$$

This, together with (10) and (17), gives

$$\begin{aligned}
\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k) &\leq 2\ell(w^*, W_C) + \frac{8c_1c_4\eta}{\epsilon} + \frac{\sqrt{32c_1c_4\eta/\epsilon}z_k}{\tau_k} + \frac{32c_1c_4\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + 2\kappa/5 \\
&\leq 2L(w^*) + \frac{8c_1c_4\eta}{\epsilon} + \frac{\sqrt{32c_1c_4\eta/\epsilon}z_k}{\tau_k} + \frac{32c_1c_4\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/2 \\
&\leq \kappa/3 + \frac{8c_1c_4\eta}{\epsilon} + \frac{\sqrt{32c_1c_4\eta/\epsilon}z_k}{\tau_k} + \frac{32c_1c_4\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/2,
\end{aligned}$$

by Lemma C.5.

Now notice that z_k/τ_k is $\Theta(1)$. Hence an $\Omega(\epsilon)$ bound on η suffices to imply that $\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k) \leq \kappa$ with probability $(1 - \frac{\delta}{k+k^2})$. □

D Proof of Theorem 4.2

Throughout this section, assume that the clean training examples are obtained by labeling data drawn according to a distribution D over $\mathbf{R}^d$ chosen from a λ-admissible sequence. The main algorithm and the outlier removal procedure remain the same with the following parameters.

D.1 Parameters for the algorithm

The parameters of the algorithm are set as follows. Let $M = \max\{\frac{2}{c_6\pi}, 2\}$, where c_6 is from Definition 4.1. Let c_1' be the value of c_5 in part 2 of Definition 4.1 corresponding to the case where c_4 is $\frac{c_6}{4M}$; then let $b_k = c_1'M^{-k}$.

Let c_6' and c_7' be c_2 and c_3 respectively, from part 1 of Definition 4.1. Let $r_k = \min\{M^{-(k-1)}/c_6, \pi/2\}$, where c_6 is from Definition 4.1 and $\kappa = \frac{1}{4c_1'c_7'M}$. Finally, let $\tau_k = \frac{c_2\min\{b_{k-1},c_1\}\kappa}{6c_3}$, where c_1, c_2 and c_3 are the values from Definition 4.1. Let $z_k^2 = (r_k^2 + b_{k-1}^2)$ and $\xi_k = c\min(\kappa, \frac{\kappa^2\tau_k^2}{z_k^2})$. The value of σ_k^2 for the outlier removal procedure is $\ln^\lambda(1 + \frac{1}{b_{k-1}})(r_k^2 + b_{k-1}^2)$

D.2 Analysis of the outlier removal subroutine

The analysis of the learning algorithm uses the following lemma about the outlier removal subroutine of Figure 2.

Theorem D.1. *For any $C > 0$, there is a constant c and a polynomial p such that, for all $\xi > 2\eta$ and all $0 < \gamma < C$, if $n \geq p(1/\eta, d, 1/\xi, 1/\delta, 1/\gamma)$, then, with probability $1 - \delta$, the output q of the algorithm in Figure 2 satisfies the following:*

- $\sum_{x\in S} q(x) \geq (1 - \xi)|S|$ *(a fraction $1 - \xi$ of the weight is retained)*
- *For all unit length w such that $\|w - u\|_2 \leq r$,*

$$\frac{1}{|S|}\sum_{x\in S} q(x)(w \cdot x)^2 \leq c\ln^\lambda(1 + \frac{1}{\gamma})(r^2 + \gamma). \tag{19}$$

Furthermore, the algorithm can be implemented in polynomial time.

Almost identical to the previous section our proof of Theorem D.1 proceeds through a series of lemmas. Again, we would like to point out that in the analysis below we will treat each element $x_i \in S$ as distinct (even if $x_i = x_j$ for some j). Obviously, a feasible q satisfies the requirements of the lemma. So all we need to show is

- there is a feasible solution q, and
- we can simulate a separation oracle: given a provisional solution $\hat{q}$, we can find a linear constraint violated by $\hat{q}$ in polynomial time.

We will start by working on proving that there is a feasible q. First of all, a Chernoff bound implies that $n \geq \text{poly}(1/\eta, 1/\delta)$ suffices for it to be the case that, with probability $1-\delta$, at most 2η members of S are noisy. Let us assume from now on that this is the case.

We will show that q^* which sets $q^*(x) = 0$ for each noisy point, and $q^*(x) = 1$ for each non-noisy point, is feasible.

First, we use VC tools to show that, if enough examples are chosen, a bound like part 4 of Definition 4.1, but averaged over the clean examples, likely holds for all relevant directions.

Lemma D.2. *If we draw ℓ times i.i.d. from D to form C, with probability $1-\delta$, we have that for any unit length a,*

$$\frac{1}{\ell}\sum_{x\in C}(a\cdot x)^2 \leq E[(a\cdot x)^2] + \sqrt{\frac{O(d\log(\ell/\delta)(d+\log(1/\delta)))}{\ell}}.$$

Proof: See Appendix H. □

Lemma D.2 and part 4 of Definition 4.1 together directly imply that

$$n = \text{poly}\left(d, 1/\eta, 1/\delta, \frac{1}{c(r^2+\gamma^2)\ln^\lambda(1+1/\gamma)}\right) = \text{poly}\left(d, 1/\eta, 1/\delta, 1/\gamma\right)$$

suffices for it to be the case that, for all $w \in B(u,r)$,

$$\frac{1}{|S|}\sum_{(x,y)} q^*(x)(a\cdot x)^2 \leq 2\mathbf{E}[(a\cdot x)^2] \leq 2c_8(r^2+\gamma^2)\ln^\lambda(1+1/\gamma),$$

so that, if $c = 2c_8$, we have that q^* is feasible.

So what is left is to prove that the convex program has a separation oracle. First, it is easy to check whether, for all x, $0 \leq q(x) \leq 1$, and whether $\sum_{x\in S} q(x) \geq (1-\xi)|S|$. An algorithm can first do that. If these pass, then it needs to check whether there is a $w \in B(u,r)$ with $||w||_2 \leq 1$ such that

$$\frac{1}{|S|}\sum_{x\in S} q(x)(w\cdot x)^2 > c\log^\lambda\left(1+\frac{1}{\gamma}\right)(r^2+\gamma^2).$$

This can be done by finding $w \in B(u,r)$ with $||w||_2 \leq 1$ that maximizes $\sum_{x\in S} q(x)(w\cdot x)^2$, and checking it.

Suppose X is a matrix with a row for each $x \in S$, where the row is $\sqrt{q(x)}x$. Then $\sum_{x\in S} q(x)(w\cdot x)^2 = w^T X^T X w$, and, maximizing this over w is an equivalent problem to minimizing $w^T(-X^TX)w$ subject to $\|w-u\|_2 \leq r$ and $||w|| \leq 1$. Since $-X^TX$ is symmetric, problems of this form are known to be solvable in polynomial time [SZ03] (see [BM13]).

D.3 The error within a band in each iteration

At each iteration, the algorithm of Figure 1 concentrates its attention on examples in the band. Our next theorem analyzes its error on these examples.

Theorem D.3. *After round k of the algorithm in Figure 1, with probability $1-\frac{\delta}{k+k^2}$, we have $err_{D_{w_{k-1},b_{k-1}}}(w_k) \leq \kappa$.*

We will prove Theorem D.3 using a series of lemmas below. First, we bound the hinge loss of the target w^* within the band $S_{w_{k-1},b_{k-1}}$. Since we are analyzing a particular round k, to reduce clutter in the formulas, for the rest of this section, let us refer to

- ℓ_{τ_k} simply as ℓ,
- $L_{\tau_k}(\cdot, D_{w_{k-1},b_{k-1}})$ as $L(\cdot)$.

Lemma D.4. $L(w^*) \leq \kappa/6$.

Proof. Notice that $y(w^* \cdot x)$ is never negative, so, on any clean example (x, y), we have

$$\ell(w^*, x, y) = \max\left\{0, 1 - \frac{y(w^* \cdot x)}{\tau_k}\right\} \leq 1,$$

and, furthermore, w^* will pay a non-zero hinge only inside the region where $|w^* \cdot x| < \tau_k$. Hence,

$$L(w^*) \leq \Pr_{D_{w_{k-1},b_{k-1}}}(|w^* \cdot x| \leq \tau_k) = \frac{\Pr_{x\sim D}(|w^* \cdot x| \leq \tau_k \,\&\, |w_{k-1} \cdot x| \leq b_{k-1})}{\Pr_{x\sim D}(|w_{k-1} \cdot x| \leq b_{k-1})}.$$

Using part 1 of Definition 4.1, for the values of c_1 and c_2 in that definition, we can lower bound the denominator:

$$\Pr_{x\sim D}(|w_{k-1} \cdot x| < b_{k-1}) \geq c_2 \min\{b_{k-1}, c_1\}.$$

part 1 of Definition 4.1 also implies that the numerator is at most

$$\Pr_{x\sim D}(|w^* \cdot x| \leq \tau_k) \leq c_3\tau_k.$$

Hence, we have

$$L(w^*) \leq \frac{c_3\tau_k}{c_2 \min\{b_{k-1}, c_1\}} = \kappa/6.$$

□

During round k we can decompose the working set W into the set of "clean" examples W_C which are drawn from $D_{w_{k-1},b_{k-1}}$ and the set of "dirty" or malicious examples W_D which are output by the adversary. We will next show that the fraction of dirty examples in round k is not too large.

Lemma D.5. *There is an absolute positive constant C_0 such that, with probability $1 - \frac{\delta}{6(k+k^2)}$,*

$$|W_D| \leq C_0\eta n_k M^k. \tag{20}$$

Proof. From Equation 1 and the setting of our parameters, the probability that an example falls in $S_{w_{k-1}}$ is at least $\Omega(M^{-k})$. Therefore, with probability $(1 - \frac{\delta}{12(k+k^2)})$, the number of examples we must draw before we encounter n_k examples that fall within $S_{w_{k-1},b_{k-1}}$ is at most $O(n_k M^k)$. The probability that each unlabeled example we draw is noisy is at most η. Applying a Chernoff bound, with probability at least $1 - \frac{\delta}{12(k+k^2)}$,

$$|W_D| \leq C_0 \eta n_k M^k.$$

completing the proof. □

Next, we bound the loss on an example in terms of the norm of x.

Lemma D.6. *There is a constant c such that, for any $w \in B(w_{k-1}, r_k)$, and all x,*

$$\ell(w, x, y) \leq c(1 + ||x||_2).$$

Proof.

$$\begin{aligned}\ell(w, x, y) &\leq 1 + \frac{|w \cdot x|}{\tau_k} \leq 1 + \frac{|w_{k-1} \cdot x| + \|w - w_{k-1}\|_2 ||x||_2}{\tau_k} \\ &\leq 1 + \frac{b_{k-1} + r_k ||x||_2}{\tau_k} = 1 + \frac{c_1' M^{-k} + \min\{M^{-(k-1)}/c_6, \pi/2\}||x||_2}{\frac{c_2 \min\{c_1' M^{-k}, c_1\}\kappa}{6c_3}}.\end{aligned}$$

□

If the support of D is bounded, Lemma D.6 gives a useful worst-case bound on the loss. Next, we give a high-probability bound that holds for all λ-admissible distributions.

Lemma D.7. *For an absolute constant c, with probability $1 - \frac{\delta}{6(k+k^2)}$,*

$$\max_{x \in W_C} ||x||_2 \leq c\sqrt{d} \ln\left(\frac{|W_C|k}{\delta}\right).$$

Proof. Applying part 5 of Definition 4.1, together with a union bound, we have

$$\Pr(\exists x \in W_C,\ ||x|| > \alpha) \leq c_9 |W_C| \exp(-\alpha/\sqrt{d}),$$

and $\alpha = \sqrt{d} \ln\left(\frac{12 c_9 |W_C| k^2}{\delta}\right)$ makes the RHS at most $\frac{\delta}{6(k+k^2)}$. □

Recall that the total variation distance between two probability distributions is the maximum difference between the probabilities that the assign to any event.

We can think of q as soft indicator functions for "keeping" examples, and so interpret the inequality $\sum_{x \in W} q(x) \geq (1 - \xi)|W|$ as roughly akin to saying that most examples are kept. This means that distribution p obtained by normalizing q is close to the uniform distribution over W. We make this precise in the following lemma.

Lemma D.8. *The total variation distance between p and the uniform distribution over W is at most ξ.*

Proof. Lemma 1 of [LS06] implies that the total variation distance ρ between p and the uniform distribution over W satisfies

$$\rho = 1 - \sum_{x \in W} \min\left\{p(x), \frac{1}{|W|}\right\}.$$

Since $q(x) \leq 1$ for all x, we have $\sum_{x \in W} q(x) \leq |W|$, so that

$$\rho \leq 1 - \frac{1}{|W|} \sum_{x \in W} \min\{q(x), 1\}.$$

Again, since $q(x) \leq 1$, we have

$$\rho \leq 1 - \frac{(1-\xi)|W|}{|W|} = \xi.$$

□

Next, we will relate the average hinge loss when examples are weighted according to p, i.e., $\ell(w, p)$ to the hinge loss averaged over clean examples W_C, i.e., $\ell(w, W_C)$. Here $\ell(w, W_C)$ and $\ell(w, p)$ are defined with respect to the true unrevealed labels that the adversary has committed to.

Lemma D.9. *There are absolute constants c_1, c_2 and c_3 such that, for large enough d, with probability $1 - \frac{\delta}{2(k+k^2)}$, if we define $z_k = \sqrt{r_k^2 + b_{k-1}^2}$, then for any $w \in B(w_{k-1}, r_k)$, we have*

$$\ell(w, W_C) \leq \ell(w, p) + \frac{c_1 \eta}{\epsilon}\left(1 + \frac{\ln^{\lambda/2}(1 + 1/b_k) z_k}{\tau_k}\right) + \kappa/32 \tag{21}$$

and

$$\ell(w, p) \leq 2\ell(w, W_C) + \kappa/32 + \frac{c_2 \eta}{\epsilon} + \frac{c_3 \sqrt{\eta/\epsilon} \ln^{\lambda/2}(1 + 1/b_k) z_k}{\tau_k} \tag{22}$$

Proof. As in the analysis of the outlier removal procedure, we will treat each element $(x, y) \in W$ as distinct. Fix an arbitrary $w \in B(w_{k-1}, r_k)$. By the guarantee of Theorem D.1, Lemma D.5, part 5 of Definition 4.1, part 4 of Definition 4.1, and Lemma D.2, we know that, with probability $1 - \frac{\delta}{2(k+k^2)}$,

$$\frac{1}{|W|} \sum_{x \in W} q(x)(w \cdot x)^2 \leq c' \ln^{\lambda}(1 + 1/b_k) z_k^2, \tag{23}$$

together with

$$|W_D| \leq C_0 \eta n_k M^{-k} \tag{24}$$

(for an absolute constant C_0) and

$$\frac{1}{|W_C|} \sum_{(x,y) \in W_C} (w \cdot x)^2 \leq c''(r^2 + \gamma^2) \ln^{\lambda}(1 + 1/b_k), \tag{25}$$

for an absolute constant c''.

Assume that (23), (24) and (25) all hold.

Since $\sum_{x \in W} q(x) \geq (1 - \xi_k)|W| \geq |W|/2$, we have that (23) implies

$$\sum_{x \in W} p(x)(w \cdot x)^2 \leq 2c' \ln^{\lambda}(1 + 1/b_k) z_k^2. \tag{26}$$

First, let us bound the weighted loss on noisy examples in the training set. In particular, we will show that

$$\sum_{(x,y)\in W_D} p(x)\ell(w,x,y) \leq C_0\eta M^{-k} + \xi_k + \frac{\sqrt{2c'C_0\eta M^{-k}}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k}. \tag{27}$$

To see this, notice that,

$$\begin{aligned}
&\sum_{(x,y)\in W_D} p(x)\ell(w,x,y) = \sum_{(x,y)\in W_D} p(x)\max\left\{0, 1-\frac{y(w\cdot x)}{\tau_k}\right\} \\
&\leq \Pr_p(W_D) + \frac{1}{\tau_k}\sum_{(x,y)\in W_D} p(x)|w\cdot x| = \Pr_p(W_D) + \frac{1}{\tau_k}\sum_{(x,y)\in W} p(x)1_{W_D}(x,y)|w\cdot x| \\
&\leq \Pr_p(W_D) + \frac{1}{\tau_k}\sqrt{\sum_{(x,y)\in W} p(x)1_{W_D}(x,y)}\sqrt{\sum_{(x,y)\in W} p(x)(w\cdot x)^2} \text{ (by the Cauchy-Shwartz inequality)} \\
&\leq \Pr_p(W_D) + \frac{\sqrt{2c'\Pr_p(W_D)}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k} \leq C_0\eta M^{-k} + \xi_k + \frac{\sqrt{2c'C_0\eta M^{-k}}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k}
\end{aligned}$$

where the second to last inequality follows by (26) and the last one by Lemma D.8 and (24).

Similarly, we will show that

$$\sum_{(x,y)\in W} p(x)\ell(w,x,y) \leq 1 + \frac{\sqrt{c'}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k}. \tag{28}$$

To see this notice that,

$$\begin{aligned}
&\sum_{(x,y)\in W} p(x)\ell(w,x,y) = \sum_{(x,y)\in W} p(x)\max\left\{0, 1-\frac{y(w\cdot x)}{\tau_k}\right\} \\
&\leq 1 + \frac{1}{\tau_k}\sum_{(x,y)\in W} p(x)|w\cdot x| \leq 1 + \frac{1}{\tau_k}\sqrt{\sum_{(x,y)\in W} p(x)(w\cdot x)^2} \\
&\leq 1 + + \frac{\sqrt{2c'}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k},
\end{aligned}$$

by (26).

Next, we have

$$\begin{aligned}
\ell(w, W_C) &= \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + (1_{W_C}(x,y) - q(x))\ell(w,x,y)\right) \\
&\le \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + \sum_{(x,y)\in W_C} (1-q(x))\ell(w,x,y)\right) \\
&\le \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + \sum_{(x,y)\in W_C} (1-q(x))\left(1 + \frac{|w\cdot x|}{\tau_k}\right)\right) \\
&\le \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + \xi_k|W| + \frac{1}{\tau_k}\sum_{(x,y)\in W_C} (1-q(x))|w\cdot x|\right) \\
&\le \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + \xi_k|W| + \frac{1}{\tau_k}\sqrt{\sum_{(x,y)\in W_C}(1-q(x))^2}\sqrt{\sum_{(x,y)\in W_C}(w\cdot x)^2}\right)
\end{aligned}$$

by the Cauchy-Shwartz inequality. Recall that $0 \le q(x) \le 1$, and $\sum_{(x,y)\in W} q(x) \ge 1 - \xi_k|W|$. Thus,

$$\begin{aligned}
\ell(w, W_C) &\le \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + \xi_k|W| + \frac{1}{\tau_k}\sqrt{\xi_k|W|}\sqrt{\sum_{(x,y)\in W_C}(w\cdot x)^2}\right) \\
&\le \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y) + \xi_k|W| + \frac{\sqrt{\xi_k|W||W_C|c''(r^2+\gamma^2)\ln^\lambda(1+1/b_k)}}{\tau_k}\right)
\end{aligned}$$

by (25). Since $|W_C| \ge |W|/2$, we have

$$\ell(w, W_C) \le \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y)\right) + 2\xi_k + \frac{\sqrt{2\xi_k c''(r^2+\gamma^2)\ln^\lambda(1+1/b_k)}}{\tau_k}.$$

We have chosen ξ_k small enough that

$$\begin{aligned}
\ell(w, W_C) &\le \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y)\right) + \kappa/32 \\
&= \frac{\sum_{(x,y)\in W} q(x)}{|W_C|}\left(\sum_{(x,y)\in W} p(x)\ell(w,x,y)\right) + \kappa/32 \\
&= \ell(w,p) + \left(\frac{\sum_{(x,y)\in W} q(x)}{|W_C|} - 1\right)\left(\sum_{(x,y)\in W} p(x)\ell(w,x,y)\right) + \kappa/32 \\
&\le \ell(w,p) + \left(\frac{|W|}{|W_C|} - 1\right)\left(\sum_{(x,y)\in W} p(x)\ell(w,x,y)\right) + \kappa/32 \\
&\le \ell(w,p) + \left(\frac{|W|}{|W_C|} - 1\right)\left(1 + \frac{\sqrt{c'}\ln^{\kappa/2}(1+1/b_k)z_k}{\tau_k}\right) + \kappa/32.
\end{aligned}$$

Applying (24) yields (21).

Also,

$$\begin{aligned}
\ell(w,p) &= \sum_{(x,y)\in W} p(x)\ell(w,x,y) \\
&= \sum_{(x,y)\in W_C} p(x)\ell(w,x,y) + \sum_{(x,y)\in W_D} p(x)\ell(w,x,y) \\
&\le \sum_{(x,y)\in W_C} p(x)\ell(w,x,y) + C_0\eta M^{-k} + \xi_k + \frac{\sqrt{2c'C_0\eta M^{-k}}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k} \quad \text{(by (27))}. \\
&= \frac{\sum_{(x,y)\in W_C} q(x)\ell(w,x,y)}{\sum_{(x,y)\in W_C} q(x)} + C_0\eta M^{-k} + \xi_k + \frac{\sqrt{2c'C_0\eta M^{-k}}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k} \\
&\le \frac{\sum_{(x,y)\in W_C} \ell(w,x,y)}{\sum_{(x,y)\in W_C} q(x)} + C_0\eta M^{-k} + \xi_k + \frac{\sqrt{2c'C_0\eta M^{-k}}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k} \quad \text{(since } \forall x, q(x) \le 1\text{)}. \\
&\le \frac{\sum_{(x,y)\in W_C} \ell(w,x,y)}{|W_C| - \xi|W|} + C_0\eta M^{-k} + \xi_k + \frac{\sqrt{2c'C_0\eta M^{-k}}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k} \\
&\le 2\ell(w, W_C) + C_0\eta M^{-k} + \xi_k + \frac{\sqrt{2c'C_0\eta M^{-k}}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k},
\end{aligned}$$

by (24), which in turn implies (22). □

Finally, we need some bounds about estimates of the hinge loss.

Lemma D.10. *With probability* $1 - \frac{\delta}{2(k+k^2)}$*, for all* $w \in B(w_{k-1}, r_k)$*,*

$$|L(w) - \ell(w, W_C)| \le \kappa/32 \tag{29}$$

and

$$|\ell(w,p) - \ell(w,T)| \leq \kappa/32. \tag{30}$$

Proof. See Appendix H. □

Proof of Theorem D.3. By Lemma D.10, with probability $1 - \frac{\delta}{2(k+k^2)}$, for all $w \in B(w_{k-1}, r_k)$, (29) and (30) hold. Also with probability $1 - \frac{\delta}{2(k+k^2)}$, both (21) and (22) hold. Let us assume from here on that all of these hold.

Then we have

$$\begin{aligned}
\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k) &= \mathrm{err}_{D_{w_{k-1},b_{k-1}}}(v_k) \\
&\leq L(v_k) \ \text{(since for each error, the hinge loss is at least 1)} \\
&\leq \ell(v_k, W_C) + \kappa/16 \ \text{(by (29))} \\
&\leq \ell(v_k, p) + \frac{c_1\eta}{\epsilon}\left(1 + \frac{\ln^{\lambda/2}(1+1/b_k))z_k}{\tau_k}\right) + \kappa/8 \ \text{(by (21))} \\
&\leq \ell(v_k, T) + \frac{c_1\eta}{\epsilon}\left(1 + \frac{\ln^{\lambda/2}(1+1/b_k))z_k}{\tau_k}\right) + \kappa/4 \ \text{(by (30))} \\
&\leq \ell(w^*, T) + \frac{c_1\eta}{\epsilon}\left(1 + \frac{\ln^{\lambda/2}(1+1/b_k))z_k}{\tau_k}\right) + \kappa/4 \ \text{(since } w^* \in B(w_{k-1}, r_k)\text{)} \\
&\leq \ell(w^*, p) + \frac{c_1\eta}{\epsilon}\left(1 + \frac{\ln^{\lambda/2}(1+1/b_k))z_k}{\tau_k}\right) + \kappa/3 \ \text{(by (30))}.
\end{aligned}$$

This, together with (22) and (29), gives

$$\begin{aligned}
\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k) &\leq 2\ell(w^*, W_C) + \frac{c_2\eta}{\epsilon} + \frac{c_3\sqrt{\eta/\epsilon}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k} + \frac{c_1\eta}{\epsilon}\left(1 + \frac{\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k}\right) + 2\kappa/5 \\
&\leq 2L(w^*) + \frac{c_2\eta}{\epsilon} + \frac{c_3\sqrt{\eta/\epsilon}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k} + \frac{c_1\eta}{\epsilon}\left(1 + \frac{\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k}\right) + \kappa/2 \\
&\leq \kappa/3 + \frac{c_2\eta}{\epsilon} + \frac{c_3\sqrt{\eta/\epsilon}\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k} + \frac{c_1\eta}{\epsilon}\left(1 + \frac{\ln^{\lambda/2}(1+1/b_k)z_k}{\tau_k}\right) + \kappa/2,
\end{aligned}$$

by Lemma D.4.

Now notice that z_k/τ_k is $\Theta(1)$. Hence an $\Omega(\frac{\epsilon}{\log^{\lambda}(1/\epsilon)})$ bound on η suffices to imply that $\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k) \leq \kappa$ with probability $(1 - \frac{\delta}{k+k^2})$. □

D.4 Putting it together

Now we are ready to put everything together. The proof of Theorem 4.2 follows the high level structure of the proof of [BBZ07]; the new element is the application of Theorem D.3 which analyzes the performance of the hinge loss minimization algorithm for learning inside the band, which in turn applies Theorem D.1, which analyzes the benefits of our new localized outlier removal procedure.

Proof (of Theorem 4.2): We will prove by induction on k that after $k \leq s$ iterations, we have $\mathrm{err}_D(w_k) \leq M^k$ with probability $1 - \delta(1 - 1/(k+1))/2$.

When $k = 0$, all that is required is $\mathrm{err}_D(w_0) \leq 1$.

Assume now the claim is true for $k-1$ ($k \geq 1$). Then by induction hypothesis, we know that with probability at least $1 - \delta(1 - 1/k)/2$, w_{k-1} has error at most $M^{-(k-1)}$. Using part 3 of Definition 4.1, this implies that $\theta(w_{k-1}, w^*) \leq M^{-(k-1)}/c_6$. This in turn implies $\theta(w_{k-1}, w^*) \leq \pi/2$. (When $k = 1$, this is by assumption, and otherwise it is implied by part 3 of Definition 4.1.)

Let us define $S_{w_{k-1},b_{k-1}} = \{x : |w_{k-1} \cdot x| \leq b_{k-1}\}$ and $\bar{S}_{w_{k-1},b_{k-1}} = \{x : |w_{k-1} \cdot x| > b_{k-1}\}$. Since w_{k-1} has unit length, and $v_k \in B(w_{k-1}, r_k)$, we have $\theta(w_{k-1}, v_k) \leq r_k$ which in turn implies $\theta(w_{k-1}, w_k) \leq \min\{M^{-(k-1)}/c_6, \pi/2\}$.

Applying part 2 of Definition 4.1 to bound the error rate outside the band, we have both:

$$\Pr_x\left[(w_{k-1} \cdot x)(w_k \cdot x) < 0, x \in \bar{S}_{w_{k-1},b_{k-1}}\right] \leq \frac{M^{-k}}{4} \quad \text{and}$$

and

$$\Pr_x\left[(w_{k-1} \cdot x)(w^* \cdot x) < 0, x \in \bar{S}_{w_{k-1},b_{k-1}}\right] \leq \frac{M^{-k}}{4}.$$

Taking the sum, we obtain $\Pr_x\left[(w_k \cdot x)(w^* \cdot x) < 0, x \in \bar{S}_{w_{k-1},b_{k-1}}\right] \leq \frac{M^{-k}}{2}$. Therefore, we have

$$\mathrm{err}(w_k) \leq (\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k)) \Pr(S_{w_{k-1},b_{k-1}}) + \frac{M^{-k}}{2}.$$

Since $\Pr(S_{w_{k-1},b_{k-1}}) \leq 2c_7' b_{k-1}$, this implies

$$\mathrm{err}(w_k) \leq (\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k))2c_7' b_{k-1} + \frac{M^{-k}}{2} \leq M^{-k}\left((\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k))2c_1' c_7' M + 1/2\right).$$

Recall that $D_{w_{k-1},b_{k-1}}$ is the distribution obtained by conditioning D on the event that $x \in S_{w_{k-1},b_{k-1}}$. Applying Theorem D.3, with probability $1 - \frac{\delta}{2(k+k^2)}$, w_k has error at most $\kappa = \frac{1}{4c_1' c_7' M}$ within $S_{w_{k-1},b_{k-1}}$, implying that $\mathrm{err}(w_k) \leq 1/M^k$, completing the proof of the induction, and therefore showing, with probability at least $1 - \delta$, $O(\log(1/\epsilon))$ iterations suffice to achieve $\mathrm{err}(w_k) \leq \epsilon$.

A polynomial number of unlabeled samples are required by the algorithm and the number of labeled examples required by the algorithm is $\sum_k m_k = O(d(d + \log\log(1/\epsilon) + \log(1/\delta))\log(1/\epsilon))$. □

E Proof of Theorem E.1

In this section, we describe an algorithm for learning λ-admissible distribution in the presence of adversarial label noise. As before, we assume that the algorithm has access to w_0 such that $\theta(w_0, w^*) < \pi/2$. This can be shown to be without loss of generality exactly as in the case of malicious noise.

Theorem E.1. *Let D be a distribution over R^d chosen from a λ-admissible sequence of distributions. Let w^* be the (unit length) target weight vector. There are absolute positive constants $c_1', ..., c_4'$ and $M > 1$ and polynomial p such that, an $\Omega\left(\frac{\epsilon}{\log^\lambda(\frac{1}{\epsilon})}\right)$ upper bound on a rate η of adversarial label noise suffices to imply that for any $\epsilon, \delta > 0$, using the algorithm from Figure 3 with cut-off values $b_k = c_1' M^{-k}$, radii $r_k = c_2' M^{-k}$, $\kappa = c_3'$, $\tau_k = c_4' M^{-k}$ for $k \geq 1$, a number $n_k = p(d, M^k, \log(1/\delta))$ of unlabeled examples*

Figure 3 COMPUTATIONALLY EFFICIENT ALGORITHM TOLERATING ADVERSARIAL LABEL NOISE

Input: allowed error rate ϵ, probability of failure δ, an oracle that returns x, for (x, y) sampled from $\mathrm{EX}_\eta(f, D)$, and an oracle for getting the label from an example; a sequence of sample sizes $m_k > 0$; a sequence of cut-off values $b_k > 0$; a sequence of hypothesis space radii $r_k > 0$; a precision value $\kappa > 0$

1. Draw m_1 examples and put them into a working set W.
2. For $k = 1, \ldots, s = \lceil \log_2(1/\epsilon) \rceil$
   (a) Find $v_k \in B(w_{k-1}, r_k)$ to approximately minimize training hinge loss over W s.t. $\|v_k\|_2 \leq 1$: $\ell_{\tau_k}(v_k, W) \leq \min_{w \in B(w_{k-1}, r_k) \cap B(0,1))} \ell_{\tau_k}(w, W) + \kappa/8$
   (b) Normalize v_k to have unit length, yielding $w_k = \frac{v_k}{\|v_k\|_2}$.
   (c) Clear the working set W.
   (d) **Until** m_{k+1} additional data points are put in W, given x for $(x, f(x))$ obtained from $\mathrm{EX}_\eta(f, D)$, **if** $|w_k \cdot x| \geq b_k$, **then** reject x **else** put into W

Output: Weight vector w_s of error at most ϵ with probability $1 - \delta$.

in round k and a number $m_k = O\left(d \log\left(\frac{d}{\epsilon\delta}\right)(d + \log(k/\delta))\right)$ of labeled examples in round $k \geq 1$, and w_0 such that $\theta(w_0, w^*) < \pi/2$, after $s = \lceil \log_2(1/\epsilon) \rceil$ iterations, we find a separator w_s satisfying $\mathrm{err}(w_s) = \Pr_{(x,y)\sim D}[\mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w^* \cdot x)] \leq \epsilon$ with probability at least $1 - \delta$.

If the support of D is bounded in a ball of radius $R(d)$, then $m_k = O\left(R(d)^2(d + \log(k/\delta))\right)$ label requests suffice.

To prove Theorem E.1, all we need is Theorem E.2 below, which bounds the error inside the band in the case of adversarial label noise. Substituting this lemma for Theorem D.3 in the proof of Theorem 4.2 suffices to prove Theorem E.1. (In particular, for the rest of this subsection, r_k, b_k, κ and τ_k are set as in the proof of Theorem 4.2.)

Theorem E.2. *During round k of the algorithm in Figure 3, with probability $1 - \frac{\delta}{k+k^2}$, we have*

$$err_{D_{w_{k-1}, b_{k-1}}}(w_k) \leq \kappa.$$

We will prove Theorem E.2 using a series of lemmas.

Define ℓ and L as in the proof of Theorem D.3.

First, Lemma C.5, that $L(w^*) \leq \kappa/6$, also applies here, using exactly the same proof.

From here, the proof is organized a little differently than before. There are two main structural differences. First, before, we analyzed a relatively large set of unlabelled examples on which the algorithm performed soft outlier removal, before subsampling and training. Here, since the algorithm will not perform outlier removal, we may analyze the underlying distribution in place of the large unlabeled sample. The second difference is that, whereas before, we separately analyzed the clean examples and the dirty examples, here, we will analyze properties of the noisy portion of the underlying distribution, but, here, instead of comparing it with the clean portion, as we did before, we will compare it with the distribution that would be obtained by fixing the incorrect labels. One reason that this is more convenient is that the marginal over the instances of this "fixed" distribution is D (whereas the marginal of the clean examples, in general, is not).

Let $\tilde{P}$ be the joint distribution used by the algorithm, which includes the noisy labels chosen by the adversary. Let $N = \{(x, y) : \mathrm{sign}(w^* \cdot x) \neq y\}$ consist of noisy examples, so that $\tilde{P}(N) \leq \eta$. Let P be the joint distribution obtained by applying the correct labels. Let $\tilde{P}_k$ be the distribution on the examples given

to the algorithm in round k (obtained by conditioning $\tilde{P}$ to examples that fall within the band), and let P_k be the corresponding joint distribution with clean labels.

The key lemma here is to relate the expected loss with respect to $\tilde{P}_k$ to the expected loss with respect to P_k.

Lemma E.3. *There is an absolute positive constant c such that, if we define $z_k = \sqrt{r_k^2 + b_{k-1}^2}$ then for any $w \in B(w_{k-1}, r_k)$, we have*

$$|\mathbf{E}_{(x,y)\in P_k}\ell(w,x,y)) - \mathbf{E}_{(x,y)\in \tilde{P}_k}\ell(w,x,y))| \leq c\frac{\sqrt{M^{-k}\eta}z_k \log^{\lambda/2}(1+1/\gamma)}{\tau_k}. \tag{31}$$

Proof. Fix an arbitrary $w \in B(w_{k-1}, r_k)$. Recalling that N is the set of noisy examples, and that the marginals of P_k and $\tilde{P}_k$ on the inputs are the same, we have

$$\begin{aligned}
&|\mathbf{E}_{(x,y)\in P_k}(\ell(w,x,y)) - \mathbf{E}_{(x,y)\in \tilde{P}_k}(\ell(w,x,y))| \\
&= |\mathbf{E}_{(x,y)\in \tilde{P}_k}(\ell(w,x,y) - \ell(w,x,\mathrm{sign}(w^*\cdot x)))| \\
&= |\mathbf{E}_{(x,y)\in \tilde{P}_k}(1_{(x,y)\in N}(\ell(w,x,y) - \ell(w,x,-y)))| \\
&\leq \mathbf{E}_{(x,y)\in \tilde{P}_k}(1_{(x,y)\in N}|\ell(w,x,y) - \ell(w,x,-y)|) \\
&\leq 2\mathbf{E}_{(x,y)\in \tilde{P}_k}\left(1_{(x,y)\in N}\left(\frac{|w\cdot x|}{\tau_k}\right)\right) \\
&= \frac{2}{\tau_k}\mathbf{E}_{(x,y)\in \tilde{P}_k}\left(1_{(x,y)\in N}|w\cdot x|\right) \\
&\leq \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim\tilde{P}_k}(N)} \times \sqrt{\mathbf{E}_{(x,y)\in\tilde{P}_k}((w\cdot x)^2)}
\end{aligned}$$

by the Cauchy-Schwartz inequality. Part 1 of Definition 4.1 implies that

$$\Pr_{(x,y)\in\tilde{P}_k}(N) \leq \frac{\Pr_{(x,y)\in\tilde{P}}(N)}{\Pr_{(x,y)\in\tilde{P}}(S_{w_{k-1},b_{k-1}})} \leq cM^{-k}\eta,$$

for an absolute constant c, and part 4 of Definition 4.1 implies $\mathbf{E}_{(x,y)\in\tilde{P}_k}((w\cdot x)^2) \leq cz_k^2\log^{\lambda}(1+1/\gamma)$. ☐

Proof of Theorem E.2. Let

$$\mathrm{cleaned}(W) = \{(x, \mathrm{sign}(w^*\cdot x)) : (x,y) \in W\}.$$

Exploiting the fact that $\ell(w,x,y) = O\left(\sqrt{d\log\left(\frac{d}{\epsilon\delta}\right)}\right)$ for all $(x,y) \in S_{w_{k-1},b_{k-1}}$ and $w \in B(w_{k-1}, r_k)$ as in the proof of Lemma D.10, with probability $1 - \frac{\delta}{k+k^2}$, for all $w \in B(w_{k-1}, r_k)$, we have

$$|\mathbf{E}_{(x,y)\in\tilde{P}}(\ell(w,x,y)) - \ell(w,W)| \leq \kappa/32, \text{ and } |\mathbf{E}_{(x,y)\in P}(\ell(w,x,y)) - \ell(w,\mathrm{cleaned}(W))| \leq \kappa/32. \tag{32}$$

Then we have, for absolute constants c_1 and c_2, the following:

$$\begin{aligned}
\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k) &\leq \mathbf{E}_{(x,y)\in P_k}(\ell(w_k,x,y)) \quad \text{(since for each error, the hinge loss is at least 1)} \\
&\leq 2\mathbf{E}_{(x,y)\in P_k}(\ell(v_k,x,y)) \quad \text{(since } \|v_k\|_2 \geq 1/2) \\
&\leq 2\mathbf{E}_{(x,y)\in \tilde{P}_k}(\ell(v_k,x,y)) + c_1\sqrt{\frac{\eta}{\epsilon}} \times \frac{z_k \log^{\lambda/2}(1+1/b_k)}{\tau_k} \quad \text{(by Lemma E.3)} \\
&\leq 2\ell(v_k,W) + c_1\sqrt{\frac{\eta}{\epsilon}} \times \frac{z_k \log^{\lambda/2}(1+1/b_k)}{\tau_k} + \kappa/16 \quad \text{(by (32))} \\
&\leq 2\ell(w^*,W) + c_1\sqrt{\frac{\eta}{\epsilon}} \times \frac{z_k \log^{\lambda/2}(1+1/b_k)}{\tau_k} + \kappa/8 \\
&\leq 2\mathbf{E}_{(x,y)\in \tilde{P}_k}(\ell(w^*,x,y)) + c_1\sqrt{\frac{\eta}{\epsilon}} \times \frac{z_k \log^{\lambda/2}(1+1/b_k)}{\tau_k} + \kappa/4 \quad \text{(by (32))} \\
&\leq 2\mathbf{E}_{(x,y)\in P}(\ell(w^*,x,y)) + c_2\sqrt{\frac{\eta}{\epsilon}} \times \frac{z_k \log^{\lambda/2}(1+1/b_k)}{\tau_k} + \kappa/4 \quad \text{(by Lemma E.3)} \\
&\leq c_2\sqrt{\frac{\eta}{\epsilon}} \times \frac{z_k \log^{\lambda/2}(1+1/b_k)}{\tau_k} + \kappa/2.
\end{aligned}$$

since $L(w^*) \leq \kappa/6$. Since $z_k/\tau_k = \Theta(1)$, there is an constant c_3 such that, $\eta \leq c_3\epsilon/\log^{\lambda}(1+1/b_k)$ suffices for $\mathrm{err}_{D_{w_{k-1},b_{k-1}}}(w_k) \leq \kappa$, completing the proof. □

F Admissibility

F.1 Uniform distribution is 0-admissible

We will show the properties in Definition 4.1 hold for the uniform distribution with $\lambda = 0$. Part 1 is an easy consequence of the corresponding known lemmas about the uniform distribution on the unit ball.

Lemma F.1 (see [Bau90, BBZ07, KKMS05])**.** *For any $C > 0$, there are $c_1, c_2 > 0$ such that, for x drawn from the uniform distribution over $\sqrt{d}S_{d-1}$ and any unit length $u \in \mathbf{R}^d$, (a) for all $a, b \in [-C, C$ for which $a \leq b$, we have $c_1|b-a| \leq \Pr(u\cdot x \in [a,b]) \leq c_2|b-a|$, and (b) if $b \geq 0$, we have $\Pr(u\cdot x > b) \leq \frac{1}{2}e^{-b^2/2}$.*

To prove part 2, we will use a lemma from [BL13] that generalizes and strengthens a key lemma from [BBZ07].

Lemma F.2 (Theorem 4 of [BL13])**.** *For any $c_1 > 0$, there is a $c_2 > 0$ such that the following holds. Let u and v be two unit vectors in R^d, and assume that $\theta(u,v) = \alpha < \pi/2$. If D is isotropic log-concave in $\mathbf{R}^d$, then $\Pr_{x\sim D}[\mathrm{sign}(u\cdot x) \neq \mathrm{sign}(v\cdot x)$ and $|v\cdot x| \geq c_2\alpha] \leq c_1\alpha$.*

This has the following corollary, which proves part 2.

Lemma F.3. *For any $c_1 > 0$, there is a $c_2 > 0$ such that the following holds for all $d \geq 4$. Let u and v be two unit vectors in R^d, and assume that $\theta(u,v) = \alpha < \pi/2$. If D is uniform over S_{d-1}, $\Pr_{x\sim D}[\mathrm{sign}(u\cdot x) \neq \mathrm{sign}(v\cdot x)$ and $|v\cdot x| \geq c_2\alpha/\sqrt{d}] \leq c_1\alpha$.*

Proof. Consider the distribution D' obtained sampling from D, and scaling the result up by a factor of $\sqrt{d}$.

We claim that the projection D'' of D' onto the space spanned by $u \cdot x$ and $v \cdot x$ is isotropic log-concave. This will imply Lemma F.3 by applying Lemma F.2, since the event in question only concerns the span of $u \cdot x$ and $v \cdot x$.

Assume without loss of generality that the span of $u \cdot x$ and $v \cdot x$ is

$$T = \{(x_1, x_2, 0, 0, ..., 0) : x \in \mathbf{R}^d\}.$$

The fact that D'' is isotropic follows from the fact that D' is isotropic and the fact that it is log-concave follows from the known fact that, if $(x_1, ..., x_d)$ is sampled uniformly from S_{d-1}, then the distribution of (x_1, x_2) is log-concave (see Corollary 4 of [BGMN05]). □

Part 3 of Definition 4.1 holds trivially in the case of the uniform distribution.

The fact that part 4 of Definition 4.1 holds in the case of the isotropic rescaling of the uniform distribution U over the surface of the unit ball follows immediately from Lemma C.2.

Part 5 follows from the fact that D is isotropic logconcave (see Lemma F.4 below).

F.2 Isotropic log-concave is 2 admissible

Part 1 of Definition 4.1 is part of the following lemma.

Lemma F.4 ([LV07])**.** *Assume that D is isotropic log-concave in R^d and let f be its density function.*

(a) We have $\Pr_{x\sim D}[||X||_2 \geq \alpha\sqrt{d}] \leq e^{-\alpha+1}$. If $d = 1$ then: $\Pr_{x\sim D}[X \in [a, b]] \leq |b - a|$.

(b) All marginals of D are isotropic log-concave.

(c) If $d = 1$ we have $f(0) \geq 1/8$ and $f(x) \leq 1$ for all x.

(d) There is an absolute constant c such that, if $d = 1$, $f(x) > c$ for all $x \in [-1/9, 1/9]$.

Part 2 is Lemma F.2.

Part 3 is implicit in [Vem10] (see Lemma 3 of [BL13]).

In order to prove part 4, we will use the following lemma.

Lemma F.5. *For any $C > 0$, there exists a constant c s.t., for any isotropic log-concave distribution D, for any a such that, $\|a\|_2 \leq 1$, and $||u - a||_2 \leq r$, for any $0 < \gamma < C$, and for any $K \geq 4$, we have*

$$\Pr_{x\sim D_{u,\gamma}}\left(|a \cdot x| > K\sqrt{r^2 + \gamma^2}\right) \leq \frac{c}{\gamma}e^{-K}.$$

Proof. W.l.o.g. we may assume that $u = (1, 0, 0, \cdots, 0)$.

Let $a' = (a_2, ..., a_d)$, and, for a random $x = (x_1, x_2, ..., x_d)$ drawn from $D_{u,\gamma}$, let $x' = (x_2, ..., x_d)$. Let

$$p = \Pr_{x\sim D_{u,\gamma}}\left(|a \cdot x| > K\sqrt{r^2 + \gamma^2}\right)$$

be the probability that we want to bound. We may rewrite p as

$$p = \frac{\Pr_{x\sim D}\left(|a \cdot x| > K\sqrt{r^2 + \gamma^2} \text{ and } |x_1| \leq \gamma\right)}{\Pr_{x\sim D}(|x_1| \leq \gamma)}. \tag{33}$$

Lemma F.4 implies that there is a positive constant c_1 such that the denominator satisfies the following lower bound:

$$\Pr_{x\sim D}\left(|x_1| \le \gamma\right) \ge c_1 \min\{\gamma, 1/9\} \ge \frac{c_1\gamma}{9C}. \tag{34}$$

So now, we just need an upper bound on the numerator. We have

$$\begin{aligned}
&\Pr_{x\sim D}\left(|a\cdot x| > K\sqrt{r^2+\gamma^2} \text{ and } |x_1| \le \gamma\right) \le \Pr_{x\sim D}\left(|a'\cdot x'| > K\sqrt{r^2+\gamma^2} - \gamma\right)\\
&\le \Pr_{x\sim D}\left(|a'\cdot x'| > (K-1)\sqrt{r^2+\gamma^2}\right) \le \Pr_{x\sim D}\left(|a'\cdot x'| > (K-1)r\right)\\
&\le \Pr_{x\sim D}\left(\left|\left(\frac{a'}{||a'||_2}\right)\cdot x'\right| > K-1\right) \le e^{-(K-1)},
\end{aligned}$$

by Lemma F.4, since the marginal distribution over x' is isotropic log-concave. Combining with (33) and (34) completes the proof. □

Now we're ready to prove Part 4.

Lemma F.6. *For any C, there is a constant c such that, for all $0 < \gamma \le C$, for all a such that $\|u-a\|_2 \le r$ and $\|a\|_2 \le 1$*

$$\mathbf{E}_{x\sim D_{u,\gamma}}((a\cdot x)^2) \le c(r^2+\gamma^2)\ln^2(1+1/\gamma).$$

Proof: Let $z = \sqrt{r^2+\gamma^2}$. Setting, with foresight, $t = 9z^2\ln^2(1+1/\gamma)$, we have

$$\begin{aligned}
&\mathbf{E}_{x\sim D_{u,\gamma}}((a\cdot x)^2)\\
&= \int_0^\infty \Pr_{x\sim D_{u,\gamma}}\left((a\cdot x)^2 \ge \alpha\right)\, d\alpha\\
&\le t + \int_t^\infty \Pr_{x\sim D_{u,\gamma}}\left((a\cdot x)^2 \ge \alpha\right)\, d\alpha.
\end{aligned} \tag{35}$$

Since $t \ge 4\sqrt{r^2+\gamma^2}$, Lemma F.5 implies that, for an absolute constant c, we have

$$\mathbf{E}_{x\sim D_{u,\gamma}}((a\cdot x)^2) \le t + \frac{c}{\gamma}\int_t^\infty \exp\left(-\left(\frac{\alpha}{r^2+\gamma^2}\right)^{1/2}\right) d\alpha.$$

Now, we want to evaluate the integral. Since $z = \sqrt{r^2+\gamma^2}$, so

$$\int_t^\infty \exp\left(-\sqrt{\frac{\alpha}{r^2+\gamma^2}}\right) d\alpha = \int_t^\infty \exp\left(-\sqrt{\alpha}/z\right)\ d\alpha.$$

Using a change of variables $u^2 = \alpha$, we get

$$\int_t^\infty \exp\left(-\sqrt{\alpha}/z\right) d\alpha = 2\int_{\sqrt{t}}^\infty u\exp\left(-u/z\right) du = 2z^2(\sqrt{t}+1)\exp\left(-\sqrt{t}/z\right).$$

Putting it together, we get

$$\mathbf{E}_{x\sim D_{u,\gamma}}((a\cdot x)^2) \le t + \frac{z^2(\sqrt{t}+1)\exp\left(-\sqrt{t}/z\right)}{\gamma} \le t + z^2,$$

since $t = 9z^2\ln^2(1+1/\gamma)$, completing the proof. □

Finally Part 5 is also part of Lemma F.4.

G Relating Adversarial Label Noise and the Agnostic Setting

In this section we study the agnostic setting of [KSS94, KKMS05] and describe how our results imply constant factor approximations in that model. In the agnostic model, data (x, y) is generated from a distribution D over $\Re^d \times \{1, -1\}$. For a given concept class C, let OPT be the error of the best classifier in C. In other words, $OPT = argmin_{f \in C} err_D(f) = argmin_{f \in C} Pr_{(x,y)\sim D}[f(x) \neq y]$. The goal of the learning algorithm is to output a hypothesis h which is nearly as good as f, i.e., given $\epsilon > 0$, we want $err_D(h) \leq c \cdot OPT + \epsilon$, where c is the approximation factor. Any result in the adversarial model that we study, translates into a result for the agnostic setting via the following lemma.

Lemma G.1. *For a given concept class C and distribution D, if there exists an algorithm in the* adversarial *noise model which runs in time poly$(d, 1/\epsilon)$ and tolerates a noise rate of $\eta = \Omega(\epsilon)$, then there exists an algorithm for (C, D) in the agnostic setting which runs in time poly$(d, 1/\epsilon)$ and achieves error $O(OPT+\epsilon)$.*

Proof. Let f^* be the optimal halfspace with error OPT. In the adversarial setting, w.r.t. f^*, the noise rate η will be exactly OPT. Set $\epsilon' = c(OPT + \epsilon)$ as input to the algorithm for the adversarial model. By the guarantee of the algorithm we will get a hypothesis h such that $Pr_{(x,y)\sim D}[h(x) \neq f^*(x)] \leq \epsilon' = c(OPT+\epsilon)$. Hence by triangle inequality, we have $err_D(h) \leq err_D(f^*)+c(OPT+\epsilon) = O(OPT+\epsilon)$. □

For the case when C is the class of origin centered halfspaces in R^d and the marginal of D is the uniform distribution over S_{d-1}, the above lemma along with Theorem 1.2 implies that we can output a halfspace of accuracy $O(OPT + \epsilon)$ in time poly$(d, 1/\epsilon)$. The work of [KKMS05] achieves a guarantee of $O(OPT + \epsilon)$ in time exponential in $1/\epsilon$ by doing L_2 regression to learn a low degree polynomial[2].

H Proof of VC lemmas

In this section, we apply some standard VC tools to establish some lemmas about estimates of expectations.

Definition H.1. *Say that a set F of real-valued functions with a common domain X* shatters *$x_1, ..., x_d \in X$ if there are thresholds $t_1, ..., t_d$ such that*

$$\{(\text{sign}(f(x_1) - t_1), ..., \text{sign}(f(x_d) - t_d)) : f \in F\} = \{-1, 1\}^d.$$

The pseudo-dimension *of F is the size of the largest set shattered by F.*

We will use the following bound.

Lemma H.2 (see [AB99])**.** *Let F be a set of functions from a common domain X to $[a, b]$ and let d be the pseudo-dimension of F, and let D be a probability distribution over X. Then, for $m = O\left(\frac{(b-a)^2}{\alpha^2}(d + \log(1/\delta))\right)$, if $x_1, ..., x_m$ are drawn independently at random according to D, with probability $1 - \delta$, for all $f \in F$,*

$$\left| \mathbf{E}_{x\sim D}(f(x)) - \frac{1}{m}\sum_{t=1}^{m} f(x_t) \right| \leq \alpha.$$

[2] They further show that L_1 regression can achieve a stronger guarantee of $OPT + \epsilon$

H.1 Proof of Lemma C.10 and Lemmas D.10

The pseudo-dimension of the set of linear combinations of d variables is known to be d [Pol11]. Since, for any non-increasing function $\psi : \mathbf{R} \to \mathbf{R}$ and any F, the pseudo-dimension of $\{\psi \circ f : f \in F\}$ is at most that of F (see [Pol11]), the pseudo-dimension of $\{\ell(w, \cdot) : w \in \mathbf{R}^d\}$ is at most d.

Let D' be the distribution obtained by conditioning D on the event that $||x|| < R$ ($||x|| < 1$ for uniform distribution). For $\ell \leq n_k$, the total variation distance between the joint distribution of ℓ draws from D' and ℓ draws from D is at most $1 - \frac{\delta}{4(k+k^2)}$, so it suffices to prove (29) and (30) with respect to D' ((17) and (18) respectively for the uniform distribution). Applying Lemma D.6 and Lemma H.2 then completes the proof.

H.2 Proof of Lemma D.2

Define f_a by $f_a(x) = (a \cdot x)^2$. The pseudo-dimension of the set of all such functions is $O(d)$ [KLS09]. As the proof of Lemma D.10, w.l.o.g., all x have $||x||_2 \leq R$, and applying Lemma H.2 completes the proof.