

# The Power of Localization for Efficiently Learning Linear Separators with Noise

Pranjal Awasthi  
Princeton University  
pawashti@cs.princeton.edu

Maria Florina Balcan<sup>\*</sup>  
Georgia Institute of  
Technology  
ninamf@cc.gatech.edu

Philip M. Long  
Microsoft  
plong@microsoft.com

## ABSTRACT

We introduce a new approach for designing computationally efficient and noise tolerant algorithms for learning linear separators. We consider the malicious noise model of Valiant [41, 32] and the adversarial label noise model of Kearns, Schapire, and Sellie [34]. For malicious noise, where the adversary can corrupt an  $\eta$  of fraction both the label part and the feature part, we provide a polynomial-time algorithm for learning linear separators in  $\mathbb{R}^d$  under the uniform distribution with nearly information-theoretically optimal noise tolerance of  $\eta = \Omega(\epsilon)$ , improving on the  $\Omega\left(\frac{\epsilon}{d^{1/4}}\right)$  noise-tolerance of [31] and the  $\Omega\left(\frac{\epsilon^2}{\log(d/\epsilon)}\right)$  of [35]. 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  $\eta$ , we give a polynomial-time algorithm for learning linear separators in  $\mathbb{R}^d$  under the uniform distribution that can also handle a noise rate of  $\eta = \Omega(\epsilon)$ . This improves over the results of [31] 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 [35]. In particular, in the case of malicious noise, unlike previous results, our noise tolerance has no dependence on the dimension  $d$  of the space.

Our algorithms are also efficient in the active learning setting, where learning algorithms 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  $\epsilon$  is polylogarithmic (and thus 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 malicious noise or adversarial label noise.

<sup>\*</sup>This work was supported in part by NSF grants CCF-0953192 and CCF-1101215, AFOSR grant FA9550-09-1-0538, and a Microsoft Research Faculty Fellowship.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

STOC '14 New York, NY, USA

Copyright 2014 ACM 978-1-4503-2710-7/14/05 ...\$15.00.

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

## Categories and Subject Descriptors

F.2.0 [Analysis of Algorithms and Problem Complexity]: [General]

## General Terms

Algorithms, Theory

## Keywords

Noise tolerant learning, Adversarial label noise, Malicious noise, Passive and active learning

## 1. INTRODUCTION

**Overview.** Dealing with noisy data is one of the main challenges in machine learning and is an active area of research. In this work we study the noise-tolerant learning of linear separators, arguably the most popular class of functions used in practice [19]. 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 [33, 42] and in the more modern active learning framework [21]. 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 [1, 27], cryptographic assumptions [14, 39], and are connected to other classic questions in learning theory (e.g., learning DNF formulas [34]), and appear as barriers in differential privacy [26].

In this paper we present new techniques for designing efficient algorithms for learning linear separators in the presence of *malicious noise* and *adversarial label noise*. These models were originally proposed for a setting in which the algorithm must work for an arbitrary, unknown distribution. As we will see, bounds on the amount of noise tolerated for this distribution-free setting were weak, and no significant progress was made for many years. This motivated research investigating the role of the distribution generating the data on the tolerable level of noise: a breakthrough result of [31] and subsequent work of [35] showed that indeed better bounds can be obtained for the uniform and isotropic log-concave

distributions. In this paper, we continue this line of research. 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 a near-optimal amount of malicious noise (within constant factor of the statistical limit) for the uniform distribution, and also improve over the previously known results for log-concave distributions. 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 class labels 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  $\epsilon$  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 [3, 37]. It also provides the first analysis showing the benefits of active learning over passive learning under the challenging malicious noise model.

Our work brings a new set of algorithmic and analysis techniques including localization (previously used for obtaining better sample complexity results) and soft outlier removal that we believe will have other applications in learning theory and optimization. Localization [9, 15, 44, 4, 16, 36, 29, 8] refers to the practice of progressively narrowing the focus of a learning algorithm to an increasingly restricted range of possibilities (which are known to be safe given the information up to a certain point in time), thereby improving the stability of estimates of the quality of these possibilities based on random data.

In the following we start by formally defining the learning models we consider. We then present the 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 learning paradigms: the classic passive learning setting and the more modern active learning scenario. As is typical [33, 42], we assume that there exists a distribution  $D$  over  $\mathbb{R}^d$  and a fixed unknown target function  $w^*$ . In the noise-free case, 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  $\text{err}_D(w) = \Pr_{x \sim D}[\text{sign}(w^* \cdot x) \neq \text{sign}(w \cdot x)] \leq \epsilon$ . In the active learning model [18, 21] 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 noise models. The first one is the malicious noise model of [41, 32] 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  $\eta$  the adversary can output an arbitrary pair  $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ . We will call  $\eta$  the noise rate. Each of the adversary’s examples can depend on the state of the learning algorithm and also the previ-

ous draws of the adversary. We will denote the malicious oracle as  $EX_\eta(D, w^*)$ . The goal remains, however, 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 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  $\eta$  the adversary can output an arbitrary pair  $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ . 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  $\mathbb{R}^d \times \{-1, 1\}$  whose marginal distribution over  $\mathbb{R}^d$  is  $D$  and such that  $\Pr_{(x, y) \sim P}(\text{sign}(w^* \cdot x) \neq y) \leq \eta$ . In the active learning version of this model, once again 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 [34] (used commonly both in passive and active learning (e.g., [31, 3, 28])). (Please see the full version [2] for the details.)

We will be interested in algorithms that run in time  $\text{poly}(d, 1/\epsilon)$  and use  $\text{poly}(d, 1/\epsilon)$  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  $\epsilon$ , 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 [32] implies that halfspaces can be efficiently learned with respect to arbitrary distributions in polynomial time while tolerating a malicious noise rate of  $\tilde{\Omega}(\frac{\epsilon}{d})$ . Kearns and Li [32] also showed that malicious noise at a rate greater than  $\frac{\epsilon}{1+\epsilon}$  cannot be tolerated (and a slight variant of their construction shows that this remains true even when the distribution is uniform over the unit sphere). The  $\tilde{\Omega}(\frac{\epsilon}{d})$  bound for the distribution-free case was not improved for many years. Kalai et al. [31] showed that,<sup>1</sup> when the distribution is uniform, the  $\text{poly}(d, 1/\epsilon)$ -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. [35] analyzed another approach to limiting the coordination of the noisy examples: they 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 machine learning applications have massive amounts of unannotated or unlabeled data, there has been significant interest in designing active learning algorithms that most efficiently utilize the available data, while mini-

<sup>1</sup>These results from [31] are most closely related to our work. We describe some of their other results, more prominently featured in their paper, later.

mizing the need for human intervention. Over the past decade there has been substantial 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 [24, 20, 3, 4, 28, 22, 17, 7, 36, 11, 43, 21, 38, 6]. However, despite many efforts, except for very simple noise models (random classification noise [5] and linear noise [23]), 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.

## 1.1 Our Results

The following are our main results.

**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  $\mathbb{R}^d$  in the presence of malicious noise such that an  $\Omega(\epsilon)$  upper bound on  $\eta$  suffices to imply that for any  $\epsilon, \delta > 0$ , the output  $w$  of  $A_{um}$  satisfies  $\Pr_{(x,y) \sim D}[\text{sign}(w \cdot x) \neq \text{sign}(w^* \cdot x)] \leq \epsilon$  with probability at least  $1 - \delta$ .*

**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  $\mathbb{R}^d$  in the presence of adversarial label noise such that an  $\Omega(\epsilon)$  upper bound on  $\eta$  suffices to imply that for any  $\epsilon, \delta > 0$ , the output  $w$  of  $A_{ul}$  satisfies  $\Pr_{(x,y) \sim D}[\text{sign}(w \cdot x) \neq \text{sign}(w^* \cdot x)] \leq \epsilon$  with probability at least  $1 - \delta$ .*

As a restatement of the above theorem, in the agnostic setting considered in [31], we can output a halfspace of error at most  $O(\eta + \alpha)$  in time  $\text{poly}(d, 1/\alpha)$ . Kalai et al. achieved error  $\eta + \alpha$  by learning a low degree polynomial in time whose dependence on the inverse accuracy is super-exponential. On the other hand, this result of [31] applies when the target halfspace does not necessary go through the origin.

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

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

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, and the most closely related previous work, in Tables 1 and 2.

## 1.2 Techniques

**Hinge Loss Minimization.** As minimizing the 0-1 loss in the presence of noise is NP-hard [30, 25], a natural approach is to minimize a surrogate convex loss that acts as a proxy for the 0-1 loss.

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

Passive Learning	Prior work	Our work
malicious	$\eta = \Omega\left(\frac{\epsilon}{d^{1/4}}\right)$ [31] $\eta = \Omega\left(\frac{\epsilon^2}{\log(d/\epsilon)}\right)$ [35]	$\eta = \Omega(\epsilon)$
adversarial	$\eta = \Omega(\epsilon/\sqrt{\log(1/\epsilon)})$ [31]	$\eta = \Omega(\epsilon)$
<b>Active Learning</b> (malicious and adversarial)	NA	$\eta = \Omega(\epsilon)$

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

Passive Learning	Prior work	Our work
malicious	$\eta = \Omega\left(\frac{\epsilon^2}{\log^2(d/\epsilon)}\right)$ [35]	$\eta = \Omega\left(\frac{\epsilon}{\log^2(1/\epsilon)}\right)$
adversarial	$\eta = \Omega\left(\frac{\epsilon^3}{\log(1/\epsilon)}\right)$ [35]	$\eta = \Omega\left(\frac{\epsilon}{\log^2(1/\epsilon)}\right)$
<b>Active Learning</b> (malicious and adversarial)	NA	$\Omega\left(\frac{\epsilon}{\log^2(1/\epsilon)}\right)$

A common choice in machine learning is to use the hinge loss:  $\max(0, 1 - y(w \cdot x))$ . In this paper, we use the slightly more general  $\ell_\tau(w, x, y) = \max\left(0, 1 - \frac{y(w \cdot x)}{\tau}\right)$ , 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  $\tau$  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  $\mathbb{R}^d$ . 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  $b_k = \Theta\left(\frac{2^{-k}}{\sqrt{d}}\right)$  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 [4] 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 distribution 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 idea was used in [4] to obtain active learning algorithms with improved label complexity ignoring computational complexity considerations<sup>2</sup>.

In this work, we build on this idea to produce polynomial time algorithms with improved noise tolerance. To obtain our results, we exploit several new ideas: (1) the performance of the rescaled

<sup>2</sup>We note that the localization considered by [4] is a more aggressive one than those considered in disagreement based active learning literature [3, 28, 36, 29, 43] and earlier in passive learning [9, 15, 44].

hinge loss minimization in smaller and smaller bands, (2) a analysis of properties of the distribution obtained after conditioning on the band that enables us to more sensitively identify cases in which the adversary concentrates the effects of noisy examples, (3) another type of localization — a novel soft outlier removal procedure.

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

In order to get better tolerance in the adversarial, or agnostic, setting, we note that examples  $x$  for which  $|w \cdot x|$  is large for  $w$  close to  $w_{k-1}$  are the most harmful, and, by analyzing the variance of  $w \cdot x$  for such directions  $w$ , we can more effectively limit 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 the algorithm’s confidence that the point is not “noisy”. We then minimize the weighted hinge loss. Combining this with the variance analysis mentioned above leads to a noise of tolerance of  $\Omega(\epsilon)$  in the malicious case.

**Soft Localized Outlier Removal.** Outlier removal techniques have been studied before in the context of learning problems [13, 35]. In [35], the goal of outlier removal was to limit the ability of the adversary to coordinate the effects of noisy examples – excessive such coordination was detected and removed. Our outlier removal procedure (see Figure 2) is similar in spirit to that of [35] with two key differences. First, as in [35], 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. As training proceeds, the band is increasingly shaped like a pancake, with  $w_{k-1}$  pointing in its flattest direction. Hypotheses that are close to  $w_{k-1}$  also point in flat directions; the variance in those directions is  $\Theta(b_k^2)$  which is much smaller than the  $\approx 1/d$  found in a generic direction. This allows us to limit the harm of the adversary to a greater extent than was possible in the analysis of [35]. The second difference is that, unlike previous outlier removal techniques, rather than making discrete remove-or-not decisions, we instead weigh the examples and then minimize the weighted hinge loss. Each weight indicates the algorithm’s confidence that an example is not noisy. 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 [40, 12].

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  $\beta$ -nearly log-concave distributions (for small enough  $\beta$ ). Such distributions, for instance, can capture mixtures of log-concave distributions [8].

## 2. PRELIMINARIES

Recall that  $\ell_{\tau}(w, x, y) = \max\left(0, 1 - \frac{y(w \cdot x)}{\tau}\right)$  and  $\ell_{\tau}(w, T) = \frac{1}{|T|} \sum_{(x,y) \in T} \ell_{\tau}(w, x, y)$ . Similarly, the expected hinge loss w.r.t.  $D$  is defined as  $L_{\tau}(w, D) = E_{x \sim D}(\ell_{\tau}(w, x, \text{sign}(w^* \cdot x)))$ . Our analysis will also consider the distribution  $D_{w, \gamma}$  obtained by conditioning  $D$  on membership in the band, i.e. the set  $\{x : \|x\|_2 = 1, |w \cdot x| \leq \gamma\}$ .

We present our algorithms in the active learning model. Since we will prove that our active algorithm only uses a polynomial number of unlabeled samples, this will imply a guarantee for passive learning setting. 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 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  $\theta(w_0, w^*) < \pi/2$ . This is without loss of generality for the types of problems we consider (see the full version [2]). We will also need the following useful properties of the uniform distribution.

- [10, 4, 31] For any  $c_1 > 0$ , there is a  $c_2 > 0$  such that, for  $x$  drawn from the uniform distribution over  $S_{d-1}$  and any unit length  $u \in \mathbf{R}^d$ , for all  $a, b \in [-c_1/\sqrt{d}, c_1/\sqrt{d}]$  for which  $a \leq b$ , we have

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

- [8] For any  $c_3 > 0$ , there is a  $c_4 > 0$  such that, for all  $d \geq 4$ , the following holds. Let  $u$  and  $v$  be two unit vectors in  $\mathbf{R}^d$ , and assume that  $\theta(u, v) = \alpha \leq \pi/2$ . Then

$$\Pr_{x \sim D} \left( \text{sign}(u \cdot x) \neq \text{sign}(v \cdot x) \text{ and } |v \cdot x| \geq c_4 \frac{\alpha}{\sqrt{d}} \right) \leq c_3 \alpha. \quad (2)$$

## 3. THE 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 distribution  $D$  is the uniform distribution over  $S_{d-1}$  and present our results for malicious noise. Theorem 1.1 is a corollary of Theorem 3.1, which follows.

**THEOREM 3.1.** *Let  $w^*$  be the (unit length) target weight vector. There are absolute positive constants  $c_1, \dots, c_5$  and a polynomial  $p$  such that an  $\Omega(\epsilon)$  upper bound on  $\eta$  suffices to imply that for any  $\epsilon, \delta > 0$ , using the algorithm from Figure 1 with cut-off values  $b_k = c_1 2^{-k} d^{-1/2}$ , radii  $r_k = c_2 2^{-k}$ ,  $\kappa = c_3$ ,  $\tau_k = c_4 2^{-k} d^{-1/2}$  for  $k \geq 1$ ,  $\xi_k = c_5$ ,  $\sigma_k^2 = 2 \left( \frac{r_k^2}{d-1} + b_{k-1}^2 \right)$ , 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  $\text{err}(w_s) = \Pr_{(x,y) \sim D}[\text{sign}(w_s \cdot x) \neq \text{sign}(w^* \cdot x)] \leq \epsilon$  with probability  $\geq 1 - \delta$ .*

### 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

---

**Figure 1** COMPUTATIONALLY EFFICIENT ALGORITHM TOLERATING MALICIOUS NOISE
 

---

**Input:** allowed error rate  $\epsilon$ , probability of failure  $\delta$ , an oracle that returns  $x$ , for  $(x, y)$  sampled from  $\text{EX}_\eta(f, D)$ , and an oracle for getting the label  $y$  from an example; a sequence of unlabeled sample sizes  $n_k > 0$ ,  $k \in \mathbb{Z}^+$ ; 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  $\xi_k$ ; a sequence of variance bounds  $\sigma_k^2$ ; precision value  $\kappa$ ; weight vector  $w_0$ .

1. Draw  $n_1$  examples and put them into a working set  $W$ .
2. For  $k = 1, \dots, s = \lceil \log_2(1/\epsilon) \rceil$ 
  - (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  $q : W \rightarrow [0, 1]$ . 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  $\text{EX}_\eta(f, D)$ , **if**  $|w_k \cdot x| \geq b_k$ , **then** reject  $x$  **else** put into  $W$

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

---

**Figure 2** LOCALIZED SOFT OUTLIER REMOVAL PROCEDURE
 

---

**Input:** a set  $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$  of samples; the reference unit vector  $u$ ; desired radius  $r$ ; a parameter  $\xi$  specifying the desired bound on the fraction of clean examples removed; a variance bound  $\sigma^2$

1. Find  $q : S \rightarrow [0, 1]$  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  $w \in B(u, r) \cap B(0, 1)$ ,  $\frac{1}{|S|} \sum_{x \in S} q(x)(w \cdot x)^2 \leq \sigma^2$

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

---

that falls outside  $S_{d-1}$  can be easily identified by the algorithm as noisy and removed, effectively lowering the noise rate.

Using techniques from [4], we may reduce our problem to a sub-problem concerning learning with respect to a distribution obtained by conditioning on membership in the band. In particular, we adapt the argument of [4] to show that, for a sufficiently small absolute constant  $\kappa$ , in order to prove Theorem 3.1, all we need is Theorem 3.2 stated below, together with the required bounds on computational, sample and label complexity.

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

We will first show how Theorem 3.2 is sufficient to prove the main result.

**Margin based analysis (Proof of Theorem 3.1):**

*Proof Sketch:* 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  $1 - \delta(1 - 1/(k+1))/2$ . See Appendix A.1 for the specific values of the constants in the statement of the theorem.

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

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 2 to bound the error rate outside the band, we have both:

$$\Pr_x [(w_{k-1} \cdot x)(w_k \cdot x) < 0, x \in \bar{S}_{w_{k-1}, b_{k-1}}] \leq 2^{-(k+3)} \quad \text{and}$$

$$\Pr_x [(w_{k-1} \cdot x)(w_k^* \cdot x) < 0, x \in \bar{S}_{w_{k-1}, b_{k-1}}] \leq 2^{-(k+3)}.$$

Taking the sum, we obtain

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

Therefore, we have

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

Equation 1 gives  $\Pr(S_{w_{k-1}, b_{k-1}}) \leq 2b_{k-1}\sqrt{d}$ , which implies

$$\begin{aligned} \text{err}(w_k) &\leq (\text{err}_{D_{w_{k-1}, b_{k-1}}}(w_k)) 2b_{k-1}\sqrt{d} + 2^{-(k+2)} \\ &\leq 2^{-(k+1)} \left( (\text{err}_{D_{w_{k-1}, b_{k-1}}}(w_k)) 4\tilde{c}_4 + 1/2 \right). \end{aligned}$$

Recall that  $D_{w_{k-1}, b_{k-1}}$  is the distribution obtained by conditioning  $D$  on the event that  $x \in S_{w_{k-1}, b_{k-1}}$ . By Theorem 3.2, with probability  $1 - \frac{\delta}{2(k+k^2)}$ ,  $w_k$  has error at most  $\kappa = \frac{1}{8\tilde{e}_4}$  within  $S_{w_{k-1}, b_{k-1}}$ , implying that  $\text{err}(w_k) \leq 2^{-(k+1)}$ , completing the proof of the induction, and therefore showing, with probability at least  $1 - \delta$ ,  $O(\log(1/\epsilon))$  iterations suffice to achieve  $\text{err}(w_k) \leq \epsilon$ .

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

### The error within a band in each iteration

In the rest of this section we will sketch the proof of Theorem 3.2 in a series of steps summarized in the lemmas below. First, we bound the expected 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  $\ell_{\tau_k}$  simply as  $\ell$  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 Eq. 1 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

$$L(w^*) \leq \frac{c'_2 \sqrt{d} \tau_k}{c'_1 \sqrt{d} b_{k-1}} = \frac{c'_2 \sqrt{d} c_4 2^{-k} / \sqrt{d}}{c'_1 \sqrt{d} c_1 2^{-k} / \sqrt{d}} < \kappa/12,$$

if we choose  $c_4$  small enough.  $\square$

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 ultimately 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 (which is applied with  $\eta' = \Theta(\eta 2^k)$ ).

THEOREM 3.4. *There is a constant  $c$  and a polynomial  $p$  such that, if  $n \geq p(1/\eta', d, 1/\xi, 1/\delta, 1/\gamma, 1/r)$  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 2 with  $\sigma^2 = c \left( \frac{r^2}{d-1} + \gamma^2 \right)$ , we have that with probability  $1 - \delta$ , the output  $q$  satisfies the following:*

(a)  $\sum_{(x, y) \in S} q(x) \geq (1 - \xi)|S|$  and (b) for all unit length  $w$  such that  $\|w - u\|_2 \leq r$ ,  $\frac{1}{|S|} \sum_{x \in S} q(x)(w \cdot x)^2 \leq \sigma^2$ . 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 show that, for all  $w$  with distance  $r$  of  $u$ , that  $E[(w \cdot x)^2]$  is at most  $\left( \frac{r^2}{d-1} + \gamma^2 \right)$ . 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 Theorem 3.4 is in Appendix A.

As explained in the introduction, the soft outlier removal procedure enables us to get a more refined bound on the extent to which the value  $\ell(w, p)$  minimized by the algorithm is a faithful proxy

for the value  $\ell(w, W_C)$  that it would minimize in the absence of noise. This is formalized in the following lemma. (Here  $\ell(w, p)$  and  $\ell(w, W_C)$  are defined with respect to the 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*

$$z_k = \sqrt{\frac{\tau_k^2}{d-1} + b_{k-1}^2}, \text{ then for any } w \in B(w_{k-1}, \tau_k), \text{ we have } \ell(w, W_C) \leq \ell(w, p) + \frac{C_1 \eta}{\epsilon} \left(1 + \frac{z_k}{\tau_k}\right) + \kappa/32 \text{ and } \ell(w, p) \leq 2\ell(w, W_C) + \kappa/32 + \frac{C_2 \eta}{\epsilon} + C_3 \sqrt{\frac{\tau_k}{\epsilon}} \times \frac{z_k}{\tau_k}.$$

A detailed proof of Lemma 3.5 is given in Appendix A. Here we 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-Schwartz 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  $z$ .

Given Lemma 3.3, Theorem 3.4, and Lemma 3.5, the proof of Theorem 3.2 can be summarized as follows. Let

$$E = \text{err}_{D_{w_{k-1}, b_{k-1}}}(w_k) = \text{err}_{D_{w_{k-1}, b_{k-1}}}(v_k)$$

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,

$$E \leq \ell(v_k, T) + \frac{C_1 \eta}{\epsilon} \left(1 + \frac{z_k}{\tau_k}\right) + \kappa/8.$$

The fact that  $v_k$  approximately minimizes the hinge loss, together with VC theory, gives  $E \leq \ell(w^*, p) + \frac{C_1 \eta}{\epsilon} \left(1 + \frac{z_k}{\tau_k}\right) + \kappa/3$ . Once again applying Lemma 3.5 and VC theory yields  $E \leq 2L(w^*) + \frac{C_1 \eta}{\epsilon} \left(1 + \frac{z_k}{\tau_k}\right) + \frac{C_2 \eta}{\epsilon} + C_3 \sqrt{\frac{\tau_k}{\epsilon}} \times \frac{z_k}{\tau_k} + \kappa/2$ . Since  $L(w^*) \leq \kappa/12$ , we get  $E \leq \kappa/6 + \frac{C_1 \eta}{\epsilon} \left(1 + \frac{z_k}{\tau_k}\right) + \frac{C_2 \eta}{\epsilon} + C_3 \sqrt{\frac{\tau_k}{\epsilon}} \times \frac{z_k}{\tau_k} + \kappa/2$ . Now notice that  $z_k/\tau_k$  is  $\Theta(1)$ . Hence an  $\Omega(\epsilon)$  bound on  $\eta$  suffices to imply, w.h.p., that  $\text{err}_{D_{w_{k-1}, b_{k-1}}}(w_k) \leq \kappa$ .

## 4. 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  $\psi$ , 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 *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  $D_4, D_5, \dots$  of probability distributions over  $\mathbf{R}^4, \mathbf{R}^5, \dots$  respectively is  $\lambda$ -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  $\mathbf{R}^d$ , and assume that  $\theta(u, v) = \alpha \leq \pi/2$ . Then  $\Pr_{x \sim D_d}[\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x) \text{ and } |v \cdot x| \geq c_5\alpha] \leq c_4\alpha$ . (3.) There is an absolute constant  $c_6$  such that, for any  $d \geq 4$ , for any two unit vectors  $u$  and  $v$  in  $\mathbf{R}^d$  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  $\mathbf{E}_{x \sim D_{d,w,\gamma}}((a \cdot x)^2) \leq c_8 \log^\lambda(1 + 1/\gamma)(r^2 + \gamma^2)$ . (5.) There is a constant  $c_9$  such that, for all  $\alpha > \sqrt{d}$ , we have  $\Pr_{x \sim D}(\|x\| > \alpha) \leq c_9 \exp(-\alpha/\sqrt{d})$ .

For the case of admissible distributions we have the following theorem, which is proved in the full version [2].

**THEOREM 4.2.** *Let a distribution  $D$  over  $\mathbf{R}^d$  be chosen from a  $\lambda$ -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  $\eta$  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 the full version [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 classifications are changed 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 for Theorem 4.2 with the facts that (a rescaling of) the uniform distribution and i.i.c. distributions are 0-admissible and 2-admissible respectively (see the full version [2]).

## 6. DISCUSSION

Recall that localization is the progressive refinement of the range of possibilities explored by an algorithm as learning proceeds. Localization in the concept space is traditionally used in statistical learning theory both in supervised and active learning for getting sharper rates [15, 16, 36]. Furthermore, the idea of localization in the instance space has been used in margin-based analysis of active learning [4, 8]. 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 (possibly non-geometric) concept spaces. Another concrete open question is to improve the logarithmic dependence on  $\epsilon$  in the noise tolerance for log-concave distributions.

## 7. REFERENCES

- [1] 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.
- [2] P. Awasthi, M.-F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise, 2014. Arxiv, 1307.8371v7.
- [3] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *ICML*, 2006.
- [4] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *COLT*, 2007.
- [5] M.-F. Balcan and V. Feldman. Statistical active learning algorithms. *NIPS*, 2013.
- [6] M.-F. Balcan and S. Hanneke. Robust interactive learning. In *COLT*, 2012.
- [7] M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *COLT*, 2008.
- [8] M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, 2013.
- [9] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.
- [10] E. B. Baum. The perceptron algorithm is fast for nonmalicious distributions. *Neural Computation*, 2:248–260, 1990.
- [11] A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *NIPS*, 2010.
- [12] D. Bienstock and A. Michalka. Polynomial solvability of variants of the trust-region subproblem, 2013. Optimization Online.
- [13] 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.
- [14] A. Blum, M. L. Furst, M. J. Kearns, and R. J. Lipton. Cryptographic primitives based on hard learning problems. In *Proceedings of the 13th Annual International Cryptology Conference on Advances in Cryptology*, 1994.
- [15] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:9:323–375, 2005.
- [16] N. H. Bshouty, Y. Li, and P. M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *JCSS*, 2009.
- [17] R. Castro and R. Nowak. Minimax bounds for active learning. In *COLT*, 2007.
- [18] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2), 1994.
- [19] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.
- [20] S. Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, volume 18, 2005.

- [21] S. Dasgupta. Active learning. *Encyclopedia of Machine Learning*, 2011.
- [22] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. *NIPS*, 20, 2007.
- [23] O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *JMLR*, 2012.
- [24] Y. Freund, H. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.
- [25] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. 1990.
- [26] A. Gupta, M. Hardt, A. Roth, and J. Ullman. Privately releasing conjunctions and the statistical query barrier. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, 2011.
- [27] V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 2006.
- [28] S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.
- [29] S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.
- [30] D. S. Johnson and F. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93 – 107, 1978.
- [31] A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, 2005.
- [32] M. Kearns and M. Li. Learning in the presence of malicious errors. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, 1988.
- [33] M. Kearns and U. Vazirani. *An introduction to computational learning theory*. MIT Press, Cambridge, MA, 1994.
- [34] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Mach. Learn.*, 17(2-3), Nov. 1994.
- [35] A. R. Klivans, P. M. Long, and R. A. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10, 2009.
- [36] V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.
- [37] C. Monteleoni. Efficient algorithms for general active learning. In *Proceedings of the 19th annual conference on Learning Theory*, 2006.
- [38] M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. In *NIPS*, 2011.
- [39] O. 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.
- [40] J. Sturm and S. Zhang. On cones of nonnegative quadratic functions. *Mathematics of Operations Research*, 28:246–267, 2003.
- [41] L. G. Valiant. Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial intelligence*, 1985.
- [42] V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
- [43] L. Wang. Smoothness, Disagreement Coefficient, and the Label Complexity of Agnostic Active Learning. *JMLR*, 2011.
- [44] T. Zhang. Information theoretical upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

## APPENDIX

### A. PROOFS FROM SECTION 3

#### A.1 Parameter choices

For easy reference throughout the proof, we first collect specifications of how parameters of the algorithm of Figure 1 are set. Let  $\tilde{c}_4$  be the value of  $c_4$  that ensures that Equation 2 holds when  $c_3 = \frac{1}{8\pi}$ , and let  $b_k = \frac{\tilde{c}_4}{\sqrt{d}} 2^{-k}$ . Let  $r_k = \pi 2^{-k}$ . Let  $\tilde{c}_2$  be the value of  $c_2$  that ensures that Equation 1 holds when  $c_1 = \tilde{c}_4$ , and let  $\kappa = \frac{1}{8\tilde{c}_4}$ ,  $\tau_k = \frac{\kappa \tilde{c}_2 b_{k-1}}{12}$ ,  $\xi_k = \min\left(\frac{\kappa}{128}, \frac{\kappa^2 \tau_k^2}{2^{14} \tilde{c}_k^2}\right)$ . Finally, let  $\sigma_k^2 = 2\left(\frac{r_k^2}{d-1} + b_{k-1}^2\right)$ .

#### A.2 The outlier removal subroutine

Before taking on the subproblem of analyzing the error within the band, we need to prove the following theorem (which is the same as Theorem 3.4 in the main body) about the outlier removal subroutine of Figure 2.

**THEOREM A.1.** *There is a polynomial  $p$  such that, if  $n \geq p(1/\eta', d, 1/\xi, 1/\delta, 1/\gamma, 1/r)$  examples are drawn from the distribution  $D_{u,\gamma}$  (each replaced with an arbitrary unit-length vector with probability  $\eta' < 1/4$ , for  $\eta' \leq \xi/2$ ), then, with probability  $1 - \delta$ , the output  $q$  of the algorithm in Figure 2 (with  $\sigma^2 = 2(r^2/(d-1) + \gamma^2)$ ) 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  $\|w - u\|_2 \leq r$ ,

$$\frac{1}{|S|} \sum_{x \in S} q(x)(w \cdot x)^2 \leq 2 \left( \frac{r^2}{d-1} + \gamma^2 \right). \quad (3)$$

Furthermore, the algorithm can be implemented in polynomial time.

Our proof of Theorem A.1 proceeds through a series of lemmas. We may assume without loss of generality that  $x_1, \dots, x_n$  are distinct.

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 infeasible 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  $2\eta'$  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  $E[(a \cdot x)^2]$  for all vectors  $a$  close to  $u$ . This is formalized in the following lemma.

**LEMMA A.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  $u = (1, 0, 0, \dots, 0)$ . We can write  $x = (x_1, x_2, \dots, x_d)$  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  $\mathbf{E}_{x \sim U_{u, \gamma}}((a \cdot x)^2)$  into parts that we can analyze separately as follows.

$$\mathbf{E}_{x \sim U_{u, \gamma}}((a \cdot x)^2) = a_1^2 \mathbf{E}_{x \sim U_{u, \gamma}}(x_1^2) + a_1 \sum_{i=2}^n a_i \mathbf{E}_{x \sim U_{u, \gamma}}(x_1 x_i) + \mathbf{E}_{x \sim U_{u, \gamma}}((x' \cdot a)^2). \quad (4)$$

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

$$\mathbf{E}_{x \sim U_{u, \gamma}}((x' \cdot a)^2) \leq \frac{1}{d-1} \sum_{i=2}^d a_i^2 \leq \frac{r^2}{d-1}. \quad (5)$$

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

$$\mathbf{E}_{x \sim U_{u, \gamma}}(x_1^2) \leq \gamma^2. \quad (6)$$

Finally, recalling that  $u = (1, 0, \dots, 0)$ ,  $\mathbf{E}_{x \sim U_{u, \gamma}}(x_1 x_i) = 0$  for all  $i$  (by symmetry). Putting this together with (6), (5) and (4) completes the proof.  $\square$

Next, using VC tools one can show that

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

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

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  $w \in B(u, r)$ ,

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

so that  $q^*$  is feasible.

So what is left is to prove is that a separation oracle for the convex program can be computed in polynomial time. 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

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

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  $\sqrt{q(x)}$ . 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 [40] (see [12]).

### A.3 The error within a band in each iteration

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 chosen by the adversary. We will next show that the fraction of dirty examples in round  $k$  is not too large.

LEMMA A.4. *With probability  $1 - \frac{\delta}{6(k+k^2)}$ ,*

$$|W_D| \leq \frac{2\eta n_k 2^k}{\tilde{c}_2 \tilde{c}_4}. \quad (7)$$

PROOF. From Equation 1 and the setting of our parameters, the probability that an example falls in  $S_{w_{k-1}, b_{k-1}}$  is at least  $2\tilde{c}_2 \tilde{c}_4 2^{-k}$ . 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  $\frac{n_k 2^k}{\tilde{c}_2 \tilde{c}_4}$ . The probability that each unlabeled example we draw is noisy is at most  $\eta$ . Applying a Chernoff bound,  $n_k = \text{poly}(1/\epsilon, 1/\eta, \log(1/\delta))$  suffices to imply that, with probability at least  $1 - \frac{\delta}{12(k+k^2)}$ ,

$$|W_D| \leq \frac{2\eta n_k 2^k}{\tilde{c}_2 \tilde{c}_4}.$$

completing the proof.  $\square$

Note that we may assume without loss of generality that  $\eta < \frac{\epsilon \tilde{c}_2 \tilde{c}_4}{8}$ , in which case Equation 7 implies  $|W_D| \leq |W|/4$ . Let us do that for the rest of the proof.

Recall that the total variation distance between two probability distributions is the maximum difference between the probabilities that they assign to any event. We can think of  $q$  as a soft indicator function for whether an example is kept, 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 easily proved lemma (see the full version [2]).

LEMMA A.5. *The total variation distance between  $p$  and the uniform distribution over  $W$  is at most  $\xi$ .*

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 (which is the same as Lemma 3.5 in the main body). Here  $\ell(w, W_C)$  and  $\ell(w, p)$  are defined with respect to the unrevealed labels that the adversary has committed to.

LEMMA A.6. *Define  $z_k = \sqrt{\frac{r^2}{d-1} + b_{k-1}^2}$ . There are absolute constants  $C_1, C_2$  and  $C_3$  such that, for large enough  $d$ , with probability  $1 - \frac{\delta}{2(k+k^2)}$ , for any  $w \in B(w_{k-1}, r_k)$ , we have*

$$\ell(w, W_C) \leq \ell(w, p) + \frac{C_1 \eta}{\epsilon} \left( 1 + \frac{z_k}{\tau_k} \right) + \kappa/32 \quad (8)$$

and

$$\ell(w, p) \leq 2\ell(w, W_C) + \kappa/32 + \frac{C_2 \eta}{\epsilon} + C_3 \sqrt{\frac{\eta}{\epsilon}} \times \frac{z_k}{\tau_k}. \quad (9)$$

PROOF. Without loss of generality, assume that each element  $(x, y) \in W$  is distinct. Fix an arbitrary  $w \in B(w_{k-1}, r_k)$ . By the guarantee of Theorem A.1, Lemma A.4, and Lemmas A.2 and A.3, we know that, with probability  $1 - \frac{\delta}{2(k+k^2)}$ ,

$$\frac{1}{|W|} \sum_{x \in W} q(x) (w \cdot x)^2 \leq 2z_k^2, \quad (10)$$

together with (for  $C_0 = \frac{2}{\tilde{c}_2 \tilde{c}_4}$ )

$$|W_D| \leq C_0 \eta n_k 2^k \quad (11)$$

and

$$\frac{1}{|W_C|} \sum_{(x,y) \in W_C} (w \cdot x)^2 \leq 2z_k^2. \quad (12)$$

Assume that (10), (11) and (12) all hold.

Since  $\sum_{x \in W} q(x) \geq (1 - \xi_k)|W| \geq |W|/2$ , we have that (10) implies

$$\sum_{x \in W} p(x)(w \cdot x)^2 \leq 4z_k^2. \quad (13)$$

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

$$\sum_{(x,y) \in W_D} p(x)\ell(w, x, y) \leq C_0\eta 2^k + \xi_k + 2\sqrt{C_0\eta 2^k + \xi_k} \times \frac{z_k}{\tau_k}. \quad (14)$$

To see this, notice that,

$$\begin{aligned} \sum_{(x,y) \in W_D} p(x)\ell(w, x, y) &= \sum_{(x,y) \in W_D} p(x) \max\left\{0, 1 - \frac{y(w \cdot x)}{\tau_k}\right\} \\ &\leq \Pr_p(W_D) + \frac{1}{\tau_k} \sum_{(x,y) \in W_D} p(x)|w \cdot x| \\ &= \Pr_p(W_D) + \frac{1}{\tau_k} \sum_{(x,y) \in W} p(x)1_{W_D}(x, y)|w \cdot x|. \end{aligned}$$

Applying the Cauchy-Shwartz inequality we get,

$$\begin{aligned} \sum_{(x,y) \in W_D} p(x)\ell(w, x, y) &\leq \Pr_p(W_D) + \frac{1}{\tau_k} \sqrt{\sum_{(x,y) \in W} p(x)1_{W_D}(x, y)} \sqrt{\sum_{(x,y) \in W} p(x)(w \cdot x)^2} \\ &\leq C_0\eta 2^k + \xi_k + 2\sqrt{C_0\eta 2^k + \xi_k} \frac{z_k}{\tau_k}, \end{aligned}$$

by (11), Lemma A.5 and (13).

Similarly, we show that

$$\sum_{(x,y) \in W} p(x)\ell(w, x, y) \leq 1 + \frac{2z_k}{\tau_k}. \quad (15)$$

Next, we have

$$\begin{aligned} \ell(w, W_C) &= \frac{1}{|W_C|} \sum_{(x,y) \in W} (q(x) + 1_{W_C}(x, y) - q(x))\ell(w, x, y) \\ &\leq \frac{1}{|W_C|} \left( \sum_{(x,y) \in W} q(x)\ell(w, x, y) + \sum_{(x,y) \in W_C} (1 - q(x))\ell(w, x, y) \right). \end{aligned}$$

We also have,

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

by the Cauchy-Shwartz inequality. Recall that  $0 \leq q(x) \leq 1$ , and  $\sum_{x \in W} q(x) \geq (1 - \xi_k)|W|$ . Combining this with (12), we get

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

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  $\xi_k$  small enough that

$$\begin{aligned} \ell(w, W_C) &\leq \frac{1}{|W_C|} \left( \sum_{(x,y) \in W} q(x)\ell(w, x, y) \right) + \kappa/32 \\ &= \frac{\sum_{(x,y) \in W} q(x)}{|W_C|} \left( \sum_{(x,y) \in W} p(x)\ell(w, x, y) \right) + \kappa/32 \\ &= \ell(w, p) + \left( \frac{\sum_{(x,y) \in W} q(x)}{|W_C|} - 1 \right) \left( \sum_{(x,y) \in W} p(x)\ell(w, x, y) \right) + \kappa/32 \\ &\leq \ell(w, p) + \left( \frac{|W|}{|W_C|} - 1 \right) \left( \sum_{(x,y) \in W} p(x)\ell(w, x, y) \right) + \kappa/32 \\ &\leq \ell(w, p) + \left( \frac{|W|}{|W_C|} - 1 \right) \left( 1 + \frac{2z_k}{\tau_k} \right) + \kappa/32 \quad (\text{by (15)}) \\ &= \ell(w, p) + \frac{|W_D|}{|W_C|} \left( 1 + \frac{2z_k}{\tau_k} \right) + \kappa/32. \end{aligned}$$

Applying (11) yields (8).

Also,

$$\begin{aligned} \ell(w, p) &= \sum_{(x,y) \in W} p(x)\ell(w, x, y) \\ &= \sum_{(x,y) \in W_C} p(x)\ell(w, x, y) + \sum_{(x,y) \in W_D} p(x)\ell(w, x, y) \\ &\leq \sum_{(x,y) \in W_C} p(x)\ell(w, x, y) + C_0\eta 2^k + \xi_k + 2\sqrt{C_0\eta 2^k + \xi_k} \frac{z_k}{\tau_k} \quad (\text{by (14)}) \\ &= \frac{\sum_{(x,y) \in W_C} q(x)\ell(w, x, y)}{\sum_{(x,y) \in W_C} q(x)} + C_0\eta 2^k + \xi_k + 2\sqrt{C_0\eta 2^k + \xi_k} \frac{z_k}{\tau_k} \\ &\leq \frac{\sum_{(x,y) \in W_C} \ell(w, x, y)}{\sum_{(x,y) \in W_C} q(x)} + C_0\eta 2^k + \xi_k + 2\sqrt{C_0\eta 2^k + \xi_k} \frac{z_k}{\tau_k} \\ &\leq \frac{\sum_{(x,y) \in W_C} \ell(w, x, y)}{|W_C| - \xi_k|W|} + C_0\eta 2^k + \xi_k + 2\sqrt{C_0\eta 2^k + \xi_k} \frac{z_k}{\tau_k} \\ &\leq 2\ell(w, W_C) + C_0\eta 2^k + \xi_k + 2\sqrt{C_0\eta 2^k + \xi_k} \frac{z_k}{\tau_k}, \end{aligned}$$

by (7), which in turn implies (9).  $\square$