A Near-optimal Algorithm for Learning Margin Halfspaces with Massart Noise

Ilias Diakonikolas
Department of Computer Sciences
UW-Madison
Madison, WI
ilias@cs.wisc.edu

Nikos Zarifis
Department of Computer Sciences
UW-Madison
Madison, WI
zarifis@wisc.edu

Abstract

We study the problem of PAC learning γ-margin halfspaces in the presence of Massart noise. Without computational considerations, the sample complexity of this learning problem is known to be $\widetilde{\Theta}(1/(\gamma^2\epsilon))$. Prior computationally efficient algorithms for the problem incur sample complexity $\tilde{O}(1/(\gamma^4\epsilon^3))$ and achieve 0-1 error of $\eta+\epsilon$, where $\eta<1/2$ is the upper bound on the noise rate. Recent work gave evidence of an information-computation tradeoff, suggesting that a quadratic dependence on $1/\epsilon$ is required for computationally efficient algorithms. Our main result is a computationally efficient learner with sample complexity $\widetilde{\Theta}(1/(\gamma^2\epsilon^2))$, nearly matching this lower bound. In addition, our algorithm is simple and practical, relying on online SGD on a carefully selected sequence of convex losses.

1 Introduction

This work studies the algorithmic task of learning margin halfspaces in the presence of Massart noise (aka bounded label noise) [MN06] with a focus on fine-grained complexity analysis. A halfspace or Linear Threshold Function (LTF) is any Boolean-valued function $h:\mathbb{R}^d\to\{\pm1\}$ of the form $h(\mathbf{x})=\mathrm{sign}\,(\mathbf{w}\cdot\mathbf{x}-\theta)$, where $\mathbf{w}\in\mathbb{R}^d$ is the weight vector and $\theta\in\mathbb{R}$ is the threshold. The function $\mathrm{sign}:\mathbb{R}\to\{\pm1\}$ is defined as $\mathrm{sign}(t)=1$ if $t\geq 0$ and $\mathrm{sign}(t)=-1$ otherwise. The problem of learning halfspaces with a margin — i.e., under the assumption that no example lies too close to the separating hyperplane — is one of the earliest algorithmic problems studied in machine learning, going back to the Perceptron algorithm [Ros58].

In the realizable PAC model [Val84] (i.e., with clean labels), the sample complexity of learning γ-margin halfspaces on the unit ball in $\mathbb{R}^d$ is $\Theta(1/(\gamma^2\epsilon))$, where $\epsilon>0$ is the desired 0-1 error; see, e.g., [SSBD14][1]. Moreover, the Perceptron algorithm is a computationally efficient learner achieving this sample complexity. That is, without label noise, there is a sample-optimal and computationally efficient learner for margin halfspaces.

In this paper, we study the same problem in the Massart noise model that we now define.

Definition 1.1 (PAC Learning with Massart Noise)**.** Let D be a distribution over $\mathcal{X}\times\{\pm1\}$, and let $\mathcal{C}$ be a class of Boolean-valued functions over $\mathcal{X}$. We say that D satisfies the η-Massart noise condition with respect to $\mathcal{C}$, for some $\eta<1/2$, if there exists a concept $f\in\mathcal{C}$ and an unknown noise function $\eta(\mathbf{x}):\mathcal{X}\mapsto[0,\eta]$ such that for $(\mathbf{x},y)\sim D$, the label y satisfies: with probability $1-\eta(\mathbf{x})$, $y=f(\mathbf{x})$; and $y=-f(\mathbf{x})$ otherwise. Given i.i.d. samples from D, the goal of the

[1] As is standard, we are assuming that $d=\Omega(1/\gamma^2)$; otherwise, a sample complexity bound of $\widetilde{O}(d/\epsilon)$ follows from standard VC-dimension arguments.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

learner is to output a hypothesis $h : \mathcal{X} \to \{\pm 1\}$ such that with high probability the 0-1 error $\mathrm{err}_D(h) \overset{\mathrm{def}}{=} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y]$ is small.

The concept class of halfspaces with a margin is defined as follows.

Definition 1.2 (γ-Margin Halfspaces)**.** Let D be a distribution over $\mathbb{S}^{d-1} \times \{\pm 1\}$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. Let $\mathbf{w}^* \in \mathbb{S}^{d-1}$ and $\gamma \in (0,1)$. We say that the distribution D satisfies the γ-margin condition with respect the halfspace $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})^2$, if (i) for $(\mathbf{x}, y) \sim D$, we have that $y = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$, and (ii) $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[|\mathbf{w}^* \cdot \mathbf{x}| < \gamma] = 0$. The parameter γ is called the margin of the halfspace $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$.

Information-theoretically, the best possible 0-1 error attainable for learning a concept class with Massart noise is $\mathrm{opt} := \mathbf{E}_{\mathbf{x}\sim D_{\mathbf{x}}}[\eta(\mathbf{x})]$. Since $\eta(\mathbf{x})$ is uniformly bounded above by η, it follows that $\mathrm{opt} \leq \eta$; also note that it may well be the case that $\mathrm{opt} \ll \eta$. Focusing on the class of γ-margin halfspaces, it follows from [MN06] that there exists a (computationally inefficient) estimator achieving error $\mathrm{opt} + \epsilon$ with sample complexity $\widetilde{O}(1/((1-2\eta)\gamma^2\epsilon))$; and moreover that this sample upper bound is nearly best possible (within a logarithmic factor) for any estimator. (That is, the sample complexity of the Massart learning problem is essentially the same as in the realizable case, as long as η is bounded from $1/2$.)

Taking computational considerations into account, the feasibility landscape of the problem changes. Prior work [DK22, NT22, DKMR22] has provided strong evidence that achieving error better than $\eta + \epsilon$ is not possible in polynomial time. Consequently, algorithmic research has been focusing on achieving the qualitatively weaker error guarantee of $\eta + \epsilon$. We note that efficiently obtaining any non-trivial guarantee had remained open since the 80s; see Appendix A.1 for a discussion. The first algorithmic progress for this problem is due to [DGT19], who gave a polynomial-time algorithm achieving error of $\eta + \epsilon$ with sample complexity $\mathrm{poly}(1/\gamma, 1/\epsilon)$. Subsequent work [CKMY20] gave an efficient algorithm with improved sample complexity of $\tilde{O}(1/(\gamma^4\epsilon^3))$. Prior to the current work, this remained the best known sample upper bound for efficient algorithms.

In summary, known computationally efficient algorithms for learning margin halfspaces with Massart noise require significantly more samples—namely, $\tilde{\Omega}(1/(\gamma^4\epsilon^3))$—than the information-theoretic minimum of $\widetilde{\Theta}_\eta(1/(\gamma^2\epsilon))$. It is thus natural to ask whether a polynomial-time algorithm with optimal (or near-optimal, i.e., within logarithmic factors) sample complexity exists. Recall that the answer to this question is affirmative in the realizable setting, where the Perceptron algorithm is optimal. Perhaps surprisingly, recent work [DDK$^+$23a] (see also [DDK$^+$23b]) gave evidence for the existence of inherent *information-computation tradeoffs* in the Massart noise model—in fact, even in the simpler model of Random Classification Noise (RCN) [AL88][3]. Specifically, they showed that any efficient Statistical Query (SQ) algorithm or low-degree polynomial tasks requires $\Omega(1/\epsilon^2)$ samples—a near quadratic blow-up compared to the $\tilde{O}(1/\epsilon)$ information-theoretic upper bound. This discussion serves as the motivation for the following question:

What is the optimal computational sample complexity *of the problem of learning γ-margin halfspaces with Massart noise?*

By the term "computational sample complexity" above, we mean the sample complexity of polynomial-time algorithms for the problem. Given the fundamental nature of this learning problem, we believe that a fine-grained sample complexity versus computational complexity analysis is interesting on its own merits. *In this work, we develop a computationally efficient algorithm with sample complexity of* $\tilde{O}(1/(\gamma^2\epsilon^2))$. Given the aforementioned information-computation tradeoffs, there is evidence that this upper bound is close to best possible. As a bonus, our algorithm is also simple and practical, relying on online SGD. (In fact, our algorithm runs in sample linear time, excluding a final testing step that slightly increases the runtime.)

1.1 Our Result and Techniques

Our main result is the following:

[2] We will henceforth assume that the threshold is $\theta = 0$, which is well-known to be no loss of generality.

[3] The RCN model is the special case of Massart noise, where $\eta(\mathbf{x}) = \eta$ for all points $\mathbf{x}$ in the domain.

Theorem 1.3 (Main Result, Informal). *Let D be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ that satisfies the η-Massart noise condition with respect to an unknown γ-margin halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. There is algorithm that draws $n = \tilde{O}(1/(\epsilon^2\gamma^2))$ samples from D, runs in time $\tilde{O}(dn/\epsilon)$, and with probability at least $9/10$ returns a vector $\hat{\mathbf{w}}$ such that $\mathrm{err}_D(\hat{\mathbf{w}}) \le \eta + \epsilon$.*

The sample upper bound of Theorem 1.3 nearly matches the computational sample complexity of the problem (for SQ algorithms and low-degree polynomial tests), which was shown to be $\Omega(1/(\epsilon^2\gamma) + 1/(\epsilon\gamma^2))$ [MN06, DDK+23a, DDK+23b]. That is, Theorem 1.3 comes close to resolving the fine-grained complexity of this basic task. Moreover, it matches known algorithmic guarantees for the easier case of Random Classification Noise [DDK+23a, KIT+23].

Independent Work Independent work [CKST24] obtained a learning algorithm for γ-margin halfspaces with essentially the same sample and computational complexity as ours.

Brief Overview of Techniques Here we provide a brief summary of our approach in tandem with a comparison to prior work. The algorithm of [DGT19] adaptively partitions the space into polyhedral regions and uses a different linear classifier in each region, each achieving error $\eta + \epsilon$ within the corresponding region. Their approach leverages the LeakyReLU loss (see (1)) as a convex proxy to the 0-1 loss. At a high-level, their approach reweights the samples in order to accurately classify a non-trivial fraction of points. [CKMY20] uses the LeakyReLU loss to efficiently identify a region where the value of the loss conditioned on this region is sub-optimal; they then use this procedure as a separation oracle along with online convex optimization (see also [DKTZ20b, DKK+21]) to output a linear classifier with 0-1 error at most $\eta + \epsilon$. Both of these approaches inherently require $\Omega(1/\epsilon^3)$ samples for the following reason: they both need to condition on a region where the probability mass of the distribution can be as small as $\Theta(\epsilon)$. Thus, even estimating the error of the loss would require at least $\Omega(1/\epsilon^2)$ conditional samples. Beyond the dependence on $1/\epsilon$, the sample complexity achieved in these prior works is also suboptimal in the margin parameter γ; namely, $\Omega(1/\gamma^4)$. This dependence follows from the facts that both of these works require estimating the loss in each iteration within error of at most $\gamma\epsilon$, and that their algorithmic approaches require $\Omega(1/\gamma^2)$ iterations.

To circumvent these issues, novel ideas are required. At a high-level, we design a uniform approach to decrease the "global" error, as opposed to the local error (as was done in prior work). Specifically, we construct a different sequence of convex loss functions, each of which attempts to accurately simulate the 0-1 objective. We note that a similar sequence of loss functions was used in the recent work [DKTZ24] in a related, but significantly different, adversarial online setting. Interestingly, a similar reweighting scheme was used in [CKMY20] for learning general Massart halfspaces. Beyond this similarity, these works have no implications for the sample complexity of our problem. (See Appendix A.2 for a detailed comparison.) Via this approach, we obtain an iterative algorithm which uses only $O_\gamma(1/\epsilon^2)$ samples in order to estimate the loss in each iterative step.

In more detail, note that the 0-1 loss can be written in the form $-\mathbf{E}[y\frac{\mathbf{w}\cdot\mathbf{x}}{|\mathbf{w}\cdot\mathbf{x}|}]$. We convexify this objective by considering, in each step, the loss $\ell(\mathbf{w}, \mathbf{u}) = -\mathbf{E}[y\frac{\mathbf{w}\cdot\mathbf{x}}{|\mathbf{u}\cdot\mathbf{x}|}]$, where $\mathbf{u}$ is independent of $\mathbf{w}$; this loss is convex with respect to $\mathbf{w}$. Observe that $\ell(\mathbf{w}, \mathbf{w})$ is proportional to the zero-one loss of $\mathbf{w}$. Unfortunately, it is possible that no optimal vector $\mathbf{w}^*$ (under 0-1 loss) minimizes $\ell(\mathbf{w}^*, \mathbf{w})$. For this reason, we consider the objective $\ell_\eta(\mathbf{w}, \mathbf{u}) = \mathbf{E}[(\mathbb{1}\{y \neq \mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\} - \eta - \epsilon)|\mathbf{w}\cdot\mathbf{x}|/|\mathbf{u}\cdot\mathbf{x}|]$. This new objective satisfies the following: $\ell_\eta(\mathbf{w}^*, \mathbf{u}) < -\epsilon\gamma$ for any vector $\mathbf{u}$ and any $\mathbf{w}^*$ that minimizes the 0-1 objective; and $\ell_\eta(\mathbf{w}, \mathbf{w}) \ge \epsilon$ as long as $\mathbf{w}$ incurs 0-1 error at least $\eta + \epsilon$. By the convexity of $\ell_\eta(\mathbf{w}, \mathbf{u})$, this allows us to construct a separation oracle. Namely, we draw enough samples so that $\widehat{\ell}_\eta(\mathbf{w}, \mathbf{w}) - \widehat{\ell}_\eta(\mathbf{w}^*, \mathbf{w}) \ge \epsilon/2$, where $\widehat{\ell}$ is the emprical version of the loss. Due to the nature of these objectives, $O_\gamma(1/\epsilon^2)$ samples per iteration suffice for this purpose. This in turn implies that the cutting planes method efficiently finds a near-optimal weight vector after $O(\log(1/\epsilon)/\gamma^2)$ iterations. Overall, this approach leads to an efficient algorithm with sample complexity $\tilde{O}_\gamma(1/\epsilon^2)$. To get the desired sample complexity of $\tilde{O}(1/(\epsilon^2\gamma^2))$, more ideas are needed.

In the previous paragraph, we hid an obstacle that makes the above approach fail. Specifically, it may be possible that, for many points $\mathbf{x}$, the value of $|\mathbf{u}\cdot\mathbf{x}|$ is arbitrarily small. To fix this issue, we consider a clipped reweighting as follows: $\ell'_\eta(\mathbf{w}, \mathbf{u}) = \mathbf{E}[(\mathbb{1}\{y \neq \mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\} - \eta - \epsilon)\frac{|\mathbf{w}\cdot\mathbf{x}|}{\max(|\mathbf{u}\cdot\mathbf{x}|, \gamma)}]$. This clipping step is not a problem for us, because the target halfspace $\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})$ was assumed to have margin γ. This guarantees that the difference between the expected (over y) pointwise losses at $(\mathbf{w}, \mathbf{w})$ and $(\mathbf{w}^*, \mathbf{w})$ is at least ϵ on the points $\mathbf{x}$ where $|\mathbf{u}\cdot\mathbf{x}| \le \gamma$. Indeed, when this is the case, then $|\mathbf{w}^*\cdot\mathbf{x}|/|\mathbf{u}\cdot\mathbf{x}| \ge 1$. Overall, this suffices to guarantee that $\ell'_\eta(\mathbf{w}, \mathbf{w}) - \ell'_\eta(\mathbf{w}^*, \mathbf{w}) \ge \epsilon$.

1.2 Notation

For $n \in \mathbb{Z}_+$, let $[n] \overset{\text{def}}{=} \{1, \ldots, n\}$. We use small boldface characters for vectors. For $\mathbf{x} \in \mathbb{R}^d$ and $i \in [d]$, $\mathbf{x}_i$ denotes the i-th coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2 \overset{\text{def}}{=} (\sum_{i=1}^d \mathbf{x}_i^2)^{1/2}$ denotes the ℓ_2-norm of $\mathbf{x}$. We will use $\mathbf{x} \cdot \mathbf{y}$ for the inner product of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. For a subset $S \subseteq \mathbb{R}^d$, we define the proj_S operator that maps a point $\mathbf{x} \in \mathbb{R}^d$ to the closest point in the set S. For $a, b \in \mathbb{R}$, we denote $W(a, b) \overset{\text{def}}{=} 1/\max(a, b)$. We will use $\mathbb{1}_A$ to denote the characteristic function of the set A, i.e., $\mathbb{1}\{\mathbf{x} \in A\} = 1$ if $\mathbf{x} \in A$, and $\mathbb{1}\{\mathbf{x} \in A\} = 0$ if $\mathbf{x} \notin A$. For $A, B \in \mathbb{R}$, we write $A \gtrsim B$ (resp. $A \lesssim B$) to denote that there exists a universal constant $C > 0$, such that $A \geq CB$ (resp. $A \leq CB$).

We use $\mathbf{E}_{x \sim D}[x]$ for the expectation of the random variable x with respect to the distribution D and $\mathbf{Pr}[\mathcal{E}]$ for the probability of event $\mathcal{E}$. For simplicity, we may omit the distribution when it is clear from the context. For $(\mathbf{x}, y) \sim D$, we use $D_\mathbf{x}$ for the marginal distribution of $\mathbf{x}$ and $D_y(\mathbf{x})$ for the distribution of y conditioned on $\mathbf{x}$. We use $\widehat{D}_N$ to denote the empirical distribution obtained by drawing N i.i.d. samples from D. We use $\mathrm{err}_D(\mathbf{w})$ to denote the 0-1 error of the halfspace defined by the weight vector $\mathbf{w}$ with respect to the distribution D, i.e., $\mathrm{err}_D(\mathbf{w}) \overset{\text{def}}{=} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y]$. We will use $\mathrm{err}(\mathbf{w}, \mathbf{x})$ for the 0-1 error of $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x})$ conditioned on $\mathbf{x}$, i.e., $\mathrm{err}(\mathbf{w}, \mathbf{x}) := \mathbf{Pr}_{y \sim D_y(\mathbf{x})}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y]$. Note that $\mathrm{err}_D(\mathbf{w}) = \mathbf{E}_{\mathbf{x} \sim D_\mathbf{x}}[\mathrm{err}(\mathbf{w}, \mathbf{x})]$. If D satisfies the η-Massart noise condition with respect to the halfspace $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x})$, then $\mathrm{err}(\mathbf{w}, \mathbf{x}) = \eta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})\} + (1 - \eta(\mathbf{x}))\mathbb{1}\{\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})\}$.

2 Our Algorithm and its Analysis: Proof of Theorem 1.3

In this section, we prove our main result. Algorithm 1 efficiently learns the class of margin halfspaces on the unit ball, in the presence of Massart noise, with sample complexity nearly matching the information-computation limit. Additionally, its runtime is linear in the sample size, excluding a final testing step to select the best hypothesis.

At a high-level, our algorithm leverages a carefully selected convex loss (or, more precisely, a sequence of convex losses) — serving as a proxy to the 0-1 error. A common loss function, introduced in this context by [DGT19] and leveraged in [DGT19, CKMY20], is the LeakyReLU function. This is the univariate function $\mathrm{LeakyReLU}_\lambda(t) = (1 - \lambda)\mathbb{1}\{t \geq 0\}t + \lambda\mathbb{1}\{t < 0\}t$, where $\lambda \in (0, 1)$ is the leakage parameter (that needs to be selected carefully). Roughly speaking, the convex function $\ell_\lambda(\mathbf{w}, \mathbf{x}, y) = \mathrm{LeakyReLU}_\lambda(-y(\mathbf{w} \cdot \mathbf{x}))$ can be viewed as a reasonable proxy to the 0-1 loss of the halfspace $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x})$ on the point $(\mathbf{x}, y)$. To see this, note that (see, e.g., Claim C.1)

$$\ell_\lambda(\mathbf{w}, \mathbf{x}, y) = (\mathbb{1}\{\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y\} - \lambda)|\mathbf{w} \cdot \mathbf{x}| \ . \tag{1}$$

Observe that a point $\mathbf{x}$ that is classified correctly by the halfspace $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x})$ will satisfy

$$\left(\mathbf{E}_{y \sim D_y(\mathbf{x})}[\mathbb{1}\{\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y\}] - \lambda\right)|\mathbf{w} \cdot \mathbf{x}| = (\eta(\mathbf{x}) - \lambda)|\mathbf{w} \cdot \mathbf{x}|$$

which is non-positive for $\lambda \geq \eta(\mathbf{x})$. Since the only guarantee we have is that $\eta(\mathbf{x}) \leq \eta$, this suggests that we need to select $\lambda \geq \eta$. It turns out that $\lambda := \eta$ is the optimal choice. We fix the choice of $\lambda := \eta$ throughout. On the other hand, if (the halfspace defined by) $\mathbf{w}$ misclassifies the point $\mathbf{x}$, this term becomes non-negative.

The factor $|\mathbf{w} \cdot \mathbf{x}|$ in Equation (1) reweights the 0-1 error so that points $\mathbf{x}$ for which $|\mathbf{w} \cdot \mathbf{x}|$ is sufficiently large (i.e., close to 1) have to be classified correctly by a minimizer of $\mathbf{E}_{(\mathbf{x},y)\sim D}[\ell_\lambda(\mathbf{w}, \mathbf{x}, y)]$. On the other hand, points closer to the separating hyperplane defined by $\mathbf{w}$, or points where $\eta(\mathbf{x})$ is close to $\lambda = \eta$, are not guaranteed to be classified correctly by the minimizer of this loss. We leverage this insight to construct a sequence of loss functions that reweight the points so that, to minimize the regret, we need to classify a large fraction of points; this leads to the desired error of $\eta + \epsilon$ with near-optimal sample complexity.

We now provide some intuition justifying our choice of surrogate loss functions. Observe that if we instead could minimize the function

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\ell_\lambda(\mathbf{w}, \mathbf{x}, y)/|\mathbf{w} \cdot \mathbf{x}|] = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(\mathbb{1}\{\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y\} - \lambda)] \ , \tag{2}$$

with respect to $\mathbf{w}$, we would obtain a halfspace with minimum 0-1 error; unfortunately, this reweighted loss is just a shift of the 0-1 loss, hence non-convex. To fix this issue, instead of reweighting by

$1/|\mathbf{w} \cdot \mathbf{x}|$, we will reweight by $W(\mathbf{v} \cdot \mathbf{x}, \gamma) \stackrel{\text{def}}{=} 1/\max(|\mathbf{v} \cdot \mathbf{x}|, \gamma)$, where γ is the margin parameter and $\mathbf{v}$ is an appropriately chosen vector that is independent of $\mathbf{w}$. The new loss is defined as follows:

$$\mathcal{L}_{\lambda,\mathbf{v}}(\mathbf{w}) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\ell_\lambda(\mathbf{w}, \mathbf{x}, y) W(\mathbf{v} \cdot \mathbf{x}, \gamma/2)] \;, \tag{3}$$

where for technical reasons we use $\gamma/2$ instead of γ in the maximum.

Since the parameter $\mathbf{v}$ is independent of $\mathbf{w}$, the loss $\mathcal{L}_{\lambda,\mathbf{v}}(\mathbf{w})$ remains convex in $\mathbf{w}$. At the same time, by carefully choosing $\mathbf{v}$, we can accurately simulate the non-convex 0-1 loss. Note that our reweighting term is a maximum over two terms. The reason for this choice is that, for some points $\mathbf{x}$, the quantity $|\mathbf{v} \cdot \mathbf{x}|$ can be arbitrarily small; taking the maximum avoids the loss becoming very large. In particular, the loss $\mathcal{L}_{\lambda,\mathbf{v}}(\mathbf{w})$ will be guaranteed to remain in a bounded length interval.

Our algorithm proceeds in a sequence of iterations. In the $(t+1)$-st iteration, it sets $\mathbf{v}$ to be $\mathbf{w}^t$, where $\mathbf{w}^t$ is the weight vector of step t. This choice attempts to simulate the 0-1 error at $\mathbf{w}^t$, as is suggested by Equation (2). Assume for simplicity that our current hypothesis is the halfspace defined by $\mathbf{w}$ and is such that $\mathbf{E}_{\mathbf{x}\sim D_\mathbf{x}}[\mathbb{1}\{|\mathbf{w} \cdot \mathbf{x}| \leq \gamma/2\}] = 0$. Note this implies that $W(\mathbf{w} \cdot \mathbf{x}, \gamma/2) = 1/|\mathbf{w} \cdot \mathbf{x}|$. By combining Equations (2) and (3), we get that $\mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w}) = \mathrm{err}_D(\mathbf{w}) - \lambda$; note that as long as $\mathrm{err}_D(\mathbf{w}) \geq \lambda + \epsilon$, we have that $\mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w}) \geq \epsilon$. On the other hand, the optimal halfspace $\mathbf{w}^*$ achieves a non-positive loss; from Equations (1) and (2), we have that

$$\begin{aligned}\mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w}^*) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(\mathbb{1}\{\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x}) \neq y\} - \lambda)|\mathbf{w}^* \cdot \mathbf{x}| W(\mathbf{w} \cdot \mathbf{x}, \gamma/2)] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim D_\mathbf{x}}[(\eta(\mathbf{x}) - \lambda)|\mathbf{w}^* \cdot \mathbf{x}| W(\mathbf{w} \cdot \mathbf{x}, \gamma/2)] \leq 0 \;,\end{aligned}$$

where the inequality follows from the fact that $\eta(\mathbf{x}) \leq \eta$. Recalling that $\mathcal{L}_{\lambda,\mathbf{v}}(\mathbf{w})$ is convex, if we run an Online Convex Optimization (OCO) algorithm, after T steps we are guaranteed to find a vector $\mathbf{w}$ such that $\mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w}) - \mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w}^*) \leq O(1/\sqrt{T})$. For $T = O(1/\epsilon^2)$, this gives that $\mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w}) < \epsilon/2$; and therefore we would have $\mathrm{err}_D(\mathbf{w}) < \lambda + \epsilon$. We provide an approach using this idea and the cutting planes algorithm in Appendix B that achieves sample complexity $\widetilde{O}(1/(\epsilon^2\gamma^4))$.

Our algorithm and its analysis work only with the gradient of $\mathcal{L}_{\lambda,\mathbf{v}}(\mathbf{w})$. The key novelty is the analysis of the sample complexity. The gradient of $\ell_\lambda(\mathbf{w}, \mathbf{x}, y) W(\mathbf{v} \cdot \mathbf{x}, \gamma)$ with respect to $\mathbf{w}$ has the following explicit form:

$$\mathbf{g}_{\lambda,\gamma}(\mathbf{w}, \mathbf{v}, \mathbf{x}, y) \stackrel{\text{def}}{=} ((1-2\lambda)\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) - y) W(\mathbf{v} \cdot \mathbf{x}, \gamma)\mathbf{x} = \frac{((1-2\lambda)\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) - y)}{\max(|\mathbf{v} \cdot \mathbf{x}|, \gamma)}\mathbf{x} \,.$$

Furthermore, we denote by $\mathbf{G}_D(\mathbf{w}, \mathbf{v}, \eta, \gamma) = \mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{g}_{\eta,\gamma}(\mathbf{w}, \mathbf{v}, \mathbf{x}, y)]$.

Before describing our algorithm and proving Theorem 2.1, we simplify our notation. We will omit the parameters η, γ from the function input (as they are fixed throughout). Therefore, we use $\mathbf{G}_{\widehat{D}_N^t}(\mathbf{w}, \mathbf{v}) \equiv \mathbf{G}_{\widehat{D}_N^t}(\mathbf{w}, \mathbf{v}, \eta, \gamma)$ and $\mathbf{g}(\mathbf{w}, \mathbf{v}, \mathbf{x}, y) \equiv \mathbf{g}_{\eta,\gamma/2}(\mathbf{w}, \mathbf{v}, \mathbf{x}, y)$.

Our algorithm is described in pseudocode below.

Algorithm 1 employs online SGD applied to a sequence of convex loss functions. We show that, after a certain number of iterations, the algorithm will find a weight vector achieving 0-1 error at most $\eta + \epsilon$. Since the desired vector may not be the last iterate, in the end, our algorithm returns the halfspace that achieves the smallest empirical 0-1 error.

We establish the following result, which implies Theorem 1.3.

Theorem 2.1 (Main Result)**.** *Let D be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ satisfying the η-Massart noise condition with respect to the γ-margin halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Given $N = \Theta(\log(1/(\gamma\delta))/\epsilon(1-2\eta))$ and $T = \Theta(\log(1/\delta)/(\epsilon^2\gamma^2))$, Algorithm 1 returns a vector $\hat{\mathbf{w}}$ such that $\mathrm{err}_D(\hat{\mathbf{w}}) \leq \eta + \epsilon$ with probability at least $1 - \delta$. The algorithm draws $n = O(N + T)$ samples from D and runs in $O(dNT)$ time.*

The rest of this section is devoted to the proof of Theorem 2.1.

Our algorithm sets $\mathbf{v} = \mathbf{w}^t$ in each round, therefore for the rest of the section we proceed by setting $\mathbf{v} = \mathbf{w}$ as arguments of $\mathbf{g}$ and $\mathbf{G}$.

```
**Input:** Sample access to a distribution $D$ supported in $\mathbb{S}^{d-1} \times \{\pm 1\}$ corrupted with $\eta$-Massart noise with respect to a halfspace $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$ that satisfies the $\gamma$-margin condition; parameters $\epsilon, \delta \in (0,1)$, and $N, T \in \mathbb{Z}_+$.
**Output:** Weight vector $\hat{\mathbf{w}}$ such that $\mathrm{err}_D(\hat{\mathbf{w}}) \leq \eta + \epsilon$ with probability at least $1 - \delta$.

    1. Let $c > 0$ be a sufficiently small universal constant.
    2. $t \leftarrow 0$, $\mathbf{w}^0 \leftarrow \mathbf{e}_1 = (1, 0, \ldots, 0)$, and $T = (1/c)\log(1/\delta)/(\epsilon^2\gamma^2)$.
    3. While $t \leq T$ do
        (a) Draw $(\mathbf{x}^{(t)}, y^{(t)})$ sample from $D$.
        (b) Set $\lambda_t \leftarrow c\gamma^2\epsilon$.
        (c) Update $\mathbf{w}^t$ as follows:        ▷ Update and project in the unit ball
                $\mathbf{v}^{t+1} \leftarrow \mathbf{w}^t - \lambda_t \mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) \qquad \mathbf{w}^{t+1} \leftarrow \frac{\mathbf{v}^{t+1}}{\max(\|\mathbf{v}^{t+1}\|_2, 1)}$
        (d) $t \leftarrow t + 1$.
    4. Draw $N$ samples from $D$ and construct the empirical distribution $\widehat{D}_N$.
    5. Return $\widehat{\mathbf{w}} = \mathrm{argmin}_{t \in [T+1]} \mathrm{err}_{\widehat{D}_N}(\mathbf{w}^t)$.
```

Algorithm 1: Learning Margin Halfspaces with Massart Noise

We decompose the stochastic gradient $\mathbf{g}(\mathbf{w}, \mathbf{w}, \mathbf{x}, y)$ into two parts: $\mathbf{g}(\mathbf{w}, \mathbf{w}, \mathbf{x}, y) = \mathbf{g}^1(\mathbf{w}, \mathbf{x}) + \mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)$, where

$$\mathbf{g}^1(\mathbf{w}, \mathbf{x}) = \left((1 - 2\eta)\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) - \mathop{\mathbf{E}}_{y \sim D_y(\mathbf{x})}[y]\right) W(\mathbf{w} \cdot \mathbf{x}, \gamma/2)\mathbf{x}$$

and

$$\mathbf{g}^2(\mathbf{w}, \mathbf{x}, y) = \left(\mathop{\mathbf{E}}_{y \sim D_y(\mathbf{x})}[y] - y\right) W(\mathbf{w} \cdot \mathbf{x}, \gamma/2)\mathbf{x} \; .$$

We also use $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w})$ and $\mathbf{G}^2_{\widehat{D}_N}(\mathbf{w})$ for the same decomposition after taking the empirical expectation, i.e., $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) = \mathbf{E}_{\mathbf{x} \sim (\widehat{D}_{\mathbf{x}})_N}[\mathbf{g}^1(\mathbf{w}, \mathbf{x})]$ and $\mathbf{G}^2_{\widehat{D}_N}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x}, y) \sim \widehat{D}_N}[\mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)]$.

This serves to decompose the gradient into two parts: one containing the population expectation over the random variable y, and the other containing the error between the empirical estimation of y and the population version of y. The vector $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w})$ contains the direction that will decrease the distance between $\mathbf{w}$ and $\mathbf{w}^*$, while $\mathbf{G}^2_{\widehat{D}_N}(\mathbf{w})$ contains the estimation error. To see this, observe that if we take the population expectation of $\mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)$, we will have:

$$\mathop{\mathbf{E}}_{(\mathbf{x}, y) \sim D}[\mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)] = \mathop{\mathbf{E}}_{\mathbf{x} \sim D_{\mathbf{x}}}\left[\left((1 - 2\eta(\mathbf{x}))\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x}) - \mathop{\mathbf{E}}_{y \sim D_y(\mathbf{x})}[y]\right) W(\mathbf{w} \cdot \mathbf{x}, \gamma/2)\mathbf{x}\right] = 0 \; ,$$

where we used that $\mathbf{E}_{y \sim D_y(\mathbf{x})}[y] = (1 - 2\eta(\mathbf{x}))\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$.

We start by bounding the contribution of $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w})$ in the direction $\mathbf{w} - \mathbf{w}^*$. We show that if instead of the corrupted label y at the point $\mathbf{x}$, we had access to $\mathbf{E}_{y \sim D_y(\mathbf{x})}[y] = (1 - 2\eta(\mathbf{x}))\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$, then the gradient has a large component in the direction of $\mathbf{w} - \mathbf{w}^*$. This effectively implies that $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w})$ can be used as a separation oracle, separating all the halfspaces with 0-1 error more than $\eta + \epsilon$ from the ones with smaller error.

Lemma 2.2 (Structural Lemma). *Let $N \in \mathbb{Z}_+$ and let D be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ satisfying the η-Massart condition with respect to the optimal classifier $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Let $\mathbf{w} \in \mathbb{R}^d$ be such that $\|\mathbf{w}\|_2 \leq 1$ and let $\{\mathbf{x}^{(i)}\}_{i=1}^N$ be a multiset of N i.i.d. samples from $D_{\mathbf{x}}$. Then, it holds $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq 2(\mathrm{err}_{\widehat{D}_N}(\mathbf{w}) - \eta)$, where $\widehat{D}_N$ is the corresponding empirical distribution.*

Proof. We partition $\mathbb{R}^d$ into two subsets R_1, R_2 as follows: R_1 contains the points that lie sufficiently far away from the separating hyperplane $\mathbf{w} \cdot \mathbf{x} = 0$, i.e., $R_1 \overset{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{w} \cdot \mathbf{x}| \geq \gamma/2\}$. R_2 contains the remaining points, i.e., $R_2 \overset{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{w} \cdot \mathbf{x}| < \gamma/2\}$.

We first show that for any $\mathbf{x} \in R_1$, the vector $\mathbf{g}^1(\mathbf{w}, \mathbf{x})$ has a large component parallel to the direction $\mathbf{w} - \mathbf{w}^*$. The proof of the claim below can be found in Appendix C.

Claim 2.3. *For any* $\mathbf{x}^{(i)} \in R_1$, *we have that* $\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq 2(\mathrm{err}(\mathbf{w}, \mathbf{x}^{(i)}) - \eta)$.

It remains to show that the same holds for all the points in R_2. The proof of the claim below can be found in Appendix C.

Claim 2.4. *For any* $\mathbf{x}^{(i)} \in R_2$, *we have that* $\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq 2(\mathrm{err}(\mathbf{w}, \mathbf{x}^{(i)}) - \eta)$.

Applying Claim 2.3 and Claim 2.4 for each sample in the set $\{\mathbf{x}^{(i)}\}_{i=1}^N$, we get that

$$\frac{1}{N}\sum_{i=1}^N \mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq \frac{2}{N}\sum_{i=1}^N (\mathrm{err}(\mathbf{w}, \mathbf{x}^{(i)}) - \eta) \ .$$

This completes the proof of Lemma 2.2. □

By Lemma 2.2, the gradient points towards the direction $\mathbf{w}^t - \mathbf{w}^*$, in the t-th iteration. This means that, in fact, the gradient is a subgradient of the potential loss $\Phi(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}^*\|_2^2$. This allows us to show convergence, even though it is generally not possible in a sequence of loss functions in the stochastic setting. We are now ready to prove our main result.

Proof of Theorem 2.1. Let T be the maximum number of iterations of Algorithm 1. Denote by $\mathcal{Z}^t := \{(\mathbf{x}^{(t)}, y^{(t)})\}$ the i.i.d. sample drawn from D in the t-th iteration, $t \in [T]$. Furthermore, let $\mathcal{F}_1, \dots, \mathcal{F}_T$ be the filtration with respect to the σ-algebra generated by $\mathcal{Z}^1, \dots, \mathcal{Z}^T$. We denote by H_t the event that $\mathrm{err}_D(\mathbf{w}^t) \geq \eta + \epsilon$.

Recall that Algorithm 1 uses the following update rule (see Step (3c)):

$$\mathbf{w}^{t+1} = \mathrm{proj}_{\{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq 1\}}(\mathbf{w}^t - \lambda_t \mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)})) \ ,$$

with $\lambda_t = c\gamma^2\epsilon$, for some sufficiently small absolute constant $c > 0$.

We begin by bounding from above the distance between $\mathbf{w}^{t+1}$ and $\mathbf{w}^*$ from the previous distance between $\mathbf{w}^t$ and $\mathbf{w}^*$. We have that

$$\begin{aligned}\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 &= \|\mathrm{proj}_{\{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq 1\}}(\mathbf{w}^t - \lambda_t \mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) - \mathbf{w}^*\|_2^2 \\ &\leq \|\mathbf{w}^t - \lambda_t \mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) - \mathbf{w}^*\|_2^2 \\ &= \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - 2\lambda_t \mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) \cdot (\mathbf{w}^t - \mathbf{w}^*) + \lambda_t^2 \|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)})\|_2^2 \ ,\end{aligned} \tag{4}$$

where in the first inequality we used the projection inequality, i.e., $\|\mathrm{proj}_B(\mathbf{v}) - \mathrm{proj}_B(\mathbf{u})\|_2 \leq \|\mathbf{v} - \mathbf{u}\|_2$ for any set B. We will decouple the mean of the random variable $\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}, y)$ and make it zero-mean.

To simplify the notation, we denote by $\xi_t := \left(\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) - \mathbf{G}_D^1(\mathbf{w}^t)\right) \cdot (\mathbf{w}^t - \mathbf{w}^*)$ and note that ξ_t is a zero-mean random variable over the sample $(\mathbf{x}^{(t)}, y^{(t)})$. Adding and subtracting $\mathbf{G}_D^1(\mathbf{w}^t)$ onto Inequality (4) a we get that

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 \underbrace{-2\lambda_t \mathbf{G}_D^1(\mathbf{w}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) + \lambda_t^2 \|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)})\|_2^2}_{I} \underbrace{-2\lambda_t \xi_t}_{\widehat{V}_t} \ . \tag{5}$$

We now outline the main steps of our analysis. Instead of accurately estimating the gradients in each round, we denote by $\widehat{V}_t$ the estimation error from which we bound above their sum. We first add and subtract the population gradient to obtain the I term, which is the decreasing direction. In this way, we decouple the expected decrease and the error of the approximation (see Claim 2.5). After that, we bound the contribution of the estimation error in Lemma 2.8. Observe that $\widehat{V}_t$ is a random variable that corresponds to the estimation error of the gradient. We will argue that with high probability the contribution of $\sum_{t=1}^{T} \widehat{V}_t$ is bounded; therefore, our algorithm will converge to an accurate solution.

Lemma 2.2 shows that $\mathbf{G}^1_{\widehat{D}^t_N}(\mathbf{w}^t)$ (and therefore the same holds for $\mathbf{G}^1_D(\mathbf{w}^t)$) contains substantial contribution towards to the direction $\mathbf{w}^t - \mathbf{w}^*$, depending of the current error. We show that our choice of step size guarantees a decreasing direction. To this end, we prove the following:

Claim 2.5. *Assume that the event H_t happens, i.e., $\mathrm{err}_D(\mathbf{w}^t) \geq \eta + \epsilon$. If $\lambda_t \leq \gamma^2\epsilon/8$, then $I \leq -\lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta)$.*

Proof of Claim 2.5. Recall that $I = -2\lambda_t \mathbf{G}^1_D(\mathbf{w}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) + \lambda_t^2 \|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)})\|_2^2$. By Lemma 2.2, we get that $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) \geq 2(\mathrm{err}_{\widehat{D}_N}(\mathbf{w}^t) - \eta)$; hence, by taking expectations over the samples, we also have $\mathbf{G}^1_D(\mathbf{w}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) \geq 2(\mathrm{err}_D(\mathbf{w}^t) - \eta)$. Furthermore, we have that $\|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)})\|_2^2 \leq 8/\gamma^2$. Hence, $I \leq -2\lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) + 8(\lambda_t^2/\gamma^2)$. The claim follows by noting that if $\lambda_t \leq \gamma^2\epsilon/8$, then $-\lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) + 8(\lambda_t^2/\gamma^2) \leq 0$. Therefore, we obtain

$$I \leq -\lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) \ .$$

This completes the proof of Claim 2.5. □

Therefore, our choice of parameters guarantees that $\lambda_t \leq \gamma^2\epsilon/8$. Using Claim 2.5 onto Inequality (5), we have that

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - \lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) + \widehat{V}_t \ . \tag{6}$$

Using Claim 2.5 and Inequality (6), we have that

$$\begin{aligned}\|\mathbf{w}^{T+1} - \mathbf{w}^*\|_2^2 &\leq \|\mathbf{w}^T - \mathbf{w}^*\|_2^2 - \lambda_T(\mathrm{err}_D(\mathbf{w}^T) - \eta) + \widehat{V}_T \\ &\leq \|\mathbf{w}^0 - \mathbf{w}^*\|_2^2 - \sum_{t=0}^{T} \lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) + \sum_{t=0}^{T} \widehat{V}_t \ . \end{aligned} \tag{7}$$

To complete the proof of Theorem 2.1, we need to bound the estimation error that corresponds to the random variable $\widehat{V}_t$. We show that $\widehat{V}_t$ does not increase the error by a lot. Recall that $\widehat{V}_t = -2\lambda_t\xi_t$.

Before proceeding, we provide some basic background on subgaussian random variables.

Definition 2.6 (Subgaussian Random Variable)**.** For $\sigma > 0$, a zero-mean random variable $X \in \mathbb{R}$ is called σ-subgaussian, if for any $\lambda \in \mathbb{R}$ it holds $\log(\mathbf{E}[\exp(\lambda X)]) \leq \lambda^2\sigma^2$.

Note that any zero-mean bounded random variable is subgaussian. Specifically, we have the following:

Fact 2.7 (Hoeffding's lemma, see, e.g., [Ver18])**.** *Let $X \in \mathbb{R}$ be a zero-mean random variable such that $|X| \leq \sigma$ for some $\sigma > 0$. Then X is $C\sigma$-subgaussian, where $C > 0$ is a universal constant.*

Equipped with the above context, we show the following:

Lemma 2.8. *With probability at least $1 - \delta$ over the random samples, it holds that $\sum_{t=0}^{T} \widehat{V}_t \leq C\gamma^2\epsilon^2 T + \log(1/\delta)$, where $C > 0$ is an absolute constant.*

Proof. We first show that ξ_t is a subgaussian random variable.

Claim 2.9. *The random vector ξ_t is $(16/\gamma)$-subgaussian.*

Proof of Claim 2.9. Note that $\xi_t = (\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) - \mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}, y)]) \cdot (\mathbf{w}^t - \mathbf{w}^*)$ and that by construction $\|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}, y)\|_2 \leq 4/\gamma$. Therefore, it holds that $|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) \cdot (\mathbf{w}^t - \mathbf{w}^*)| \leq 8/\gamma$, where we used that $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq 2$ as both of these vectors lie in the unit ball. Hence, by Fact 2.7, we have that ξ_t is $(16/\gamma)$-subgaussian. □

Using Claim 2.9 and Definition 2.6 with parameter $\lambda = -2\lambda_t$ and $X = \xi_t$, we have that

$$\log \mathbf{E}[\exp(\widehat{V}_t)] = \log \mathbf{E}[\exp(-2\lambda_t \xi_t)] \le C(\lambda_t^2/\gamma^2) ,$$

where $C > 0$ is a universal constant. To bound the contribution of $\sum_{t=0}^{T} \widehat{V}_t$, we use Markov's inequality with respect to the filtration $\mathcal{F}_1, \ldots, \mathcal{F}_T$. We have that for any $Z \in \mathbb{R}$, it holds that

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\mathcal{Z}^1,\ldots,\mathcal{Z}^T\sim D}\left[\sum_{t=0}^{T} \widehat{V}_t \ge Z\right] &= \mathop{\mathbf{Pr}}_{\mathcal{Z}^1,\ldots,\mathcal{Z}^T\sim D}\left[\exp\left(\sum_{t=0}^{T} \widehat{V}_t\right) \ge \exp(Z)\right] \\
&\le \mathop{\mathbf{E}}_{\mathcal{Z}^1,\ldots,\mathcal{Z}^T\sim D}\left[\exp\left(\sum_{t=0}^{T} \widehat{V}_t\right)\right]\exp(-Z) \\
&= \prod_{t=1}^{T} \mathop{\mathbf{E}}_{\mathcal{Z}^t\sim D}\left[\exp \widehat{V}_t \mid \mathcal{F}_t\right]\exp(-Z) \le \exp\left(C\sum_{t=0}^{T}\frac{\lambda_t^2}{\gamma^2} - Z\right) ,
\end{aligned}$$

where in the second inequality we use the independence of $\widehat{V}_t$ with $\{\widehat{V}_k\}_{k=1}^{t-1}$ with respect to the filtration $\mathcal{F}_t$. Recalling that $\lambda_t = c\gamma^2\epsilon$, where $c > 0$ is a sufficiently small universal constant, we have that

$$\mathop{\mathbf{Pr}}_{\mathcal{Z}^1,\ldots,\mathcal{Z}^T\sim D}\left[\sum_{t=0}^{T} \widehat{V}_t \ge Z\right] \le \exp\left(Cc^2\gamma^2\epsilon^2 T - Z\right) \le \exp\left(Cc^2\gamma^2\epsilon^2 T - Z\right) .$$

Setting $Z = Cc^2\gamma^2\epsilon^2 T + \log(1/\delta)$ and taking c to be a sufficiently small absolute constant (as is done in our algorithm), we get that $\mathbf{Pr}_{\mathcal{Z}^1,\ldots,\mathcal{Z}^T\sim D}\left[\sum_{t=0}^{T} \widehat{V}_t \ge Z\right] \le \delta$. This completes the proof of Lemma 2.8. □

Assume that until the round T the event H_T holds, i.e., for all $i \in [T]$ we have that $\mathrm{err}_D(\mathbf{w}^i) \ge \eta + \epsilon$. Using Lemma 2.8 onto Inequality (7), with probability at least $1 - \delta$, we have that:

$$\begin{aligned}
\|\mathbf{w}^{T+1} - \mathbf{w}^*\|_2^2 &\le \|\mathbf{w}^0 - \mathbf{w}^*\|_2^2 - \sum_{t=0}^{T} \lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) + \sum_{t=0}^{T} \widehat{V}_t \\
&\le \|\mathbf{w}^0 - \mathbf{w}^*\|_2^2 - cT\epsilon^2\gamma^2 + \log(1/\delta) .
\end{aligned}$$

Running the algorithm for $T = \Theta(\log(1/\delta)/(\epsilon^2\gamma^2))$ iterations guarantees that with probability at least $1 - \delta$, we will have that $\|\mathbf{w}^{T+1} - \mathbf{w}^*\|_2^2 \le 0$, which means $\mathbf{w}^{T+1} = \mathbf{w}^*$. In that case, i.e., in the case where all the events H_i for $i \in [T]$ hold, $\mathbf{w}^{T+1}$ achieves the same error as the optimal halfspace, thus it has 0-1 error of at most $\eta + \epsilon$. Therefore, at least one vector $\mathbf{w}^{t'}$ with $t' \in [T + 1]$ achieves 0-1 error of at most $\eta + \epsilon$. The algorithm, in Step (5), returns a vector $\widehat{\mathbf{w}}$ that has 0-1 error at most $\mathrm{err}_D(\widehat{\mathbf{w}}) \le \min_{t\in[T+1]} \mathrm{err}_D(\mathbf{w}^t) + \epsilon \le \eta + 2\epsilon$. The algorithm requires $N = O(\log(T/\delta)/(\epsilon(1 - 2\eta)))$ samples for Step (5), due to [MN06]. The algorithm draws a sample in each round and runs for at most T rounds. Therefore, Algorithm 1 draws $n = N + T = \widetilde{O}(\log(1/\delta)/(\epsilon^2\gamma^2))$ samples. The algorithm needs to test each of the T hypotheses with N samples to find the closest one. Therefore, the total runtime is $O(dTN)$ (as in the other subroutines the algorithm uses the samples only to estimate the gradients $\mathbf{g}$, which requires $O(1)$ additions of d-dimenional vectors). This completes the proof of Theorem 2.1. □

3 Conclusions and Open Problems

In this paper, we give the first sample near-optimal and computationally efficient algorithm for learning margin halfspaces in the presence of Massart noise. Specifically, the sample complexity of our algorithm nearly matches the computational sample complexity of the problem and its computational complexity is polynomial in the sample size. An interesting direction for future work is to develop a sample near-optimal and computationally efficient learner for general halfspaces (i.e., without the margin assumption). While our approach can likely be leveraged to obtain an efficient algorithm with sample complexity $\mathrm{poly}(d)/\epsilon^2$, the sample dependence on the dimension d would be suboptimal. Obtaining the right dependence on the dimension seems to require novel ideas, as prior works rely on fairly sophisticated methods [DV04, DKT21, DTK23] to effectively reduce to the large margin case.

Acknowledgments

ID was supported in part by NSF Medium Award CCF-2107079 and an H.I. Romnes Faculty Fellowship. NZ was supported in part by NSF Medium Award CCF-2107079.

References

[ABHU15] P. Awasthi, M. F. Balcan, N. Haghtalab, and R. Urner. Efficient learning of linear separators under bounded noise. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pages 167–190, 2015.

[ABHZ16] P. Awasthi, M. F. Balcan, N. Haghtalab, and H. Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, pages 152–192, 2016.

[AL88] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[Blu03] A. Blum. Machine learning: My favorite results, directions, and open problems. In *44th Symposium on Foundations of Computer Science (FOCS 2003)*, pages 11–14, 2003.

[CKMY20] S. Chen, F. Koehler, A. Moitra, and M. Yau. Classification under misspecification: Halfspaces, generalized linear models, and connections to evolvability. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

[CKMY21] S. Chen, F. Koehler, A. Moitra, and M. Yau. Online and distribution-free robustness: Regression and contextual bandits with huber contamination. In *FOCS*, 2021.

[CKST24] G. Chandrasekaran, V. Kontonis, K. Stavropoulos, and K. Tian. Learning noisy halfspaces with a margin: Massart is no harder than random. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[DDK$^+$23a] I. Diakonikolas, J. Diakonikolas, D. M. Kane, P. Wang, and N. Zarifis. Information-computation tradeoffs for learning margin halfspaces with random classification noise. In *COLT*, 2023.

[DDK$^+$23b] I. Diakonikolas, J. Diakonikolas, D. M. Kane, P. Wang, and N. Zarifis. Near-optimal bounds for learning gaussian halfspaces with random classification noise. In *NeurIPS*, 2023.

[DGT19] I. Diakonikolas, T. Gouleakis, and C. Tzamos. Distribution-independent PAC learning of halfspaces with Massart noise. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4751–4762. Curran Associates, Inc., 2019.

[DIK$^+$21] I. Diakonikolas, R. Impagliazzo, D. M. Kane, R. Lei, J. Sorrell, and C. Tzamos. Boosting in the presence of Massart noise. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021.

[DK22] I. Diakonikolas and D. Kane. Near-optimal Statistical Query hardness of learning halfspaces with Massart noise. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4258–4282. PMLR, 2022. Preliminary Version 2021: Arxiv eprint: 2012.09720.

[DKK$^+$20] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. A polynomial time algorithm for learning halfspaces with Tsybakov noise. *arXiv*, 2020.

[DKK$^+$21] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Efficiently learning halfspaces with Tsybakov noise. *STOC*, 2021.

[DKK$^+$22] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Learning general halfspaces with general Massart noise under the gaussian distribution. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, 2022*, pages 874–885. ACM, 2022.

[DKMR22] I. Diakonikolas, D. Kane, P. Manurangsi, and L. Ren. Cryptographic hardness of learning halfspaces with Massart noise. In *Advances in Neural Information Processing Systems*, 2022.

[DKRS22] I. Diakonikolas, D. Kane, L. Ren, and Y. Sun. SQ lower bounds for learning single neurons with Massart noise. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022.

[DKT21] I. Diakonikolas, D. Kane, and C. Tzamos. Forster decomposition and learning halfspaces with noise. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 7732–7744, 2021.

[DKTZ20a] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning halfspaces with Massart noise under structured distributions. In *Conference on Learning Theory, COLT*, 2020.

[DKTZ20b] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning halfspaces with Tsybakov noise. *arXiv*, 2020.

[DKTZ24] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Online Linear Classification with Massart Noise, 2024. Arxiv eprint: 2405.12958.

[DPT21] I. Diakonikolas, J. Park, and C. Tzamos. Relu regression with Massart noise. In *Advances in Neural Information Processing Systems*, 2021.

[DTK23] I. Diakonikolas, C. Tzamos, and D. M. Kane. A strongly polynomial algorithm for approximate forster transforms and its application to halfspace learning. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023*, pages 1741–1754. ACM, 2023.

[DV04] J. Dunagan and S. Vempala. Optimal outlier removal in high-dimensional spaces. *J. Computer & System Sciences*, 68(2):335–373, 2004.

[JL84] W. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[KIT$^+$23] V. Kontonis, F. Iliopoulos, K. Trinh, C. Baykal, G. Menghani, and E. Vee. Slam: Student-label mixing for distillation with unlabeled examples. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023.

[MN06] P. Massart and E. Nedelec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 10 2006.

[NT22] R. Nasser and S. Tiegel. Optimal SQ lower bounds for learning halfspaces with Massart noise. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1047–1074. PMLR, 2022.

[Ros58] F. Rosenblatt. The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958.

[Slo88] R. H. Sloan. Types of noise in data for concept learning. In *Proceedings of the First Annual Workshop on Computational Learning Theory*, COLT ’88, pages 91–96, San Francisco, CA, USA, 1988. Morgan Kaufmann Publishers Inc.

[Slo92] R. H. Sloan. Corrigendum to types of noise in data for concept learning. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992*, page 450, 1992.

[SSBD14] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[SZ07] S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[Val84] L. G. Valiant. A theory of the learnable. In *Proc. 16th Annual ACM Symposium on Theory of Computing (STOC)*, pages 436–445. ACM Press, 1984.

[Ver18] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[YZ17] S. Yan and C. Zhang. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 1056–1066, 2017.

[ZLC17] Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 1980–2022, 2017.

Supplementary Material

Organization The structure of this appendix is as follows: In Appendix A, we provide additional summary and comparison with related and prior work. In Appendix B, we provide a polynomial time cutting-planes based algorithm with sample complexity $\widetilde{O}(1/(\epsilon^2\gamma^4))$. Finally, in Appendix C, we provide the proofs omitted from Section 2.

A Related and Prior Work

A.1 Additional Related Work

The computational problem of learning halfspaces with Massart noise has been extensively studied, both in the distribution-specific and the distribution-free settings.

In the distribution-specific setting, the first efficient algorithm for homogeneous Massart halfspaces was given in [ABHU15]. Subsequent work generalized this result in various directions [ABHZ16, ZLC17, YZ17, DKTZ20a, DKTZ20b, DKK$^+$20, DKK$^+$21, DKK$^+$22].

The first algorithmic progress in the distribution-free setting was made by [DGT19], answering a longstanding open problem [Slo88, Slo92, Blu03]. Subsequent work gave an algorithm with improved sample complexity [CKMY20] and provided strong evidence that an error of $\eta + \epsilon$ is the best to hope for in polynomial time [DK22, NT22, DKMR22] (in both the Statistical Query model and under plausible cryptographic assumptions). In a related direction, [DIK$^+$21] gave the first efficient boosting algorithm in the presence of Massart noise, which can boost a weak learner to one with error $\eta + \epsilon$. Finally, we note that natural generalizations of the Massart model to learning real-valued functions (in an essentially distribution-free setting) have also been studied [CKMY21, DPT21, DKRS22].

Very recent work [DDK$^+$23a] gave SQ (and low-degree polynomial testing) lower bounds for learning γ-margin halfspaces with RCN [AL88], which is a special case of Massart noise. Specifically, [DDK$^+$23a] showed that any efficient SQ algorithm for the problem requires sample complexity $\Omega(1/(\gamma^{1/2}\epsilon^2))$. Subsequently, [DDK$^+$23b] showed a related SQ lower bound under the Gaussian distribution, which can be adapted to obtain a lower bound of $\Omega(1/(\gamma\epsilon^2))$ for the margin setting.

A.2 Comparison with [DKTZ24]

The work [DKTZ24] uses a similar sequence of loss functions for the problem of "online learning" Massart margin halfspaces. Intuitively, their goal is to minimize regret in an adversarial online setting. In their online setting, the adversary in each round commits to covariates $\mathbf{x}^1, \mathbf{x}^2 \in \mathbb{R}^d$ and distribution D^t over $\mathbb{R}_+ \times \mathbb{R}_+$. Then the algorithm observes the covariates, chooses an action $a \in \{1, 2\}$, and observes a reward $r_a \in \mathbb{R}_+$. It is only guaranteed that there exists a unit vector $\mathbf{w}^*$ so that $\mathbf{E}_{(r_1,r_2)\sim D^t}[\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x}^1 - \mathbf{w}^* \cdot \mathbf{x}^2)(r_a - r_b)] \geq \Delta$ for some $\Delta > 0$.

Despite this superficial similarity, the work of [DKTZ24] has no new implications on the sample complexity of PAC learning Massart halfspaces with a margin. Specifically, they achieve a regret bound of $O(T^{3/4}/\gamma)$. If one translates this bound to a sample complexity upper bound for PAC learning, one would obtain a bound of $\Omega(1/(\epsilon^4\gamma^8))$ — which is quantitatively worse than prior work of [DGT19, CKMY20].

At a technical level, our work leverages this sequence of loss functions as subgradients of the potential function $\Phi(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}^*\|_2^2$. Via a novel analysis, we show that these subgradients $\Omega(\epsilon)$-correlate with the direction of $\mathbf{w} - \mathbf{w}^*$. This in turn means that we can expect a decrease of order $\Omega(\lambda\epsilon)$ in each iteration, where λ is the corresponding step-size, as long as we get 0-1 error more than $\eta + \epsilon$. This structural understanding suffices for obtaining an algorithm, based on a separation oracle, that achieves a sample complexity of $\widetilde{O}(1/(\gamma^4\epsilon^2))$. In order to obtain an algorithm with near-optimal sample complexity (and runtime), we required additional new ideas as elaborated in the body of the paper.

B Learning Margin Massart Halfspaces via Cutting Planes

In this section, we show how to use the cutting-planes method along with Lemma 2.2 to efficiently learning margin Massart Halfspaces using $\widetilde{O}(1/(\gamma^4\epsilon^2))$ samples.

Specifically, we establish the following result:

Theorem B.1 (Learning Margin Massart Halfspaces with Cutting Planes)**.** *Let D be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ which satisfies the η-Massart noise condition with respect to the γ-margin halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Given $N = \Theta(\log(1/(\gamma\delta)/(\gamma^4\epsilon^2))$ i.i.d. samples from D, there is a $\mathrm{poly}(d, N)$ time algorithm that returns a vector $\hat{\mathbf{w}}$ such that $\mathrm{err}_D(\hat{\mathbf{w}}) \leq \eta + \epsilon$ with probability at least $1 - \delta$.*

Remark B.2. We can always assume that $d = \widetilde{O}(1/\gamma^2)$. This holds since we can efficiently preprocess the data, using the Johnson-Lindenstrauss transform [JL84]. Similar dimension-reduction steps have been use in prior work, e.g., [CKMY20, DDK$^+$23a].

Given the above remark, it suffices to establish the following:

Theorem B.3. *Let D be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ which satisfies the η-Massart noise condition with respect to the γ-margin halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Given $N = \Theta(d\log(1/(\gamma\delta)/(\gamma^2\epsilon^2))$ i.i.d. samples from D, there is a $\mathrm{poly}(d, N)$ time algorithm that returns a vector $\hat{\mathbf{w}}$ such that $\mathrm{err}_D(\hat{\mathbf{w}}) \leq \eta + \epsilon$ with probability at least $1 - \delta$.*

The idea of using the cutting plane method is slightly adapted from [CKMY20]. Given access to a separation oracle for a convex set $\mathcal{K}$, we can find a point inside the set $\mathcal{K}$ by querying the separation oracle $O(d \log d)$ times. The difference with [CKMY20] is that we are using a more sophisticated (and sample efficient) separation oracle. This allows us to use $O(1/\epsilon^2)$ samples, instead of $O(1/\epsilon^3)$ samples, and leads to the optimal sample complexity as a function of ϵ (but not γ).

Fact B.4. *Suppose that $\mathcal{K}$ is an (unknown) convex body in $\mathbb{R}^d$ which contains a Euclidean ball of radius $r > 0$ and contained in a Euclidean ball centered at the origin of radius $R > 0$. There exists an algorithm which, given access to a separation oracle for $\mathcal{K}$, finds a point $\mathbf{x}^* \in \mathcal{K}$, runs in time $\mathrm{poly}(\log(R/r), d)$, and makes $O(d\log(Rd/r))$ calls to the separation oracle.*

We first show that if we get enough samples, we can efficiently approximate the gradients $\mathbf{G}(\mathbf{w}, \mathbf{w})$. Formally, we have:

Proposition B.5 (Separation Oracle)**.** *Let $\epsilon, \delta \in (0, 1)$ and let D be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ satisfying the η-Massart noise condition with respect to the halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Fix $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 \leq 1$. Let $N \gtrsim \log(1/(\gamma\delta))/(\epsilon^2\gamma^2))$ and $\widehat{D}_N$ be the corresponding empirical distribution. Then, with probability at least $1 - \delta$, it holds that*

$$\mathbf{G}_{\widehat{D}_N}(\mathbf{w}, \mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq 2(\mathrm{err}_D(\mathbf{w}) - \eta) - \epsilon \,.$$

Proof. By construction, $\mathbf{G}_{\widehat{D}_N}(\mathbf{w}, \mathbf{w}) = \mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) + \mathbf{G}^2_{\widehat{D}_N}(\mathbf{w})$ and by Lemma 2.2 we have that $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq 2(\mathrm{err}_{\widehat{D}_N}(\mathbf{w}) - \eta)$. By definition, we have $\mathbf{E}_{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)}) \sim D}[\mathbf{G}^2_{\widehat{D}_N}(\mathbf{w})] = 0$, where the expectation is taken with respect to the sample set. Note that the norm of $\mathbf{g}^1(\mathbf{w}, \mathbf{x}), \mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)$, i.e., $\|\mathbf{g}^1(\mathbf{w}, \mathbf{x})\|_2, \|\mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)\|_2$, is bounded pointwise from above by $4/\gamma$ for all $\mathbf{w} \in \mathbb{R}^d$. This can be seen as $\|\mathbf{x}\|_2 \leq 1$, $W(\cdot, \gamma/2) \leq 2/\gamma$, and $(1 - 2\eta), (1 - 2\eta(\mathbf{x})) \leq 1$.

We use the following concentration inequality to show that our sample size is enough to guarantee that the estimated gradient is close to its population version.

Fact B.6 ([SZ07], Lemma 1)**.** *Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \in \mathbb{R}^d$ be random vectors such that for each $i \in [n]$ it holds $\|\mathbf{Z}_i\|_2 \leq M < \infty$ almost surely and let $\sigma^2 = \sum_{i=1}^n \mathbf{E}[\|\mathbf{Z}_i\|_2^2]$. Then, we have that for any $\epsilon > 0$,*

$$\mathbf{Pr}\left[\left\|\frac{1}{n}\sum_{i=1}^n (\mathbf{Z}_i - \mathbf{E}[\mathbf{Z}_i])\right\|_2 \geq \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon}{2M}\log\left(1 + \frac{nM\epsilon}{\sigma^2}\right)\right) \,.$$

Using Fact B.6, along with the inequality $\log(1+z) \geq z/2$, for $z \in (0,1)$, we get that if $N \geq \Theta(\frac{\log(1/\delta)}{(\epsilon\gamma)^2})$, with probability at least $1-\delta$, we have

$$\left\|\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{g}^1(\mathbf{w},\mathbf{x})]\right\|_2 \leq \epsilon \;, \tag{8}$$

and

$$\left\|\mathbf{G}^2_{\widehat{D}_N}(\mathbf{w}) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{g}^2(\mathbf{w},\mathbf{x},y)]\right\|_2 \leq \epsilon \;. \tag{9}$$

To complete the proof, recall that by Lemma 2.2 it holds $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \geq 2(\mathrm{err}_{\widehat{D}_N}(\mathbf{w})-\eta)-\epsilon$. Therefore, by taking the expectation over $D_\mathbf{x}$, we get that

$$\mathbf{G}^1_D(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \geq 2(\mathrm{err}_D(\mathbf{w})-\eta) \;.$$

The proof is completed by recalling that $\|\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) - \mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{g}^1(\mathbf{w},\mathbf{x})]\|_2 \leq \epsilon$ from Inequality (8) and that $\mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{g}^2(\mathbf{w},\mathbf{x},y)] = 0$. □

Equipped with Proposition B.5, we are ready to prove a weaker version of Theorem 2.1 using separation oracles and the cutting plane algorithm. Formally, we show that

Proof of Theorem B.3. Our convex set $\mathcal{K}$ is a Euclidean ball of radius $\gamma/2$ centered at $\mathbf{w}^*$. To see that, note that for any $\mathbf{v}$ such that $\|\mathbf{w}^*-\mathbf{v}\|_2 \leq \gamma/2$, we have that $|(\mathbf{w}^*-\mathbf{v})\cdot\mathbf{x}| \leq \gamma/2$ for any $\mathbf{x}$ with $\|\mathbf{x}\|_2 = 1$. This implies that $\gamma/2 + \mathbf{w}^*\cdot\mathbf{x} \geq \mathbf{v}\cdot\mathbf{x} \geq \mathbf{w}^*\cdot\mathbf{x} - \gamma/2$. Moreover, by definition we have that $\mathbf{w}^*\cdot\mathbf{x} \geq \gamma$. Hence, if $\mathbf{w}^*\cdot\mathbf{x} \geq 0$, we have that $\mathbf{v}\cdot\mathbf{x} \geq \gamma/2$; and if $\mathbf{w}^*\cdot\mathbf{x} \leq 0$, we have that $\mathbf{v}\cdot\mathbf{x} \leq -\gamma/2$. Therefore, this ball contains all the vectors $\mathbf{w}$ with margin $\gamma/2$ and separates the points in the same way as $\mathbf{w}^*$.

Therefore, as long as we are not in the set $\mathcal{K}$ or the 0-1 error is more than $\eta+\epsilon$, we can use Proposition B.5 to construct a new separation oracle. By Fact B.4, the maximum number of calls to the separation oracle is $T = O(d\log(d/\gamma))$. By Proposition B.5, in each round we need $n = O(\log(T/\delta))/(\epsilon^2\gamma^2)$ samples from D to construct a separation oracle. Therefore, the maximum number of samples is $O(nT) = O(d\log(T/\delta))/(\epsilon^2\gamma^2)$. This completes the proof. □

C Omitted Proofs from Section 2

C.1 Proof of Claim C.1

Claim C.1 (Claim 2.1 [DGT19]). *For any* $\mathbf{w},\mathbf{x}$, *we have that*

$$\ell_\lambda(\mathbf{w},\mathbf{x},y) = \big(\mathbb{1}\{y(\mathbf{w}\cdot\mathbf{x}) \leq 0\} - \lambda\big)|\mathbf{w}\cdot\mathbf{x}| \;.$$

Proof. Recall that

$$\ell_\lambda(\mathbf{w},\mathbf{x},y) = \mathrm{LeakyReLU}_\lambda(-y(\mathbf{w}\cdot\mathbf{x})) = (1-\lambda)\mathbb{1}\{y(\mathbf{w}\cdot\mathbf{x}) \leq 0\}(-y\mathbf{w}\cdot\mathbf{x}) + \lambda\mathbb{1}\{y(\mathbf{w}\cdot\mathbf{x}) > 0\}(-y\mathbf{w}\cdot\mathbf{x}) \;.$$

Therefore, we have that

$$\begin{aligned}\ell_\lambda(\mathbf{w},\mathbf{x},y) &= (1-\lambda)\mathbb{1}\{y(\mathbf{w}\cdot\mathbf{x}) \leq 0\}|y\mathbf{w}\cdot\mathbf{x}| - \lambda\mathbb{1}\{y(\mathbf{w}\cdot\mathbf{x}) > 0\}|y\mathbf{w}\cdot\mathbf{x}| \\ &= \mathbb{1}\{y(\mathbf{w}\cdot\mathbf{x}) \leq 0\}|\mathbf{w}\cdot\mathbf{x}| - \lambda|\mathbf{w}\cdot\mathbf{x}| = \Bigg(\mathbb{1}\{y(\mathbf{w}\cdot\mathbf{x}) \leq 0\} - \lambda\Bigg)|\mathbf{w}\cdot\mathbf{x}| \;,\end{aligned}$$

where we used that $y \in \{\pm 1\}$. □

C.2 Proof of Claim 2.3

We restate and prove the following claim:

Claim 2.3. *For any* $\mathbf{x}^{(i)} \in R_1$, *we have that* $\mathbf{g}^1(\mathbf{w},\mathbf{x}^{(i)})\cdot(\mathbf{w}-\mathbf{w}^*) \geq 2(\mathrm{err}(\mathbf{w},\mathbf{x}^{(i)})-\eta)$.

Proof of Claim 2.3. For any $\mathbf{x}^{(i)} \in R_1$, we have that

$$\begin{aligned}
\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot \mathbf{w} &= \Big((1-2\eta)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}^{(i)}) - (1-2\eta(\mathbf{x}^{(i)}))\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}^{(i)})\Big)\mathbf{w}\cdot\mathbf{x}^{(i)}W(\mathbf{w}\cdot\mathbf{x}^{(i)}) \\
&= \Big((1-2\eta)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}^{(i)}) - (1-2\eta(\mathbf{x}^{(i)}))\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}^{(i)})\Big)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}^{(i)}) \\
&= 2(\mathrm{err}(\mathbf{w}, \mathbf{x}^{(i)}) - \eta)\ , \qquad (10)
\end{aligned}$$

where we used that for any $\mathbf{x}^{(i)} \in R_1$, $W(\mathbf{w}\cdot\mathbf{x}^{(i)}) = 1/|\mathbf{w}\cdot\mathbf{x}^{(i)}|$, and hence $W(\mathbf{w}\cdot\mathbf{x}^{(i)}, \gamma/2)\mathbf{w}\cdot\mathbf{x}^{(i)} = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x}^{(i)})$; and that $\mathrm{err}(\mathbf{w}, \mathbf{x}^{(i)}) = \eta(\mathbf{x}^{(i)})$ if $\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}^{(i)}) = \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$ and $1-\eta(\mathbf{x}^{(i)})$ otherwise.

We now bound the contribution of $\mathbf{w}^*$. Since $\eta(\mathbf{x}) \leq \eta$, we have

$$(1-2\eta(\mathbf{x})) - (1-2\eta)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}) {\geq} 0\ .$$

Therefore, we have that

$$\begin{aligned}
\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot \mathbf{w}^* &= \Big((1-2\eta)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) - (1-2\eta(\mathbf{x}))\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\Big)\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})|\mathbf{w}^*\cdot\mathbf{x}|W(\mathbf{w}\cdot\mathbf{x}^{(i)}) \\
&= -\Big((1-2\eta(\mathbf{x})) - (1-2\eta)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\Big)|\mathbf{w}^*\cdot\mathbf{x}|W(\mathbf{w}\cdot\mathbf{x}^{(i)}) \leq 0\ ,
\end{aligned}$$

which gives that $-\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot \mathbf{w}^* \geq 0$. This completes the proof of Claim 2.3. □

C.3 Proof of Claim 2.4

We restate and prove the following:

Claim 2.4. *For any* $\mathbf{x}^{(i)} \in R_2$*, we have that* $\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq 2(\mathrm{err}(\mathbf{w}, \mathbf{x}^{(i)}) - \eta)$.

Proof of Claim 2.4. We have that

$$\begin{aligned}
\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot (\mathbf{w} - \mathbf{w}^*) &= \Big((1-2\eta)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}^{(i)}) - (1-2\eta(\mathbf{x}^{(i)}))\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}^{(i)})\Big)\left(\frac{\mathbf{w}\cdot\mathbf{x}^{(i)} - \mathbf{w}^*\cdot\mathbf{x}^{(i)}}{\max(\gamma/2, |\mathbf{w}\cdot\mathbf{x}^{(i)}|)}\right) \\
&= \Big((1-2\eta)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}^{(i)}) - (1-2\eta(\mathbf{x}^{(i)}))\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}^{(i)})\Big)\left(\frac{\mathbf{w}\cdot\mathbf{x}^{(i)} - \mathbf{w}^*\cdot\mathbf{x}^{(i)}}{\gamma/2}\right)\ ,
\end{aligned}$$

where we used that $\max(\gamma/2, |\mathbf{w}\cdot\mathbf{x}^{(i)}|) = \gamma/2$ for any $\mathbf{x}^{(i)} \in R_2$. Since $\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})$ has γ-margin, we have that $|\mathbf{w}^*\cdot\mathbf{x}^{(i)}| \geq \gamma$. Since $\mathbf{x}^{(i)} \in R_2$, it holds $|\mathbf{w}\cdot\mathbf{x}^{(i)}| < \gamma/2$. Therefore, $-\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}^{(i)})(\mathbf{w}\cdot\mathbf{x}^{(i)} - \mathbf{w}^*\cdot\mathbf{x}^{(i)}) = \big(|\mathbf{w}^*\cdot\mathbf{x}^{(i)}| - \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}^{(i)})\mathbf{w}\cdot\mathbf{x}^{(i)}\big) \geq \gamma/2$. This in turn implies that

$$\begin{aligned}
\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot (\mathbf{w} - \mathbf{w}^*) &\geq (1-2\eta(\mathbf{x}^{(i)}) - (1-2\eta)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}^{(i)})\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}^{(i)})) \\
&= 2(\mathrm{err}(\mathbf{w}, \mathbf{x}^{(i)}) - \eta)\ ,
\end{aligned}$$

completing the proof of Claim 2.4. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract summarizes the result provided in Theorem 1.3 (and Theorem 2.1). The introduction describes how this contribution resolves an open problem in the literature by summarizing the motivation for the model and describing prior work's contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are clearly stated in the statements of each theorem and are discussed in the introduction of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Each theorem statement provides all the assumptions and we provide a complete proof for all statements that is either in the main body of the paper or in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: The paper is theoretical in nature and does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper is theoretical in nature and does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The work is theoretical and we do not see any major or immediate implications on society.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The work is theoretical.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This work does not use any assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This work does not use any assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve any crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.