Efficient Learning of Linear Separators under Bounded Noise

Pranjal Awasthi PAWASHTI@CS.PRINCETON.EDU
Princeton University
Maria-Florina Balcan NINAMF@CS.CMU.EDU
Nika Haghtalab NHAGHTAL@CS.CMU.EDU
Carnegie Mellon University
Ruth Urner RURNER@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems

Abstract

We study the learnability of linear separators in $\Re^d$ in the presence of bounded (a.k.a Massart) noise. This is a realistic generalization of the random classification noise model, where the adversary can flip each example x with probability $\eta(x) \leq \eta$. We provide the first polynomial time algorithm that can learn linear separators to arbitrarily small excess error in this noise model under the uniform distribution over the unit sphere in $\Re^d$, for some constant value of η. While widely studied in the statistical learning theory community in the context of getting faster convergence rates, computationally efficient algorithms in this model had remained elusive. Our work provides the first evidence that one can indeed design algorithms achieving arbitrarily small excess error in polynomial time under this realistic noise model and thus opens up a new and exciting line of research.

We additionally provide lower bounds showing that popular algorithms such as hinge loss minimization and averaging cannot lead to arbitrarily small excess error under Massart noise, even under the uniform distribution. Our work, instead, makes use of a margin based technique developed in the context of active learning. As a result, our algorithm is also an active learning algorithm with label complexity that is only logarithmic in the desired excess error ϵ.

1. Introduction

Overview Linear separators are the most popular classifiers studied in both the theory and practice of machine learning. Designing noise tolerant, polynomial time learning algorithms that achieve arbitrarily small excess error rates for linear separators is a long-standing question in learning theory. In the absence of noise (when the data is realizable) such algorithms exist via linear programming (Cristianini and Shawe-Taylor, 2000). However, the problem becomes significantly harder in the presence of label noise. In particular, in this work we are concerned with designing algorithms that can achieve error $\mathrm{OPT} + \epsilon$ which is arbitrarily close to OPT, the error of the best linear separator, and run in time polynomial in $\frac{1}{\epsilon}$ and d (as usual, we call ϵ the *excess error*). Such strong guarantees are only known for the well studied random classification noise model (Blum et al., 1998). In this work, we provide the first algorithm that can achieve arbitrarily small excess error, in truly polynomial time, for bounded noise, also called Massart noise (Massart and Nédélec, 2006), a much more realistic and widely studied noise model in statistical learning theory (Bousquet et al., 2005). We additionally show strong lower bounds under the same noise model for two other computationally efficient learning algorithms (hinge loss minimization and the averaging algorithm), which could be of independent interest.

Motivation The work on computationally efficient algorithms for learning halfspaces has focused on two different extremes. On one hand, for the very stylized random classification noise model (RCN), where each example x is flipped independently with equal probability η, several works have provided computationally efficient algorithms that can achieve arbitrarily small excess error in polynomial time (Blum et al., 1998; Servedio, 2001; Balcan and Feldman, 2013) — note that all these results crucially exploit the high amount of symmetry present in the RCN noise. At the other extreme, there has been significant work on much more difficult and adversarial noise models, including the agnostic model (Kearns et al., 1992) and malicious noise models (Kearns and Li, 1988). The best results here however, not only require additional distributional assumptions about the marginal over the instance space, but they only achieve much weaker multiplicative approximation guarantees (Kalai et al., 2008b; Klivans et al., 2009; Awasthi et al., 2014); for example, the best results of this form for the case of uniform distribution over the unit sphere S_{d-1} include an efficient algorithm to achieve excess error $c\mathrm{OPT}$ (Awasthi et al., 2014) for a large constant c, and a subsequent PTAS to achieve excess error $(1+\mu)\mathrm{OPT}+\epsilon$ for any ϵ in run time that is exponential in $\frac{1}{\mu}$ and polynomial in $\frac{1}{\epsilon}$ (Daniely, 2015). While interesting from a technical point of view, guarantees of this form are somewhat troubling from a statistical point of view, as they are inconsistent, in the sense there is a barrier $O(\mathrm{OPT})$, after which we cannot prove that the excess error further decreases as we get more and more samples.

Our Results In this work we identify a realistic and widely studied noise model in the statistical learning theory, the so called Massart noise (Massart and Nédélec, 2006; Bousquet et al., 2005), for which we can prove much stronger guarantees. Massart noise can be thought of as a generalization of the random classification noise model where the label of each example x is flipped independently with probability $\eta(x) \leq \eta < 1/2$. The adversary has control over choosing a different noise rate $\eta(x) \leq \eta$ for every example x with the only constraint that $\eta(x) \leq \eta$. From a statistical point of view, it is well known that under this model, we can get faster rates compared to worst case joint distributions (Bousquet et al., 2005). In computational learning theory, this noise model was also studied, but under the name of malicious misclassification noise (Rivest and Sloan, 1994; Sloan, 1996). However due to its highly asymmetric nature, til date, computationally efficient learning algorithms in this model have remained elusive. In this work, we provide the first computationally efficient algorithm achieving arbitrarily small excess error for learning linear separators.

Formally, we show that there exists a polynomial time algorithm that can learn linear separators to error $\mathrm{OPT}+\epsilon$ and run in $\mathrm{poly}(d, \frac{1}{\epsilon})$ when the underlying distribution is the uniform distribution over the unit ball in $\Re^d$ and the noise of each example is upper bounded by a constant η (independent of the dimension).

As mentioned earlier, a result of this form was only known for random classification noise. From a technical point of view, as opposed to random classification noise, where the error of each classifier scales uniformly under the observed labels, the observed error of classifiers under Masasart noise could change drastically in a non-monotonic fashion. This is due to the fact that the adversary has control over choosing a different noise rate $\eta(x) \leq \eta$ for every example x. As a result, as we show in our work (see Section 4), standard algorithms such as the averaging algorithm (Servedio, 2001) which work for random noise can only achieve a much poorer excess error (as a function of η) under Massart noise. Technically speaking, this is due to the fact that Massart noise can introduce high correlations between the observed labels and the component orthogonal to the direction of the best classifier.

In face of these challenges, we take an entirely different approach than previously considered for random classification noise. Specifically, we analyze a recent margin based algorithm of Awasthi et al. (2014). This algorithm was designed for learning linear separators under agnostic and malicious noise models, and it was shown to achieve an excess error of $c\mathrm{OPT}$ for a constant c. By using new structural insights, we show that there exists a *constant* η (independent of the dimension), we can use a modification of the algorithm in Awasthi et al. (2014) and achieve arbitrarily small excess error under Massart noise bounded by η. One way to think about this result is that we define an adaptively chosen sequence of hinge loss minimization problems around smaller and smaller bands around the current guess for the target. We show by relating the hinge loss and 0/1-loss together with a careful localization analysis that these will direct us closer and closer to the optimal classifier, allowing us to achieve arbitrarily small excess error rates in polynomial time.

Given that our algorithm is an adaptively chosen sequence of hinge loss minimization problems, one might wonder what guarantee one-shot hinge loss minimization could provide. In Section 5, we show a strong negative result: for every τ, and $\eta \leq 1/2$, there is a noisy distribution $\tilde{D}$ over $\Re^d \times \{0, 1\}$ satisfying Massart noise with parameter η and an $\epsilon > 0$, such that τ-hinge loss minimization returns a classifier with excess error $\Omega(\epsilon)$. This result could be of independent interest. While there exists earlier work showing that hinge loss minimization can lead to classifiers of large $0/1$-loss (Ben-David et al., 2012), the lower bounds in that paper employ distributions with *significant mass* on *discrete points* with *flipped label* (which is not possible under Massart noise) at a very *large distance* from the optimal classifier. Thus, that result makes strong use of the hinge loss's sensitivity to errors at large distance. Here, we show that hinge loss minimization is bound to fail under much more benign conditions.

One appealing feature of our result is the algorithm we analyze is in fact naturally adaptable to the active learning or selective sampling scenario (intensively studied in recent years (Hanneke, 2007; Dasgupta, 2005; Hanneke, 2014), where the learning algorithms only receive the classifications of examples when they ask for them. We show that, in this model, our algorithms achieve a label complexity whose dependence on the error parameter ϵ is polylogarithmic (and thus exponentially better than that of any passive algorithm). This provides the first polynomial-time active learning algorithm for learning linear separators under Massart noise. We note that prior to our work only inefficient algorithms could achieve the desired label complexity under Massart noise (Balcan et al., 2007; Hanneke, 2014).

Related Work Agnostic learning is notoriously hard to deal with computationally and there is significant evidence that achieving arbitrarily small excess error in polynomial time is hard in this model (Arora et al., 1993; Guruswami and Raghavendra, 2006; Daniely et al., 2014). For this model, under our distributional assumptions, (Kalai et al., 2008b) provides an algorithm that learns linear separators in $\Re^d$ to excess error at most ϵ, but with running time $poly(d^{\exp(1/\epsilon)})$. Recent work shows evidence that the exponential dependence on $1/\epsilon$ is unavoidable in this case (Klivans and Kothari, 2014). We side-step this by considering a more structured, yet realistic noise model. Another line of work gives runtimes exponential in a margin parameter under different distributional assumptions (Ben-David and Simon, 2000; Shalev-Shwartz et al., 2011).

Motivated by the fact that many modern machine learning applications have massive amounts of unannotated or unlabeled data, there has been significant interest in designing active learning algorithms that most efficiently utilize the available data, while minimizing the need for human intervention. Over the past decade there has been substantial progress on understanding the underlying statistical principles of active learning, and several general characterizations have been developed for describing when active learning could have an advantage over the classical passive supervised learning paradigm both in the noise free settings and in the agnostic case (Freund et al., 1997; Dasgupta, 2005; Balcan et al., 2006, 2007; Hanneke, 2007; Dasgupta et al., 2007; Castro and Nowak, 2007; Dasgupta, 2011; Hanneke, 2014; Urner et al., 2013). However, despite many efforts, except for very simple noise models (random classification noise (Balcan and Feldman, 2013) and linear noise (Dekel et al., 2012)), to date there are no known computationally efficient algorithms with provable guarantees in the presence of Massart noise that can achieve arbitrarily small excess error.

We note that work of Hanneke and Yang (2014) provides computationally efficient algorithms for both passive and active learning under the assumption that the hinge loss (or other surrogate loss) minimizer aligns with the minimizer of the 0-1 loss. In our work (Section 5), we show that this is not the case under

Massart noise even when the marginal over the instance space is uniform, but still provide a computationally efficient algorithm for this much more challenging setting.

2. Preliminaries

We consider the binary classification problem; that is, we work on the problem of predicting a binary label y for a given instance x. We assume that the data points (x, y) are drawn from an unknown underlying distribution $\tilde{D}$ over $X \times Y$, where $X = \Re^d$ is the instance space and $Y = \{-1, 1\}$ is the label space. For the purpose of this work, we consider distributions where the marginal of $\tilde{D}$ over X is a uniform distribution on a d-dimensional unit ball. We work with the class of all homogeneous halfspaces, denoted by $\mathcal{H} = \{\mathrm{sign}(w \cdot x) : \ w \in \Re^d\}$. For a given halfspace $w \in \mathcal{H}$, we define the error of w with respect to $\tilde{D}$, by $\mathrm{err}_{\tilde{D}}(w) = \Pr_{(x,y)\sim\tilde{D}}[\mathrm{sign}(w \cdot x) \neq y]$.

We examine learning halfspaces in the presence of Massart noise. In this setting, we assume that the Bayes optimal classifier is a linear separator w^*. Note that w^* can have a non-zero error. Then Massart noise with parameter $\beta > 0$ is a condition such that for all x, the conditional label probability is such that

$$|\Pr(y = 1|x) - \Pr(y = -1|x)| \geq \beta. \tag{1}$$

Equivalently, we say that $\tilde{D}$ satisfies Massart noise with parameter β, if an adversary construct $\tilde{D}$ by first taking the distribution D over instances $(x, \mathrm{sign}(w^* \cdot x))$ and then flipping the label of an instance x with probability *at most* $\frac{1-\beta}{2}$. [1] Also note that under distribution $\tilde{D}$, w^* remains the Bayes optimal classier. In the remainder of this work, we refer to $\tilde{D}$ as the "noisy" distribution and to distribution D over instances $(x, \mathrm{sign}(w^* \cdot x))$ as the "clean" distribution.

Our goal is then to find a halfspace w that has small excess error, as compared to the Bayes optimal classifier w^*. That is, for any $\epsilon > 0$, find a halfspace w, such that $\mathrm{err}_{\tilde{D}}(w) - \mathrm{err}_{\tilde{D}}(w^*) \leq \epsilon$. Note that the excess error of any classifier w only depends on the points in the region where w and w^* disagree. So, $\mathrm{err}_{\tilde{D}}(w) - \mathrm{err}_{\tilde{D}}(w^*) \leq \frac{\theta(w,w^*)}{\pi}$, where $\theta(w, w^*) = \arccos(w \cdot w^*)$, is the angle between w and w^*. Additionally, under Massart noise the amount of noise in the disagreement region is also bounded by $\frac{1-\beta}{2}$. It is not difficult to see that under Massart noise,

$$\beta \, \frac{\theta(w, w^*)}{\pi} \leq \mathrm{err}_{\tilde{D}}(w) - \mathrm{err}_{\tilde{D}}(w^*). \tag{2}$$

In our analysis, we frequently examine the region within a certain margin of a halfspace. For a halfspace w and margin b, let $S_{w,b}$ be the set of all points that fall within a margin b from w, i.e., $S_{w,b} = \{x : \ |w \cdot x| \leq b\}$. For distributions $\tilde{D}$ and D, we indicate the distribution conditioned on $S_{w,b}$ by $\tilde{D}_{w,b}$ and $D_{w,b}$, respectively. In the remainder of this work, we refer to the region $S_{w,b}$ as "the band".

In our analysis, we use hinge loss, as a convex surrogate function for the 0/1-loss. For a halfspace w, we use τ-normalized hinge loss that is defined as $\ell(w, x, y) = \max\{0, 1 - \frac{(w \cdot x)y}{\tau}\}$. For a labeled sample set W, let $\ell(w, W) = \frac{1}{|W|} \sum_{(x,y)\in W} \ell(w, x, y)$ be the empirical hinge loss of a vector w with respect to W.

3. Computationally Efficient Algorithm for Massart Noise

In this section, we prove our main result for learning half-spaces under Massart noise. We focus on the case where D is the uniform distribution on the d-dimensional unit ball. Our main Theorem is as follows.

[1] Note that the relationship between Massart noise parameter β, and the maximum flipping probability discussed in the introduction η, is $\eta = \frac{1-\beta}{2}$.

Theorem 1 *Let the bayes classifier be a half-space denoted by w^*. Assume that the massart noise condition holds for some $\beta > 1 - 3.6 \times 10^{-6}$ with the marginal over $\Re^d$ being the uniform distribution over the unit ball S_{d-1} and $d > 20$. Then for any ϵ, $\delta > 0$, Algorithm 1 with $\lambda = 10^{-8}$, $\alpha_k = 0.038709\pi(1-\lambda)^{k-1}$, $b_{k-1} = \frac{2.3463\alpha_k}{\sqrt{d}}$, and $\tau_k = \sqrt{2.50306}\ (3.6 \times 10^{-6})^{1/4} b_{k-1}$, 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 $m_k = O(d(d + \log(k/\delta)))$ labels and with probability $(1-\delta)$ returns a linear separator that has excess error (compared to w^*) of at most ϵ.*

Note that in the above theorem and Algorithm 1, the value of β is unknown to the algorithm, and therefore, our results are adaptive to values of β within the acceptable range defined by the theorem.

The algorithm described below is similar to that of Awasthi et al. (2014) and uses an iterative margin-based approach. The algorithm runs for $s = \log_{\frac{1}{1-\lambda}}(\frac{1}{\epsilon})$ rounds for a constant $\lambda \in (0, 1]$. By induction assume that our algorithm produces a hypothesis w_{k-1} at round $k-1$ such that $\theta(w_{k-1}, w^*) \leq \alpha_k$. We satisfy the base case by using an algorithm of Klivans et al. (2009). At round k, we sample m_k labeled examples from the conditional distribution $\tilde{D}_{w_{k-1}, b_{k-1}}$ which is the uniform distribution over $\{x : |w_{k-1} \cdot x| \leq b_{k-1}\}$. We then choose w_k from the set of all hypothesis $B(w_{k-1}, \alpha_k) = \{w : \ \theta(w, w_{k-1}) \leq \alpha_k\}$ such that w_k minimizes the empirical hinge loss over these examples. Subsequently, as we prove in detail later, $\theta(w_k, w^*) \leq \alpha_{k+1}$. Note that for any w, the excess error of w is at most the error of w on $\tilde{D}$ when the labels are corrected according to w^*, i.e., $\mathrm{err}_{\tilde{D}}(w) - \mathrm{err}_{\tilde{D}}(w^*) \leq \mathrm{err}_D(w)$. Moreover, when D is uniform, $\mathrm{err}_D(w) = \frac{\theta(w^*, w)}{\pi}$. Hence, $\theta(w_s, w^*) \leq \pi\epsilon$ implies that w_s has excess error of at most ϵ.

```
Algorithm 1 EFFICIENT ALGORITHM FOR ARBITRARILY SMALL EXCESS ERROR FOR MASSART NOISE
Input: An oracle that returns $x$ and an oracle that returns $y$ for a $(x, y)$ sampled from an unknown distribution $\tilde{D}$. Permitted excess error $\epsilon$ and probability of failure $\delta$.
Parameters: A learning rate $\lambda$; a sequence of sample sizes $m_k$; a sequence of angles of the hypothesis space $\alpha_k$; a sequence of widths of the labeled space $b_k$; a sequence of thresholds of hinge-loss $\tau_k$.
Algorithm:
  1. Take $\mathrm{poly}(d, \frac{1}{\delta})$ samples and run $\mathrm{poly}(d, \frac{1}{\delta})$-time algorithm by Klivans et al. (2009) to find a half-space $w_0$ with excess error $0.0387089$ such that $\theta(w^*, w_0) \leq 0.038709\pi$ (Refer to Appendix D)
  2. Draw $m_1$ examples $(x, y)$ from $\tilde{D}$ and put them into a working set $W$.
  3. For $k = 1, \ldots, \log_{(\frac{1}{1-\lambda})}(\frac{1}{\epsilon}) = s$.
     (a) Find $v_k$ such that $\|v_k - w_{k-1}\| < \alpha_k$ (so $v_k \in B(w_{k-1}, \alpha_k)$), that minimizes the empirical hinge loss over $W$ using threshold $\tau_k$, i.e., $\ell_{\tau_k}(v_k, W) \leq \min_{w \in B(w_{k-1}, \alpha_k)} \ell_{\tau_k}(w, W) + 10^{-8}$.
     (b) Clear the working set $W$.
     (c) Normalize $v_k$ to $w_k = \frac{v_k}{\|v_k\|_2}$. Until $m_{k+1}$ additional examples are put in $W$: Draw an example $(x, y)$ from $\tilde{D}$. If $|w_k \cdot x| \geq b_k$, then reject $x$, else put $(x, y)$ into $W$.
Output: Return $w_s$, which has excess error $\epsilon$ with probability $1 - \delta$.
```

This algorithm was originally introduced to achieve an error of $c \cdot \mathrm{err}(w^*)$ for some constant c in presence of adversarial noise. Achieving arbitrary small excess error $\mathrm{err}(w^*) + \epsilon$ is a much more ambitious goal – one that requires new technical insights. Our two crucial technical innovations are as follow: We first make a key observation that under Massart noise, the noise rate over any conditional distribution $\tilde{D}$ is still at most $\frac{1-\beta}{2}$. Therefore, as we focus on the distribution within the band, our noise rate does not increase. Our second technical contribution is a careful choice of parameters. Indeed the choice of parameters, upto a constant, plays an important role in tolerating a constant amount of Massart noise. Using these insights, we show that the algorithm by Awasthi et al. (2014) can indeed achieve a much stronger guarantee, namely arbitrarily small excess error in presence of Massart noise. That is, for any ϵ, this algorithm can achieve error of $\mathrm{err}(w^*) + \epsilon$ in the presence of Massart noise.

Overview of our analysis: Similar to Awasthi et al. (2014), we divide $\mathrm{err}_D(w_k)$ to two categories; error in the band, i.e., on $x \in S_{w_{k-1},b_{k-1}}$, and error outside the band, on $x \notin S_{w_{k-1},b_{k-1}}$. We choose b_{k-1} and α_k such that, for every hypothesis $w \in B(w_{k-1}, \alpha_k)$ that is considered at step k, the probability mass outside the band such that w and w^* also disagree is very small (Lemma 8). Therefore, the error associated with the region outside the band is also very small. This motivates the design of the algorithm to only minimize the error in the band. Furthermore, the probability mass of the band is also small enough such that for $\mathrm{err}_D(w_k) \leq \alpha_{k+1}$ to hold, it suffices for w_k to have a small constant error over the clean distribution restricted to the band, namely $D_{w_{k-1},b_{k-1}}$.

This is where minimizing hinge loss in the band comes in. As minimizing the 0/1-loss is NP-hard, an alternative method for finding w_k with small error in the band is needed. Hinge loss that is a convex loss function can be efficiently minimized. So, we can efficiently find w_k that minimizes the empirical hinge loss of the sample drawn from $\tilde{D}_{w_{k-1},b_{k-1}}$. To allow the hinge loss to remain a faithful proxy of 0/1-loss as we focus on bands with smaller widths, we use a normalized hinge loss function defined by $\ell_\tau(w, x, y) = \max\{0, 1 - \frac{w \cdot xy}{\tau}\}$.

A crucial part of our analysis involves showing that if w_k minimizes the empirical hinge loss of the sample set drawn from $\tilde{D}_{w_{k-1},b_{k-1}}$, it indeed has a small 0/1-error on $D_{w_{k-1},b_{k-1}}$. To this end, we first show that when τ_k is proportional to b_k, the hinge loss of w^* on $D_{w_{k-1},b_{k-1}}$, which is an upper bound on the 0/1-error of w_k in the band, is itself small (Lemma 2). Next, we notice that under Massart noise, the noise rate in any marginal of the distribution is still at most $\frac{1-\beta}{2}$. Therefore, focusing the distribution in the band does not increase the probability of noise in the band. Moreover, the noise points in the band are close to the decision boundary so intuitively speaking, they can not increase the hinge loss too much. Using these insights we can show that the hinge loss of w_k on $\tilde{D}_{w_{k-1},b_{k-1}}$ is close to its hinge loss on $D_{w_{k-1},b_{k-1}}$ (Lemma 3).

Proof of Theorem 1 and related lemmas

To prove Theorem 1, we first introduce a series of lemmas concerning the behavior of hinge loss in the band. These lemmas build up towards showing that w_k has error of at most a fixed small constant in the band.

For ease of exposition, for any k, let D_k and $\tilde{D}_k$ represent $D_{w_{k-1},b_{k-1}}$ and $\tilde{D}_{w_{k-1},b_{k-1}}$, respectively, and $\ell(\cdot)$ represent $\ell_{\tau_k}(\cdot)$. Furthermore, let $c = 2.3463$, such that $b_{k-1} = \frac{c\alpha_k}{\sqrt{d}}$.

Our first lemma, whose proof appears in Appendix B, provides an upper bound on the true hinge error of w^* on the clean distribution in the band.

Lemma 2 $\mathbb{E}_{(x,y)\sim D_k} \ell(w^*, x, y) \leq 0.665769 \frac{\tau}{b}$.

The next Lemma compares the true hinge loss of any $w \in B(w_{k-1}, \alpha_k)$ on two distributions, $\tilde{D}_k$ and D_k. It is clear that the difference between the hinge loss on these two distributions is entirely attributed to the noise points and their margin from w. A key insight in the proof of this lemma is that as we concentrate in the band, the probability of seeing a noise point remains under $\frac{1-\beta}{2}$. This is due to the fact that under Massart noise, each label can be changed with probability at most $\frac{1-\beta}{2}$. Furthermore, by concentrating in the band all points are close to the decision boundary of w_{k-1}. Since w is also close in angle to w_{k-1}, then points in the band are also close to the decision boundary of w. Therefore the hinge loss of noise points in the band can not increase the total hinge loss of w by too much. This lemma sheds light on the importance of concentrating the distribution in the band together with restricting the search for the hinge minimizer to $B(w_{k-1}, \alpha_k)$.

Lemma 3 *For any w such that $w \in B(w_{k-1}, \alpha_k)$, we have*

$$|\mathbb{E}_{(x,y)\sim D_k}\ell(w,x,y) - \mathbb{E}_{(x,y)\sim \tilde{D}_k}\ell(w,x,y)| \leq 1.092\sqrt{2}\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k}.$$

Proof Let N be the set of noise points. We have,

$$\begin{aligned}
|\mathbb{E}_{(x,y)\sim \tilde{D}_k}\ell(w,x,y) - \mathbb{E}_{(x,y)\sim D_k}\ell(w,x,y)| &= |\mathbb{E}_{(x,y)\in \tilde{D}_k}\left(\ell(w,x,y) - \ell(w,x,\mathrm{sign}(w^*\cdot x))\right)| \\
&\leq \mathbb{E}_{(x,y)\sim \tilde{D}_k}\left(\mathbf{1}_{x\in N}(\ell(w,x,y) - \ell(w,x,-y))\right) \\
&\leq 2\mathbb{E}_{(x,y)\sim \tilde{D}_k}\left(\mathbf{1}_{x\in N}\frac{|w\cdot x|}{\tau_k}\right) \\
&\leq \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim \tilde{D}_k}(x\in N)} \times \sqrt{\mathbb{E}_{(x,y)\sim \tilde{D}_k}(w\cdot x)^2} \quad \text{(By Cauchy Shwarz)} \\
&\leq \frac{2}{\tau_k}\sqrt{\frac{1-\beta}{2}}\sqrt{\frac{\alpha_k^2}{d-1} + b_{k-1}^2} \quad \text{(By Lemma A.2 of (Awasthi et al., 2014) for uniform)} \\
&\leq \sqrt{2}\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k}\sqrt{\frac{d}{(d-1)c^2} + 1} \\
&\leq 1.092\sqrt{2}\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k} \qquad \text{(for } d > 20, c > 1\text{)}
\end{aligned}$$

■

For a labeled sample set W drawn at random from $\tilde{D}_k$, let $\mathrm{cleaned}(W)$ be the set of samples with the labels corrected by w^*, i.e., $\mathrm{cleaned}(W) = \{(x, \mathrm{sign}(w^*\cdot x)) : \text{ for all } (x,y) \in W\}$. Then by standard VC-dimension bounds (Proof included in Appendix C) there is $m_k \in O(d(d + \log(k/d)))$ such that for any randomly drawn set W of m_k labeled samples from $\tilde{D}_k$, with probability $1 - \frac{\delta}{2(k+k^2)}$, for any $w \in B(w_{k-1}, \alpha_k)$,

$$|\mathbb{E}_{(x,y)\sim \tilde{D}_k}\ell(w,x,y) - \ell(w,W)| \leq 10^{-8}, \tag{3}$$

$$|\mathbb{E}_{(x,y)\sim D_k}\ell(w,x,y) - \ell(w,\mathrm{cleaned}(W))| \leq 10^{-8}. \tag{4}$$

The next lemma is a crucial step in our analysis of Algorithm 1. This lemma proves that if $w_k \in B(w_{k-1}, \alpha_k)$ minimizes the empirical hinge loss on the sample drawn from the noisy distribution in the band, namely $\tilde{D}_{w_{k-1},b_{k-1}}$, then with high probability w_k also has a small 0/1-error with respect to the clean distribution in the band, i.e., $D_{w_{k-1},b_{k-1}}$.

Lemma 4 *There exists $m_k \in O(d(d+\log(k/d)))$, such that for a randomly drawn labeled sampled set W of size m_k from $\tilde{D}_k$, and for w_k such that w_k has the minimum empirical hinge loss on W between the set of all hypothesis in $B(w_{k-1}, \alpha_k)$, with probability $1 - \frac{\delta}{2(k+k^2)}$,*

$$\mathrm{err}_{D_k}(w_k) \leq 0.757941\frac{\tau_k}{b_{k-1}} + 3.303\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k} + 3.28\times 10^{-8}.$$

Proof Sketch 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. Lemma 2 provides an upper bound on the true hinge loss on distribution D_k. Therefore, it remains to create a connection between the empirical hinge loss of w_k on the sample drawn from $\tilde{D}_k$ to its true hinge loss on distribution D_k. This, we achieve by using the generalization bounds of Equations 3 and 4 to connect the empirical and true hinge loss of w_k and w^*, and using Lemma 3 to connect the hinge of w_k and w^* in the clean and noisy distributions. The details of this derivation are deferred to Appendix B. ■

Proof of Theorem 1 Recall that $c = 2.3463$, $\lambda = 10^{-8}$, $\alpha_k = 0.038709\pi(1-\lambda)^{k-1}$, $b_{k-1} = \frac{c\alpha_k}{\sqrt{d}}$, $\tau_k = \sqrt{2.50306}\,(3.6 \times 10^{-6})^{1/4} b_{k-1}$, and $\beta > 1 - 3.6 \times 10^{-6}$.

Note that for any w, the excess error of w is at most the error of w on the clean distribution D, i.e., $\mathrm{err}_{\tilde{D}}(w) - \mathrm{err}_{\tilde{D}}(w^*) \leq \mathrm{err}_D(w)$. Moreover, for uniform distribution D, $\mathrm{err}_D(w) = \frac{\theta(w^*, w)}{\pi}$. Hence, to show that w has ϵ excess error, it suffices to show that $\mathrm{err}_D(w) \leq \epsilon$.

Our goal is to achieve excess error of $0.038709(1-\lambda)^k$ at round k. This we do indirectly by bounding $\mathrm{err}_D(w_k)$ at every step. We use induction. For $k = 0$, we use the algorithm for adversarial noise model by Klivans et al. (2009), which can achieve excess error of ϵ if $\mathrm{err}_{\tilde{D}}(w^*) < \frac{\epsilon^2}{256 \log(1/\epsilon)}$ (Refer to Appendix D for more details). For Massart noise, $\mathrm{err}_{\tilde{D}}(w^*) \leq \frac{1-\beta}{2}$. So, for our choice of β, this algorithm can achieve excess error of 0.0387089 in $\mathrm{poly}(d, \frac{1}{\delta})$ samples and run-time. Furthermore, using Equation 2, $\theta(w_0, w^*) < 0.038709\pi$.

Assume that at round $k-1$, $\mathrm{err}_D(w_{k-1}) \leq 0.038709(1-\lambda)^{k-1}$. We will show that w_k, which is chosen by the algorithm at round k, also has $\mathrm{err}_D(w_k) \leq 0.038709(1-\lambda)^k$.

First note that $\mathrm{err}_D(w_{k-1}) \leq 0.038709(1-\lambda)^{k-1}$ implies $\theta(w_{k-1}, w^*) \leq \alpha_k$. Let $S = S_{w_{k-1}, b_{k-1}}$ indicate the band at round k. We divide the error of w_k to two parts, error outside the band and error inside of the band. That is

$$\mathrm{err}_D(w_k) = \Pr_{x \sim D}[x \notin S \text{ and } (w_k \cdot x)(w^* \cdot x) < 0] + \Pr_{x \sim D}[x \in S \text{ and } (w_k \cdot x)(w^* \cdot x) < 0].$$

For the first part, by the application of Lemma 8 and the fact that $\theta(w_{k-1}, w_k) \leq \alpha_k$ and $\theta(w_{k-1}, w^*) \leq \alpha_k$, the error outside of the band, i.e., $\Pr_{x \sim D}[x \notin S \text{ and } (w_k \cdot x)(w^* \cdot x) < 0]$, is at most

$$\Pr_{x \sim D}[x \notin S \text{ and } (w_k \cdot x)(w_{k-1} \cdot x) < 0] + \Pr_{x \sim D}[x \notin S \text{ and } (w_{k-1} \cdot x)(w^* \cdot x) < 0] \leq \frac{2\alpha_k}{\pi} e^{-\frac{c^2(d-2)}{2d}}.$$

For the second part, i.e., error inside the band

$$\begin{aligned}
\Pr_{x \sim D}[x \in S \text{ and } (w_k \cdot x)(w^* \cdot x) < 0] &= \mathrm{err}_{D_k}(w_k) \Pr_{x \sim D}[x \in S] \leq \mathrm{err}_{D_k}(w_k) \frac{V_{d-1}}{V_d} 2\, b_{k-1} \quad \text{(By Lemma 7)} \\
&\leq \mathrm{err}_{D_k}(w_k)\, c\, \alpha_k \sqrt{\frac{2(d+1)}{\pi d}},
\end{aligned}$$

where V_d is the volume of the d-dimensional unit ball and the last transition holds by the fact that $\frac{V_{d-1}}{V_d} \leq \sqrt{\frac{d+1}{2\pi}}$ (Borgwardt, 1987). Replacing an upper bound on $\mathrm{err}_{D_k}(w_k)$ from Lemma 4, to show that $\mathrm{err}_D(w_k) \leq \frac{\alpha_{k+1}}{\pi}$, it suffices to show that the following inequality holds.

$$\left(0.757941 \frac{\tau_k}{b_{k-1}} + 3.303\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k} + 3.28 \times 10^{-8}\right) c\, \alpha_k \sqrt{\frac{2(d+1)}{\pi d}} + \frac{2\alpha_k}{\pi} e^{-\frac{c^2(d-2)}{2d}} \leq \frac{\alpha_{k+1}}{\pi}.$$

We simplify this inequality as follows.

$$\left(0.757941 \frac{\tau_k}{b_{k-1}} + 3.303\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k} + 3.28 \times 10^{-8}\right) c \sqrt{\frac{2\pi(d+1)}{d}} + 2e^{-\frac{c^2(d-2)}{2d}} \leq 1 - \lambda.$$

Replacing in the r.h.s., the values of $c = 2.3463$, and $\tau_k = \sqrt{2.50306}(3.6 \times 10^{-6})^{1/4} b_{k-1}$, we have

$$\begin{aligned}
&\left(\sqrt{2.50306}(3.6 \times 10^{-6})^{1/4} + \sqrt{2.50306}\frac{\sqrt{1-\beta}}{(3.6 \times 10^{-6})^{1/4}} + 3.28 \times 10^{-8}\right) c \sqrt{\frac{2\pi(d+1)}{d}} + 2e^{-\frac{c^2(d-2)}{2d}} \\
&\leq 5.88133 \left(2\sqrt{2.50306}(3.6 \times 10^{-6})^{1/4} + 3.28 \times 10^{-8}\right) \sqrt{\frac{21}{20}} + 0.167935 \qquad \text{(For } d > 20\text{)} \\
&\leq 0.998573 < 1 - \lambda
\end{aligned}$$

Therefore, $\mathrm{err}_D(w_k) \leq 0.038709(1-\lambda)^k$.

Sample complexity analysis: We require m_k labeled samples in the band $S_{w_{k-1},b_{k-1}}$ at round k. By Lemma 7, the probability that a randomly drawn sample from $\tilde{D}$ falls in $S_{w_{k-1},b_{k-1}}$ is at least $O(b_{k-1}\sqrt{d}) = O((1-\lambda)^{k-1})$. Therefore, we need $O((1-\lambda)^{k-1} m_k)$ unlabeled samples to get m_k examples in the band with probability $1 - \frac{\delta}{8(k+k^2)}$. So, the total unlabeled sample complexity is at most

$$\sum_{k=1}^{s} O\left((1-\lambda)^{k-1} m_k\right) \leq s \sum_{k=1}^{s} m_k \in O\left(\frac{1}{\epsilon} \log\left(\frac{d}{\epsilon}\right)\left(d + \log\frac{\log(1/\epsilon)}{\delta}\right)\right).$$

■

As noted above, our analysis of the algorithm can tolerate $\beta \geq 1 - 3.6 \times 10^{-6}$, which is equivalent to maximum probability of flipping the label $\eta \leq 1.8 \times 10^{-6}$. This is due to a tradeoff involved in ensuring that hinge loss remains a good proxy for $0/1$ loss at every round. As we show in Section 5, algorithms that do not take such measures fail under the bounded noise model. Two factors play a role in our analysis. First, the hinge loss of w^* in the band, and second, the degree to which hinge loss minimization on noisy and clean distributions differs (because minimization is done over the noisy distribution). By Lemma 2, the first factor is proportional to $\frac{\tau}{b}$, while, by Lemma 3, the second factor is proportional to $\sqrt{1-\beta}\frac{b}{\tau}$. Therefore, the noise tolerated by our method has to balance out the quality of hinge loss as a surrogate loss function in these two factors, as shown in the statement of Lemma 4. The optimal choice of parameters for this tradeoff leads to the value of $\eta \leq 1.8 \times 10^{-6}$.

4. Average Does Not Work

Our algorithm described in the previous section uses convex loss minimization (in our case, hinge loss) in the band as an efficient proxy for minimizing the $0/1$ loss. The Average algorithm introduced by Servedio (2001) is another computationally efficient algorithm that has provable noise tolerance guarantees under certain noise models and distributions. For example, it achieves arbitrarily small excess error in the presence of random classification noise and monotonic noise when the distribution is uniform over the unit sphere. Furthermore, even in the presence of a small amount of malicious noise and less symmetric distributions, Average has been used to obtain a weak learner, which can then be boosted to achieve a non-trivial noise tolerance (Klivans et al., 2009). Therefore it is natural to ask, *whether the noise tolerance that Average exhibits could be extended to the case of Massart noise under the uniform distribution?* We answer this question in the negative. We show that the lack of symmetry in Massart noise presents a significant barrier for the one-shot application of Average, even when the marginal distribution is completely symmetric. Additionally, we also discuss obstacles in incorporating Average as a weak learner with the margin-based technique.

In a nutshell, Average takes m sample points and their respective labels, $W = \{(x^1, y^1), \ldots, (x^m, y^m)\}$, and returns $\frac{1}{m}\sum_{i=1}^{m} x^i y^i$. Our main result in this section shows that for a wide range of distributions that are very symmetric in nature, including the Gaussian and the uniform distribution, there is an instance of Massart noise under which Average can not achieve an arbitrarily small excess error.

Theorem 5 *For any continuous distribution D with a p.d.f. that is a function of the distance from the origin only, there is a noisy distribution $\tilde{D}$ over $X \times \{0,1\}$ that satisfies Massart noise condition in Equation 1 for some parameter $\beta > 0$ and Average returns a classifier with excess error $\Omega(\frac{\beta(1-\beta)}{1+\beta})$.*

Proof Let $w^* = (1, 0, \ldots, 0)$ be the target halfspace. Let the noise distribution be such that for all x, if $x_1 x_2 < 0$ then we flip the label of x with probability $\frac{1-\beta}{2}$, otherwise we keep the label. Clearly, this satisfies Massart noise with parameter β. Let w be expected vector returned by Average. We first show that w is far from w^* in angle. Then, using Equation 2 we show that w has large excess error.

First we examine the expected component of w that is parallel to w^*, i.e., $w \cdot w^* = w_1$. For ease of exposition, we divide our analysis to two cases, one for regions with no noise (first and third quadrants) and second for regions with noise (second and fourth quadrants). Let E be the event that $x_1 x_2 > 0$. By symmetry, it is easy to see that $\Pr[E] = 1/2$. Then $\mathbb{E}[w \cdot w^*] = \Pr(E)\, \mathbb{E}[w \cdot w^* | E] + \Pr(\bar{E})\, \mathbb{E}[w \cdot w^* | \bar{E}]$.

For the first term, for $x \in E$ the label has not changed. So, $\mathbb{E}[w \cdot w^* | E] = \mathbb{E}[|x_1| \; |E] = \int_0^1 z f(z)$. For the second term, the label of each point stays the same with probability $\frac{1+\beta}{2}$ and is flipped with probability $\frac{1-\beta}{2}$. Hence, $\mathbb{E}[w \cdot w^* | E] = \beta\, \mathbb{E}[|x_1| \; |E] = \beta \int_0^1 z f(z)$. Therefore, the expected parallel component of w is $\mathbb{E}[w \cdot w^*] = \frac{1+\beta}{2} \int_0^1 z f(z)$

Next, we examine w_2, the orthogonal component of w on the second coordinate. Similar to the previous case for the clean regions $\mathbb{E}[w_2 | E] = \mathbb{E}[|x_2| \; |E] = \int_0^1 z f(z)$. Next, for the second and forth quadrants, which are noisy, we have

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim\tilde{D}}[x_2 y | x_1 x_2 < 0] &= (\frac{1+\beta}{2}) \int_{-1}^{0} z \frac{f(z)}{2} + (\frac{1-\beta}{2}) \int_{-1}^{0} (-z) \frac{f(z)}{2} && \text{(Fourth quadrant)} \\
&\quad + (\frac{1+\beta}{2}) \int_{0}^{1} (-z) \frac{f(z)}{2} + (\frac{1-\beta}{2}) \int_{0}^{1} z \frac{f(z)}{2} && \text{(Second quadrant)} \\
&= -(\frac{1+\beta}{2}) \int_{0}^{1} z \frac{f(z)}{2} + (\frac{1-\beta}{2}) \int_{0}^{1} z \frac{f(z)}{2} \\
&\quad - (\frac{1+\beta}{2}) \int_{0}^{1} z \frac{f(z)}{2} + (\frac{1-\beta}{2}) \int_{0}^{1} z \frac{f(z)}{2} && \text{(By symmetry)} \\
&= -\beta \int_0^1 z f(z).
\end{aligned}$$

So, $w_2 = \left(\frac{1-\beta}{2}\right) \int_0^1 z f(z)$. Therefore $\theta(w, w^*) = \arctan(\frac{1-\beta}{1+\beta}) \geq \frac{1-\beta}{(1+\beta)}$. By Equation 2, we have $\text{err}_{\tilde{D}}(w) - \text{err}_{\tilde{D}}(w^*) \geq \beta\, \frac{\theta(w,w^*)}{\pi} \geq \beta \frac{1-\beta}{\pi(1+\beta)}$. ■

Our margin-based analysis from Section 3 relies on using hinge-loss minimization in the band at every round to efficiently find a halfspace w_k that is a weak learner for D_k, i.e., $\text{err}_{D_k}(w_k)$ is at most a small constant, as demonstrated in Lemma 4. Motivated by this more lenient goal of finding a weak learner, one might ask whether Average, as an efficient algorithm for finding low error halfspaces, can be incorporated with the margin-based technique in the same way as hinge loss minimization? We argue that the margin-based technique is inherently incompatible with Average.

The Margin-based technique maintains two key properties at every step: First, the angle between w_k and w_{k-1} and the angle between $w_{k-1}$and w^* are small, and as a result $\theta(w^*, w_k)$ is small. Second, w_k

is a weak learner with $\mathrm{err}_{D_{k-1}}(w_k)$ at most a small constant. In our work, hinge loss minimization in the band guarantees both of these properties simultaneously by limiting its search to the halfspaces that are close in angle to w_{k-1} and limiting its distribution to $D_{w_{k-1},b_{k-1}}$. However, in the case of Average as we concentrate in the band $D_{w_{k-1},b_{k-1}}$ we bias the distributions towards its orthogonal component with respect to w_{k-1}. Hence, an upper bound on $\theta(w^*, w_{k-1})$ only serves to assure that most of the data is orthogonal to w^* as well. Therefore, informally speaking, we lose the signal that otherwise could direct us in the direction of w^*. More formally, consider the construction from Theorem 5 such that $w_{k-1} = w^* = (1, 0, \dots, 0)$. In distribution $D_{w_{k-1},b_{k-1}}$, the component of w_k that is parallel to w_{k-1} scales down by the width of the band, b_{k-1}. However, as most of the probability stays in a band passing through the origin in any log-concave (including Gaussian and uniform) distribution, the orthogonal component of w_k remains almost unchanged. Therefore, $\theta(w_k, w^*) = \theta(w_k, w_{k-1}) \in \Omega(\frac{1-\beta}{b_{k-1}(1+\beta)}) \geq \left(\frac{(1-\beta)\sqrt{d}}{(1+\beta)\alpha_{k-1}}\right)$.

5. Hinge Loss Minimization Does Not Work

Hinge loss minimization is a widely used technique in Machine Learning. In this section, we show that, perhaps surprisingly, hinge loss minimization does not lead to arbitrarily small excess error even under very small noise condition, that is it is not consistent. (Note that in our setting of Massart noise, consistency is the same as achieving arbitrarily small excess error, since the Bayes optimal classifier is a member of the class of halfspaces).

It has been shown earlier that hinge loss minimization can lead to classifiers of large $0/1$-loss (Ben-David et al., 2012). However, the lower bounds in that paper employ distributions with *significant mass* on *discrete points* with *flipped label* (which is not possible under Massart noise) at a very *large distance* from the optimal classifier. Thus, that result makes strong use of the hinge loss's sensitivity to errors at large distance. Here, we show that hinge loss minimization is bound to fail under much more benign conditions. More concretely, we show that for every parameter τ, and arbitrarily small bound on the probability of flipping a label, $\eta = \frac{1-\beta}{2}$, hinge loss minimization is not consistent even on distributions with a uniform marginal over the unit ball in $\Re^2$, with the Bayes optimal classifier being a halfspace and the noise satisfying the Massart noise condition with bound η. That is, there exists a constant $\epsilon \geq 0$ and a sample size $m(\epsilon)$ such that hinge loss minimization returns a classifier of excess error at least ϵ with high probability over sample size of at least $m(\epsilon)$.

Hinge loss minimization does approximate the optimal hinge loss. We show that this does not translate into an agnostic learning guarantee for halfspaces with respect to the $0/1$-loss even under very small noise conditions. Let $\mathcal{P}_\beta$ be the class of distributions $\tilde{D}$ with uniform marginal over the unit ball $B_1 \subseteq \Re^2$, the Bayes classifier being a halfspace w, and satisfying the Massart noise condition with parameter β. Our lower bound for hinge loss minimization is stated as follows.

Theorem 6 *For every hinge-loss parameter $\tau \geq 0$ and every Massart noise parameter $0 \leq \beta < 1$, there exists a distribution $\tilde{D}_{\tau,\beta} \in \mathcal{P}_\beta$ (that is, a distribution over $B_1 \times \{-1, 1\}$ with uniform marginal over $B_1 \subseteq \Re^2$ satisfying the β-Massart condition) such that τ-hinge loss minimization is not consistent on $\tilde{D}_{\tau,\beta}$ with respect to the class of halfspaces. That is, there exists an $\epsilon \geq 0$ and a sample size $m(\epsilon)$ such that hinge loss minimization will output a classifier of excess error larger ϵ (with high probability over samples of size at least $m(\epsilon)$).*

Proof idea To prove the above result, we define a subclass of $\mathcal{P}_{\alpha,\eta} \subseteq \mathcal{P}_\beta$ consisting of well structured distributions. We then show that for every hinge parameter τ and every bound on the noise $\eta = \frac{1-\beta}{2}$, there is a distribution $\tilde{D} \in \mathcal{P}_{\alpha,\eta}$ on which τ-hinge loss minimization is not consistent.

In the remainder of this section, we use the notation h_w for the classifier associated with a vector $w \in B_1$, that is $h_w(x) = \text{sign}(w \cdot x)$, since for our geometric construction it is convenient to differentiate between the two. We define a family $\mathcal{P}_{\alpha,\eta} \subseteq \mathcal{P}_\beta$ of distributions $\tilde{D}_{\alpha,\eta}$, indexed by an angle α and a noise parameter η as follows. Let the Bayes optimal classifier be linear $h^* = h_{w^*}$ for a unit vector w^*. Let h_w be the classifier that is defined by the unit vector w at angle α from w^*. We partition the unit ball into areas A, B and D as in the Figure 5. That is A consists of the two wedges of disagreement between h_w and h_{w^*} and the wedge where the two classifiers agree is divided into B (points that are closer to h_w than to h_{w^*}) and D (points that are closer to h_{w^*} than to h_w). We now flip the labels of all points in A and B with probability $\eta = \frac{1-\beta}{2}$ and leave the labels deterministic according to h_{w^*} in the area D.

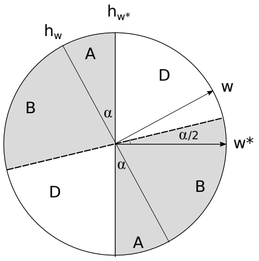

Figure 1: $P_{\alpha,\eta}$

More formally, points at angle between $\alpha/2$ and $\pi/2$ and points at angle between $\pi + \alpha/2$ and $-\pi/2$ from w^* are labeled per $h_{w^*}(x)$ with conditional label probability 1. All other points are labeled $-h_{w^*}(x)$ with probability η and $h_{w^*}(x)$ with probability $(1 - \eta)$. Clearly, this distribution satisfies Massart noise conditions in Equation 1 with parameter β.

The goal of the above construction is to design distributions where vectors along the direction of w have smaller hinge loss than those along the direction of w^*. Observe that the noise in the are A will tend to "even out" the difference in hinge loss between w and w^* (since are A is symmetric with respect to these two directions). The noise in area B however will "help w": Since all points in area B are closer to the hyperplane defined by w than to the one defined by w^*, vector w^* will pay more in hinge loss for the noise in this area. In the corresponding area D of points that are closer to the hyperplane defined by w^* than to the one defined by w we do not add noise, so the cost for both w and w^* in this area is small.

We show that for every α, from a certain noise level η on, w^*(or any other vector in its direction) is not the expected hinge minimizer on $\tilde{D}_{\alpha,\eta}$. We then argue that thereby hinge loss minimization will not approximate w^* arbitrarily close in angle and can therefore not achieve arbitrarily small excess $0/1$-error (see Equation (2)). Overall, we show that for every (arbitrarily small) bound on the noise η_0 and hinge parameter τ_0, we can choose an angle α such that τ_0-hinge loss minimization is not consistent for distribution $\tilde{D}_{\alpha,\eta_0}$. The details of the proof can be found in the Appendix E.

6. Conclusions

Our work is the first to provide a computationally efficient algorithm under Massart noise – a distributional assumption that has been identified in statistical learning to yield fast (statistical) rates of convergence – for a value of β that is a constant independent of the dimension. It remains an important open question whether computationally efficient algorithms exist under Massart noise for *any* constant β, or for more general noise models such as the *Tsybakov* noise model (Tsybakov (2004); Bousquet et al. (2005)).

While both computational and statistical efficiency are crucial in machine learning applications, computational and statistical complexity have been studied under disparate sets of assumptions and models. We view our results on the computational complexity of learning under Massart noise also as a step towards bringing these two lines of research closer together. We hope that this will spur more work identifying situations that lead to both computational and statistical efficiency to ultimately shed light on the underlying connections and dependencies of these two important aspects of automated learning.

Acknowledgments This work was supported in part by NSF grants CCF-1302518, CCF-0953192, CCF-1451177, CCF- 1422910, a Sloan Research Fellowshp, a Microsoft Research Faculty Fellowship, and a Google Research Award.

References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning - Theoretical Foundations*. Cambridge University Press, 2002.

Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. In *Proceedings of the 34th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 1993.

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)*, 2014.

Maria-Florina Balcan and Vitaly Feldman. Statistical active learning algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

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

Shai Ben-David and Hans-Ulrich Simon. Efficient learning of linear perceptrons. In *Advances in Neural Information Processing Systems (NIPS)*, pages 189–195, 2000.

Shai Ben-David, David Loker, Nathan Srebro, and Karthik Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

Avrim Blum, Alan Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1-2):35–52, 1998.

Karl-Heinz Borgwardt. *The simplex method*, volume 1 of *Algorithms and Combinatorics: Study and Research Texts*. Springer-Verlag, Berlin, 1987.

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

Rui M. Castro and Robert D. Nowak. Minimax bounds for active learning. In *Proceedings of the 20th Annual Conference on Learning Theory, (COLT)*, 2007.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

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

Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. From average case complexity to improper learning complexity. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, 2014.

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

Sanjoy Dasgupta. Active learning. *Encyclopedia of Machine Learning*, 2011.

Sanjoy Dasgupta, Daniel Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

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

Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

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.

Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24rd International Conference on Machine Learning (ICML)*, 2007.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2-3):131–309, 2014.

Steve Hanneke and Liu Yang. Surrogate losses in passive and active learning. *CoRR*, abs/1207.3772, 2014.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008a.

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)*, 2008b.

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

Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. In *Proceedings of the 5th Annual Conference on Computational Learning Theory (COLT)*, 1992.

Adam R. Klivans and Pravesh Kothari. Embedding hard learning problems into gaussian space. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, (APPROX/RANDOM)*, 2014.

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

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5): 2326–2366, 10 2006.

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

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

Shai Shalev-Shwartz, Ohad Shamir, and Karthik Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM J. Comput.*, 40(6):1623–1646, 2011.

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.

Ruth Urner, Sharon Wulff, and Shai Ben-David. PLAL: cluster-based active learning. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 376–397, 2013.

Appendix A. Probability Lemmas For The Uniform Distribution

The following probability lemmas are used throughout this work. Variation of these lemmas are presented in previous work in terms of their asymptotic behavior (Awasthi et al., 2014; Balcan et al., 2007; Kalai et al., 2008a). Here, we focus on finding bounds that are tight even when the constants are concerned. Indeed, the improved constants in these bounds are essential to tolerating Massart noise with $\beta > 1 - 3.6 \times 10^{-6}$.

Throughout this section, let D be the uniform distribution over a d-dimensional ball. Let $f(\cdot)$ indicate the p.d.f. of D. For any d, let V_d be the volume of a d-dimensional unit ball. Ratios between volumes of the unit ball in different dimensions are commonly used to find the probability mass of different regions under the uniform distribution. Note that for any d

$$\frac{V_{d-2}}{V_d} = \frac{d}{2\pi}.$$

The following bound due to Borgwardt (1987) proves useful in our analysis.

$$\sqrt{\frac{d}{2\pi}} \leq \frac{V_{d-1}}{V_d} \leq \sqrt{\frac{d+1}{2\pi}}$$

The next lemma provides an upper and lower bound for the probability mass of a band in uniform distribution.

Lemma 7 *Let u be any unit vector in $\Re^d$. For all $a, b \in [-\frac{C}{\sqrt{d}}, \frac{C}{\sqrt{d}}]$, such that $C < d/2$, we have*

$$|b-a|2^{-C}\frac{V_{d-1}}{V_d} \leq \Pr_{x\sim D}[u \cdot x \in [a,b]] \leq |b-a|\frac{V_{d-1}}{V_d}.$$

Proof We have

$$\Pr_{x\sim D}[u \cdot x \in [a,b]] = \frac{V_{d-1}}{V_d}\int_a^b (1-z^2)^{(d-1)/2}\, dz.$$

For the upper bound, we note that the integrant is at most 1, so $\Pr_{x\sim D}[u \cdot x \in [a,b]] \leq \frac{V_{d-1}}{V_d}|b-a|$. For the lower bound, note that since $a, b \in [-\frac{C}{\sqrt{d}}, \frac{C}{\sqrt{d}}]$, the integrant is at least $(1-\frac{C}{d})^{(d-1)/2}$. We know that for any $x \in [0, 0.5]$, $1 - x > 4^{-x}$. So, assuming that $d > 2C$, $(1-\frac{C}{d})^{(d-1)/2} \geq 4^{-\frac{C}{d}(d-1)/2} \geq 2^{-C}$ $\Pr_{x\sim D}[u \cdot x \in [a,b]] \geq |b-a|2^{-C}\frac{V_{d-1}}{V_d}$. ■

Lemma 8 *Let u and v be two unit vectors in $\Re^d$ and let $\alpha = \theta(u, v)$. Then,*

$$\Pr_{x\sim D}[\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(w \cdot x) \textit{ and } |u \cdot x| > \frac{c\,\alpha}{\sqrt{d}}] \leq \frac{\alpha}{\pi}e^{-\frac{c^2(d-2)}{2d}}$$

Proof Without the loss of generality, we can assume $u = (1, 0, \dots, 0)$ and $w = (\cos(\alpha), \sin(\alpha), 0, \dots, 0)$. Consider the projection of D on the first 2 coordinates. Let E be the event we are interested in. We first show that for any $x = (x_1, x_2) \in E$, $\|x\|_2 > c/\sqrt{d}$. Consider $x_1 \geq 0$ (the other case is symmetric). If $x \in E$, it must be that $\|x\|_2 \sin(\alpha) \geq \frac{c\alpha}{\sqrt{d}}$. So, $\|x\|_2 = \frac{c\ \alpha}{\sin(\alpha)\sqrt{d}} \geq \frac{c}{\sqrt{d}}$.

Next, we consider a circle of radius $\frac{c}{\sqrt{d}} < r < 1$ around the center, indicated by $S(r)$. Let $A(r) = S(r) \cap E$ be the arc of such circle that is in E. Then the length of such arc is the arc-length that falls in the disagreement region, i.e., $r\alpha$, minus the arc-length that falls in the band of width $\frac{c\alpha}{\sqrt{d}}$. Note, that for every $x \in A(r)$, $\|x\|_2 = r$, so $f(x) = \frac{V_{d-2}}{V_d}(1 - \|x\|^2)^{(d-2)/2} = \frac{V_{d-2}}{V_d}(1 - r^2)^{(d-2)/2}$.

$$
\begin{aligned}
\Pr_{x\sim D}[\text{sign}(u\cdot x) \neq \text{sign}(w \cdot x) \text{ and } |u\cdot x| > \frac{\alpha}{\sqrt{d}}] &= 2\int_{\frac{c}{\sqrt{d}}}^{1} (r\alpha - \frac{c\alpha}{\sqrt{d}}) f(r)\ dr \\
&= 2\int_{1}^{\sqrt{d}/c} (\frac{rc}{\sqrt{d}}\alpha - \frac{c\alpha}{\sqrt{d}}) f(\frac{cr}{\sqrt{d}}) \frac{c}{\sqrt{d}}\ dr \qquad \text{(change of variable } z = r\sqrt{d}/c\text{)} \\
&= 2\frac{V_{d-2}}{V_d}\frac{c^2\alpha}{d}\int_{1}^{\sqrt{d}/c} (r-1)(1 - \frac{c^2r^2}{d})^{(d-2)/2}\ dr \\
&= \frac{c^2\alpha}{\pi}\int_{1}^{\sqrt{d}/c} (r-1) e^{-\frac{r^2(d-2)}{2d}}\ dr \\
&\leq \frac{c^2\alpha}{\pi}\int_{1}^{\sqrt{d}} \frac{(r-1)}{\frac{(d-2)c^2r}{d}} (-1)(\frac{-(d-2)c^2r}{d}) e^{-\frac{(d-2)c^2r^2}{2d}}\ dr \\
&\leq \frac{\alpha}{\pi}\int_{1}^{\sqrt{d}/c} (-1)(\frac{-(d-2)c^2r}{d}) e^{-\frac{(d-2)c^2r^2}{2d}}\ dr \\
&\leq \frac{\alpha}{\pi}\Big[- e^{-\frac{(d-2)r^2}{2d}} \Big]_{r=1}^{r=\sqrt{d}/c} \\
&\leq \frac{\alpha}{\pi}(e^{-\frac{c^2(d-2)}{2d}} - e^{-(d-2)/2}) \\
&\leq \frac{\alpha}{\pi} e^{-\frac{c^2(d-2)}{2d}}
\end{aligned}
$$

■

Appendix B. Proofs of Margin-based Lemmas

Proof of Lemma 2 Let $L(w^*) = \mathbb{E}_{(x,y)\sim D_k}\ell(w^*, x, y)$, $\tau = \tau_k$, and $b = b_{k-1}$. First note that for our choice of $b \leq 2.3463 \times 0.0121608\frac{1}{\sqrt{d}}$, using Lemma 7 we have that

$$
\Pr_{x\sim D}[|w_{k-1}\cdot x| < b] \geq 2\ b \times 2^{-0.285329}.
$$

Note that $L(w^*)$ is maximized when $w^* = w_{k-1}$. Then

$$
L(w^*) \leq \frac{2\int_0^\tau (1 - \frac{a}{\tau}) f(a)\ da}{\Pr_{x\sim D}[|w_{k-1}\cdot x| < b]} \leq \frac{\int_0^\tau (1 - \frac{a}{\tau})(1 - a^2)^{-(d-1)/2}\ da}{b\ 2^{-0.285329}}.
$$

For the numerator:

$$\begin{aligned}
&\int_0^\tau (1-\frac{a}{\tau})(1-a^2)^{-(d-1)/2}\,da \le \int_0^\tau (1-\frac{a}{\tau})e^{-a^2(d-1)/2}\,da\\
&\quad\le \frac{1}{2}\int_{-\tau}^{\tau} e^{-a^2(d-1)/2}\,da - \frac{1}{\tau}\int_0^\tau a e^{-a^2(d-1)/2}\,da\\
&\quad\le \sqrt{\frac{\pi}{2(d-1)}}\,\mathrm{erf}\left(\tau\sqrt{\frac{d-1}{2}}\right) - \frac{1}{(d-1)\tau}(1-e^{-(d-1)\tau^2/2})\\
&\quad\le \sqrt{\frac{\pi}{2(d-1)}}\sqrt{1-e^{-\tau^2(d-1)}} - \frac{1}{(d-1)\tau}\left(\frac{(d-1)\tau^2}{2} - \frac{1}{2}(\frac{(d-1)\tau^2}{2})^2\right) \qquad \text{(By Taylor expansion)}\\
&\quad\le \tau\sqrt{\frac{\pi}{2}} - \frac{\tau}{2} + \frac{1}{8}(d-1)\tau^3\\
&\quad\le \tau(0.5462 + \frac{1}{8}(d-1)\tau^2)\\
&\quad\le 0.5463\tau \qquad \text{(By } \frac{1}{8}(d-1)\tau^2 < 2\times 10^{-4}\text{)}
\end{aligned}$$

Where the last inequality follows from the fact that for our choice of parameters $\tau \le \frac{\sqrt{2.50306}(3.6\times 10^{-6})^{1/4}b}{\sqrt{d}} < \frac{0.003}{\sqrt{d}}$, so $\frac{1}{8}(d-1)\tau^2 < 10^{-5}$. Therefore,

$$L(w^*) \le 0.5463 \times 2^{0.285329}\frac{\tau}{b} \le 0.665769\frac{\tau}{b}.$$

■

Proof of Lemma 4 Note that the convex loss minimization procedure returns a vector v_k that is not necessarily normalized. To consider all vectors in $B(w_{k-1},\alpha_k)$, at step k, the optimization is done over all vectors v (of any length) such that $\|w_{k-1} - v\| < \alpha_k$. For all k, $\alpha_k < 0.038709\pi$ (or 0.0121608), so $\|v_k\|_2 \ge 1 - 0.0121608$, and as a result $\ell(w_k, W) \le 1.13844\,\ell(v_k, W)$. We have,

$$\begin{aligned}
&\mathrm{err}_{D_k}(w_k) \le \mathbb{E}_{(x,y)\sim D_k}\ell(w_k,x,y)\\
&\le \mathbb{E}_{(x,y)\sim \tilde{D}_k}\ell(w_k,x,y) + \left(1.092\sqrt{2}\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k}\right) \qquad \text{(By Lemma 3)}\\
&\le \ell(w_k,W) + 1.092\sqrt{2}\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k} + 10^{-8} \qquad \text{(By Equation 3)}\\
&\le 1.13844\,\ell(v_k,W) + 1.092\sqrt{2}\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k} + 10^{-8} \qquad \text{(By } \|v_k\|_2 \ge 1-0.0121608\text{)}\\
&\le 1.13844\,\ell(w^*,W) + 1.092\sqrt{2}\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k} + 2.14\times 10^{-8} \qquad \text{(By } v_k \text{ minimizing the hinge-loss)}\\
&\le 1.13844\,\mathbb{E}_{(x,y)\sim \tilde{D}_k}\ell(w^*,x,y) + 1.092\sqrt{2}\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k} + 3.28\times 10^{-8} \qquad \text{(By Equation 3)}\\
&\le 1.13844\,\mathbb{E}_{(x,y)\sim D_k}\ell(w^*,x,y) + 2.13844\left(1.092\sqrt{2}\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k}\right) + 3.28\times 10^{-6} \qquad \text{(By Lemma 3)}\\
&\le 0.757941\frac{\tau_k}{b_{k-1}} + 3.303\sqrt{1-\beta}\frac{b_{k-1}}{\tau_k} + 3.28\times 10^{-8} \qquad \text{(By Lemma 2)}
\end{aligned}$$

■

Appendix C. VC Dimension Tools

In this section, we apply generalized VC dimension tools to prove that the empirical hinge loss of a vector is closely concentrated around its expectation. Similar lemmas appear in the work of Awasthi et al. (2014), here we provide the details for completeness.

Definition 9 *A set F of real-valued functions, all from a domain X, shatters the set $S := \{x_1, \dots, x_d\} \subseteq X$, if there are thresholds $t_1, \dots, t_d$ such that*

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

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

Next, we provide a well-known result for the pseudo-dimension of a function class, which is analogous to VC dimension bounds.

Lemma 10 (See (Anthony and Bartlett, 2002)) *Let F be a set of functions from a common domain $X \to [a, b]$, let d be the pseudo-dimension of F, and let D be any probability distribution over X. Then if $x_1, \dots, x_m$ are drawn i.i.d. from D, where $m = O\left(\frac{(b-a)^2}{\epsilon^2}(d + \log(\frac{1}{\delta}))\right)$, then with probability $1 - \delta$, for all $f \in F$,*

$$\left| \mathbb{E}_{x \sim D}(f(x)) - \frac{1}{m} \sum_{i=1}^{m} f(x_i) \right| \leq \epsilon$$

Lemma 11 *For any constant c', there is $m_k \in O(d(d + \log(k/d)))$ such that for a randomly drawn set W of m_k labeled samples from $\tilde{D}_k$, with probability $1 - \frac{\delta}{k+k^2}$, for any $w \in B(w_{k-1}, \alpha_k)$,*

$$|\mathbb{E}_{(x,y) \sim \tilde{D}_k} (\ell(w, x, y) - \ell(w, W)) | \leq c',$$
$$|\mathbb{E}_{(x,y) \sim D_k} (\ell(w, x, y) - \ell(w, \mathit{cleaned}(W))) | \leq c'.$$

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

$$\begin{aligned}
\ell_{\tau_k}(w, x, y) &\leq 1 + \frac{|w \cdot x|}{\tau_k} \\
&\leq 1 + \frac{|w_{k-1} \cdot x| + \|w - w_{k-1}\|}{\tau_k} \\
&\leq 1 + \frac{b_{k-1} + \alpha_k}{\tau_k} \\
&\in O(\sqrt{d}).
\end{aligned}$$

The claim follows by using Lemma 10 and the fact that $\{\ell_{\tau_k}(w, \cdot) : w \in \Re^d\}$ is a set of functions from X to the range of size $O(\sqrt{d})$ with pseudo dimension d.

■

Appendix D. Initialization

We initialize our margin based procedure with the algorithm from Klivans et al. (2009). The guarantees mentioned in Klivans et al. (2009) hold as long as the noise rate is $\eta \leq c\frac{\epsilon^2}{\log 1/\epsilon}$. Klivans et al. (2009) do not explicitly compute the constant but it is easy to check that $c \leq \frac{1}{256}$. This can be computed from inequality 17 in the proof of Lemma 16 in Klivans et al. (2009). We need the l.h.s. to be at least $\epsilon^2/2$. On the r.h.s., the first term is lower bounded by $\epsilon^2/512$. Hence, we need the second term to be at most $\frac{255}{512}\epsilon^2$. The second term is upper bounded by $4c^2\epsilon^2$. This implies that $c \leq 1/256$.

Appendix E. Hinge Loss Minimization

In this section, we show that hinge loss minimization is not consistent in our setup, that is, that it does not lead to arbitrarily small excess error. We let B_1^d denote the unit ball in R^d. In this section, we will only work with $d = 2$, thus we set $B_1 = B_1^2$.

Recall that the τ-hinge loss of a vector $w \in \Re^d$ on an example $(x, y) \in \Re^d \times \{-1, 1\}$ is defined as follows:

$$\ell_\tau(w, x, y) = \max\left\{0,\ 1 - \frac{y(w \cdot x)}{\tau}\right\}$$

For a distribution $\tilde{D}$ over $\Re^d \times \{-1, 1\}$, we let $\mathcal{L}_\tau^{\tilde{D}}$ denote the expected hinge loss over D, that is

$$\mathcal{L}_\tau^{\tilde{D}}(w) = \mathbb{E}_{(x,y)\sim\tilde{D}}\ell_\tau(w, x, y).$$

If clear from context, we omit the superscript and write $\mathcal{L}_\tau(w)$ for $\mathcal{L}_\tau^{\tilde{D}}(w)$.

Let $\mathcal{A}_\tau$ be the algorithm that minimizes the empirical τ-hinge loss over a sample. That is, for $W = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, we have

$$\mathcal{A}_\tau(W) \in \mathrm{argmin}_{w\in B_1}\frac{1}{|W|}\sum_{(x,y)\in W}\ell_\tau(w, x, y).$$

Hinge loss minimization over halfspaces converges to the optimal hinge loss over all halfspace (it is "hinge loss consistent"). That is, for all $\epsilon > 0$ there is a sample size $m(\epsilon)$ such that for all distributions $\tilde{D}$, we have

$$\mathbb{E}_{W\sim\tilde{D}^m}[\mathcal{L}_\tau^{\tilde{D}}(\mathcal{A}_\tau(W))] \leq \min_{w\in B_1}\mathcal{L}_\tau^{\tilde{D}}(w) + \epsilon.$$

In this section, we show that this does not translate into an agnostic learning guarantee for halfspaces with respect to the $0/1$-loss. Moreover, hinge loss minimization is not even consistent with respect to the $0/1$-loss even when restricted to a rather benign classes of distributions $\mathcal{P}$. Let $\mathcal{P}_\beta$ be the class of distributions $\tilde{D}$ with uniform marginal over the unit ball in $\Re^2$, the Bayes classifier being a halfspace w, and satisfying the Massart noise condition with parameter β. We show that there is a distribution $\tilde{D} \in \mathcal{P}_\beta$ and an $\epsilon \geq 0$ and a sample size m_0 such that hinge loss minimization will output a classifier of excess error larger than ϵ on expectation over samples of size larger than m_0. More precisely, for all $m \geq m_0$:

$$\mathbb{E}_{W\sim\tilde{D}^m}[\mathcal{L}_\tau^{\tilde{D}}(\mathcal{A}_\tau(W))] > \min_{w\in B_1}\mathrm{err}_{\tilde{D}}(w) + \epsilon.$$

Formally, our lower bound for hinge loss minimization is stated as follows.

Theorem 6 (Restated). *For every hinge-loss parameter $\tau \geq 0$ and every Massart noise parameter $0 \leq \beta < 1$, there exists a distribution $\tilde{D}_{\tau,\beta} \in \mathcal{P}_\beta$ (that is, a distribution over $B_1 \times \{-1, 1\}$ with uniform marginal over $B_1 \subseteq \Re^2$ satisfying the β-Massart condition) such that τ-hinge loss minimization is not consistent on $P_{\tau,\beta}$ with respect to the class of halfspaces. That is, there exists an $\epsilon \geq 0$ and a sample size $m(\epsilon)$ such that hinge loss minimization will output a classifier of excess error larger than ϵ (with high probability over samples of size at least $m(\epsilon)$).*

In the section, we use the notation h_w for the classifier associated with a vector $w \in B_1$, that is $h_w(x) = \operatorname{sign}(w \cdot x)$, since for our geometric construction it is convenient to differentiate between the two. The rest of this section is devoted to proving the above theorem.

A class of distributions

Let $\eta = \frac{1-\beta}{2}$. We define a family $\mathcal{P}_{\alpha,\eta} \subseteq \mathcal{P}_\beta$ of distributions $\tilde{D}_{\alpha,\eta}$, indexed by an angle α and a noise parameter η as follows. We let the marginal be uniform over the unit ball $B_1 \subseteq \Re^2$ and let the Bayes optimal classifier be linear $h^* = h_{w^*}$ for a unit vector w^*. Let h_w be the classifier that is defined by the unit vector w at angle α from w^*. We partition the unit ball into areas A, B and D as in the Figure 2. That is A consists of the two wedges of disagreement between h_w and h_{w^*} and the wedge where the two classifiers agree is divided in B (points that are closer to h_w than to h_{w^*}) and D (points that are closer to h_{w^*} than to h_w). We now "add noise η" at all points in areas A and B and leave the labels deterministic according to h_{w^*} in the area D.

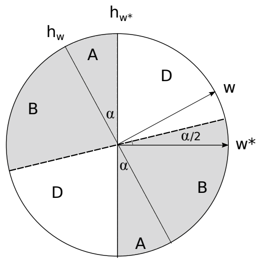

Figure 2: $\tilde{D}_{\alpha,\eta}$

More formally, points at angle between $\alpha/2$ and $\pi/2$ and points at angle between $\pi + \alpha/2$ and $-\pi/2$ from w^* are labeled with $h_{w^*}(x)$ with (conditional) probability 1. All other points are labeled $-h_{w^*}(x)$ with probability η and $h_{w^*}(x)$ with probability $(1 - \eta)$.

Useful lemmas

The following lemma relates the τ-hinge loss of unit length vectors to the hinge loss of arbitrary vectors in the unit ball. It will allow us to focus our attention to comparing the τ-hinge loss of unit vectors for all $\tau > \tau_0$, instead of having to argue about the τ_0 hinge loss of vectors of arbitrary norms in B_1.

Lemma 12 *Let $\tau > 0$ and $0 < \lambda \leq 1$. Let w and w^* be two vectors of unit length. Then $\mathcal{L}_\tau(\lambda w) < \mathcal{L}_\tau(\lambda w^*)$ if and only if $\mathcal{L}_{\tau/\lambda}(w) < \mathcal{L}_{\tau/\lambda}(w^*)$.*

Proof By the definition of the hinge loss, we have

$$\ell_\tau(\lambda w, x, y) = \max\left(0,\ 1 - \frac{y(\lambda w \cdot x)}{\tau}\right) = \max\left(0,\ 1 - \frac{y(w \cdot x)}{\tau/\lambda}\right) = \ell_{\tau/\lambda}(w, x, y).$$

■

Lemma 13 *Let $\tau > 0$, for any $\tilde{D} \in \mathcal{P}_{\alpha,\eta}$ let w_τ denote the halfspace that minimizes the τ-hinge loss with respect to $\tilde{D}$. If $w_\tau \neq \lambda w^*$ for all $0 < \lambda \leq 1$, then hinge loss minimization is not consistent for the $0/1$-loss.*

Proof

First we show that the hinge loss minimizer is never the vector 0. Note that $\mathcal{L}_\tau^{\tilde{D}}(0) = 1$ (for all $\tau > 0$). Consider the case $\tau \geq 1$, we show that w^* has τ-hinge loss strictly smaller than 1. Integrating the hinge loss over the unit ball using polar coordinates, we get

$$\begin{aligned}
\mathcal{L}_\tau^{\tilde{D}}(w^*) &< \frac{2}{\pi}\left((1-\eta)\int_0^1\int_0^\pi (1 - \frac{z}{\tau}\sin(\varphi))\, z\, d\varphi\, dz + \eta\int_0^1\int_0^\pi (1 + \frac{z}{\tau}\sin(\varphi))\, z\, d\varphi\, dz\right) \\
&= \frac{2}{\pi}\left((1-\eta)\int_0^1\int_0^\pi z - \frac{z^2}{\tau}\sin(\varphi)\, d\varphi\, dz + \eta\int_0^1\int_0^\pi z + \frac{z^2}{\tau}\sin(\varphi)\, d\varphi\, dz\right) \\
&= 1 + \frac{2}{\pi}\left((1-2\eta)\int_0^1\int_0^\pi -\frac{z^2}{\tau}\sin(\varphi)\, d\varphi\, dz\right) \\
&= 1 - \frac{2}{\pi}\left((1-2\eta)\int_0^1\int_0^\pi \frac{z^2}{\tau}\sin(\varphi)\, d\varphi\, dz\right) < 1.
\end{aligned}$$

For the case of $\tau < 1$, we have

$$\mathcal{L}_\tau^{\tilde{D}}(\tau w^*) = \mathcal{L}_1^{\tilde{D}}(w^*) < 1.$$

Thus, $(0, 0)$ is not the hinge-minimizer. Then, by the assumption of the lemma w_τ has some positive angle γ to the w^*. Furthermore, for all $0 \leq \lambda \leq 1$, $\mathcal{L}_\tau^{\tilde{D}}(w_\tau) < \mathcal{L}_\tau^{\tilde{D}}(\lambda w^*)$. Since $w \mapsto \mathcal{L}_\tau^{\tilde{D}}(w)$ is a continuous function we can choose an $\epsilon > 0$ such that

$$\mathcal{L}_\tau^{\tilde{D}}(w_\tau) + \epsilon/2 < \mathcal{L}_\tau^{\tilde{D}}(\lambda w^*) - \epsilon/2.$$

for all $0 \leq \lambda \leq 1$ (note that the set $\{\lambda w^* \mid 0 \leq \lambda \leq 1\}$ is compact). Now, we can choose an angle $\mu < \gamma$ such that for all vectors v at angle at most μ from w^*, we have

$$\mathcal{L}_\tau^{\tilde{D}}(v) \geq \min_{0 \leq \lambda \leq 1} \mathcal{L}_\tau^{\tilde{D}}(\lambda w^*) - \epsilon/2$$

Since hinge loss minimization will eventually (in expectation over large enough samples) output classifiers of hinge loss strictly smaller than $\mathcal{L}_\tau^{\tilde{D}}(w_\tau) + \epsilon/2$, it will then not output classifiers of angle smaller than μ to w^*. By Equation 2, for all w, $\mathrm{err}_{\tilde{D}}(w) - \mathrm{err}_{\tilde{D}}(w^*) > \beta\frac{\theta(w,w^*)}{\pi}$, therefore, the excess error of a the classfier returned by hinge loss minimization is lower bounded by a constant $\beta\frac{\mu}{\pi}$. Thus, hinge loss minimization is not consistent with respect to the $0/1$-loss. ■

Proof of Theorem 6

We will show that, for every bound on the noise η_0 and for every every $\tau_0 \geq 0$ there is an $\alpha_0 > 0$, such that the unit length vector w has strictly lower τ-hinge loss than the unit length vector w^* for all $\tau \geq \tau_0$. By Lemma 12, this implies that for every bound on the noise η_0 and for every τ_0 there is an $\alpha_0 > 0$ such that for all $0 < \lambda \leq 1$ we have $\mathcal{L}_{\tau_0}(\lambda w) < \mathcal{L}_{\tau_0}(\lambda w^*)$. This implies that the hinge minimizer is not a multiple of w^*, that is $w_\tau \neq \lambda w^*$ for all $0 < \lambda \leq 1$. Now Lemma 13 tells us that hinge loss minimization is not consistent for the $0/1$-loss.

In the sequel, we will now focus on the unit length vectors w and w^* and show how to choose α_0 as a function of τ_0 and η_0. We let cA denote the hinge loss of h_{w^*} on one wedge (one half of) area A when the labels are correct and dA that hinge loss on that same area when the labels are not correct. Analogously, we define $\mathrm{cB}, \mathrm{dB}, \mathrm{cD}$ and dD. For example, for $\tau \geq 1$, we have (integrating the hinge loss over the unit ball using polar coordinates)

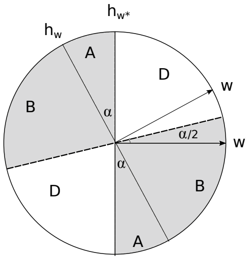

Figure 3: $\tilde{D}_{\alpha,\eta}$

$$\mathrm{cA} = \frac{1}{\pi}\int_0^1 \int_0^{\alpha} (1 - \frac{z}{\tau}\sin(\varphi)) z \, d\varphi \, dz,$$
$$\mathrm{dA} = \frac{1}{\pi}\int_0^1 \int_0^{\alpha} (1 + \frac{z}{\tau}\sin(\varphi)) z \, d\varphi \, dz,$$
$$\mathrm{cB} = \frac{1}{\pi}\int_0^1 \int_{\alpha}^{\frac{\pi+\alpha}{2}} (1 - \frac{z}{\tau}\sin(\varphi)) z \, d\varphi \, dz,$$
$$\mathrm{dB} = \frac{1}{\pi}\int_0^1 \int_{\alpha}^{\frac{\pi+\alpha}{2}} (1 + \frac{z}{\tau}\sin(\varphi)) z \, d\varphi \, dz,$$
$$\mathrm{cD} = \frac{1}{\pi}\int_0^1 \int_0^{\frac{\pi-\alpha}{2}} (1 - \frac{z}{\tau}\sin(\varphi)) z \, d\varphi \, dz,$$
$$\text{and} \qquad \mathrm{dD} = \frac{1}{\pi}\int_0^1 \int_0^{\frac{\pi-\alpha}{2}} (1 + \frac{z}{\tau}\sin(\varphi)) z \, d\varphi \, dz.$$

Now we can express the hinge loss of both h_{w^*} and h_w in terms of these quantities. For h_{w^*} we have

$$\mathcal{L}_\tau(h_{w^*}) = 2 \cdot [\eta(\mathrm{dA} + \mathrm{dB}) + (1-\eta)(\mathrm{cA} + \mathrm{cB}) + \mathrm{cD}] .$$

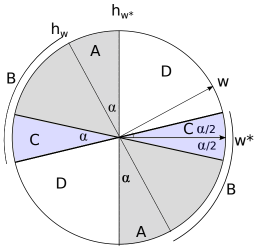

Figure 4: Area C

For h_w, note that area B relates to h_w as area D relates to h_{w^*} (and vice versa). Thus, the roles of B and D are exchanged for h_w. That is, for example, for the noisy version of area B the classifier h_w pays dD. We have

$$\mathcal{L}_\tau(h_w) = 2 \cdot [\eta(\mathrm{cA} + \mathrm{dD}) + (1-\eta)(\mathrm{dA} + \mathrm{cD}) + \mathrm{cB}] .$$

This yields

$$\mathcal{L}_\tau(h_w) - \mathcal{L}_\tau(h_{w^*}) = 2 \cdot [(1-2\eta)(\mathrm{dA} - \mathrm{cA}) - \eta((\mathrm{dB} - \mathrm{cB}) - (\mathrm{dD} - \mathrm{cD}))] .$$

We now define area C as the points at angle between $\pi - \alpha/2$ and $\pi + \alpha/2$ from w^* (See Figure 3). We let cC and dC be defined analogously to the above.

Note that $\mathrm{dA} + \mathrm{dB} - \mathrm{dD} = \mathrm{dC}$ and $\mathrm{cA} + \mathrm{cB} - \mathrm{cD} = \mathrm{cC}$. Thus we get

$$\begin{aligned}
&\mathcal{L}_\tau(h_w) - \mathcal{L}_\tau(h_{w^*}) \\
=&2 \cdot [(1-2\eta)(\mathrm{dA} - \mathrm{cA}) - \eta((\mathrm{dB} - \mathrm{cB}) - (\mathrm{dD} - \mathrm{cD}))] \\
=&2 \cdot [(1-\eta)(\mathrm{dA} - \mathrm{cA}) - \eta((\mathrm{dB} - \mathrm{cB}) + (\mathrm{dA} - \mathrm{cA}) - (\mathrm{dD} - \mathrm{cD}))] \\
=&2 \cdot [(1-\eta)(\mathrm{dA} - \mathrm{cA}) - \eta((\mathrm{dC} - \mathrm{cC}))] .
\end{aligned}$$

If $\eta > \eta(\alpha, \tau) := \frac{(\mathrm{dA}-\mathrm{cA})}{(\mathrm{dA}-\mathrm{cA})+(\mathrm{dC}-\mathrm{cC})}$, then we get $\mathcal{L}_\tau(h_w) - \mathcal{L}_\tau(h_{w^*}) < 0$ and thus h_w having smaller hinge loss than h_{w^*}. Thus, $\eta(\alpha, \tau)$ signifies the amount of noise from which onward, w will have smaller hinge loss than w^*

Given $\tau_0 \geq 0$, choose α small enough (we can always choose the angle α sufficiently small for this) so that the area A is included in the τ_0-band around h_{w^*}. We have for all $\tau \geq \tau_0$:

$$\begin{aligned}(\mathrm{dA} - \mathrm{cA}) &= \frac{2}{\pi} \int_0^1 \int_0^\alpha \frac{z^2}{\tau} \sin(\varphi)\, d\varphi\, dz \\ &= \frac{2}{3\pi} \int_0^\alpha \frac{1}{\tau} \sin(\varphi)\, d\varphi \\ &= \frac{2}{3\pi\tau} \left[-\cos(\varphi)\right]_0^\alpha \\ &= \frac{2}{3\pi\tau}(1 - \cos(\alpha)).\end{aligned}$$

For the area C we now consider the case of $\tau \geq 1$ and $\tau < 1$ separately. For $\tau \geq 1$ we get

$$\begin{aligned}(\mathrm{dC} - \mathrm{cC}) &= \frac{4}{\pi} \int_0^1 \int_{\frac{\pi-\alpha}{2}}^{\frac{\pi}{2}} \frac{z^2}{\tau} \sin(\varphi)\, d\varphi\, dz \\ &= \frac{4}{3\pi} \int_{\frac{\pi-\alpha}{2}}^{\frac{\pi}{2}} \frac{1}{\tau} \sin(\varphi)\, d\varphi \\ &= \frac{4}{3\pi\tau} \cos\left(\frac{\pi - \alpha}{2}\right) \\ &= \frac{4}{3\pi\tau} \sin\left(\frac{\alpha}{2}\right).\end{aligned}$$

Thus, for $\tau \geq 1$ we get

$$\eta(\alpha, \tau) = \frac{(\mathrm{dA} - \mathrm{cA})}{(\mathrm{dA} - \mathrm{cA}) + (\mathrm{dC} + \mathrm{cC})} = \frac{1 - \cos(\alpha)}{1 - \cos(\alpha) + 2\sin(\alpha/2)}.$$

We call this quantity $\eta_1(\alpha)$ since, given that $\tau \geq 1$, it does not depend on τ:

$$\eta_1(\alpha) = \frac{(\mathrm{dA} - \mathrm{cA})}{(\mathrm{dA} - \mathrm{cA}) + (\mathrm{dC} + \mathrm{cC})} = \frac{1 - \cos(\alpha)}{1 - \cos(\alpha) + 2\sin(\alpha/2)}.$$

Observe that $\lim_{\alpha \to 0} \eta_1(\alpha) = 0$. This will yield the first condition on the angle α: Given some bound on the allowed noise η_0, we can choose an α small enough so that $\eta_1(\alpha) \leq \eta_0/2$. Then, for the distribution $\tilde{D}_{\alpha,\eta_0}$ we have $\mathcal{L}_\tau(w) < \mathcal{L}_\tau(w^*)$ for all $\tau \geq 1$.

We now consider the case $\tau < 1$. For this case we lower bound $(\mathrm{dC} - \mathrm{cC})$ as follows. We have

$$\begin{aligned}\mathrm{dC} &= \frac{2}{\pi} \int_0^1 \int_{\frac{\pi-\alpha}{2}}^{\frac{\pi}{2}} z + \frac{z^2}{\tau} \sin(\varphi)\, d\varphi\, dz \\ &= \frac{\alpha}{2\pi} + \frac{2}{\pi} \int_0^1 \int_{\frac{\pi-\alpha}{2}}^{\frac{\pi}{2}} \frac{z^2}{\tau} \sin(\varphi)\, d\varphi\, dz \\ &= \frac{\alpha}{2\pi} + \frac{2}{3\tau\pi} \sin\left(\frac{\alpha}{2}\right).\end{aligned}$$

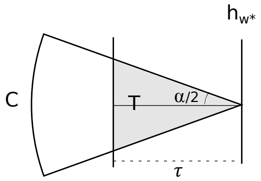

Figure 5: Area T

We now provide an upper bound on cC by integrating over a the triangular shape T (see Figure 4). Note that this bound on cC is actually exact if $\tau \leq \cos(\alpha/2)$ and only a strict upper bound for $\cos(\alpha/2) < \tau < 1$. We have

$$\begin{aligned}\mathrm{cC} \leq (cT) &= \frac{2}{\pi} \cdot \int_0^\tau (1 - \frac{z}{\tau})(z \tan(\alpha/2))\, dz \\ &= \frac{2}{\pi} \cdot \int_0^\tau z \tan(\alpha/2) - \frac{z^2}{\tau} \tan(\alpha/2)\, dz \\ &= \frac{\tau^2}{3\pi} \tan\left(\frac{\alpha}{2}\right).\end{aligned}$$

Thus we get

$$(\mathrm{dC} - \mathrm{cC}) \geq (\mathrm{dC} - (cT)) = \frac{1}{\pi}\left(\frac{\alpha}{2} + \frac{2}{3\tau}\sin\left(\frac{\alpha}{2}\right) - \frac{\tau^2}{3}\tan\left(\frac{\alpha}{2}\right)\right).$$

This yields, for the case $\tau \leq 1$

$$\eta(\alpha, \tau) = \frac{\frac{2}{3}(1 - \cos(\alpha))}{\frac{2}{3}(1 - \cos(\alpha)) + \frac{2}{3}\sin(\alpha) + \frac{\alpha\tau}{2} - \frac{\tau^3}{3}\tan(\frac{\alpha}{2})}$$

We call this quantity $\eta_2(\alpha, \tau)$ to differentiate it from $\eta_1(\alpha)$. Again, we have $\lim_{\alpha \to 0} \eta_2(\alpha, \tau) = 0$ for every τ. Thus, for a fixed τ_0, we can choose an angle α small enough so that $\mathcal{L}_{\tau_0}(w) \leq \mathcal{L}_{\tau_0}(w^*)$.

To argue that we will then also have $\mathcal{L}_\tau(w) \leq \mathcal{L}_\tau(w^*)$ for all $\tau \geq \tau_0$, we show that, for a fixed angle α, the function $\eta(\alpha, \tau)$ gets smaller as τ grows. For this, it suffices to show that $g(\tau) = \tau\frac{\alpha}{2} - \frac{\tau^3}{3}\tan(\frac{\alpha}{2})$ is monotonically increasing with τ for $\tau \leq 1$. We have

$$g'(\tau) = \frac{\alpha}{2} - \frac{\tau^2}{2}\tan\left(\frac{\alpha}{2}\right).$$

Since we have $\tau^2 \leq 1$ and $\frac{2\alpha}{\tan(\frac{\alpha}{2})} \geq 1$ for $0 \leq \alpha \leq \pi/3$, we get that (for sufficiently small α) $g'(\tau) \geq 0$ and thus $g(\tau)$ is monotonically increasing for $0 \leq \tau \leq 1$ as desired.

Summarizing, for a given τ_0 and η_0, we can always choose α_0 sufficiently small so that both $\eta_1(\alpha_0) < \frac{\eta_0}{2}$ and $\eta_2(\alpha_0, \tau) < \frac{\eta_0}{2}$ for all $\tau \geq \tau_0$ and thus $\mathcal{L}_\tau^{\tilde{D}_{\alpha_0,\eta_0}}(w) < \mathcal{L}_\tau^{\tilde{D}_{\alpha_0,\eta_0}}(w^*)$ for all $\tau \geq \tau_0$. This completes the proof.