JMLR: Workshop and Conference Proceedings vol 49:1–39, 2016

Learning and 1-bit Compressed Sensing under Asymmetric Noise

Pranjal Awasthi PRANJAL.AWASTHI@RUTGERS.EDU
Rutgers University

Maria-Florina Balcan NINAMF@CS.CMU.EDU
Nika Haghtalab NHAGHTAL@CS.CMU.EDU
Hongyang Zhang HONGYANZ@CS.CMU.EDU
Carnegie Mellon University

Abstract

We study the *approximate recovery* problem: Given corrupted 1-bit measurements of the form $\mathrm{sign}(w^* \cdot x_i)$, recover a vector w that is a good approximation to $w^* \in \mathbb{R}^d$. This problem has been studied by both the learning theory and signal processing communities. In learning theory, this is known as the problem of *learning halfspaces with noise*, and in signal processing, as *1-bit compressed sensing*, in which there is an additional assumption that w^* is t-sparse. The challenge in both cases is to design computationally efficient algorithms that are tolerant to large amounts of noise under realistic noise models. Furthermore, in the case of 1-bit compressed sensing, we require the number of measurements x_i to scale polynomially in t and only polylogarithmically in d, the ambient dimension. In this work, we introduce algorithms with nearly optimal guarantees for both problems under two realistic noise models, *bounded (Massart) noise* and *adversarial (agnostic) noise*, when the measurements x_i's are drawn from any isotropic log-concave distribution.

In bounded (Massart) noise, an adversary can flip the measurement of each point x with probability $\eta(x) \leq \eta < 1/2$. For this problem, we present an efficient algorithm that returns w such that $\|w - w^*\|_2 \leq \epsilon$ in time $\mathrm{poly}(d, \frac{1}{\epsilon})$ for *any* constant $\eta < 1/2$. This improves significantly over the best known result of Awasthi et al. (2015a) in this space that required the noise to be as small as $\eta \approx 10^{-6}$. We then introduce an attribute-efficient variant of this algorithm for 1-bit compressed sensing that achieves the same guarantee with $\mathrm{poly}(t, \log(d), \frac{1}{\epsilon})$ measurements when $\|w^*\|_0 \leq t$. For adversarial (agnostic) noise, where any ν fraction of measurements can be corrupted, we provide an algorithm that returns w such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$, with $\tilde{\Omega}(\frac{t}{\epsilon^3}\mathrm{polylog}(d))$ measurements. Our results improve on the best known approximation results in this space and under some regimes improve on the sample complexity of the existing results. Furthermore, this is the first result of its kind in 1-bit compressed sensing that goes beyond the Gaussian marginal distribution and works for any isotrpic log-concave distribution.

1. Introduction

Linear models are a central object of study in machine learning, statistics, signal processing, and many other domains (Cristianini and Shawe-Taylor, 2000; Freund and Schapire, 1997; Kearns and Vazirani, 1994; Valiant, 1984; Vapnik, 1998). In machine learning and statistics, study of such models has led to significant advances in both the theory and practice of prediction and regression problems. In signal processing, linear models are used to recover sparse signals via a few linear measurements. This is known as compressed sensing or sparse recovery. In both cases, the problem can be stated as approximately recovering a vector $w^* \in \mathbb{R}^d$ given information about $w^* \cdot x_i$, where the x_i's are drawn from a distribution. The feedback typically comes in the form of the value of $w^* \cdot x_i$ or just the sign of the value. The focus of this work is on the latter setting known as classification or 1-bit compressed sensing in the respective communities. That is, given noisy 1-bit measurements of the form $\mathrm{sign}(w^* \cdot x_i)$, how to efficiently recover a vector w that is a good approximation to $w^* \in \mathbb{R}^d$, in terms of the value $\|w - w^*\|_2$. Furthermore, in the context of 1-bit compressed sensing, where w^* is t-sparse, we must use a number of measurements x_i's that scale polynomially in t and only polylogarithmically in d, the ambient dimension.

Despite a large amount of work on linear models, many fundamental questions remain unresolved. In learning theory, one of the long-standing questions is designing efficient noise-tolerant learning algorithms that can approximate the unknown target vector w^* to any arbitrary accuracy. Here noise corresponds to the corruption in the observations $\mathrm{sign}(w^* \cdot x_i)$. In the absence of noise, the recovery problem can be solved efficiently via linear programming. Several other algorithms such as Support Vector Machines (Vapnik, 1998), Perceptron (Minsky and Papert, 1987) and Winnow (Littlestone, 1988) exist that provide better guarantees when the target vector has low L_2 or L_1 norm. This problem becomes more challenging in the context of 1-bit compressed sensing, as in addition to computational efficiency, one has to approximately recover w^* given a number of measurements $\mathrm{poly}(t, log(d))$. In the absence of noise, methods of this type are known only for Gaussian marginal distribution (Gopi et al., 2013; Plan and Vershynin, 2013a,b) or when the data has a large L_1 margin. However, this problem is left open for general distributions even in the absence of noise.

When measurements are noisy, this problem becomes more challenging in both its classification and 1-bit compressed sensing forms. This is due to the fact that direct formulations of the approximate recovery problem are non-convex and are NP-hard to optimize (Guruswami and Raghavendra, 2006). There is significant evidence to indicate that without assumptions on the noise and the distribution of x_i, such recovery might not be computationally possible (Daniely, 2015a; Klivans and Kothari, 2014). When no assumptions are made on the nature of the noise (agnostic model), the best known result shows that when the distribution is uniform over the unit ball, one can achieve an $O(\nu) + \epsilon$ approximation, where ν is the fraction of the noisy labels (Awasthi et al., 2014). An exciting work of Plan and Vershynin (2013a) considers 1-bit compressed sensing under the challenging agnostic noise model and provides the best known result in approximately recovering a t-sparse w^* efficiently with a number of samples $\mathrm{poly}(t \log d)$, albeit with an approximation factor $(11\nu\sqrt{\log \frac{e}{\nu}} + \epsilon\sqrt{\log \frac{e}{\epsilon}})^{1/2}$ that does not match that of its non-sparse counterpart (Awasthi et al., 2014).

Due to the difficulty of the most general form of the problem, most positive results for obtaining arbitrarily good approximation have focused on the case of symmetric noise. A noise process is called *symmetric* if the probability that $\mathrm{sign}(w^* \cdot x_i)$ is corrupted only depends on the magnitude $|w^* \cdot x_i|$ (Plan and Vershynin, 2013a). Symmetric noise has many structural properties that one can exploit. For instance, when samples x_i's are generated from a symmetric distribution, it can be shown that the sign weighted average of the samples is enough to approximate w^*. This is the main insight behind some existing works on classification and 1-bit compressed sensing algorithms that are concerned with symmetric noise, such as (Servedio, 2001; Plan and Vershynin, 2013a). When 1-bit compressed sensing is considered, the more challenging aspect is to show that the number of samples scale linearly with the sparsity of w^*. Even when x_i's are not generated from a "nice" distribution, one can show that the weighted average is not far from w^*[1] However, these observations and techniques break down when the noise is not symmetric.

Our Results: Our work tackles the problem of approximate recovery under highly asymmetric noise and advances the state-of-the-art results in multiple aspects. We first study a natural asymmetric noise model known as the *bounded noise (a.k.a Massart noise)* model. In this model, the probability of corrupting the $\mathrm{sign}(w^* \cdot x_i)$ is upper bounded by a constant $\frac{1}{2} - \frac{\beta}{2}$, i.e., an adversary flips the label of each point x_i with probability $\eta(x_i) \leq \frac{1}{2} - \frac{\beta}{2}$. This is a natural generalization of the well known *random classification noise model* of (Kearns and Vazirani, 1994), where the probability of flipping the label of each example is $\eta = \frac{1}{2} - \frac{\beta}{2}$. Bounded noise model has been widely studied in statistical learning theory (Bousquet et al., 2005) in the context of achieving improved convergence rate. However, except for very simple classes with constant VC dimension, computationally efficient results in this space had remained unknown until recently.[2] We provide the first polynomial time algorithm for approximate recovery to arbitrary accuracy in this model for *any constant noise level*. Our work improves over that of Awasthi et al. (2015a) that required β to be very close to 1 (noise of order 10^{-7}). In this work, we introduce a novel algorithm that goes beyond this value of β and efficiently approximates linear separators to arbitrary accuracy ϵ for any constant value of $\beta > 0$ in time $\mathrm{poly}(d, \frac{1}{\epsilon})$, when the marginal distribution is isotropic log-concave in $\mathbb{R}^d$. We also introduce an attribute-efficient variant of this algorithm and perform 1-bit compressed sensing with number of samples scaling only polynomially in the sparsity parameter and polylogarithmic in the ambient dimension. This is the first such result demonstrating that efficient 1-bit compressed sensing to any desired level of accuracy is possible under highly asymmetric noise. Below, we state our main theorems informally:

Theorems 3 and 8 (informal). *Let $x_1, x_2, \ldots x_m \in \mathbb{R}^d$ be generated i.i.d. from an isotropic log-concave distribution. Let $y_1, y_2, \ldots y_m$ be the corresponding labels generated as $\mathcal{N}_\beta(\mathrm{sign}(w^* \cdot x_i))$, where $\mathcal{N}_\beta$ is the Massart noise process with a constant β. There is an efficient algorithm that for any $\epsilon > 0$, runs in time polynomial in $m, d, \frac{1}{\epsilon}$, and with probability $1 - \delta$ outputs a vector w such that $\|w - w^*\|_2 \leq \epsilon$, provided that $m \geq \mathrm{poly}(d, \frac{1}{\epsilon}, \log(\frac{1}{\delta}))$. Furthermore, if w^* is t-sparse then the algorithm only needs $m \geq \mathrm{poly}(t, \log(d), \frac{1}{\epsilon})$.*

We also consider a more challenging noise model known as *adversarial (a.k.a. agnostic)* noise. Here, no assumptions are made about the nature of the noise and as a result, even information theoretically, approximate recovery within arbitrarily small error is not possible (Kearns and Li, 1988). However, one can still recover w such that $\|w - w^*\|_2 \leq c\nu + \epsilon$, where $\epsilon > 0$ can be arbitrarily small and ν is the fraction of examples that are adversarially corrupted. One would like to keep c as small as possible, ideally a constant[3]. We provide a polynomial time algorithm that can approximately recover w^* in this model with $c = O(1)$ and the dependence on the number of samples is $O(\frac{t}{\epsilon^3}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$. Below, we state our main theorems informally:

[1] This needs additional assumption on the nature of noise. The most widely studied among them is the *random classification* noise model where the sign of each observation is flipped i.i.d. with probability $\eta < \frac{1}{2}$. This can then be boosted in polynomial time to obtain a vector that is arbitrarily close (Blum et al., 1997).

[2] A variant of bounded noise, where the flipping probability for each point is either $\eta(x) = 0$ or $\eta(x) = \eta$ has been also considered as an important open problem in learning theory with the hope that understanding the complexities involved in this type of noise could shed light on the problem of learning disjunctions in the presence of noise (Blum, 2003).

[3] This is the information theoretic limit.

Theorem 11 (informal). *Let $x_1, x_2, \dots x_m \in \mathbb{R}^d$ be generated i.i.d. from an isotropic log-concave distribution. Let w^* be a t-sparse vector and $y_1, y_2, \dots y_m$ be the measurements generated by $\mathcal{N}_{adversarial}(\mathrm{sign}(w^* \cdot x_i))$, where $\mathcal{N}_{adversarial}$ is the adversarial noise process that corrupts a ν fraction of the measurements. There is an efficient algorithm that for any $\epsilon > 0$, runs in time polynomial in $m, d, \frac{1}{\epsilon}$, and with probability $1 - \delta$ outputs a vector w such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$, provided that $m = \Omega(\frac{t}{\epsilon^3}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$ or the number of actively labeled samples is $\Omega(\frac{t}{\epsilon^2}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$.*

1-bit compressed sensing under adversarial noise is also considered under a stronger requirement of *uniformity*, where the approximate recovery guarantee is required to hold with high probability over all sparse signals w^* and all possible corruption of ν fraction of the samples. In other words, in the non-uniform case (Theorem 11) an unknown sparse target vector w^* and a noisy distribution $\tilde{D}$ are fixed in advance before the samples (x_i, y_i) are drawn from $\tilde{D}$, while in the uniform case, the adversary first observes x_i's and then chooses a w^* and noisy labels y_i's. In the uniform case, one typically needs more samples to achieve the same accuracy as in the non-uniform case. In this work, when uniformity is considered our algorithm returns w such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$ when the number of samples is $O(\frac{t}{\epsilon^4}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$.

Theorem 12 (informal). *Let $x_1, x_2, \dots x_m \in \mathbb{R}^d$ be generated i.i.d. from an isotropic log-concave distribution. With probability $1 - \delta$ the following holds. For any signal w^* such that $\|w^*\|_0 \leq t$ and measurements $y_1, y_2, \dots y_m$ generated by $\mathcal{N}_{adversarial}(\mathrm{sign}(w^* \cdot x_i))$, where $\mathcal{N}_{adversarial}$ is the adversarial noise process that corrupts a ν fraction of the measurements, there is an efficient algorithm that for any ϵ, such that $\nu \in O(\epsilon / \log(d/\epsilon)^2)$, runs in time polynomial in $m, d, \frac{1}{\epsilon}$ and outputs a vector w such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$, provided that $m = \Omega(\frac{t}{\epsilon^4}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$.*

Our work on 1-bit compressed sensing provides the first result in non-uniform 1-bit compressed sensing under adversarial noise. Under the uniform case when ν is small, we considerably improve the best known approximation results of Plan and Vershynin (2013a) from $\|w - w^*\|_2 \leq (11\nu\sqrt{\log\frac{e}{\nu}} + \epsilon\sqrt{\log\frac{e}{\epsilon}})^{1/2}$ to $\|w - w^*\|_2 \leq O(\nu) + \epsilon$. Furthermore, we improve the dependence of the sample complexity on ϵ from $\frac{1}{\epsilon^6}$ in the case of the results of Plan and Vershynin (2013a) to $\frac{1}{\epsilon^4}$. While prior work on 1-bit compressed sensing only handles the special case when the distribution is Gaussian, our results hold when the distribution of x_i is any isotropic log-concave distribution.

1.1. Techniques

In this section, we discuss the techniques used for achieving our results.

Iterative Polynomial Regression: Our algorithm follows a localization technique inspired by the work of Balcan et al. (2007). Our algorithm is initialized by a classifier w_0 with a $0/1$ error that is at most an appropriate small constant more than the error of w^* w.r.t. the observed labels. This difference is known as the *excess error*. The algorithm then proceeds in rounds, aiming to cut down the excess error by half in each round. By the properties of bounded noise (Lemma 1) and the log-concave distribution (Lemma 2, Part 2), excess error of a classifier is a linear function of its angle to w^*. Therefore, our algorithm aims to cut the angle by half at each round and eventually will output a w that is close to w^*.

Consider w_{k-1} with angle $\leq \alpha_k$ to w^*. It can be shown that for a band of width $\gamma_{k-1} = \Theta(\alpha_k)$ around the separator w_{k-1}, w_{k-1} makes most of its error in this band. Therefore, improving the accuracy of w_{k-1} in the band significantly improves the accuracy of w_{k-1} overall. When considering vectors that are at angle $\leq \alpha_k$ to w_{k-1}, it can be shown that any vector w_k that achieves *a small enough constant excess error with respect to the distribution in the band*, indeed, enjoys a much stronger guarantee of having *excess error that is half of w_{k-1} overall.* Therefore, if such a vector w_k can be found efficiently in the presence of bounded noise, a classifier of excess error ϵ can be learned in $O(\log(\frac{1}{\epsilon}))$ steps. In order to make the above method work we need to achieve two goals: a) achieve a constant excess error while tolerating noise rate of $\frac{1}{2} - \frac{\beta}{2}$ and b) the hypothesis output should be a halfspace.

On one hand, efficient proper learning methods, such as surrogate loss minimization in the band, readily achieve goal (b). However, convex surrogate loss functions are only a good approximation of the 0/1 loss when the noise is small enough. Since the noise in the band can be as high as $\frac{1}{2} - \frac{\beta}{2}$, this directly restricts the noise rate of bounded noise that can be tolerated with such methods. Indeed, Awasthi et al. (2015a) demonstrated that when hinge-loss minimization is used in the band, such a method only works if the probability of flipping the label is as small as $\approx 10^{-6}$, i.e., when β is very close to 1. On the other hand, the polynomial regression approach of Kalai et al. (2008) learns linear separators to an arbitrary excess error of ϵ with runtime $\mathrm{poly}(d, \exp(\mathrm{poly}(\frac{1}{\epsilon}))$ when the marginal distribution is log-concave, requiring no additional assumption on noise. Since the distribution in the band is also log-concave, this method can achieve *an arbitrarily small constant* excess error in the band thereby achieving goal (a). However, this algorithm outputs the sign of a polynomial $p(\cdot)$ as a hypothesis, which is not necessarily a halfspace.

Instead, our algorithm takes a novel two-step approach to find w_k for *any amount of noise*. This is done by first finding a polynomial p_k that has a small constant excess error in the band. To obtain such a polynomial, we choose $\mathrm{poly}(d, \log(\frac{\log(1/\epsilon)}{\delta}))$ labeled samples from the distribution in the band and use the algorithm by Kalai et al. (2005) to find a polynomial with a small enough but, importantly, *a constant excess error*, e_{KKMS}, in the band. Note that at this point p_k already satisfies goal (a) but it does not satisfy goal (b) as it is not a halfspace. At a high level, since p_k has a small excess error with respect to w^* in the band, using a structural property of bounded noise that connects the excess error and disagreement of a classifier with respect to w^* (Lemma 1), we can show that p_k is also close in classification to w^*. Therefore, it suffices to agnostically learn a halfspace w_k to a constant error for samples in the band that are labeled based on $\mathrm{sign}(p(\cdot))$. To achieve this, we use localized hinge loss minimization in the band over a set of samples that are labeled based on predictions of p_k to find w_k. Therefore, w_k is close in classification to p_k in the band, which is in turn close to w^* in the band. As a result, w_k also has a small error in the band as desired [4].

1-bit compressed sensing: Notice that the techniques mentioned above involve minimizing a convex loss function over a suitably chosen convex set, i.e., the band. When the target vector is sparse, we show that it is enough to perform the minimization task over the set of separators (or polynomials) of small L_1 norm. Since we focus on a smaller candidate set than that of the general case, we can hope to achieve tighter concentration bounds and thus obtain better sample complexity.

Specifically, in the case of Massart noise we extend the polynomial regression algorithm of Kalai et al. (2005) to the sparse case by adding L_1 constraint for polynomials. The target polynomial can then be found using L_1 regression over the convex set of low degree polynomials with small L_1

[4] The recent work of Daniely (2015b) also combines the margin-based approach with polynomial regression. However, in (Daniely, 2015b) polynomial regression is only used once in the end as opposed to the iterative application of polynomial regression used in this work.

norm. To prove the correctness of this algorithm, we show that when w^* is sparse, there exists a low degree polynomial of small L_1 norm that approximates w^*. This is due to the fact that the target polynomial can be represented by a linear combination of sparse Hermite polynomials. To derive the sample complexity, we use a concentration result of Zhang (2002) on the covering number of linear functions of L_1-constrained vectors that satisfy a certain margin property. We analyze such margin property by extending the random thresholding argument of Kalai et al. (2005). The sample complexity of our method follows by combining the two techniques together.

For non-uniform 1-bit compressed in presence of adversarial noise, we build on the algorithm of Awasthi et al. (2014) for learning halfspaces. Similarly as in the previous procedure, this algorithm relies on hinge loss minimization in the band for computing a halfspace of a constant error. However, this algorithm does not use the polynomial regression as an intermediate step, rather, it directly minimizes the hinge loss on a set of points drawn from the noisy distribution in the band. To make this algorithm attribute-efficient, we constrain the hinge loss minimization step to the set of vectors with L_1 norm of at most $\sqrt{t}$. The challenge here is to derive the sample complexity under L_1 constraint. To do this, we use tools from Rademacher theory that exploit the L_1 bound of the linear separators. The improved sample complexity follows from stronger upperbounds on the L_∞ norm of x_i's and the value of hinge loss.

In the uniform case, we build on the techniques described above and show that for a larger number of samples, the analysis would hold *uniformly over all possible noisy measurements on the samples obtained from a choice of sparse w^* and any ν fraction of points corrupted*. First, we show that when the number of samples is $m = \Omega(\frac{t}{\epsilon^4}\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$, then *every band* that could be considered by the algorithm has a sufficient number of samples. This can be proved using covering number and uniform convergence bounds for a class of bands around halfspaces whose L_1 norm is bounded by $\sqrt{t}$. Next, we show that at round k, the empirical hinge loss is concentrated around its expectation uniformly over all choices of w^*, w_{k-1}, and a ν fraction of the samples whose labels differ from the labels of w^*. We note that w_{k-1} is uniquely determined by the labeled samples used by the algorithm in the previous rounds. Therefore, by arguing about the number of possible labelings that can be produced by a sparse w^* and adversarial noise only on the samples that have been used by the algorithm, we can derive a concentration bound that holds uniformly.

2. Related Work

Learning linear classifiers under noise has been extensively studied in the past. One of the noise models considered in the past is the random classification noise (RCN) model (Kearns and Vazirani, 1994). Blum et al. (1997) provided the first polynomial time algorithm capable of learning halfspaces in $\mathbb{R}^d$ to an arbitrary accuracy ϵ under the RCN model. The algorithm works under any data distribution and runs in time polynomial in d, $\frac{1}{\epsilon}$ and $\frac{1}{1-2\eta}$, where $\eta < \frac{1}{2}$ is the probability of flipping a label under the RCN model. A simpler algorithm was later proposed by Dunagan and Vempala (2008). At the other extreme is the agnostic noise model where no assumption is made on the nature of the noise. In other words, the label of each example can be flipped in a completely adversarial fashion (Kearns et al., 1994). The goal is to output a hypothesis of error at most $OPT + \epsilon$, where OPT is the error of the best hypothesis in the class. Kalai et al. (2005) designed an algorithm for learning halfspaces in this model under log-concave distributions. The algorithm relies on a structural result that under log-concave distributions, halfspaces are approximated in L_2 norm to ϵ accuracy, by a polynomial of degree $f(1/\epsilon)$. Here, $f(\cdot)$ is an exponentially growing function.

Hence, by minimizing the absolute loss between the observed labels and a degree $f(1/\epsilon)$ polynomial, one can get arbitrarily close to the error of the best halfspace. Because the analysis does not rely on the existence of a good margin, the algorithm runs in time $d^{f(1/\epsilon)}$ and hence, is efficient only if ϵ is a constant. Shalev-Shwartz et al. (2010) extended the work of Kalai et al. to design a learning algorithm for halfspaces which works for any distribution with a good *margin*. The run time however, still has a mild exponential dependence on $1/\epsilon$. In the agnostic model, algorithms with run time polynomial in d and $\frac{1}{\epsilon}$ are known if one is allowed to output a hypothesis of error multiplicatively worse than OPT. The simple *averaging algorithm* achieves a multiplicative error of $O(\sqrt{\log \frac{1}{OPT}})$ (Kalai et al., 2005). This was improved to an error of $O(OPT)$ using a different algorithm by Awasthi et al. (2014). Later, Daniely (2015b) showed how to get error of $(1+\mu)OPT+\epsilon$ in time inverse exponential in μ. The last two results mentioned above for the agnostic model hold for isotopic log-concave distributions. There are computational lower bounds suggesting that such multiplicative guarantees cannot be obtained under arbitrary distributions (Daniely, 2015a).

A family of interesting noise models lie between the RCN model and the agnostic noise model. The two most popular are the *bounded (a.k.a Massart)* noise model and the more general *Tsybakov noise* model (Tsybakov, 2004; Bousquet et al., 2005). These models can be viewed as *semi-random* adversarial. For instance, in the bounded noise model, an adversary can decide, for each example x, the probability $\eta(x) \leq \frac{1}{2} - \frac{\beta}{2}$ of flipping the label of x. The actual label of x is then generated by flipping a coin of the given bias $\eta(x)$. The computational study of bounded noise in the learning theory community dates back to early 90's with the work of Rivest and Sloan (1994); Sloan (1996) who studied this model under the name *Malicious Misclassification Noise*. However, except for very simple cases, such as intervals on the line or other classes of constant VC-dimension, efficient algorithms in this model had remained unknown until recently. A variant of bounded noise, where the flipping probability for each point is either $\eta(x) = 0$ or $\eta(x) = \eta$ has been considered as an important open problem in learning theory with the hope that understanding the complexities involved in this type of noise could shed light on the problem of learning disjunctions in the presence of noise (Blum, 2003). From the statistical point of view, it is also known that under this models, it is possible to get faster learning rates (Bousquet et al., 2005). However, Computationally efficient algorithms were not known until recently. The recent work of Awasthi et al. (2015a) provide the first evidence that efficiently learning linear separators may be possible by providing an efficient algorithm for learning linear separators under bounded noise provided that β is very close to 1 ($\beta > 1 - 3.6 \times 10^{-6}$) and the marginal distribution is uniform over the unit ball S^d. Our work goes beyond this value of β and efficiently approximates linear separators to arbitrary accuracy ϵ for any constant value of $\beta > 0$ in time $\mathrm{poly}(d, \frac{1}{\epsilon})$, when the marginal distribution is isotropic log-concave in $\mathbb{R}^d$.

Many other noise models are studied in the literature as well. The most popular among them is the linear noise model, where one assumes that the probability of a flipping the label of x is proportional to $|w^* \cdot x|$, where w^* is the optimal classifier. Because of the highly symmetric nature of the noise, efficient algorithms for halfspaces are known under this model (Dekel et al., 2012). The recent work of Feige et al. (2015) studies a noise model where one is allowed to perturb inputs and models the problem as a zero-sum game between a learner, minimizing the expected error, and an adversary, maximizing the expected error.

Attribute-efficient Learning and 1-Bit Compressed Sensing: Attribute-efficient learning, which is learning in the presence of a large number of irrelevant features, was formally introduced by Blum (1990) and Blum and Langley (1997). Mossel et al. (2003) considered this problem applied to learning an arbitrary Boolean function which depends on an unknown set of k out of n Boolean variables, and introduced the first (non-efficient) algorithm with runtime that is better than the naïve algorithm, which tries every k subset of n variables with runtime $O(n^k)$. As for efficient algorithms in this space, some progress has been made for special cases of decision lists (Klivans and Servedio, 2006; Long and Servedio, 2006; Servedio et al., 2012). For halfspaces in the absence of noise, when the distribution has large L_1 margin or Gaussian marginal, efficient algorithms with sample complexity of $\mathrm{poly}(t, \log(d))$ exists (Gopi et al., 2013; Plan and Vershynin, 2013a,b). However, in the general case, the problem of attribute-efficient learning for general distributions is left open even in the absence of noise.

Attribute-efficient learnability of halfspaces under isotropic log-concave distributions has a natural connection to 1-bit compressed sensing (Gopi et al., 2013; Plan and Vershynin, 2013a,b), which was first introduced by Boufounos and Baraniuk (2008). In this framework, one is given information about a t-sparse, unit length vector w^* in $\mathbb{R}^d$ in terms of sign measurements $y_i = \mathrm{sign}(w^* \cdot x_i)$, where x_i is typically a standard random Gaussian in $\mathbb{R}^d$. The goal is to use $m = O(\frac{1}{\epsilon^2} t \log(2d/t))$ measurements to output an approximation w such that $\|w - w^*\|_2 \leq \epsilon$. This problem has received significant attention in recent years, as it is a natural variation of the traditional compressed sensing framework (Donoho, 2006; Candes and Tao, 2006). In the noiseless case, Boufounos and Baraniuk (2008) and Jacques et al. (2013) proposed non-convex models to solve 1-bit compressed sensing problem, while Plan and Vershynin (2013b) further modified these models to a convex one. The 1-bit compressed sensing problem becomes significantly harder under noise. Existing work on 1-bit compressed sensing studies linear noise models, where the probability of flipping a bit of an example x_i is a function of the distance from the target vector, i.e., $|w^* \cdot x_i|$. In a recent work, Plan and Vershynin (2013a) study 1-bit compressed sensing under adversarial noise and find a $(11\nu\sqrt{\log\frac{e}{\nu}}+\epsilon\sqrt{\log\frac{e}{\epsilon}})^{1/2}$ approximate recovery guarantee when the marginal distribution is Gaussian. More recently, Zhang et al. (2014) propose learning 1-bit compressed sensing in an adaptive way with sample complexity $\mathrm{poly}(\frac{1}{\epsilon}, t, \log d)$. However, they use a variant of linear noise model, that is still highly symmetric. Furthermore, they assume oracle access to the flipping probability of each point. In contrast, our work uses a noise model that is highly asymmetric and furthermore, does not require access to such an oracle.

3. Preliminaries

We use X to denote the domain of the samples and Y to denote the label set. In this work X is $\mathbb{R}^d$ and Y is the set $\{+1, -1\}$. We define the *sign* function as $\mathrm{sign}(x) = 1$ if $x \geq 0$ and -1 otherwise. The problem of interest in this paper is approximate recovery: *Given $\epsilon > 0$ and m i.i.d. samples $x_1, x_2, \ldots x_m$ drawn from a distribution over $\mathbb{R}^d$, and labeled as $y_i = \mathcal{N}(\mathrm{sign}(w^* \cdot x_i))$, design a polynomial time algorithm to recover a vector w such that $\|w - w^*\|_2 \leq \epsilon$. Furthermore, if $\|w^*\|_0 = t$, we require m to grow as poly$(t, \log(d), \frac{1}{\epsilon})$.* Here $\mathcal{N}$ is a noise process that corrupts the measurements/labels. We study two asymmetric noise models in this work. The first is $\mathcal{N}_\beta$, the *bounded (a.k.a Massart) noise* model. A joint distribution over (X, Y) satisfies the bounded noise condition with parameter $\beta > 0$, if

$$|\Pr(Y = +1|x) - \Pr(Y = -1|x)| \geq \beta, \, \forall x \in X.$$

In other words, bounded noise is equivalent to the setting where an adversary constructs the distribution by flipping the label of each point x from $\mathrm{sign}(w^* \cdot x)$ to $-\mathrm{sign}(w^* \cdot x)$ with a probability

$\eta(x) \le \frac{1-\beta}{2}$. As is customary, we will use *Bayes optimal classifier* to refer to w^*, the vector generating the uncorrupted measurements. The other noise model that we study is $\mathcal{N}_{adversarial}$, the *adversarial* noise model. Here the adversary can corrupt the labels in any fashion. In this model, the goal of approximate recovery will be to get a vector w such that $\|w - w^*\|_2 \le O(\nu) + \epsilon$, where ν is the fraction of examples corrupted by the adversary.

For any halfspace w, we denote the resulting classifier $h_w = \text{sign}(w \cdot x)$. For any classifier $h : X \mapsto \{+1, -1\}$, we define the error w.r.t. distribution P as $\text{err}_P(h) = \Pr_{(x,y)\sim P}[h(x) \neq y]$. We define the *excess* error of h as $\text{err}_P(h) - \text{err}_P(h_{w^*})$. We use OPT to denote the error of the Bayes classifier, i.e., $\text{err}_P(h_{w^*})$. When the distribution is clear from the context, we use $\text{err}(h_{w^*})$ instead of $\text{err}_P(h_{w^*})$. The next lemma demonstrates an important relation between the excess error of a classifier h and its "closeness" to w^* in terms of classification (or its disagreement). Refer to Appendix A for the proof.

Lemma 1 *Given a classifier $h : X \mapsto \{+1, -1\}$ and distribution P satisfying bounded noise condition with parameter β, let w^* be the Bayes optimal classifier. Then we have*

$$\beta \Pr_{(x,y)\sim P}[h(x) \neq h_{w^*}(x)] \le \text{err}_P(h) - \text{err}_P(h_{w^*}) \le \Pr_{(x,y)\sim P}[h(x) \neq h_{w^*}(x)]. \tag{1}$$

We frequently examine the region within a specified margin of a given halfspace. For distribution P, halfspace w, and margin γ, we denote by $P_{w,\gamma}$ the conditional distribution over the set $S_{w,\gamma} = \{x : |w \cdot x| \le \gamma\}$. We define the *$\tau$-hinge loss* of a halfspace w over an example-label pair (x, y) as $\ell_\tau(w, x, y) = \max\left(0, 1 - \frac{y(w\cdot x)}{\tau}\right)$. When τ is clear from the context, we simply refer to the above quantity as the hinge loss. For a given set T of examples, we use $L_\tau(w, T)$ to denote the empirical hinge loss over the set, i.e., $L_\tau(w, T) = \frac{1}{|T|}\sum_{(x,y)\in T} \ell_\tau(w, x, y)$. For a classifier $w \in \mathbb{R}^d$ and a value r, we use $B(w, r)$ to denote the set $\{v \in \mathbb{R}^d : \|w - v\|_2 \le r\}$. Moreover, for two unit vectors u and v, we use $\theta(u, v) = \arccos(u \cdot v)$ to denote the angle between the two vectors.

In this work, we focus on distributions whose marginal over X is an *isotropic log-concave* distribution. A distribution over d-dimensional vectors $x = \{x_1, x_2, \ldots, x_d\}$ with density function $f(x)$ is log-concave if $\log f(x)$ is concave. In addition, the distribution is isotropic if it is centered at the origin, and its covariance matrix is the identity, i.e., $\mathbb{E}[x_i] = 0$, $\mathbb{E}[{x_i}^2] = 1$, $\forall i$ and $\mathbb{E}[x_i x_j] = 0$, $\forall i \neq j$. Below we state useful properties of such distributions. See (Balcan and Long, 2013; Lovász and Vempala, 2007; Awasthi et al., 2015b) for a proof of Lemma 2.

Lemma 2 *Let P be an isotropic log-concave distribution in $\mathbb{R}^d$. Then there exist absolute constants C_1, C_2 and C_3 such that*

1. *All marginals of P are isotropic log-concave.*
2. *For any two unit vectors u and v in $\mathbb{R}^d$, $C_1\theta(v, u) \le \Pr_{x\sim P}[\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x)]$.*
3. *For any unit vectors w and any γ, $C_3\gamma \le \Pr_{x\sim P}[|w \cdot x| \le \gamma] \le C_2\gamma$.*
4. *For any constant C_4, there exists a constant C_5 such that for two unit vectors u and v in $\mathbb{R}^d$ with $\|u-v\|_2 \le r$ and $\theta(u, v) \le \pi/2$, we have that $\Pr_{x\sim P}[\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x)$ and $|v \cdot x| \ge C_5 r] \le C_4 r$.*
5. *For any constant C_6, there exists another constant C_7, such that for any unit vectors v and u in $\mathbb{R}^d$ such that $\|u - v\|_2 \le r$ and any $\gamma \le C_6$, $\mathbb{E}_{x\sim P_{u,\gamma}}[(v \cdot x)^2] \le C_7(r^2 + \gamma^2)$.*

4. Learning in Presence of Bounded Noise for Any Constant β

In this section, we introduce efficient algorithms for recovering the true classifier in the presence of bounded noise for any constant β. We first consider the non-sparse case and show how our algorithm can return a classifier that is arbitrarily close to w^*. Building on this, we introduce an attribute-efficient variation of this algorithm that is applicable to 1-bit compressed sensing and recovers a t-sparse w^* from few measurements.

4.1. Algorithm for the general case

Here, we describe an efficient algorithm for learning in the presence of bounded noise for any constant β. At a high level, our algorithm proceeds in $\log(\frac{1}{\epsilon})$ rounds and returns a linear separator w_k at round k whose disagreement with respect to w^* is halved at every step. Refer to Appendix C for the procedure that finds an appropriate initial classifier w_0. By induction, consider w_{k-1} whose disagreement with w^* is at most $\Pr[\mathrm{sign}(w^* \cdot x) \neq \mathrm{sign}(w_{k-1} \cdot x)] \leq \frac{\alpha_k}{\pi}$. First, we draw samples from the distribution of points that are at distance at most γ_{k-1} to w_{k-1}. We call this region *the band* at round k and indicate it by $S_{w_{k-1},\gamma_{k-1}}$. Next we apply the polynomial regression algorithm of Kalai et al. (2005) to get a polynomial $p(\cdot)$ of error a constant e_{KKMS} in the band. We draw additional samples from the band, label them based on $\mathrm{sign}(p(\cdot))$, and minimize hinge loss with respect to these labels to get w_k. We then show that w_k that is obtained using this procedure has disagreement at most $\frac{\alpha_{k+1}}{\pi}$ with the target classifier. We can then use w_k as the classifier for the next iteration. The detailed procedure is presented in Algorithm 1. The main result of this section is that Algorithm 1 efficiently learns halfspaces under log-concave distributions in the presence of bounded noise for any constant parameter β that is independent of the dimension. The small excess error implies arbitrarily small approximation rate to the optimal classifier w^* under bounded noise model.

Theorem 3 *Let the optimal Bayes classifier be a halfspace denoted by w^*. Assume that the bounded noise condition holds for some constant $\beta \in (0, 1]$. For any $\epsilon > 0$, $\delta > 0$, there exist absolute constants $e_0, C, C_1, C_2, c_1, c_2$ such that Algorithm 1 with parameters $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = C r_k$, $\lambda = \frac{3C_1}{8CC_2}$, $e_{KKMS} = \beta(\lambda/(4c_1+4c_2+2))^4$, and $\tau_k = \lambda\,\gamma_{k-1}/(4c_1+4c_2+2)$ runs in polynomial time, proceeds in $s = O(\log\frac{1}{\epsilon})$ rounds, where in round k it takes $n_k = \mathrm{poly}(d, \exp(k), \log(\frac{1}{\delta}))$ unlabeled samples and $m_k = \mathrm{poly}(d, \log(s/\delta))$ labels and with probability $1 - \delta$ returns a vector $w \in \mathbb{R}^d$ such that $\|w - w^*\|_2 \leq \epsilon$.*

For the remainder of this paper, we denote by $\tilde{D}$ the noisy distribution and by D the distribution with labels corrected according to w^*. Furthermore, we refer to $\tilde{D}_{w_{k-1},\gamma_{k-1}}$ and $D_{w_{k-1}}, \gamma_{k-1}$, the noisy and clean distributions in the band, by $\tilde{D}_k$ and D_k, respectively.

Proof Outline Here, we provide an outline of the analysis of Algorithm 1 and defer the detailed proof of the Theorem 3 and the related lemmas to Appendix B. Consider a halfspace w_{k-1} at angle α_k to w^* and consider the band of width γ_{k-1} around w_{k-1}. In a log-concave distribution, a $\Theta(\gamma_{k-1})$ fraction of the distribution falls in the band $S_{w_{k-1},\gamma_{k-1}}$ (Property 3). Moreover, the probability that w_{k-1} makes a mistake outside of the band is a small constant fraction of α_k (Property 4). So, w_{k-1} makes most of its mistakes in the band $S_{w_{k-1},\gamma_{k-1}}$. Therefore, if we can find a w_k that has a small (constant) error in the band and, similarly as in w_{k-1}, is close to w^*, then the overall error of w_k is a constant times better than that of w_{k-1}. This is the underlying analysis of the

Algorithm 1 LEARNING HALFSPACES UNDER ARBITRARILY BOUNDED NOISE

```
Input: An initial classifier $w_0$, a sequence of values $\gamma_k$, $\tau_k$ and $r_k$ for $k = 1, \dots, \log(1/\epsilon)$. An error value $e_{\text{KKMS}}$.
1. Let $w_0$ be the initial classifier as describe in Appendix C.
2. For $k = 1, \dots, \log(1/\epsilon) = s$.
    (a) Take $\text{poly}(d, \log(\frac{s}{\delta}))$ labeled samples from $\tilde{D}_k$, the conditional distribution within the band $\{x : |w_{k-1} \cdot x| \leq \gamma_{k-1}\}$, and place them in the set $T$. Run the polynomial regression algorithm (Kalai et al., 2005) over $T$ to find a polynomial $p_k$ such that $\text{err}_{\tilde{D}_k}(\text{sign}(p_k)) \leq \text{err}_{\tilde{D}_k}(h_{w^*}) + e_{\text{KKMS}}$.
    (b) Take $d(d + \log(k/\delta))$ unlabeled samples from $\tilde{D}_k$ and label them according to $\text{sign}(p_k(\cdot))$. Call this set of labeled samples $T'$.
    (c) Find $v_k \in B(w_{k-1}, r_{k-1})$ that approximately minimizes the empirical hinge loss over $T'$ using threshold $\tau_k$, i.e., $L_{\tau_k}(v_k, T') \leq \min_{w \in B(w_{k-1}, r_{k-1})} L_{\tau_k}(w, T') + \frac{\lambda}{12}$.
    (d) Let $w_k = \frac{v_k}{\|v_k\|_2}$.
Output: Return $w_s$, which has excess error $\epsilon$ with probability $1 - \delta$.
```

margin-based technique (Balcan et al., 2007). It suffices to show that w_k, indeed, has a small error rate (of a constant) in the band $S_{w_{k-1}, \gamma_{k-1}}$. That is, $\text{err}_{D_k}(h_{w_k}) \leq \lambda$ for the small constant λ. At each step of the algorithm, we first consider a polynomial $p(\cdot)$ obtained at Step 2a such that $\text{err}(\text{sign}(p(\cdot)) - \text{err}(h_{w^*}) \leq e_{\text{KKMS}}$. Since the distribution in the band is also log concave, we can use the polynomial regression algorithm of Kalai et al. (2005) to find such a polynomial.

Theorem 4 (Kalai et al. (2005)) *Let D be a joint distribution over $X \subseteq \mathbb{R}^d$ and $Y \in \{+1, -1\}$, such that the marginal over X is log-concave. Let OPT be the classification error of the best halfspace w^* w.r.t. D. Then there exists an algorithm which, for any $\epsilon > 0$, outputs a polynomial $p(\cdot)$ such that $\text{err}(\text{sign}(p(\cdot)) \leq \text{err}(h_{w^*}) + \epsilon$. The running time and the number of samples needed by the algorithm is $d^{\exp(1/\epsilon^4)}$.*

Note that, $\text{err}_{D_k}(h_{w_k}) \leq \Pr_{(x,y)\sim D_k}[\text{sign}(p_k(x)) \neq h_{w^*}(x)] + \Pr_{(x,y)\sim D_k}[h_{w_k}(x) \neq \text{sign}(p_k(x))]$. By the relation between the excess error and disagreement of a classifier under bounded noise (Lemma 1), polynomial p is e_{KKMS}/β close in classification to w^*. Therefore, the first part of this inequality is at most e_{KKMS}/β. For the second part of this inequality we need to argue that w_k is close in classification to $p(\cdot)$ inside the band. Recall that at an intuitive level, we choose w_k so as to learn the labels of $p(\cdot)$. For this purpose, we draw a sample from inside the band, label them based on $\text{sign}(p(x))$, and then choose w_k that minimizes the hinge loss over these labels. Since this hinge loss is an upper bound on the disagreement of $p(\cdot)$ and w_k, it suffices to show that it is small. We prove this in Lemma 5, where D'_k denotes the distribution D_k where the labels are predicted based on $\text{sign}(p_k(\cdot))$.

Lemma 5 *There exists an absolute constant c_2 such that with probability $1 - \frac{\delta}{2(k+k^2)}$,*

$$\text{err}_{D'_k}(h_{w_k}) \leq 2\mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)] + 2c_2 \frac{\gamma_{k-1}}{\tau_k} \sqrt{\Pr_{(x,y)\sim D_k}[\text{sign}(p_k(x)) \neq h_{w^*}(x)]} + \frac{\lambda}{2}.$$

We prove this lemma by first showing that the expected hinge loss of w_k on distribution D'_k is at least approximately as good as the expected hinge loss of w^* on D'_k. This is true because w_k minimizes the empirical hinge loss on a large enough sample set from distribution D'_k, so w_k approximately minimizes the *expected hinge* on distribution D'_k (See Lemma 17 for the convergence bound of hinge loss). Next in Lemma 6, the expected hinge of w^* with respect to D' is close to the expected hinge of w^* on the clean distribution D because p and w^* have small disagreement.

Lemma 6 *There exists an absolute constant c_2 such that*

$$|\mathbb{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(w^*,x,y)] - \mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*,x,y)]| \leq c_2 \frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)]}.$$

Finally, we show that hinge loss of w^* on the clean distribution can be upper bounded by the parameters of the algorithm. Together with the result of Lemma 5 this shows that $\mathrm{err}_{D'_k}(w_k) \leq \lambda$ as desired.

Lemma 7 *There exists an absolute constant c_1 such that $\mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*,x,y)] \leq c_1 \frac{\tau_k}{\gamma_{k-1}}$.*

■

4.2. 1-bit Compressed Sensing in Presence of Bounded Noise

We consider the true classifier w^* to be t-sparse and build upon our previous Algorithm 1 to return a vector w such that $\|w - w^*\|_2 \leq \epsilon$, given a number of samples $m \geq \mathrm{poly}(t, \frac{\log(d)}{\epsilon})$. Our main result is the following:

Theorem 8 (Bounded Noise) *Let the optimal Bayes classifier be a halfspace denoted by w^* such that $\|w^*\|_0 = t$. Assume that the bounded noise condition holds for some constant $\beta \in (0,1]$. For any $\epsilon > 0$, $\delta > 0$, there exist absolute constants e_0, C, C_1, C_2, c_1, c_2 such that Algorithm 2 with parameters $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = C r_k$, $\lambda = \frac{3C_1}{8CC_2}$, $e_{KKMS} = \beta(\lambda/(4c_1 + 4c_2 + 2))^4$, and $\tau_k = \lambda\, \gamma_{k-1}/(4c_1 + 4c_2 + 2)$ runs in polynomial time, proceeds in $s = O(\log \frac{1}{\epsilon})$ rounds, where in round k it takes $n_k = \mathrm{poly}(t\log(d), \exp(k), \log(\frac{1}{\delta}))$ unlabeled samples and $m_k = \mathrm{poly}(t, \log(sd/\delta), \exp(k))$ labels and with probability $1-\delta$ returns a vector $w \in \mathbb{R}^d$ such that $\|w - w^*\|_2 \leq \epsilon$.*

Proof Outline The procedure is presented in Algorithm 2. The analysis of correctness of our algorithm is similar to that of Algorithm 1 and Theorem 3, with two exceptions. First, in Step 2a, when we use the polynomial regression of Kalai et al. (2005), we impose an additional constraint that the polynomial $p(\cdot)$ belongs to the convex set $S = \{q : \|q\|_1 = O((\frac{t}{\epsilon})^{\mathrm{poly}(1/e_{\mathrm{KKMS}})})\}$ and $\mathrm{degree}(q) \leq \mathrm{poly}(1/e_{\mathrm{KKMS}})$. Here $\|q\|_1$ is the L_1 norm of the coefficients of q in the monomial basis. This is possible because of the following result about approximation of halfspaces by polynomials:

Theorem 9 (Kalai et al. (2005)) *Let w^* be a halfspace in $\mathbb{R}^d$. Then, for every log-concave distribution D over $\mathbb{R}^d$, there exists a degree $\frac{1}{\epsilon^2}$ polynomial $p(\cdot)$ such that $\mathbb{E}_{x\sim D}[(p(x) - \mathrm{sign}(w^* \cdot x))^2] \leq \epsilon$.*

For the purpose of our algorithm, we need error of at most e_{KKMS} in the band. Here, we define a polynomial $p(\cdot)$ to be t-sparse if $p(\cdot)$ is supported on at most t monomials. If w^* is a t-sparse halfspace, then the $\frac{1}{e^2_{\mathrm{KKMS}}}$-degree polynomial $p(\cdot)$ will be $t^{1/e^2_{\mathrm{KKMS}}}$-sparse. This is due to the fact that the log-concavity of the distribution is preserved when considering the projection of the instance

Algorithm 2 LEARNING SPARSE HALFSPACES UNDER ARBITRARILY BOUNDED NOISE

```
Input: An initial classifier $w_0$, a sequence of values $\gamma_k, \tau_k$ and $r_k$ for $k = 1, \dots, \log(1/\epsilon)$. An error value $e_{\text{KKMS}}$.
1. Let $w_0$ be the initial classifier.
2. For $k = 1, \dots, \log(1/\epsilon) = s$.
    (a) Take $\text{poly}(\frac{t}{\gamma_k}, \log(\frac{ds}{\delta}))$ labeled samples from $\tilde{D}_k$, the conditional distribution within the band $\{x : |w_{k-1} \cdot x| \leq \gamma_{k-1}\}$, and place them in the set $T$. Run the polynomial regression algorithm (Kalai et al., 2005) over $T$ to find a polynomial $p_k$ such that $\text{err}_{\tilde{D}_k}(\text{sign}(p_k)) \leq \text{err}_{\tilde{D}_k}(h_{w^*}) + e_{\text{KKMS}}$ and $\|p\|_1 = O((\frac{t}{\epsilon})^{\text{poly}(1/e_{\text{KKMS}})})$.
    (b) Take $m_k = \Omega(\frac{t}{\tau_k^2}\text{polylog}(d, \frac{1}{\delta}, \frac{1}{\epsilon}))$ unlabeled samples from $\tilde{D}_k$ and label them according to $\text{sign}(p_k(\cdot))$. Call this set of labeled samples $T'$.
    (c) Find $v_k \in B(w_{k-1}, r_{k-1})$ such that $\|v_k\|_1 \leq \sqrt{t}$ and $v_k$ approximately minimizes the empirical hinge loss over $T'$ using threshold $\tau_k$, i.e., $L_{\tau_k}(v_k, T') \leq \min_{w \in B(w_{k-1}, r_{k-1}) \text{ and } ||w||_1 \leq \sqrt{t}} L_{\tau_k}(w, T') + \frac{\lambda}{12}$.
    (d) Let $w_k = \frac{v_k}{\|v_k\|_2}$.
Output: Return $w_s$, which has excess error $\epsilon$ with probability $1 - \delta$.
```

space on the relevant t variable. Since there are only t^{1/e_{KKMS}^2} monomials in the lower dimension, the $\frac{1}{e_{\text{KKMS}}^2}$-degree polynomial $p(\cdot)$ that satisfies the theorem in this lower dimension also satisfies the requirement in the original space and is t^{1/e_{KKMS}^2}-sparse. The analysis of Kalai et al. (2005) also shows that the polynomial $p(\cdot)$ is a linear combination of normalized Hermite polynomials up to degree $deg = \frac{1}{e_{\text{KKMS}}^2}$, $\sum_{i=0}^{deg} c_i \bar{H}_i$, where $\sum_{i=0}^{deg} c_i^2 < 1$. Since $\bar{H}_i$s are normalized, we have that $\|p\|_1 \leq O(t^{1/e_{\text{KKMS}}^2})$. This bound holds when D is an isotropic log-concave distribution. The distribution we consider at Step k of the algorithm is a conditional of isotropic log-concave distribution over $\{x : |w \cdot x| \leq \gamma_k\}$ and as a result is not isotropic. For such distributions, the coefficients of p blow up by a factor of $O((\frac{1}{\gamma_k})^{poly(1/e_{\text{KKMS}})})$. Since $\gamma_k = \exp(k) \geq \epsilon$ for all k, we will get a bound of $\|p\|_1 \leq O((\frac{t}{\epsilon})^{poly(1/e_{\text{KKMS}})})$ at each time step. Thus, by enforcing that the polynomial $p(\cdot)$ belongs to S, we only need to argue about polynomials in the set S as opposed to general $\text{poly}(\frac{1}{e_{\text{KKMS}}})$-degree polynomials.

Second, in Step 2c, we minimize hinge loss with respect to the polynomial $p(\cdot)$ under an additional constraint that the resulting linear separator has L_1 norm bounded by $\sqrt{t}$. Here again, by the analysis from the previous section, we know that the induced polynomial $p(\cdot)$ is $O(e_{\text{KKMS}}/\beta)$-close to a t-sparse polynomial w^*. Hence, when looking for a separator v_k, we can safely limit our search to linear separators with L_1 norm bounded $\sqrt{t}$ because of the fact $\|w^*\|_1 \leq \sqrt{t}\|w^*\|_2 = \sqrt{t}$. This shows that the algorithm will indeed output a halfspace of error at most $OPT + \epsilon$.

Next, we need to argue about the sample complexity of the algorithm. The sample complexity is dominated by the polynomial regression step. Notice, that during hinge loss minimization, we are generating samples labeled by $p(\cdot)$ and hence they do not contribute to the sample complexity. In order to argue about the polynomial regression algorithm, we prove the following extension of the result of Kalai et al. (2005).

Theorem 10 *Let (X, Y) be drawn from a distribution over $\mathbb{R}^d \times \{+1, -1\}$ with isotropic log-concave marginal over X. Let OPT be the error of the best t-sparse halfspace, i.e., $OPT =$*

$\min\limits_{w\in\mathbb{R}^d, \|w\|_0\le t} \Pr_{(x,y)}[\mathrm{sign}(w\cdot x) \neq y]$. *Then, for every* $\epsilon > 0$, *there is an algorithm that runs in time* $d^{\mathrm{poly}(\frac{1}{\epsilon})}$ *and uses a number of samples* $m = t^{\mathrm{poly}(\frac{1}{\epsilon})}\mathrm{polylog}(d)$ *from the distribution and outputs a polynomial* $p(\cdot)$ *such that* $\mathrm{err}(p) \le OPT + \epsilon$. *Here,* $\mathrm{err}(p) = \Pr_{(x,y)}[\mathrm{sign}(p(x)) \neq y]$. *Furthermore, the polynomial* $p(\cdot)$ *satisfies* $\|p\|_1 \le t^{\mathrm{poly}(\frac{1}{\epsilon})}$.

In order to prove the theorem above, we follow the same outline as in (Kalai et al., 2005). The proof of correctness follows immediately using Theorem 9 above. See Appendix D for the details. In order to argue about sample complexity, we need to argue that for every polynomial $q \in S$, the empirical 0/1 error with margin γ is within $\epsilon/4$ of the expected 0/1 error. In order to do this, we use a concentration result of Zhang (2002) (Lemma 18) on the covering number of linear functions of L_1-constrained vectors that satisfy a certain margin property. We analyze such margin property by extending the random thresholding argument of Kalai et al. (2005). This leads to the desired sample complexity. We defer the details of this analysis to Appendix D. ■

5. 1-bit Compressed Sensing in Presence of Adversarial Noise

In this section, we first consider 1-bit compressed sensing of linear separators under adversarial noise. In this noise model, the adversary can choose any distribution $\tilde{D}$ over $\mathbb{R}^d \times \{+1, -1\}$ such that the marginal over $\mathbb{R}^d$ is unchanged but a ν fraction of the labels are flipped adversarially. We introduce an attribute-efficient variant of the algorithm of Awasthi et al. (2014) for noise-tolerant learning that given $O(t\ \mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})/\epsilon^3)$ samples from a given $\tilde{D}$ distribution, with probability $1-\delta$ returns a vector w, such that $\|w - w^*\|_2 \le O(\nu) + \epsilon$. To the best of our knowledge this is the first result in non-uniform 1-bit compressed sensing under adversarial noise. Furthermore, the approximation factor of this result almost matches the information theoretic bound.

Theorem 11 (Adversarial Noise – Non-uniform) *Assume that the noise is adversarial and let the optimal linear classifier be a halfspace denoted by* w^* *such that* $\|w^*\|_0 = t$. *Let* $\nu > 0$ *be the error of* w^*. *For any* $\epsilon > 0$, $\delta > 0$, *there exist absolute constants* $e_0, C, C_1, C_2, c_1, c_2$ *such that Algorithm 3 with parameters* $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = C r_k$, $\lambda = \frac{3C_1}{8CC_2}$, *and* $\tau_k = \lambda\ \gamma_{k-1}/(4c_1 + 4c_2 + 2)$ *runs in polynomial time, proceeds in* $s = \log\frac{1}{\epsilon}$ *rounds, where in round* k *it takes* $n_k = \mathrm{poly}(t, \log(d), \exp(k), \log(\frac{1}{\delta}))$ *unlabeled samples and* $m_k = O(t\ \mathrm{polylog}(sd/\delta)2^{2k})$ *labels and with probability* $1-\delta$ *returns a vector* $w \in \mathbb{R}^d$ *such that* $\|w - w^*\|_2 \le O(\nu) + \epsilon$. *That is the total number of unlabeled samples is* $m = O(\frac{t}{\epsilon^3}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$ *and at every round* $m_k \le O(\frac{t}{\epsilon^2}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$ *labels are requested.*

We also consider the stronger requirements of *uniform* 1-bit compressed sensing. In this setting, we show that given $O(t\ \mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})/\epsilon^4)$ samples x_i, with probability $1-\delta$, *uniformly over all possible noisy measurements on* x_i*'s obtained from a choice of sparse* w^* *and any* ν *fraction of measurements corrupted*, the algorithm returns a vector w such that $\|w - w^*\|_2 \le O(\nu) + \epsilon$, when ν is small with respect to ϵ. When ν is small, this result considerably improves the best known approximation results of Plan and Vershynin (2013a) from $\|w - w^*\|_2 \le (11\nu\sqrt{\log\frac{e}{\nu}} + \epsilon\sqrt{\log\frac{e}{\epsilon}})^{1/2}$ to $\|w - w^*\|_2 \le O(\nu) + \epsilon$. Furthermore, we improve the dependence of the sample complexity on ϵ from $\frac{1}{\epsilon^6}$ in the case of the results of Plan and Vershynin (2013a) to $\frac{1}{\epsilon^4}$[5]. Our result for this setting is as follows.

Theorem 12 (Adversarial Noise – uniform) *Let $x_1, x_2, \ldots x_m \in \mathbb{R}^d$ be drawn i.i.d. from an isotropic log-concave distribution. With probability $1-\delta$ the following holds. For all signals w^* such that $\|w^*\|_0 \le t$ and measurements $y_1, y_2, \ldots y_m$ generated by $\mathcal{N}_{adversarial}(\mathrm{sign}(w^* \cdot x_i))$, where $\mathcal{N}_{adversarial}$ is the adversarial noise process that corrupts a ν fraction of the measurements, and for any ϵ such that $\nu \in O(\epsilon/\log(d/\epsilon)^2)$, there exist absolute constants e_0, C, C_1, C_2, c_1, c_2 such that Algorithm 5 with parameters $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = Cr_k$, $\lambda = \frac{3C_1}{8CC_2}$, and $\tau_k = \lambda\, \gamma_{k-1}/(4c_1 + 4c_2 + 2)$ runs in time poly$(d, \frac{1}{\epsilon})$ and returns a vector $w \in \mathbb{R}^d$ such that $\|w - w^*\|_2 \le O(\nu) + \epsilon$ if $m = \Omega(\frac{t}{\epsilon^4}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})$. Furthermore, the number of labeled samples at every round k is $m_k \le O(\frac{t}{\epsilon^3}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})$.*

We build on the algorithm of Awasthi et al. (2014) for learning halfspaces under adversarial noise. Much like our procedure for bounded noise, this algorithm also relies on hinge loss minimization in the band for computing a halfspace of a constant error. However, this algorithm does not make use of polynomial regression as an intermediate step, rather, it directly minimizes the hinge loss on a labeled set of points drawn from the noisy distribution in the band, $\tilde{D}_k$.

To make this algorithm attribute-efficient, we constrain the hinge loss minimization step to the set of vectors with L_1 norm of at most $\sqrt{t}$. See Algorithm 3. Since w^* is t-sparse, $\|w^*\|_1 \le \sqrt{t}$ and therefore the comparison between the outcome of every step and w^* remains valid. This shows that such a change preserves the correctness of the algorithm. We prove the correctness of this algorithm and its sample complexity in Appendix E.

Algorithm 3 NON-UNIFORM 1-BIT COMPRESSED SENSING UNDER ADVERSARIAL NOISE

```
Input: An initial classifier $w_0$, a sequence of values $\gamma_k, \tau_k$ and $r_k$ for $k = 1, \ldots, \log(1/\epsilon)$.
1. Let $w_0$ be the initial classifier.
2. For $k = 1, \ldots, \log(1/\epsilon) = s$.
   (a) Take $m_k = \Omega(\frac{t}{\epsilon^2}\mathrm{polylog}(d, \frac{1}{\delta}, \frac{1}{\epsilon}))$ samples from $\tilde{D}_k$ and request the labels. Call this set $T'$.
   (b) Find $v_k \in B(w_{k-1}, r_{k-1})$ such that $\|v_k\|_1 \le \sqrt{t}$ and $v_k$ approximately minimizes the empirical hinge loss over $T'$ using threshold $\tau_k$, that is, $L_{\tau_k}(v_k, T') \le \min_{w \in B(w_{k-1}, r_{k-1}) \text{ and } ||w||_1 \le \sqrt{t}} L_{\tau_k}(w, T') + \frac{\lambda}{12}$.
   (c) Let $w_k = \frac{v_k}{\|v_k\|_2}$.
Output: Return $w_s$, which has excess error $O(\nu) + \epsilon$ with probability $1 - \delta$.
```

Let us briefly discuss the sample and label complexity of this approach. Labeled samples are only needed in Step 2a of the algorithm for the purpose of minimizing the hinge loss. The crux of the argument is to show that when the number of labeled samples in the band is large enough, the empirical hinge loss of w_k and w^* in the band is close to their expectation. To prove this we use the following tool from VC dimension and Rademacher complexity theory that show that linear functions (such as hinge loss) of vectors with L_1 norm bounded by $\sqrt{t}$ are nicely concentrated

[5] The sample complexity of the method of Plan and Vershynin (2013a) is expressed as $O(t \text{ polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})/\epsilon^6)$ for achieving error $\left(11\nu\sqrt{\log \frac{e}{\nu}} + \epsilon\sqrt{\log \frac{e}{\epsilon}}\right)^{1/2}$. When ν is small compared to ϵ, this in fact compares to a method with sample complexity $\frac{1}{\epsilon^{12}}$ for achieving excess error $\tilde{O}(\epsilon)$ which is the regime we work in.

around their expectation within a gap that is determined by the smoothness of the function, $\|x_i\|_\infty$, and the hinge loss on any sample (x, y).

Lemma 13 (Shalev-Shwartz and Ben-David (2014)) *Let $\mathcal{F}$ be the class of linear predictors with the L_1 norm of the weights bounded by W_1. Assume that the infinity norm of all instances is bounded by X_∞. Then for the ρ-Lipschitz loss ℓ such that $\max_{w\cdot x\in[-W_1X_\infty, W_1X_\infty]} |\ell(w, x, y)| \leq U$ and the choice of an i.i.d. sample T of size m,*

$$\forall w,\ s.t.,\ \|w\|_1 \leq W_1,\ \Pr\left[|\mathbb{E}\ell(w, x, y) - \ell(w, T)| \geq 2\rho W_1 X_\infty \sqrt{\frac{2\log(2d)}{m}} + s\right] \leq 2\exp\left(-\frac{ms^2}{2U^2}\right).$$

Note that the smoothness of hinge loss functions $\ell_{\tau_k}(w, x, y) = \max(0, 1 - y(w\cdot x)/\tau_k)$, is given by $\tau_k < \frac{1}{\epsilon}$. Furthermore, for $x \sim D$, each coordinate of x represent a 1-dimensional isotropic log-concave distribution and as a result is concentrated around 0. Therefore, with high probability $\|x\|_\infty$, which is the maximum of d draws from one-dimensional isotropic log-concave distributions, is at most $\log(d)$ (See Lemma 20). As for the value of hinge loss, it can be represented as

$$\ell_{\tau_k}(w, x, y) \leq 1 + \frac{|w\cdot x|}{\tau_k} \leq 1 + \frac{|w_k\cdot x|}{\tau_k} + \frac{|(w - w_k)\cdot x|}{\tau_k}.$$

Note that for $x \sim \tilde{D}_k$, $w_k\cdot x \leq \gamma_k$ and therefore, $\frac{|w_k\cdot x|}{\tau_k} \leq O(1)$. For the second part of the inequality, since $\|w_k - w\|_2 \leq r_k$ and $(w_k - w)\cdot x$ is a one-dimensional log-concave distribution, we can show that with high probability $|(w_k - w)\cdot x| \leq r_k\mathrm{polylog}(d, \frac{1}{\epsilon})$. Therefore, $\frac{|(w-w_k)\cdot x|}{\tau_k} \leq \mathrm{polylog}(d, \frac{1}{\epsilon})$ (See Lemma 21). Using these bounds together with Lemma 13 immediately implies the sample complexity of our algorithm. See Appendix E for details of the omitted proofs.

For uniform 1-bit compressed sensing, Our algorithm is similar to the case of non-uniform 1-bit compressed sensing with the exception of requiring more samples (See Algorithm 5). We build on the analysis of the non-uniform case and show that for a larger number of samples, the analysis would hold *uniformly over all possible noisy measurements on the samples obtained from any choice of sparse w^* and any ν fraction of measurements corrupted.*

First, we show that when the number of samples is $m = \Omega(\frac{t}{\epsilon^4}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$, then *every band* around a halfspace that can be considered by the algorithm as a candidate for w_k for $k \leq \log(\frac{1}{\epsilon})$ has sufficient samples (See Lemma 26). This fact follows from the analysis of covering number (or VC dimension) of bands. Since bands are simple structures such as intersection of two halfsapces with bounded L_1, it can be shown that the covering number of the class of all bands around vectors with L_1 bounded by $O(\sqrt{t})$ is small. Using uniform convergence bounds for covering numbers, we see that *every band* that can be considered by the algorithm has sufficient number of samples.

Next, we build up on the concentration results from the previous section and show that the hinge loss is concentrated around its expectation uniformly over all choice of w^* and ν fraction of the samples whose labels differ from the labels of w^*. The proof of this claim follows from applying a union bound over all possible labelings produced by a sparse w^* and adversarial noise. The proof of Theorem 12 follows similarly as in Theorem 11, using this new concentration result. See Appendix E.1 for more details.

6. Lower Bound under Bounded Noise

In this section, we show that one-shot minimization does not work for a large family of loss functions that include any continuous loss with a natural property that points at the same distance from the separator have the same loss. This generalizes the result of Awasthi et al. (2015a) who showed that one-shot minimization of hinge loss does not lead to an arbitrarily small 0/1 error even under bounded noise with small flipping probability, and justifies why minimizing a sequence of carefully designed losses, as we did in the last few sections, is indispensable to achieving an arbitrarily small excess error.

Without loss of generality, we discuss the lower bound in $\mathbb{R}^2$. Formally, let $\mathcal{P}_\beta$ be the class of noisy distribution $\tilde{D}$ with uniform marginal over the unit ball, and let (z_w, φ_w) represent the polar coordinate of a point P in the instance space, where φ_w represents the angle between the linear separator h_w and the vector from origin to P, and z_w is the L_2 distance of the point P and the origin. Let $\ell_+^w(z_w, \varphi_w)$ and $\ell_-^w(z_w, \varphi_w)$ denote the loss functions on point P with correct and incorrect classification by h_w, respectively. The loss functions we study here satisfy the following properties.

Definition 14 *Continuous loss functions $\ell_+^w(z_w, \varphi_w)$ and $\ell_-^w(z_w, \varphi_w)$ are called* proper, *if and only if*

1. *$\ell_+^w(z_w, \varphi_w) = \ell_+^w(z_w, k\pi \pm \varphi_w)$ and $\ell_-^w(z_w, \varphi_w) = \ell_-^w(z_w, k\pi \pm \varphi_w)$, for $k \in N$;*
2. *For $z_w > 0$, $\ell_-^w(z_w, \varphi_w) \geq \ell_+^w(z_w, \varphi_w)$; The equality holds if and only if $\varphi_w = k\pi$, $\forall k \in N$.*

Figure 1(*a*) in the Appendix is a visualization of Property 1 in Definition 14, which states that the loss $\ell_+^w(z_w, \varphi_w)$ (or $\ell_-^w(z_w, \varphi_w)$) on the points with the same angle to the separator (indicated by points of the same color) are the same. Note that all losses that are functions of the distance to the classifier, e.g. the hinge-loss and logistic loss, etc., satisfy Property 1, since the distance of a point to classifier w is $|z_w \sin \varphi_w| = |z_w \sin(k\pi \pm \varphi_w)|$. However, Property 1 only requires the symmetry of the loss w.r.t. the linear separator, and is not limited to distance-based losses, that is, the losses on the points with the same distance can be different (See Figure 1(*a*) for the red and light blue points). Moreover, this property does not require the loss to be monotonically increasing in the distance. Property 2 is a very natural assumption since to achieve low error, it is desirable to penalize misclassification more. Note that we equally penalize correct and incorrect classifications if and only if points lie exactly on the linear separator.

In fact, most of the commonly used loss functions (Bartlett et al., 2006) satisfy our two properties in Definition 14, e.g., the (normalized) hinge loss, logistic loss, square loss, exponential loss, and truncated quadratic loss (See Figure 2 and Table 1 for more details), because they are all functions of the distance to classifier. Furthermore, we highlight that Definition 14 covers the loss even with regularized term on w. A concrete example is 1-bit compressed sensing, with loss function formulated as $\ell_+(z_w, \varphi_w) = -|z_w \sin \varphi_w| + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2$ and $\ell_-(z_w, \varphi_w) = |z_w \sin \varphi_w| + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2$. Thus our lower bound demonstrates that one-shot 1-bit compressed sensing cannot always achieve arbitrarily small excess error under the Massart noise.

Our lower bound for any proper function is stated as follows. See Appendix F for details.

Theorem 15 *For every bounded noise parameter $0 \leq \beta < 1$, there exists a distribution $\tilde{D}_\beta \in \mathcal{P}_\beta$ (that is, a distribution over $\mathbb{R}^2 \times \{+1, -1\}$, where the marginal distribution on $\mathbb{R}^2$ is uniform over the unit ball, and the labels $\{+1, -1\}$ satisfies bounded noise condition with parameter β) such that any proper loss minimization is not consistent on $\tilde{D}_\beta$ w.r.t. the class of halfspaces. That is, there exists an $\epsilon \geq 0$ and a sample size $m(\epsilon)$ such that any proper loss minimization will output a classifier of excess error larger than ϵ by a high probability over sample size at least $m(\epsilon)$.*

7. Conclusion

Our work improves the state of the art results on classification and 1-bit compressed in presence of asymmetric noise in multiple directions. For the general non-sparse case, our work provides the first algorithm for finding a halfspace that is arbitrarily close to w^* in presence of bounded noise for any constant maximum flipping probability. Our analysis and algorithm combine the strengths of two algorithms that are individually insufficient: polynomial regression, with runtime that is inverse exponential in the required accuracy, and margin-based localization technique that only achieves a multiplicative approximation to the optimum. We show how using ideas from the localization technique helps us boost the performance of polynomial regression method. That is, by applying polynomial regression, which only guarantees a *constant* excess error in polynomial time, iteratively on the conditional distributions within the margin of the previous classifiers, we can achieve an *arbitrarily small* excess error *while maintaining computational efficiency*. It would be interesting to see if similar ideas can be applied to more general decision boundaries.

Furthermore, we extend the margin-based platform used for approximate recovery in presence of bounded or adversarial noise to an attribute-efficient algorithm. Our work improves on the best known result of Plan and Vershynin (2013a) on 1-bit compressed sensing in presence of adversarial noise, and achieves an improved approximation factor while allowing broader class of distribution. We also improve on the sample complexity of existing results when ν is small. Our hope is that this first application of the margin-based technique to compressed sensing will lead to improved results for wider class of problems in compressed sensing.

Acknowledgments This work was supported in part by grants NSF-CCF 1535967, NSF CCF-1422910, NSF CCF-145117, a Sloan Fellowship, and an IBM Ph.D. fellowship.

References

Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, pages 449–458. ACM, 2014.

Pranjal Awasthi, Maria Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, 2015a.

Pranjal Awasthi, Maria Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. *Submitted to Journal of the ACM*, 2015b.

Maria Florina Balcan and Phillip M. Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, 2013.

Maria Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT)*, 2007.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Avrim Blum. Learning boolean functions in an infinite attribute space. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing (STOC)*, pages 64–72. ACM, 1990.

Avrim Blum. Machine learning: a tour through some favorite results, directions, and open problems. *FOCS 2003 tutorial slides, available at http://www.cs.cmu.edu/~avrim/Talks/FOCS03/tutorial.ppt*, 2003.

Avrim Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1):245–271, 1997.

Avrim Blum, Alan M. Frieze, Ravi Kannan, and Santosh Vempala. A polynomial time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1997.

Petros T Boufounos and Richard G Baraniuk. 1-bit compressive sensing. In *Proceedings of the 42nd Annual Conference on Information Sciences and Systems (CISS)*, pages 16–21. IEEE, 2008.

Olivier Bousquet, Stéphane Boucheron, and Gabor Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:9:323–375, 2005.

Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

Amit Daniely. Complexity theoretic limitations on learning halfspaces. *CoRR*, abs/1505.05800, 2015a.

Amit Daniely. A PTAS for agnostically learning halfspaces. In *Proceedings of the 28th Annual Conference on Learning Theory*, 2015b.

Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Selective sampling and active learning from single and multiple teachers. *The Journal of Machine Learning Research*, 13(1):2655–2697, 2012.

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

John Dunagan and Santosh Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. *Mathematical Programming*, 114(1):101–114, 2008.

Uriel Feige, Yishay Mansour, and Robert Schapire. Learning and inference in the presence of corrupted inputs. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, pages 637–657, 2015.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Sivakant Gopi, Praneeth Netrapalli, Prateek Jain, and Aditya Nori. One-bit compressed sensing: Provable support and vector recovery. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 154–162, 2013.

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

Laurent Jacques, Jason N Laska, Petros T Boufounos, and Richard G Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013.

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 (FOCS)*, 2005.

Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. On agnostic boosting and parity learning. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 629–638. ACM, 2008.

Michael Kearns and Ming Li. Learning in the presence of malicious errors. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing (STOC)*, 1988.

Michael Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3), November 1994.

Michael J Kearns and Umesh Virkumar Vazirani. *An introduction to computational learning theory*. MIT Press, Cambridge, MA, 1994.

Adam Klivans and Pravesh Kothari. Embedding hard learning problems into gaussian space. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*, 28:793–809, 2014.

Adam R Klivans and Rocco A Servedio. Toward attribute efficient learning of decision lists and parities. *The Journal of Machine Learning Research*, 7:587–602, 2006.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Philip M Long and Rocco Servedio. Attribute-efficient learning of decision lists and linear threshold functions under unconcentrated distributions. In *Advances in Neural Information Processing Systems (NIPS)*, pages 921–928, 2006.

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

Marvin L Minsky and Seymour A Papert. *Perceptrons - Expanded Edition: An Introduction to Computational Geometry*. MIT press Boston, MA:, 1987.

Elchanan Mossel, Ryan O'Donnell, and Rocco P Servedio. Learning juntas. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC)*, pages 206–212. ACM, 2003.

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

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

Ronald L Rivest and Robert Sloan. A formal model of hierarchical concept-learning. *Information and Computation*, 114(1):88–114, 1994.

Rocco A Servedio, Li-Yang Tan, and Justin Thaler. Attribute-efficient learning and weight-degree tradeoffs for polynomial threshold functions. In *COLT*, volume 23, pages 14–1, 2012.

Rocco Anthony Servedio. *Efficient algorithms in computational learning theory*. Harvard University, 2001.

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

Shai Shalev-Shwartz, Ohad Shamir, and Karthik Sridharan. Learning kernel-based halfspaces with the zero-one loss. *arXiv preprint arXiv:1005.3681*, 2010.

Robert H Sloan. PAC learning, noise, and geometry. In *Learning and Geometry: Computational Approaches*, pages 21–41. Springer, 1996.

Alexandre B Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166, 2004.

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

Vladimir Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Lijun Zhang, Jinfeng Yi, and Rong Jin. Efficient algorithms for robust one-bit compressive sensing. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 820–828, 2014.

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

Appendix A. Linear Dependence of Disagreement and Excess Error

Lemma 1 (restated). *Given a classifier $h : X \mapsto \{+1, -1\}$ and distribution P satisfying Massart noise with parameter β, let w^* be the Bayes optimal classifier. Then we have,*

$$\beta \Pr_{(x,y)\sim P}[h(x) \neq h_{w^*}(x)] \leq \mathrm{err}_P(h) - \mathrm{err}_P(h_{w^*}) \leq \Pr_{(x,y)\sim P}[h(x) \neq h_{w^*}(x)].$$

Proof Here, we prove that the following equation holds for distribution P with Massart noise parameter $\beta > 0$.

$$\beta \Pr_{(x,y)\sim P}[h(x) \neq h_{w^*}(x)] \leq \mathrm{err}_P(h) - \mathrm{err}_P(h_{w^*}) \leq \Pr_{(x,y)\sim P}[h(x) \neq h_{w^*}(x)].$$

The right hand side inequality holds by the following.

$$\mathrm{err}_P(h) \leq \Pr_{(x,y)\sim P}[h(x) \neq h_{w^*}(x)] + \Pr_{(x,y)\sim P}[h_{w^*}(x) \neq y] = \Pr_{(x,y)\sim P}[h(x) \neq h_{w^*}(x)] + \mathrm{err}_P(h_{w^*}).$$

Let $A = \{x : h(x) \neq h_{w^*}(x)\}$ be the region where h and h_{w^*} disagree in their predictions. Note that $\Pr(A) = \Pr_{(x,y)\sim P}[h(x) \neq h_{w^*}(x)]$. Then,

$$\mathrm{err}_P(h) - \mathrm{err}_P(h_{w^*}) = \Pr(A)[\mathrm{err}_P(h|A) - \mathrm{err}_P(h_{w^*}|A)] + \Pr(\bar{A})[\mathrm{err}_P(h|\bar{A}) - \mathrm{err}_P(h_{w^*}|\bar{A})].$$

Classifiers h and h_{w^*} agree over the set $\bar{A}$, i.e., either both make mistakes or neither does, simultaneously. Hence the second term is zero. On the other hand, the two classifiers disagree over A, so exactly one of them is making an incorrect prediction. Hence, $\mathrm{err}_P(h|A) + \mathrm{err}_P(h_{w^*}|A) = 1$. We have

$$\mathrm{err}_P(h) - \mathrm{err}_P(h_{w^*}) = \Pr(A)[1 - 2\mathrm{err}_P(h_{w^*}|A)].$$

Since the labels are each flipped with probability at most $\frac{1-\beta}{2}$, we have that $\mathrm{err}_P(h_{w^*}|A) \leq \frac{1-\beta}{2}$. Re-arranging the above inequality proves the claim. ■

Appendix B. Proofs of Section 4

We use the following upper bound on the density of isotropic log-concave distributions.

Lemma 16 (Lovász and Vempala (2007)) *Let P be a 1-dimensional isotropic log-concave distribution over $\mathbb{R}$. Then $\Pr_{x\sim P}[x \geq \alpha] \leq \exp(-\alpha + 1)$.*

Lemma 7 (restated). *There exists an absolute constant c_1, such that $\mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)] \leq c_1 \frac{\tau_k}{\gamma_{k-1}}$.*

Proof Notice that w^* never makes a mistake on distribution D_k, so the hinge loss of w^* on D_k is entirely attributed to the points of D_k that are within distance τ_k from w^*. We have,

$$\begin{aligned}\mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)] &\leq \Pr_{(x,y)\sim D_k}[|w^* \cdot x| < \tau_k] \\ &= \frac{\Pr_{(x,y)\sim D}[|w^* \cdot x| < \tau_k]}{\Pr_{(x,y)\sim D}[|w_{k-1} \cdot x| \leq \gamma_{k-1}]}\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{C_2 \tau_k}{C_3 \gamma_{k-1}} && \text{(By Part 3 of Lemma 2)}\\
&\leq c_1 \frac{\tau_k}{\gamma_{k-1}}.
\end{aligned}
$$

■

The next lemma uses VC dimension tools to show that for linear classifiers that are considered in Step 2c (the ones with angle α_k to w_k), the empirical and expected hinge loss are close. Let D'_k denote the distribution D_k where the labels are predicted based on $\mathrm{sign}(p_k(\cdot))$. Note that T' is drawn from distribution D'_k.

Lemma 17 *There is $m_k = O(d(d + \log(k/d)))$ such that for a randomly drawn set T' of m_k labeled samples from D'_k, with probability $1 - \frac{\delta}{4(k+k^2)}$, for any $w \in B(w_{k-1}, r_{k-1})$,*

$$
\left| \mathbb{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T') \right| \leq \frac{\lambda}{12}.
$$

Proof The pseudo-dimension of the set of hinge loss values, i.e., $\{\ell_{\tau_k}(w,\cdot) : w \in \mathbb{R}^d\}$ is known to be at most d. Next, we prove that for any halfspace $w \in B(w_{k-1}, r_{k-1})$ and for any point $(x,y) \sim D'_k$, $\ell_{\tau_k}(w,x,y) \in O(\sqrt{d})$. We have,

$$
\begin{aligned}
\ell_{\tau_k}(w,x,y) &\leq 1 + \frac{|w \cdot x|}{\tau_k}\\
&\leq 1 + \frac{|w_{k-1} \cdot x| + \|w - w_{k-1}\|_2 \|x\|_2}{\tau_k}\\
&\leq 1 + \frac{\gamma_{k-1} + r_{k-1}\|x\|_2}{\tau_k}\\
&\leq c(1 + \|x\|_2).
\end{aligned}
$$

By Lemma 16, for any $(x,y) \in T'$, $\Pr_{(x,y)\sim D'_k}[\|x\|_2 > \alpha] \leq c \exp(-\alpha/\sqrt{d})$. Using union bound and setting $\alpha = \Theta(\sqrt{d}\ln(|T'|k^2/\delta))$ we have that with probability $1 - \frac{\delta}{8(k+k^2)}$, $\max_{x\in T'} \|x\|_2 \in O(\sqrt{d}\ln(|T'|k^2/\delta))$. Using standard pseudo-dimension rule we have that for $|T'| > \tilde{O}(d(d + \log \frac{k}{\delta}))$, with probability $1 - \frac{\delta}{4(k+k^2)}$,

$$
\left| \mathbb{E}_{(x,y)\sim D'_k}[\ell(w,x,y)] - \ell(w,T') \right| \leq \frac{\lambda}{12}.
$$

■

Lemma 6 (restated). *There exists an absolute constant c_2 such that*

$$
|\mathbb{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(w^*,x,y)] - \mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*,x,y)]| \leq c_2 \frac{\gamma_{k-1}}{\tau_k} \sqrt{\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)]}.
$$

Proof Let N indicate the set of points (x, y) such that p_k and h_{w^*} disagree. We have,

$$\begin{aligned}
&\left|\mathbb{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(w^*, x, y)] - \mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)]\right| \\
&\quad \leq \left|\mathbb{E}_{(x,y)\sim D'_k}\left[\mathbf{1}_{x\in N}\left(\ell_{\tau_k}(w^*, x, y) - \ell_{\tau_k}(w^*, x, \mathrm{sign}(w^* \cdot x))\right)\right]\right| \\
&\quad \leq 2\,\mathbb{E}_{(x,y)\sim D'_k}\left[\mathbf{1}_{x\in N}\left(\frac{|w^* \cdot x|}{\tau_k}\right)\right] \\
&\quad \leq \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim D'_k}[x \in N]} \times \sqrt{\mathbb{E}_{(x,y)\sim D'_k}[(w^* \cdot x)^2]} \qquad \text{(By Cauchy Schwarz)} \\
&\quad \leq \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)]} \times \sqrt{\mathbb{E}_{(x,y)\sim D_k}[(w^* \cdot x)^2]} \qquad \text{(By definition of } N\text{)} \\
&\quad \leq \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)]} \times \sqrt{C_7(r_{k-1}^2 + \gamma_{k-1}^2)} \qquad \text{(By Lemma 5)} \\
&\quad \leq c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)]}.
\end{aligned}$$

■

Lemma 5 (restated). *Let c_1 and c_2 be the absolute constants from Lemmas 6 and 7, respectively. Then with probability $1 - \frac{\delta}{2(k+k^2)}$,*

$$\mathrm{err}_{D'_k}(h_{w_k}) \leq 2c_1\frac{\tau_k}{\gamma_{k-1}} + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)]} + \frac{\lambda}{2}.$$

Proof First, we note that the true 0/1 error of w_k on any distribution is at most its true hinge loss on that distribution. So, it suffices to bound the hinge loss of w_k on D'_k. Moreover, v_k approximately minimizes the hinge loss on distribution D'_k, so in particular, it performs better than w^* on D'_k. On the other hand, Lemma 6 shows that the difference between hinge loss of w^* on D'_k and D_k is small. So, we complete the proof by using Lemma 7 and bounding the hinge of w^* on D_k. The following equations show the process of derivation of this bound as we explained.

$$\begin{aligned}
&\mathrm{err}_{D'_k}(h_{w_k}) \leq \mathbb{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(w_k, x, y)] \qquad \text{(Since hinge loss larger than 0/1 loss)} \\
&\leq 2\mathbb{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(v_k, x, y)] \qquad \text{(Since } \|v_k\|_2 > 0.5\text{)} \\
&\leq 2L_{\tau_k}(v_k, T') + 2(\frac{\lambda}{12}) \qquad \text{(By Lemma 17)} \\
&\leq 2L_{\tau_k}(w^*, T') + 4(\frac{\lambda}{12}) \qquad (v_k \text{ was an approximate hinge loss minimizer}) \\
&\leq 2\mathbb{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(w^*, x, y)] + 6(\frac{\lambda}{12}) \qquad \text{(By Lemma 17)} \\
&\leq 2\mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)] + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)]} + \frac{\lambda}{2} \quad \text{(By Lemma 6)} \\
&\leq 2c_1\frac{\tau_k}{\gamma_{k-1}} + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)]} + \frac{\lambda}{2}. \qquad \text{(By Lemma 7)}
\end{aligned}$$

■

We are now ready to prove our main theorem.

Proof of Theorem 3 Recall that we use the following parameters in Algorithm 1: $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = Cr_k$, where we defer the choice of C to later in the proof, $\lambda = \frac{3C_1}{8CC_2}$, $e_{\text{KKMS}} = \beta(\lambda/(4c_1 + 4c_2 + 2))^4$, and $\tau_k = \lambda\gamma_{k-1}/(4c_1 + 4c_2 + 2)$. Note, that by Equation 1, for any classifier h the excess error of h is upper bounded by the probability that h disagrees with h_{w^*}, i.e., $\text{err}_D(h)$. Here, we show that Algorithm 1 returns w_s such that $\text{err}_D(h_{w_s}) = \Pr_{(x,y)\sim D}[h_{w_s}(x) \neq h_{w^*}(x)] \leq \epsilon$, and in turn, the excess error of h_{w_s} is also at most ϵ.

We use induction to show that at the k^{th} step of the algorithm, $\theta(w_k, w^*) \leq \frac{e_0}{C_1 2^k}$. Since Part 4 of Lemma 2 and other Lemmas that build on it require $\theta(w, w^*) \leq \frac{\pi}{2}$ for any considered halfspace, we need to choose e_0 such that $\theta(w_0, w^*) \leq \frac{\pi}{2}$. Using Part 2 of Lemma 2, we have that $e_0 \leq \frac{\pi}{2C_1}$. Refer to Appendix C for the procedure that finds w_0 with this error value.

Assume by the induction hypothesis that at round $k-1$, $\text{err}_D(h_{w_{k-1}}) \leq e_0/2^{k-1}$. We will show that w_k, which is chosen by the algorithm at round k, also has the property that $\text{err}_D(h_{w_k}) \leq e_0/2^k$. Let $S_k = \{x : |w_{k-1} \cdot x| \leq \gamma_{k-1}\}$ indicate the band at round k. We divide the error of w_k to two parts, error outside the band and error inside the band. That is,

$$\text{err}_D(h_{w_k}) = \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \neq h_{w^*}(x)] + \Pr_{(x,y)\sim D}[x \in S_k \text{ and } h_{w_k}(x) \neq h_{w^*}(x)]. \tag{2}$$

By Part 2 of Lemma 2, $\theta(w_{k-1}, w^*) \leq r_{k-1}$. So, for the first part of the above inequality, which is the error of w_k outside the band, we have that

$$\begin{aligned} &\Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \neq h_{w^*}(x)] \\ &\qquad \leq \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \neq h_{w_{k-1}}(x)] + \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_{k-1}}(x) \neq h_{w^*}(x)] \\ &\qquad \leq 2\frac{C_1 r_{k-1}}{16} \leq \frac{e_0}{4 \times 2^k}, \end{aligned} \tag{3}$$

where the penultimate inequality follows from the fact that by the choice of $w_k \in B(w_{k-1}, r_{k-1})$ and the induction hypothesis, respectively, $\theta(w_{k-1}, w_k) < r_{k-1}$ and $\theta(w_{k-1}, w^*) < r_{k-1}$; By choosing large enough constant C in $\gamma_{k-1} = Cr_{k-1}$, using Part 4 of Lemma 2, the probability of disagreement outside of the band is $C_1 r_{k-1}/16$.

For the second part of Equation 2 we have that

$$\Pr_{(x,y)\sim D}[x \in S_k \text{ and } h_{w_k}(x) \neq h_{w^*}(x)] = \text{err}_{D_k}(h_{w_k}) \Pr_{(x,y)\sim D}[x \in S_k], \tag{4}$$

and

$$\text{err}_{D_k}(h_{w_k}) \Pr_{(x,y)\sim D}[x \in S_k] \leq \text{err}_{D_k}(h_{w_k}) C_2 \gamma_{k-1} \leq \text{err}_{D_k}(h_{w_k}) \frac{2C_2 Ce_0}{C_1 2^k}, \tag{5}$$

where the penultimate inequality is based on Part 3 of Lemma 2. Therefore, by replacing Equations 3 and 5 with Equation 2, we see that in order to have $\text{err}_D(h_{w_k}) < \frac{e_0}{2^k}$, it suffices to show that $\text{err}_{D_k}(h_{w_k}) \leq \frac{3C_1}{8CC_2} = \lambda$. The rest of the analysis is contributed to proving this bound. We have $\text{err}_{D_k}(h_{w_k}) = \Pr_{(x,y)\sim D_k}[h_{w_k}(x) \neq h_{w^*}(x)] \leq \Pr_{(x,y)\sim D_k}[\text{sign}(p_k(x)) \neq h_{w^*}(x)] +$

$\Pr_{(x,y)\sim D_k}[h_{w_k}(x) \neq \mathrm{sign}(p_k(x))]$. For the first part, using the assumption in Equation 1, we have that

$$\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)] \leq \frac{1}{\beta}\left(\mathrm{err}_{\tilde{D}_k}(\mathrm{sign}(p_k)) - \mathrm{err}_{\tilde{D}_k}(h_{w^*})\right) \leq \frac{e_{\mathrm{KKMS}}}{\beta}. \tag{6}$$

For the second part, using Lemma 5, we have

$$\Pr_{(x,y)\sim D_k}[h_{w_k}(x) \neq \mathrm{sign}(p_k(x))] = \mathrm{err}_{D'_k}(h_{w_k}) \leq 2c_1\frac{\tau_k}{\gamma_{k-1}} + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\frac{e_{\mathrm{KKMS}}}{\beta}} + \frac{\lambda}{2}.$$

Therefore, by the choice of parameter $\tau_k = \lambda\gamma_{k-1}/(4c_1+4c_2+2) = \gamma_{k-1}\left(e_{\mathrm{KKMS}}/\beta\right)^{1/4}$, we have

$$\begin{aligned}\mathrm{err}_{D_k}(h_{w_k}) &\leq \frac{e_{\mathrm{KKMS}}}{\beta} + 2c_1\frac{\tau_k}{\gamma_{k-1}} + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\frac{e_{\mathrm{KKMS}}}{\beta}} + \frac{\lambda}{2} \\ &\leq \frac{e_{\mathrm{KKMS}}}{\beta} + 2c_1\left(\frac{e_{\mathrm{KKMS}}}{\beta}\right)^{1/4} + 2c_2\left(\frac{e_{\mathrm{KKMS}}}{\beta}\right)^{1/4} + \frac{\lambda}{2} \\ &\leq (2c_1+2c_2+1)\left(\frac{e_{\mathrm{KKMS}}}{\beta}\right)^{1/4} + \frac{\lambda}{2} \leq \frac{\lambda}{2} + \frac{\lambda}{2} \leq \lambda.\end{aligned}$$

Sample Complexity and Runtime: To get error of e_{KKMS} with probability $1-\frac{s}{\delta}$ at every round, we need a labeled set of size $\mathrm{poly}(d, \log\frac{s}{\delta})$. The sample set T' is labeled based on p_k, so it does not contribute to the label complexity. So, at each round, we need $m_k = \mathrm{poly}(d, \log(\frac{\log(1/\epsilon)}{\delta}))$ labels. At each round, to get $\mathrm{poly}(d, \log(\frac{\log(1/\epsilon)}{\delta}))$ labels for the polynomial regression algorithm in the band of S_k we need $O(2^k m_k)$ samples from $\tilde{D}$. To get $d(d+\log(k/\delta))$ unlabeled samples in the band for Step 2b, we need $O(2^k(d(d+\log(k/\delta))) = \mathrm{poly}(d, \exp(k), \log(\frac{1}{\delta}))$ unlabeled samples. So, overall, we need $n_k = \mathrm{poly}(d, \exp(k), \log(\frac{1}{\delta}))$ unlabeled samples at each round. The running time is dominated by the polynomial regression algorithm which takes time $d^{\exp(\frac{1}{\beta^4})}$. However, since β is a constant, this is a polynomial in d. ■

Appendix C. Initializing w_0

We need to find w_0 such that $\mathrm{err}_D(h_{w_0}) \leq \frac{\pi}{2C_1}$. To find such w_0, we first take a labeled sample of size $m_0 = \mathrm{poly}(d, \log(\log(1/\epsilon)/\delta))$ from $\tilde{D}$ and run the polynomial regression algorithm of Kalai et al. (2008) on this set to find a polynomial $p_0(\cdot)$ with excess error $e'_{\mathrm{KKMS}} = \beta\left(\frac{\pi}{4(1+C'_1+C'_2)C_1}\right)^4$, where we defer the choice of C'_1 and C'_2 to later in the proof.

Then we take a sample of size $\mathrm{poly}(d, \log(1/\delta))$ from D and label it based on $\mathrm{sign}(p_0(\cdot))$. We find w_0 that approximately minimizes the empirical hinge loss over this sample. Similar to the proof of Theorem 3, we have

$$\mathrm{err}_D(h_{w_0}) = \Pr_{(x,y)\sim D}[p_0(x) \neq h_{w^*}(x)] + \Pr_{(x,y)\sim D}[p_0(x) \neq h_{w_0}(x)].$$

For the second part of this equation, similar to the analysis of Lemma 5, there exists τ large enough for the hinge loss threshold such that for any constant $\kappa > 0$.

$$\mathrm{err}_{D'}(h_{w_0}) \leq 2C'_1\left(\Pr_{(x,y)\sim D_k}[p(x) \neq h_{w^*}(x)]\right)^{1/4} + 2C'_2\left(\Pr_{(x,y)\sim D_k}[p(x) \neq h_{w^*}(x)]\right)^{1/4} + \kappa.$$

For $\kappa = e'_{\text{KKMS}}$ we have

$$\begin{aligned}\operatorname{err}_D(h_{w_0}) &\leq \Pr_{(x,y)\sim D}[p_0(x) \neq h_{w^*}(x)] + \Pr_{(x,y)\sim D}[p_0(x) \neq h_{w_0}(x)] \\ &\leq (2 + 2C'_1 + 2C'_2)\left(\Pr_{(x,y)\sim D}[p_0(x) \neq h_{w^*}(x)]\right)^{1/4} \\ &\leq (2 + 2C'_1 + 2C'_2)(\frac{1}{\beta}e'_{\text{KKMS}})^{1/4} \\ &\leq \frac{\pi}{2C_1}.\end{aligned}$$

Appendix D. Analysis of Sparse Case under Bounded Noise

In this section, we prove Theorem 8 for efficiently learning halfspaces under isotropic log-concave distributions in presence of bounded noise with parameter β that is independent of the dimension. We will assume that the target vector w^* is t-sparse.

We will first argue the proof of correctness and then the sample complexity. To argue correctness, we need to show that the new polynomial regression based algorithm in Step 2a of Algorithm 2 will indeed output a polynomial of excess error at most e_{KKMS}. Secondly, we need to argue that the hinge loss minimization w.r.t. the polynomial $p(\cdot)$ will output a vector v_k that is close to $p(\cdot)$. The second part is easy to see, since the vector w^* itself has L_1 norm at most $\sqrt{t}$. By restricting to vectors of small L_1 norm we still have to find a v_k with L_1 norm at most $\sqrt{t}$ that does well in the class (in comparison to w^*) in learning labels of $p(\cdot)$. For the first part, we prove the following extension of Kalai et al. (2005).

Theorem 10 (restated). *Let (X, Y) be drawn from a distribution over $\mathbb{R}^d \times \{+1, -1\}$ with isotropic log-concave marginal, constrained to the set $\{x : |w \cdot x| \leq \gamma\}$ for some w and γ. Let OPT be the error of the best t-sparse halfspace, i.e., $OPT = \min_{w\in\mathbb{R}^d,\|w\|_0\leq t} \Pr_{(x,y)\sim D}[\operatorname{sign}(w \cdot x) \neq y]$. Then, for every $\epsilon > 0$, there is an algorithm that runs in time $d^{\operatorname{poly}(\frac{1}{\epsilon})}$ and uses $m = O_\epsilon\left((\frac{t}{\gamma})^{\operatorname{poly}(\frac{1}{\epsilon})}\operatorname{polylog}(d)\right)$ samples from the distribution and outputs a polynomial $p(\cdot)$ such that $\operatorname{err}(p) \leq OPT + \epsilon$. Here, $\operatorname{err}(p) = \Pr_{(x,y)}[\operatorname{sign}(p(x)) \neq y]$. Furthermore, the polynomial $p(\cdot)$ satisfies $\|p\|_1 \leq (\frac{t}{\gamma})^{\operatorname{poly}(\frac{1}{\epsilon})}$.*

Note that the claimed sample complexity of our approach is an immediate consequence of the above theorem, since we require error of e_{KKMS} in the band and the subsequent hinge loss minimization step of our algorithm only uses examples labeled by $p(\cdot)$ and, therefore, does not affect the overall sample complexity of our algorithm. In order to prove the theorem, we need the following result about approximation of sign of halfspaces by polynomials.

Theorem 9 (Kalai et al. (2005)). *Let w^* be a halfspace in $\mathbb{R}^d$. Then, for every log-concave distribution over $\mathbb{R}^d$, there exists a degree $\frac{1}{\epsilon^2}$ polynomial $p(\cdot)$ such that $\mathbb{E}[(p(x) - \operatorname{sign}(w^* \cdot x))^2] \leq \epsilon$. Here the expectation is over a random x drawn from the distribution.*

Proof of Theorem 10 First, consider an isotropic log-concave distribution. Notice that if w^* is t-sparse, then the polynomial $p(\cdot)$ referred to in Theorem 9 will have support size at most $t^{\frac{1}{\epsilon^2}}$. This is due to the fact that the isotropicity and log-concavity of the distribution is preserved when considering the projection of the instance space on the relevant t variables. Since there are only t^{1/ϵ^2} monomials in the lower dimension of degree at most $\frac{1}{\epsilon^2}$, the $\frac{1}{\epsilon^2}$-degree polynomial $p(\cdot)$ that satisfies the theorem in this lower dimension also satisfies the requirement in the original space and is t^{1/ϵ^2}-sparse. The analysis of Kalai et al. (2005) also shows that $p(\cdot)$ is $\sum_{i=0}^{deg} c_i \bar{H}_i(\cdot)$, the linear combination of up to degree $deg = \frac{1}{\epsilon^2}$ normalized Hermite polynomials, where $\sum_{i=0}^{deg} c_i^2 < 1$ and $\bar{H}_i(x) = H_i(x)/\sqrt{2^i i!}$ refers to the normalized Hermite polynomial with degree i. By a naïve bound of $\sqrt{i!2^i}$ on the coefficients of $\bar{H}_i(x)$ and the fact that $i < \frac{1}{\epsilon^2}$, we know that the L_1 norm of each of the Hermite polynomials is bounded by $O_\epsilon(t^{1/\epsilon^2})$, where O_ϵ considers ϵ to be a constant. Moreover, since $\sum_{i=0}^{deg} c_i \le \sqrt{deg}\sum_{i=0}^{deg} c_i^2 < \sqrt{deg}$, the L_1 norm of p is also bounded by $t^{O(\frac{1}{\epsilon^2})}$.

This holds when the distribution is isotropic log-concave. However, the distributions we consider are conditionals of isotropic log-concave distribution over $\{|w \cdot x| \le \gamma_k\}$. These distributions are log-concave but not isotropic. To put them in the isotropic position, we transform each instance x to x' by a factor $O(\frac{1}{\gamma})$ along the direction of w. Then applying the above procedure on the transformed distribution we get a polynomial $p'(x') = \sum_{i=0}^{deg} p'_i \prod_{j=1}^{d} (x'_j)^{a_j}$. Since $x'_i \le O(\frac{1}{\gamma})x_i$ for every i, this polynomial can be formed in terms of x as $p(x) = \sum_{i=0}^{deg} p_i \prod_{j=1}^{d}(x_j)^{a_j}$, where $p_i \le O((\frac{1}{\gamma})^i)p'_i$. Therefore, for such distributions, the coefficients of the polynomial blow up by a factor of $O((\frac{1}{\gamma})^{\mathrm{poly}(1/\epsilon)})$ and as a result $\|p\|_1 \le O((\frac{t}{\gamma})^{\mathrm{poly}(1/\epsilon)})$. Thus, by enforcing that the polynomial $p(\cdot)$ belongs to $S = \{q : \|q\|_1 = O((\frac{t}{\gamma})^{\mathrm{poly}(1/\epsilon)})$ and degree$(q) \le \mathrm{poly}(1/\epsilon)\}$, we only need to argue about polynomials in the set S as opposed to general $\mathrm{poly}(\frac{1}{\epsilon})$-degree polynomials. Hence, as in (Kalai et al., 2005), we run the L_1 regression algorithm, but we also ensure that the L_1 norm of the induced polynomial is bounded by $\|q\|_1 = O((\frac{t}{\gamma})^{\mathrm{poly}(1/\epsilon)})$. This can be done via constrained L_1 norm minimization. The analysis of this algorithm is similar as that of Kalai et al. (2005). For the self-completeness of the paper, we show a complete proof here. Denote by $\mathcal{Z} = (x^1, y^1), \cdots, (x^m, y^m)$ the samples. Firstly, we have

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m} I(q(x^j)y^j < \gamma) &= \frac{1}{m}\sum_{j=1}^{m} I(\mathrm{sign}(q(x^j)) \neq y^j) + \frac{1}{m}\sum_{j=1}^{m} I(\mathrm{sign}(q) = y^j \;\&\; q(x^j)y^j < \gamma)\\ &\le \frac{1}{2m}\sum_{j=1}^{m}|y^j - p(x^j)| + \frac{\gamma}{2},\end{aligned} \tag{7}$$

where $q(x) = p(x) - T$. The above inequality holds because of a standard argument on the randomized threshold T: Note that $\mathrm{sign}(q(x^j)) \neq y^j$ iff the threshold T lies between $p(x^j)$ and y^j; Similarly, $\mathrm{sign}(q(x^j)) = y^j \;\&\; q(x^j)y^j < \gamma$ iff the threshold T lies between $p(x^j)$ and $p(x^j) - y^j\gamma$. So if we choose T uniformly at random on $[-1, 1]$, Equation 7 holds in expectation. Since we select T to minimize the LHS of Equation 7, the inequality holds with certainty. Then by the L_1 polynomial regression algorithm which fits the labels by polynomial in the sense of L_1 norm, we have

$$\frac{1}{m}\sum_{j=1}^{m}|y^j - p(x^j)| \le \frac{1}{m}\sum_{j=1}^{m}|y^j - p^*(x^j)| \le \frac{1}{m}\sum_{j=1}^{m}|y^j - c(x^j)| + |c(x^j) - p^*(x^j)|,$$

where c is the optimal classifier and p^* is a polynomial satisfying Theorem 9. Thus

$$\mathbb{E}_{\mathcal{Z}}\left[\frac{1}{m}\sum_{j=1}^{m} I(q(x^j)y^j < \gamma)\right] \le OPT + \frac{\epsilon}{2} + \frac{\gamma}{2}.$$

Let $S = \{q : \text{degree}(q) \leq \frac{1}{\epsilon^2}, \|q\|_1 \leq (\frac{t}{\gamma})^{O(\frac{1}{\epsilon^2})}\}$ and let $\hat{L}(q) = \frac{1}{m}\sum_{(x^j,y^j)} I(q(x^j)y^j < \gamma)$ be the empirical $0/1$ loss of the polynomial q with margin γ. In order to complete the proof, we need to argue that if m is large enough then for all $q \in S$, we have, with high probability, $|\hat{L}(q) - err(q)| \leq \epsilon/4$. To see this, we need the following lemma of Zhang (2002).

Lemma 18 (Zhang (2002)) *Let the instance space be bounded as $\|x\|_\infty \leq X_\infty$, and consider the class of hyperplane w such that $\|w\|_1 \leq W_1$. Denote by $err(w)$ the expected 0/1 error of w. Then there is a constant C such that with probability $1-\delta$, for all γ, we have*

$$err(w) \leq \frac{1}{m}\sum_{j=1}^{m} I(y^j(w \cdot x^j) < \gamma) + \sqrt{\frac{C}{m}\left(\frac{X_\infty^2 W_1^2(\log d + 1)}{\gamma^2}\log m + \log\frac{1}{\delta}\right)}.$$

Setting γ as $\epsilon/2$, W_1 as $\left(\frac{t}{\gamma}\right)^{O(\frac{1}{\epsilon^2})}$, and X_∞ as $O\left((\log(md))^{O(\frac{1}{\epsilon^2})}\right)$ (see Lemma 20), viewing the polynomial q as a $d^{O(1/\epsilon^2)}$-dimensional vector, Lemma 18 gives the desired sample complexity $m = O_\epsilon\left((\frac{t}{\gamma})^{\text{poly}(\frac{1}{\epsilon})}\text{polylog}(d)\right)$. ■

In the above, we explicitly suppressed the dependence on ϵ, because for the purpose of our algorithm, we use a constant value e_{KKMS} for the desired value of the error in Theorem 10. Moreover, the distribution at every round is restricted to the set $\{x : |w_{k-1} \cdot x| \leq \gamma_{k-1}\}$. Since $\gamma_k \geq \epsilon$, for all k, we use the value of $\gamma = \epsilon$ in Theorem 10 and achieve the results of Theorem 8 as a consequence.

Appendix E. 1-bit Compressed Sensing under Adversarial Noise

In this section, we first consider the case of non-uniform 1-bit compressed sensing under adversarial noise and provide a proof of Theorem 11. Then, we discuss an extension of our analysis that holds for uniform 1-bit compressed sensing under adversarial noise and provide a proof of Theorem 12.

We start with the following result of Awasthi et al. (2014).

Theorem 19 (Awasthi et al. (2014)) *Let (x, y) be drawn from a distribution over $\mathbb{R}^d \times \{+1, -1\}$ such that the marginal over x is isotropic log-concave. Let OPT be the $0/1$ error of the best halfspace, i.e., $OPT = \min_{w:\|w\|_2=1} \Pr[\text{sign}(w \cdot x) \neq y]$ and let w^* be the halfspace that achieves OPT. Then, there exists an algorithm that, for every $\epsilon > 0$, runs in time polynomial in d and $\frac{1}{\epsilon}$ and outputs a halfspace w such that $\|w - w^*\|_2 \leq O(OPT) + \epsilon$.*

We extend the algorithm of Awasthi et al. (2014) for 1-bit compressed sensing. The main difference between our algorithm and the algorithm of Awasthi et al. (2014) is in the hinge-loss minimization step and the sample complexity. In this case, when minimizing hinge loss at each step, we restrict the search to vectors of L_1 norm bounded by $\sqrt{t}$. Note that this does not affect the correctness of the algorithm, as w^* itself is t-sparse and $\|w^*\|_1 \leq \sqrt{t}$. The crux of the argument is in showing that when $\|w\|_1 \leq \sqrt{t}$, the empirical hinge loss of w is nicely concentrated around its expectation. This is proved in Lemma 22. Using this new concentration results, the proof of Theorem 11 follows immediately by the analysis of Awasthi et al. (2014). For completeness, here we provide a complete proof of Theorem 11.

Algorithm 4 NON-UNIFORM 1-BIT COMPRESSED SENSING UNDER ADVERSARIAL NOISE

```
Input: An initial classifier $w_0$, a sequence of values $\gamma_k, \tau_k$ and $r_k$ for $k = 1, \dots, \log(1/\epsilon)$.
1. Let $w_0$ be the initial classifier.
2. For $k = 1, \dots, \log(1/\epsilon) = s$.
    (a) Take $m_k = \Omega(\frac{t}{\epsilon^2}\mathrm{polylog}(d, \frac{1}{\delta}, \frac{1}{\epsilon}))$ samples from $\tilde{D}_k$ and request the labels. Call this set of labeled samples $T'$.
    (b) Find $v_k \in B(w_{k-1}, r_{k-1})$ such that $\|v_k\|_1 \leq \sqrt{t}$ and $v_k$ approximately minimizes the empirical hinge loss over $T'$ using threshold $\tau_k$, i.e., $L_{\tau_k}(v_k, T') \leq \min_{w \in B(w_{k-1}, r_{k-1}) \text{ and } \|w\|_1 \leq \sqrt{t}} L_{\tau_k}(w, T') + \frac{\lambda}{12}$.
    (c) Let $w_k = \frac{v_k}{\|v_k\|_2}$.
Output: Return $w_s$, which has excess error $O(\nu) + \epsilon$ with probability $1 - \delta$.
```

To achieve desirable concentration result, we use the tools from VC and Rademacher complexity theory to obtain a sample complexity that is polynomial in t and only logarithmic in the ambient dimension d. The following lemma helps us in achieving such concentration result.

Lemma 13 (Shalev-Shwartz and Ben-David (2014)) *Let $\mathcal{F}$ be the class of linear predictors with the L_1 norm of the weights bounded by W_1. Assume that the infinity norm of all instances is bounded by X_∞. Then for the ρ-Lipschitz loss ℓ such that $\max_{w \cdot x \in [-W_1 X_\infty, W_1 X_\infty]} |\ell(w, x, y)| \leq U$ and the choice of an i.i.d. sample T of size m,*

$$\forall w, \text{ s.t., } \|w\|_1 \leq W_1, \ \Pr\left[|\mathbb{E}\ell(w, x, y) - \ell(w, T)| \geq 2\rho W_1 X_\infty \sqrt{\frac{2\log(2d)}{m}} + s \right] \leq 2\exp\left(-\frac{ms^2}{2U^2} \right).$$

In preparation to use Lemma 13, we bound the infinity norm of the instances used by our algorithm in the next lemma.

Lemma 20 *Let S be the set of all (unlabeled) samples drawn from D. With probability $1 - \delta$ for all $x \in S$, $\|x\|_\infty \leq O(\log \frac{|S|d}{\delta})$.*

Proof Since D is an isotropic log-concave distribution, the marginal distribution on any coordinate is a one-dimensional isotropic log-concave distribution. Therefore, by concentration results of Lovász and Vempala (2007), we have

$$\Pr_{x \sim D}\left[\|x\|_\infty \geq c' \log \frac{d}{\delta} \right] \leq \sum_{i \in [d]} \Pr_{x \sim D}\left[x_i \geq c' \log \frac{d}{\delta} \right] \leq \delta.$$

Taking union bound over all elements of S, with probability $1 - \delta$, $\|x\|_\infty \leq O(\log \frac{|S|d}{\delta})$. ■

Next, we bound the value of hinge loss on any instance (x, y) used by our algorithm. Let H be a class of halfspaces w, with $\|w\|_1 \leq \sqrt{t}$ and $\|w\|_2 = 1$.

Lemma 21 *For a given k and $v \in H$, let T' be the set of m_k samples drawn from $\tilde{D}_{v, \gamma_k}$. For any halfspace u such that $\|u\|_2 = 1$, $\|u\|_1 \leq \sqrt{t}$ and $u \in B(v, r_k)$, with probability $1 - \delta$, for all $x \in T'$, $\ell_{\tau_k}(u, x, y) \leq O(\log \frac{m_k}{\gamma_k \delta})$.*

Proof We have

$$\ell_{\tau_k}(u,x,y) \leq 1 + \frac{|u\cdot x|}{\tau_k} \leq 1 + \frac{|v\cdot x|}{\tau_k} + \frac{|(u-v)\cdot x|}{\tau_k}.$$

By the choice of $x \sim D_{v,\gamma_k}$, we know that $v \cdot x \leq \gamma_k$. Therefore, $\frac{|v\cdot x|}{\tau_k} \leq O(1)$. For the second part of the inequality, $|(u-v)\cdot x|$, first consider all $x \sim D$. Since, D is an isotropic log-concave distribution and $\|u - v\| \leq r_k$, without loss of generality, we can assume that $u - v = (r, 0, \ldots, 0)$ for some $r \leq r_k$. Moreover, $(u-v)\cdot x = r|x_1|$, and x_1 is a one-dimensional isotropic log-concave distribution. Therefore,

$$\Pr_{x\sim D}\left[|(u-v)\cdot x| \geq r_k(1+\log\frac{1}{\delta})\right] \leq \Pr_{x\sim D}\left[r|x_1| \geq r_k(1+\log\frac{1}{\delta})\right] \leq \Pr_{x\sim D}\left[|x_1| \geq 1+\log\frac{1}{\delta}\right] \leq \delta.$$

So,

$$\begin{aligned}\Pr_{x\sim D_k}\left[|(u-v)\cdot x| \geq r_k(1+\log\frac{1}{\gamma_k\delta})\right] &= \frac{\Pr_{x\sim D}\left[|(u-v)\cdot x| \geq r_k(1+\log\frac{1}{\gamma_k\delta}) \text{ \& } |v\cdot x| \leq \gamma_k\right]}{\Pr_{x\sim D}[|v\cdot x| \leq \gamma_k]}\\ &\leq \frac{\Pr_{x\sim D}\left[|(u-v)\cdot x| \geq r_k(1+\log\frac{1}{\gamma_k\delta})\right]}{\Pr_{x\sim D}[|v\cdot x| \leq \gamma_k]}\\ &\leq \Theta(\frac{1}{\gamma_k})\Pr_{x\sim D}\left[|(u-v)\cdot x| \geq r_k(1+\log\frac{1}{\gamma_k\delta})\right]\\ &\leq \delta.\end{aligned}$$

So for a fixed v and k, and for all m_k samples T' with probability $1-\delta$, $\frac{|(u-v)\cdot x|}{\tau_k} \leq \frac{r_k}{\tau_k}\log\frac{m_k}{\gamma_k\delta} \leq O(\log\frac{m_k}{\gamma_k\delta})$.

■

Lemma 22 *Let $m_k = \Omega(\frac{t}{\epsilon^2}\text{polylog}(d, \frac{1}{\delta}, \frac{1}{\epsilon}))$ and T' be the samples drawn from $\tilde{D}_k$ and T to be the corresponding samples when their labels are corrected based on w^*. With probability $1-\delta$,*

$$\sup_w \left|\mathbb{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T')\right| \leq \frac{\lambda}{12},$$

and

$$\sup_w \left|\mathbb{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T)\right| \leq \frac{\lambda}{12}.$$

where $w \in B(w_{k-1}, r_{k-1})$ such that $\|w\|_1 \leq \sqrt{t}$.

Proof Using Lemma 13 we have that

$$\Pr\left[\sup_w \left|\mathbb{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T')\right| \geq 2\rho W_1 X_\infty\sqrt{\frac{2\log(2d)}{m_k}} + s\right] \leq 2\exp\left(-\frac{m_k s^2}{2U^2}\right), \tag{8}$$

where U, ρ, W_1 and X_∞ are defined as Lemma 13, and the supremum is taken over all w in $K = \{w \in \mathbb{R}^d : \|w\|_1 \le W_1, \|w\|_2 \le 1\}$. Note that $W_1 \le \sqrt{t}$ and $\rho = \frac{1}{\tau_k} \le \frac{1}{\epsilon}$ and by Lemma 20 and 21 for any δ, with probability δ, $X_\infty \le O(\log \frac{md}{\delta})$ and $U \le O(\log \frac{m_k}{\gamma_k \delta})$.

Assume that these bounds hold for X_∞ and U. For $m = \Theta(\frac{t}{\epsilon^3}\text{polylog}(d, \frac{1}{\delta}, \frac{1}{\epsilon}))$ and $m_k \ge \Omega(\frac{t}{\epsilon^2}\log(md/\delta)\log d)$, and for appropriate choice of constant s, with probability at most δ,

$$\sup_w \left| \mathbb{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w, x, y)] - \ell_{\tau_k}(w, T') \right| \ge 2\frac{\sqrt{t}}{\epsilon}\log(\frac{md}{\delta})\sqrt{\frac{2\log(2d)}{m_k}} + s \ge \lambda/12.$$

The proof for the case of T is similar to the above. ■

The rest of the proof follows a similar outline as that of Section B. We would need the following lemmas:

Lemma 23 *There exists an absolute constant c_1 such that in round k of Algorithm 3, we have $\mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)] \le c_1 \frac{\tau_k}{\gamma_{k-1}}$.*

Proof Notice that w^* never makes a mistake on distribution D_k, so the hinge loss of w^* on D_k is entirely attributed to the points of D_k that are within distance τ_k from w^*. We have,

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)] &\le \Pr_{(x,y)\sim D_k}[|w^* \cdot x| < \tau_k] \\
&= \frac{\Pr_{(x,y)\sim D}[|w^* \cdot x| < \tau_k]}{\Pr_{(x,y)\sim D}[|w_{k-1} \cdot x| \le \gamma_{k-1}]} \\
&\le \frac{C_2 \tau_k}{C_3 \gamma_{k-1}} \qquad \text{(By Part 3 of Lemma 2)} \\
&\le c_1 \frac{\tau_k}{\gamma_{k-1}}.
\end{aligned}$$

■

Lemma 24 *There exists an absolute constant c_2 such that*

$$\left| \mathbb{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w^*, x, y)] - \mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)] \right| \le c_2 \frac{\gamma_{k-1}}{\tau_k}\sqrt{C2^k\nu}.$$

Proof Let N indicate the set of points where w^* makes a mistake. We have,

$$\begin{aligned}
&\left| \mathbb{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w^*, x, y)] - \mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)] \right| \\
&\quad \le \left| \mathbb{E}_{(x,y)\sim\tilde{D}_k}\left[\mathbf{1}_{x\in N}\left(\ell_{\tau_k}(w^*, x, y) - \ell_{\tau_k}(w^*, x, \text{sign}(w^* \cdot x))\right)\right] \right| \\
&\quad \le 2\,\mathbb{E}_{(x,y)\sim\tilde{D}_k}\left[\mathbf{1}_{x\in N}\left(\frac{|w^* \cdot x|}{\tau_k}\right)\right] \\
&\quad \le \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim\tilde{D}_k}[x \in N]} \times \sqrt{\mathbb{E}_{(x,y)\sim\tilde{D}_k}[(w^* \cdot x)^2]} \qquad \text{(By Cauchy Schwarz)}
\end{aligned}$$

$$
\begin{aligned}
&\le \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[x\in N]} \times \sqrt{\mathbb{E}_{(x,y)\sim D_k}[(w^*\cdot x)^2]} \qquad \text{(By definition of } N\text{)}\\
&\le \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[x\in N]} \times \sqrt{C_7(r_{k-1}^2+\gamma_{k-1}^2)} \qquad \text{(By Lemma 5)}\\
&\le c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[x\in N]}\\
&\le c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{C2^k\nu}. \qquad \text{(The noise rate within the band can go up by a factor of } 2^k\text{)}
\end{aligned}
$$

■

Lemma 25 *Let c_1 and c_2 be the absolute constants from Lemmas 23 and 24, respectively. Then with probability $1-\frac{\delta}{2(k+k^2)}$,*

$$
\mathrm{err}_{\tilde{D}_k}(h_{w_k}) \le 2c_1\frac{\tau_k}{\gamma_{k-1}} + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{C\nu 2^k} + \frac{\lambda}{2}.
$$

Proof First, we note that the true 0/1 error of w_k on any distribution is at most its true hinge loss on that distribution. So, it suffices to bound the hinge loss of w_k on $\tilde{D}_k$. Moreover, v_k approximately minimizes the hinge loss on distribution $\tilde{D}_k$, so in particular, it performs better than w^* on $\tilde{D}_k$. On the other hand, Lemma 24 shows that the difference between hinge loss of w^* on $\tilde{D}_k$ and D_k is small. So, we complete the proof by using Lemma 23 and bounding the hinge of w^* on D_k. The following equations show the process of derivation of this bound as we explained.

$$
\begin{aligned}
&\mathrm{err}_{\tilde{D}_k}(h_{w_k}) \le \mathbb{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w_k,x,y)] \qquad \text{(Since hinge loss larger than 0-1 loss)}\\
&\le 2\mathbb{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(v_k,x,y)] \qquad \text{(Since } \|v_k\|_2 > 0.5\text{)}\\
&\le 2L_{\tau_k}(v_k,T') + 2(\frac{\lambda}{12}) \qquad \text{(By Lemma 22)}\\
&\le 2L_{\tau_k}(w^*,T') + 4(\frac{\lambda}{12}) \qquad (v_k \text{ was an approximate hinge loss minimizer})\\
&\le 2\mathbb{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w^*,x,y)] + 6(\frac{\lambda}{12}) \qquad \text{(By Lemma 22)}\\
&\le 2\mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*,x,y)] + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{C2^k\nu} + \frac{\lambda}{2} \quad \text{(By Lemma 24)}\\
&\le 2c_1\frac{\tau_k}{\gamma_{k-1}} + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{C2^k\nu} + \frac{\lambda}{2}. \qquad \text{(By Lemma 23)}
\end{aligned}
$$

■

We are now ready to prove our main theorem.

Proof of Theorem 11 We use induction to show that at the k^{th} step of the algorithm, $\theta(w_k, w^*) \le \frac{e_0}{C_1 2^k}$ where e_0 is the initial error of w_0. Assume by the induction hypothesis that at round $k-1$, $\mathrm{err}_D(h_{w_{k-1}}) \le e_0/2^{k-1}$. We will show that w_k, which is chosen by the algorithm at round k, also has the property that $\mathrm{err}_D(h_{w_k}) \le e_0/2^k$. Let $S_k = \{x : |w_{k-1}\cdot x| \le \gamma_{k-1}\}$ indicate the band at round k. We divide the error of w_k to two parts, error outside the band and error inside the band. That is,

$$\mathrm{err}_D(h_{w_k}) = \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \neq h_{w^*}(x)] + \Pr_{(x,y)\sim D}[x \in S_k \text{ and } h_{w_k}(x) \neq h_{w^*}(x)]. \tag{9}$$

By Part 2 of Lemma 2, $\theta(w_{k-1}, w^*) \leq r_{k-1}$. So, for the first part of the above inequality, which is the error of w_k outside the band, we have that

$$\begin{aligned} &\Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \neq h_{w^*}(x)] \\ &\quad\leq \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \neq h_{w_{k-1}}(x)] + \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_{k-1}}(x) \neq h_{w^*}(x)] \\ &\quad\leq 2\frac{C_1 r_{k-1}}{16} \leq \frac{e_0}{4 \times 2^k}, \end{aligned}$$

where the penultimate inequality follows from the fact that by the choice of $w_k \in B(w_{k-1}, r_{k-1})$ and the induction hypothesis, respectively, $\theta(w_{k-1}, w_k) < r_{k-1}$ and $\theta(w_{k-1}, w^*) < r_{k-1}$; By choosing large enough constant in $\gamma_{k-1} = C r_{k-1}$, using Part 4 of Lemma 2, the probability of disagreement outside of the band is $C_1 r_{k-1}/16$.

For the second part of Equation 9, using the same derivation as in Lemma 25 we get that

$$\mathrm{err}_{D_k}(h_{w_k}) \leq 2c_1 \frac{\tau_k}{\gamma_{k-1}} + 2c_2 \frac{\gamma_{k-1}}{\tau_k}\sqrt{C\nu 2^k} + \frac{\lambda}{2}.$$

By our choice of parameters, we know that the ratio of τ_k and γ_{k-1} is bounded by $\leq \frac{\lambda}{12}$. Hence, for the sum to be bounded by λ, we need $C 2^k \nu$ to be bounded by a constant. But this is true since $k \geq \log \frac{1}{c\nu+\epsilon}$ for an appropriate constant c. ■

E.1. Uniform 1-bit Compressed Sensing under Adversarial Noise

Next, we provide a proof for Theorem 12. We extend our analysis from the previous section to hold for the case of uniform 1-bit compressed sensing. The main difference between the results of this section and the analysis of the previous section is that we need to obtain a concentration result that holds uniformly over all choice of underlying noisy distribution. In other words, they hold uniformly over the choice of w^* and the ν fraction of the samples whose labels differ from the labels of w^*.

First, we introduce Lemma 26 that shows that for a large enough number of unlabeled samples, *every band* around a halfspace that can be considered by the algorithm has sufficient samples. In contrast, the results of the previous section only show that the bands around $w_1, \dots, w_k$, which are uniquely determined by the samples and the fixed (but unknown) distribution $\tilde{D}$, have sufficient samples. Next, we build on the concentration results from the previous section and show that the hinge loss is concentrated around its expectation uniformly over all choice of all w^* and ν fraction of the samples whose labels differ from the labels of w^*. Using this new concentration result, the proof of Theorem 12 follows immediately by the analysis of the non-uniform case.

Note that at every step of the algorithm, vector v_k that is chosen by the hinge loss minimization step is such that $\|v_k\|_1 \leq \sqrt{t}$. As Awasthi et al. (2014) argue, $\|v_k\|_2 \geq 1/2$. Therefore, the outcome of step 3c also satisfies $\|w_k\|_1 \leq O(\sqrt{t})$. The following lemma shows that when the number of unlabeled samples is large enough, every possible band around every such w_k considered by the

```
**Algorithm 5** UNIFORM 1-BIT COMPRESSED SENSING UNDER ADVERSARIAL NOISE
Input: An initial classifier $w_0$, a sequence of values $\gamma_k, \tau_k$ and $r_k$ for $k = 1, \dots, \log(1/\epsilon)$.
1. Let $w_0$ be the initial classifier.
2. For $k = 1, \dots, \log(1/\epsilon) = s$.
3. Take $m = O(t\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})/\epsilon^4)$ unlabeled samples from $\tilde{D}$, in set $S$.
    (a) Take $m_k = O(t\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})/\epsilon^2)$ of the samples in $S \cap S_k$ request the labels. Call this set of labeled samples $T'$.
    (b) Find $v_k \in B(w_{k-1}, r_{k-1})$ such that $\|v_k\|_1 \leq \sqrt{t}$ and $v_k$ approximately minimizes the empirical hinge loss over $T'$ using threshold $\tau_k$, i.e., $L_{\tau_k}(v_k, T') \leq \min_{w \in B(w_{k-1}, r_{k-1}) \text{ and } \|w\|_1 \leq \sqrt{t}} L_{\tau_k}(w, T') + \frac{\lambda}{12}$.
    (c) Let $w_k = \frac{v_k}{\|v_k\|_2}$.
Output: Return $w_s$, which has excess error $O(\nu) + \epsilon$ with probability $1 - \delta$.
```

algorithm contains a number of points that is at least a multiplicative approximation to the number of points expected to be in that band. Therefore, in every step of the algorithm, there is a sufficient number of samples in the band.

Lemma 26 *Let S be a set of $m \geq \frac{t}{\epsilon^4}$ polylog$(d) \log(\frac{1}{\delta})$ samples drawn from D. With probability $1 - \delta$ for all $\gamma \in \Omega(\epsilon)$ and w such that $\|w\|_1 \leq O(\sqrt{t})$ and $\|w\|_2 = 1$,*

$$\frac{1}{m}\Big|\{x \mid x \in S \text{ and } |w \cdot x| \leq \gamma\}\Big| \geq c\gamma,$$

where c is a constant.

Proof Let H be a class of (non-homogeneous) halfspaces w, with $\|w\|_1 \leq O(\sqrt{t})$ and $\|w\|_2 = 1$. Let B be a class of hypothesis defined by bands around homogeneous halfspaces, w, such that $\|w\|_1 \leq O(\sqrt{t})$ and $\|w\|_2 = 1$ with arbitrary width.

The covering number of H is at most $\log N(\gamma, H) = O\left(t \text{ polylog}(d)/\gamma^2\right)$ (Plan and Vershynin, 2013a). Since every band is an intersection of two halfspaces, each band of B can be represented by the intersection of two halfspaces from H. Therefore, $\log N(\gamma, B) = O\left(t \text{ polylog}(d)/\gamma^2\right)$. Furthermore, by Lemma 2, $\mathbb{E}\left[\frac{1}{m}\left|\{x \mid x \in S \text{ and } |w \cdot x| \leq \gamma\}\right|\right] = \Theta(\gamma)$. Therefore, by the uniform convergence results for covering number, we have that

$$\begin{aligned}\Pr\left[\sup_{w,\gamma} \ \frac{1}{m}\Big|\{x \mid x \in S \text{ and } |w \cdot x| \leq \gamma\}\Big| - \Theta(\gamma) \leq \gamma\right] &\leq N(\gamma, H) e^{-\frac{\gamma^2 m}{8}} \\ &\leq e^{-\frac{\epsilon^2 m}{8} + \frac{t \text{ polylog}(d)}{\epsilon^2}} \\ &\leq \delta.\end{aligned}$$

■

Lemma 27 *For $\nu \in O(\epsilon/\log(\frac{d}{\epsilon})^2)$ and S of size $m = \Theta(\frac{t}{\epsilon^4}\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$, with probability $1 - \delta$,*

$$\sup_{w^*, \{y_i\}_{i=1}^m, w} \left|\mathbb{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w, x, y)] - \ell_{\tau_k}(w, T')\right| \leq \frac{\lambda}{12},$$

and

$$\sup_{w^*,\{y_i\}_{i=1}^m,w} \left|\mathbb{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T)\right| \le \frac{\lambda}{12},$$

where w^* is a t-sparse halfspace, $\{y_i\}_{i=1}^m$ are the labels of the set of samples S such at most ν fraction of them differs from the labels of w^*, and $w_{k-1} \in H$ is the unique halfspace determined by the outcome of step k of the algorithm given w^* and $\{y_k\}_{k=1}^m$ (labels used in the previous round), and $w \in B(w_{k-1}, r_{k-1})$ such that $\|w\|_1 \le \sqrt{t}$.

Proof Using Lemma 13 we have that

$$\Pr\left[\sup_{w}\left|\mathbb{E}_{(x,y)\sim \tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T')\right| \ge 2\rho W_1 X_\infty\sqrt{\frac{2\log(2d)}{m_k}} + s\right] \le 2\exp\left(-\frac{m_k s^2}{2U^2}\right), \tag{10}$$

where U, ρ, W_1 and X_∞ are defined as Lemma 13, and the supremum is taken over all w in $K = \{w \in \mathbb{R}^d : \|w\|_1 \le W_1, \|w\|_2 \le 1\}$. Note that $W_1 = \sqrt{t}$ and $\rho = \frac{1}{\tau_k} \le \frac{1}{\epsilon}$ and by Lemma 20 and 21 for any δ, with probability δ, $X_\infty \le O(\log\frac{md}{\delta})$ and $U \le O(\log\frac{m_k}{\gamma_k\delta})$.

Assume that these bounds hold for X_∞ and U. Then for a fixed w_{k-1} considering $m_k = \frac{t}{\epsilon^3}\operatorname{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$ of the samples in the band around it (there are $O(m\gamma_k) \ge m_k$ such samples by Lemma 26), and for an appropriate choice of constant s, with probability at most $2\exp\left(-\frac{m_k s^2}{2U^2}\right)$,

$$\sup_{w}\left|\mathbb{E}_{(x,y)\sim \tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T')\right| \ge 2\frac{\sqrt{t}}{\epsilon}\operatorname{polylog}(d,\frac{1}{\epsilon},\frac{1}{\delta})\sqrt{\frac{2\log(2d)}{m_k}} + s \ge \lambda/12$$

Next, we show how to achieve a similar concentration result over all choices of w^* and choices of νm corrupted measurements and the resulted w_{k-1}. Note that w_{k-1} depends only on the samples of S and their labels used in previous steps. Since, we only use labels of $m_k = O(\frac{t}{\gamma^3}\operatorname{polylog}(d,\frac{1}{\epsilon},\frac{1}{\delta}))$ points in every step, overall, Equation 10 only depends on the labels of these sample points. This is uniquely determined by the choice of w^* and the ν fraction of the samples that do not agree with labels of w^*. Therefore, we can restrict our attention to the different labelings that can be produced by such w^* and adversarial corruption on the sample of size $\sum_i m_i \le \frac{t}{\epsilon^3}\operatorname{polylog}(d,\frac{1}{\epsilon},\frac{1}{\delta})$.

Let $K' = \{w \in \mathbb{R}^d : \|w\|_0 \le t, \|w\|_2 \le 1\}$ be the set of all possible true signals w^*. It is known that the VC dimension of the set K is $t\log d$, therefore there are $O((\sum_i m_i)^{t\log d})$ possible labeling that can be produced by some $w^* \in K'$. Moreover, because $\sum_{i\le k} m_i = \Theta(\gamma_k m)$. Therefore, the adversary can corrupt a $\frac{\nu}{\gamma_k}$ fraction of the $\sum_{i\le k} m_i$ samples. This is in the worst case, $(\frac{\nu}{\epsilon})\frac{t}{\epsilon^3}\operatorname{polylog}(d,\frac{1}{\epsilon},\frac{1}{\delta}))$. Let $m' = \sum_{i\le k} m_i$. By taking the union bound over choices of w^* and $\frac{\nu}{\epsilon}m'$ corrupted points, we have

$$\begin{aligned}\Pr\left[\sup_{w^*,\{y_k\}_{k=1}^{m_k},w}\left|\mathbb{E}_{(x,y)\sim \tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T')\right| \ge \lambda/12\right] &\le \exp\left(-\frac{m_k s^2}{2U^2}\right)\; c' m'^{t\log d}\binom{m'}{\frac{\nu}{\epsilon}m'}\\ &\le c\;\exp\left(-\frac{m_k s^2}{2U^2} + t\log d\log(m') + \frac{\nu}{\epsilon}\log(\frac{\epsilon}{\nu})m'\right)\\ &\le c\;\exp\left(-\frac{m_k s^2}{2U^2} + t\log d\log(m') + \frac{\nu}{\epsilon}\log(\frac{\epsilon}{\nu})\log(\frac{1}{\epsilon})m_k\right)\end{aligned}$$

$$\leq \exp\left(-O\left(\frac{m_k}{\log \frac{m_k}{\gamma_k \delta}}\right)\right)$$

where the last inequality follows from $\nu \in O(\epsilon/\log(\frac{d}{\epsilon})^2)$. Therefore, with probability at least $1-\delta$,

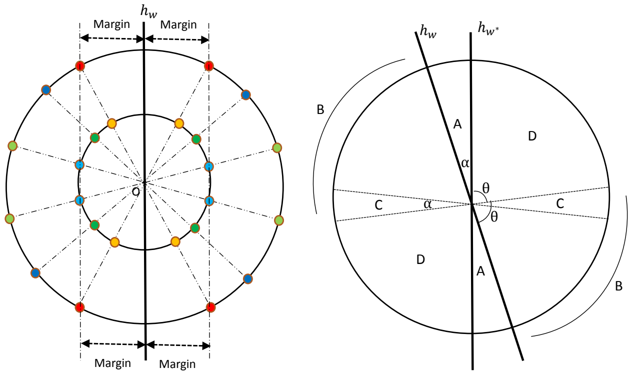
$$\sup_{w^*,\{y_k\}_{k=1}^{m_k},w} \left|\mathbb{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T')\right| \leq \lambda/12.$$

■

Appendix F. Lower Bound under Bounded Noise

Theorem 15 (restated). *For every bounded noise parameter $0 \leq \beta < 1$, there exists a distribution $\tilde{D}_\beta \in \mathcal{P}_\beta$ (that is, a distribution over $\mathbb{R}^2 \times \{+1,-1\}$, where the marginal distribution on $\mathbb{R}^2$ is uniform over the unit ball, and the labels $\{+1,-1\}$ satisfies β-bounded noise condition) such that any proper loss minimization is not consistent on $\tilde{D}_\beta$ w.r.t. the class of halfspaces. That is, there exists an $\epsilon \geq 0$ and a sample size $m(\epsilon)$ such that any proper loss minimization will output a classifier of excess error larger than ϵ by a high probability over sample size at least $m(\epsilon)$.*

Proof We prove the theorem by constructing a distribution $\tilde{D}_\beta \in \mathcal{P}_\beta$ that is consistent with our conclusion. Since we have assumed that the marginal distribution over the instance space X is uniform over the unit ball, we now construct a noisy distribution on the label space for our purpose. To do so, given the Bayes optimal classifier h_{w^*} and some linear classifier h_w such that $\theta(h_{w^*}, h_w) = \alpha$, we first divide the instance space X into four areas A, B, C, and D, as shown in Figure 1(*b*). Namely, area A is the disagreement region between h_w and h_{w^*} with angle α, and the agreement region consists of areas B (points closer to h_w) and D (points closer to h_{w^*}). Area C, a wedge with an angle of α, is a part of area B. We flip the labels of all points in areas A and B with probability $\eta = (1-\beta)/2$, and retain the original labels of instances in area D. This setting naturally satisfies β-bounded noise condition. As we will show later, when the angle α is small enough, the expected value of proper loss of h_w over the whole instance space will be smaller than that of h_{w^*}. Then by the standard analysis of Awasthi et al. (2015a), we conclude that there exists an $\epsilon \geq 0$ and a sample



(*a*) Property 1. (*b*) Partition.

Figure 1: Demonstrating the construction for the lower bound.

size $m(\epsilon)$ such that any proper loss minimization will output a classifier of excess error larger than ϵ by a high probability over sample size at least $m(\epsilon)$.

We now show the key steps in our analysis. We consider here unit vectors w^* and w. Let cA, cB, cC, and cD be the proper loss of h_{w^*} on areas A, B, C, and D when the labels are correct, and let dA, dB, dC, and dD be the loss of h_{w^*} on areas A, B, C, and D when the labels are incorrect. By the symmetry property 1 in Definition 14, we have

$$cA = \frac{2}{\pi}\int_0^{\alpha}\int_0^1 \ell_+^{w^*}(z,\varphi)zdzd\varphi. \tag{11}$$

Similarly, we can calculate cB, cC, cD, dA, dB, dC, dD, and can check that

$$cA + cB = \frac{2}{\pi}\int_0^{\frac{\pi+\alpha}{2}}\int_0^1 \ell_+^{w^*}(z,\varphi)zdzd\varphi = cC + cD, \tag{12}$$

$$dA + dB = \frac{2}{\pi}\int_0^{\frac{\pi+\alpha}{2}}\int_0^1 \ell_-^{w^*}(z,\varphi)zdzd\varphi = dC + dD. \tag{13}$$

On the other side, according to the noisy distribution $\tilde{D}$ designed by us, the expected loss of h_{w^*} is

$$\mathcal{L}(h_{w^*}) = \eta(dA + dB) + (1-\eta)(cA + cB) + cD. \tag{14}$$

For h_w, as the role of B to h_w is the same as the role D to h_{w^*} by Property 1 in Definition 14, we have

$$\mathcal{L}(h_w) = \eta(cA + dD) + (1-\eta)(dA + cD) + cB. \tag{15}$$

Therefore, combining with Equations 12 and 13, we have

$$\mathcal{L}(h_w) - \mathcal{L}(h_{w^*}) = (1-\eta)(dA - cA) - \eta(dC - cC). \tag{16}$$

That is to say, once $\eta > \eta(\alpha) \triangleq \frac{dA-cA}{dA-cA+dC-cC}$, we will have $\mathcal{L}(h_w) < \mathcal{L}(h_{w^*})$. We now show that $\frac{dA-cA}{dA-cA+dC-cC}$ can be arbitrarily small when α approaches to zero, i.e., $\lim_{\alpha\to 0}\frac{dA-cA}{dA-cA+dC-cC} = 0$. To see this, let $f_{w^*}(z,\varphi) = \ell_-^{w^*}(z,\varphi) - \ell_+^{w^*}(z,\varphi)$, then

$$\begin{aligned}
&\lim_{\alpha\to 0}\frac{dA - cA}{dA - cA + dC - cC}\\
&= \lim_{\alpha\to 0}\frac{\frac{2}{\pi}\int_0^{\alpha}\int_0^1 f_{w^*}(z,\varphi)zdzd\varphi}{\frac{2}{\pi}\int_0^{\alpha}\int_0^1 f_{w^*}(z,\varphi)zdzd\varphi + \frac{4}{\pi}\int_{\frac{\pi-\alpha}{2}}^{\frac{\pi}{2}}\int_0^1 f_{w^*}(z,\varphi)zdzd\varphi}\\
&= \lim_{\alpha\to 0}\frac{\frac{2}{\pi}\int_0^1 f_{w^*}(z,\alpha)zdz}{\frac{2}{\pi}\int_0^1 f_{w^*}(z,\alpha)zdz + \frac{2}{\pi}\int_0^1 f_{w^*}(z,\frac{\pi-\alpha}{2})zdz} \qquad \text{(By L'Hospital's rule)}\\
&= \frac{\lim_{\alpha\to 0}\frac{2}{\pi}\int_0^1 f_{w^*}(z,\alpha)zdz}{\lim_{\alpha\to 0}\frac{2}{\pi}\int_0^1 f_{w^*}(z,\alpha)zdz + \frac{2}{\pi}\int_0^1 f_{w^*}(z,\frac{\pi-\alpha}{2})zdz} \qquad \text{(By existence of the limit)}\\
&= \frac{\int_0^1 f_{w^*}(z,0)zdz}{\int_0^1 f_{w^*}(z,0)zdz + \int_0^1 f_{w^*}(z,\frac{\pi}{2})zdz} \qquad \left(\text{By continuity of } \int_0^1 f_{w^*}(z,\alpha)zdz\right)\\
&= \frac{0}{0 + \int_0^1 f_{w^*}(z,\frac{\pi}{2})zdz}\\
&= 0. \qquad \left(\text{Since } \int_0^1 f_{w^*}\left(z,\frac{\pi}{2}\right)zdz > 0, \text{see Lemma 28}\right)
\end{aligned} \tag{17}$$

The following lemma guarantees the denominator of the last equation is non-zero:

Lemma 28 *For any continuous function $f_{w^*}(z, \varphi)$, we have*

$$\int_0^1 f_{w^*}\left(z, \frac{\pi}{2}\right) z dz > 0. \tag{18}$$

Proof In the close interval $[1/2, 1]$, since function $f_{w^*}(z, \frac{\pi}{2})$ is continuous, by extreme value theorem, there exists $\xi \in [1/2, 1]$ such that $\min_z f_{w^*}(z, \frac{\pi}{2})z = f_{w^*}(\xi, \frac{\pi}{2})\xi > 0$ (By Property 2 in Definition 14). So

$$\begin{aligned}
\int_0^1 f_{w^*}\left(z, \frac{\pi}{2}\right) z dz &= \int_0^{\frac{1}{2}} f_{w^*}\left(z, \frac{\pi}{2}\right) z dz + \int_{\frac{1}{2}}^1 f_{w^*}\left(z, \frac{\pi}{2}\right) z dz \\
&\geq \int_{\frac{1}{2}}^1 f_{w^*}\left(z, \frac{\pi}{2}\right) z dz \\
&\geq \frac{1}{2} \min_z f_{w^*}\left(z, \frac{\pi}{2}\right) z \\
&\geq \frac{1}{2} f_{w^*}\left(\xi, \frac{\pi}{2}\right) \xi \\
&> 0.
\end{aligned} \tag{19}$$

■

This completes our proof.

■

F.1. Proper Loss

In this section, we show that most of the commonly used loss functions, e.g., the (normalized) hinge loss, the logistic loss, the square loss, the exponential loss, the truncated quadratic loss, etc., together with their regularized versions, satisfy our conditions in Theorem 15. Thus one-shot minimization does not work for all these loss functions.

In particular, Theorem 15 requires the loss functions ℓ_+ and ℓ_- to be continuous and satisfy two properties in Definition 14. The first property in Definition 14 holds because the commonly used surrogate losses are functions of the distance to the classifier. The second property in Definition 14 are natural since to achieve low error, it is desirable to penalized misclassification more.

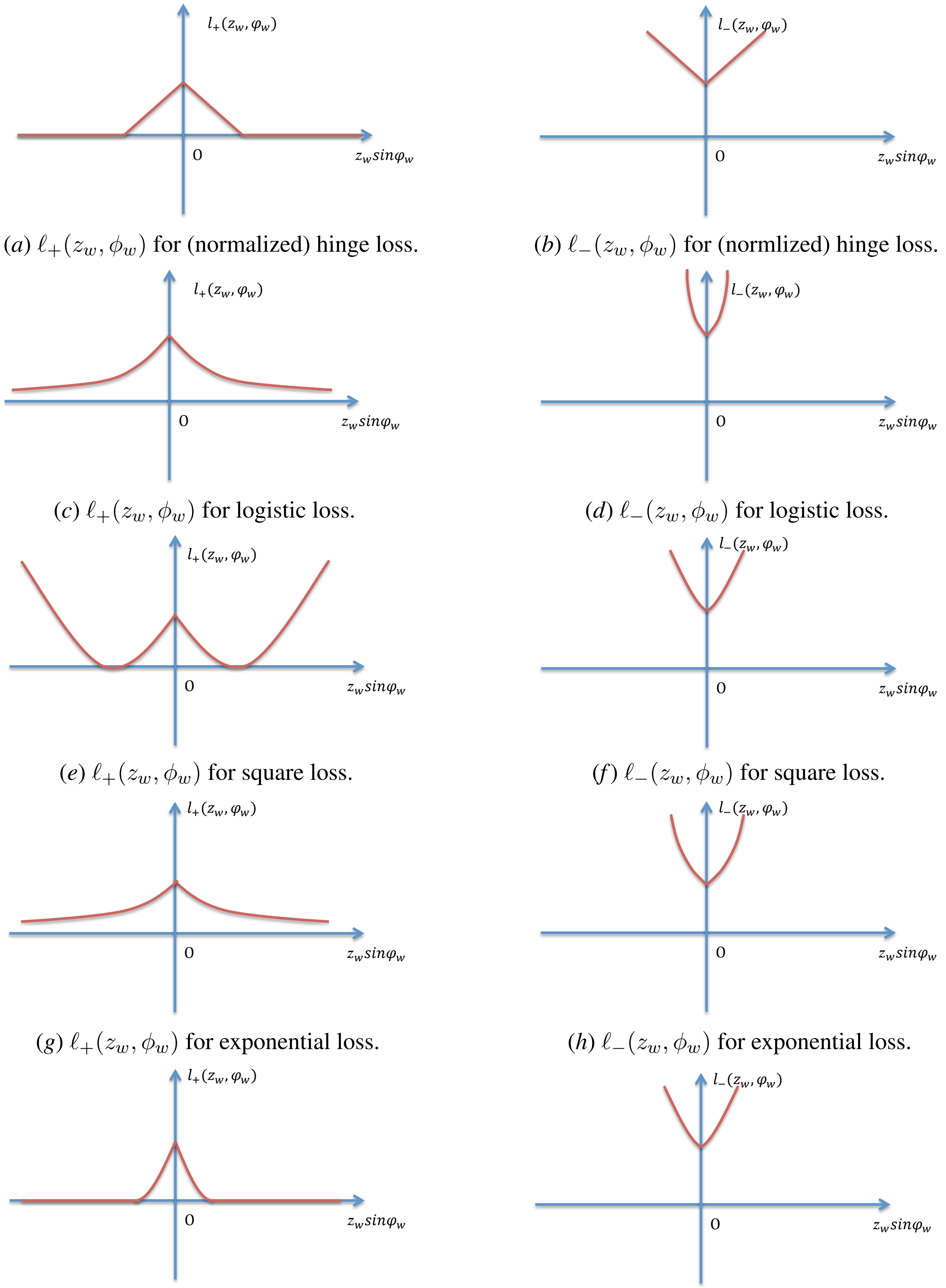
The following tables list some of the commonly used losses in both Cartesian coordinates and polar coordinates which are proper losses, with graphs shown in Figure 2. Here the polar coordinates are established in the sense that z_w represents the L_2 norm of the point and φ_w denotes the angle between the linear separator h_w and the vector from origin to that point. The coordinates for the same point in the two systems have the relation $w \cdot x = z_w \sin \varphi_w$.

Table 1: Definitions of $\ell_+(x)$ and $\ell_-(x)$ for commonly used loss functions in Cartesian coordinates.

	$\ell_+(x)$		$\ell_-(x)$	
	$w \cdot x \geq 0$	$w \cdot x < 0$	$w \cdot x \geq 0$	$w \cdot x < 0$
hinge	$(1 - w \cdot x/\tau)_+$	$(1 + w \cdot x/\tau)_+$	$w \cdot x/\tau$	$-w \cdot x/\tau$
logistic	$\ln(1 + e^{-2w \cdot x})$	$\ln(1 + e^{2w \cdot x})$	$\ln(1 + e^{2w \cdot x})$	$\ln(1 + e^{-2w \cdot x})$
square	$(1 - w \cdot x)^2$	$(1 + w \cdot x)^2$	$(1 + w \cdot x)^2$	$(1 - w \cdot x)^2$
exponential	$e^{-w \cdot x}$	$e^{w \cdot x}$	$e^{w \cdot x}$	$e^{-w \cdot x}$
trun-quadratic	$(1 - w \cdot x)_+^2$	$(1 + w \cdot x)_+^2$	$(1 + w \cdot x)^2$	$(1 - w \cdot x)^2$

Table 2: Definitions of $\ell_+(z_w, \varphi_w)$ and $\ell_-(z_w, \varphi_w)$ for commonly used loss functions in polar coordinates.

	$\ell_+(z_w, \varphi_w)$		$\ell_-(z_w, \varphi_w)$	
	$z_w \sin \varphi_w \geq 0$	$z_w \sin \varphi_w < 0$	$z_w \sin \varphi_w \geq 0$	$z_w \sin \varphi_w < 0$
hinge	$(1 - z_w \sin \varphi_w/\tau)_+$	$(1 + z_w \sin \varphi_w/\tau)_+$	$z_w \sin \varphi_w/\tau$	$-z_w \sin \varphi_w/\tau$
logistic	$\ln(1 + e^{-2z_w \sin \varphi_w})$	$\ln(1 + e^{2z_w \sin \varphi_w})$	$\ln(1 + e^{2z_w \sin \varphi_w})$	$\ln(1 + e^{-2z_w \sin \varphi_w})$
square	$(1 - z_w \sin \varphi_w)^2$	$(1 + z_w \sin \varphi_w)^2$	$(1 + z_w \sin \varphi_w)^2$	$(1 - z_w \sin \varphi_w)^2$
exponential	$e^{-z_w \sin \varphi_w}$	$e^{z_w \sin \varphi_w}$	$e^{z_w \sin \varphi_w}$	$e^{-z_w \sin \varphi_w}$
trun-quadratic	$(1 - z_w \sin \varphi_w)_+^2$	$(1 + z_w \sin \varphi_w)_+^2$	$(1 + z_w \sin \varphi_w)^2$	$(1 - z_w \sin \varphi_w)^2$



(*a*) $\ell_+(z_w, \phi_w)$ for (normalized) hinge loss.

(*b*) $\ell_-(z_w, \phi_w)$ for (normlized) hinge loss.

(*c*) $\ell_+(z_w, \phi_w)$ for logistic loss.

(*d*) $\ell_-(z_w, \phi_w)$ for logistic loss.

(*e*) $\ell_+(z_w, \phi_w)$ for square loss.

(*f*) $\ell_-(z_w, \phi_w)$ for square loss.

(*g*) $\ell_+(z_w, \phi_w)$ for exponential loss.

(*h*) $\ell_-(z_w, \phi_w)$ for exponential loss.

(*i*) $\ell_+(z_w, \phi_w)$ for truncated quadratic loss.

(*j*) $\ell_-(z_w, \phi_w)$ for truncated quadratic loss.

Figure 2: The graphs of $\ell_+(z_w, \varphi_w)$ and $\ell_-(z_w, \varphi_w)$ for (normalized) hinge loss, logistic loss, square loss, exponential loss, and truncated quadratic loss.