UC San Diego

UC San Diego Previously Published Works

Title

Efficiently learning halfspaces with Tsybakov noise

Permalink

https://escholarship.org/uc/item/2hv0x2hr

Authors

Diakonikolas, Ilias

Kane, Daniel M

Kontonis, Vasilis

et al.

Publication Date

2021-06-15

DOI

10.1145/3406325.3450998

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Efficiently Learning Halfspaces with Tsybakov Noise

Ilias Diakonikolas
University of Wisconsin-Madison
USA
ilias@cs.wisc.edu

Daniel M. Kane
University of California at San Diego
USA
dakane@ucsd.edu

Vasilis Kontonis
University of Wisconsin-Madison
USA
kontonis@wisc.edu

Christos Tzamos
University of Wisconsin-Madison
USA
tzamos@wisc.edu

Nikos Zarifis
University of Wisconsin-Madison
USA
zarifis@wisc.edu

ABSTRACT

We study the problem of PAC learning homogeneous halfspaces in the presence of Tsybakov noise. In the Tsybakov noise model, the label of every sample is independently flipped with an adversarially controlled probability that can be arbitrarily close to $1/2$ for a fraction of the samples. *We give the first polynomial-time algorithm for this fundamental learning problem.* Our algorithm learns the true halfspace within any desired accuracy ϵ and succeeds under a broad family of well-behaved distributions including log-concave distributions. Prior to our work, the only previous algorithm for this problem required quasi-polynomial runtime in $1/\epsilon$.

Our algorithm employs a recently developed reduction [29] from learning to certifying the non-optimality of a candidate halfspace. This prior work developed a quasi-polynomial time certificate algorithm based on polynomial regression. *The main technical contribution of the current paper is the first polynomial-time certificate algorithm.* Starting from a non-trivial warm-start, our algorithm performs a novel "win-win" iterative process which, at each step, either finds a valid certificate or improves the angle between the current halfspace and the true one. Our warm-start algorithm for isotropic log-concave distributions involves a number of analytic tools that may be of broader interest. These include a new efficient method for reweighting the distribution in order to recenter it and a novel characterization of the spectrum of the degree-2 Chow parameters.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

PAC learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
STOC '21, June 21–25, 2021, Virtual, Italy
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8053-9/21/06.
https://doi.org/10.1145/3406325.3450998

ACM Reference Format:
Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. 2021. Efficiently Learning Halfspaces with Tsybakov Noise. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC '21), June 21–25, 2021, Virtual, Italy.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3406325.3450998

1 INTRODUCTION

The main result of this paper is the first polynomial-time algorithm for learning halfspaces in the presence of Tsybakov noise under a broad family of distributions. Before we explain our contributions in detail, we provide some context and motivation for this work.

1.1 Background

Learning in the presence of noise is a central challenge in machine learning. In this paper, we study the (supervised) binary classification setting, where the goal is to learn a Boolean function from random labeled examples with noisy labels. In more detail, we focus on the problem of learning *homogeneous halfspaces* in Valiant's PAC learning model [58] when the labels have been corrupted by *Tsybakov noise* [57].

A (homogeneous) halfspace is any function $h_{\mathbf{w}} : \mathbb{R}^d \to \{\pm 1\}$ of the form $h_{\mathbf{w}}(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$, where the vector $\mathbf{w} \in \mathbb{R}^d$ is called the weight vector of $h_{\mathbf{w}}$ and $\mathrm{sign} : \mathbb{R} \to \{\pm 1\}$ is defined by $\mathrm{sign}(t) = 1$ if $t \geq 0$ and $\mathrm{sign}(t) = -1$ otherwise. Halfspaces (or Linear Threshold Functions) are arguably the most fundamental and extensively studied concept class in the learning theory and machine learning literature, starting with early work in the 1950s and 60s [53–55] and leading to fundamental and practically important techniques [33, 59].

Halfspaces are known to be efficiently learnable without noise, i.e., when the labels are consistent with a halfspace, see, e.g., [49]. In the presence of noisy labels, the picture is more muddled. In the agnostic model [38, 41] (when a constant fraction of the labels can be adversarially chosen), learning halfspaces is computationally hard [18, 31, 35], even under the Gaussian distribution [26, 34]. This motivates the study of "benign" noise models, where positive results may be possible. The most basic such model, known as Random Classification Noise (RCN) [1], prescribes that each label is flipped independently with probability *exactly* $\eta < 1/2$. In the RCN model, halfspaces are known to be learnable in polynomial time [10].

The uniform noise assumption in the RCN model is accepted to be unrealistic. To address this issue, various natural noise models have been proposed and studied, capturing a number of realistic

noise sources. The two most prominent such models are, in order of increasing difficulty, the Massart (or bounded) noise model [51], and the Tsybakov noise model [57]. In the Massart model, the label of a datapoint $\mathbf{x}$ is flipped independently with probability $\eta(\mathbf{x})$ *at most* $\eta < 1/2$. Importantly, the flipping probability depends on the datapoint $\mathbf{x}$ (instance specific noise).

Motivation for Tsybakov Noise Model. The bounded (Massart) noise assumption, i.e., that the probability that labels are flipped is *globally* bounded away from $1/2$, fails to accurately capture a number of practically relevant noise sources, including the *human annotator noise* [9, 15, 42, 43]. In particular, the humans responsible for labeling the training data are much more prone to incorrectly classify points closer to the decision boundary (where "cats" and "dogs" look almost the same), than points far from the boundary. For example, it was empirically shown in [44] that when non-expert annotators (Amazon Mechanical Turk) were used to annotate the RTE-1 dataset [16], roughly 20% of the datapoints were classified almost at random, i.e., had $\eta(\mathbf{x}) \approx 1/2$. More broadly, a long line of research (both applied and theoretical) [12, 13, 28, 32, 39, 52, 63] focuses on noise models that do not restrict the flipping probability globally, but allow it to be arbitrarily close to $1/2$ near the decision boundary. On the other hand, since datapoints from low-density regions are also likely to be classified almost randomly (see, e.g., [32] and references therein), assuming that high noise rates occur only close to the decision boundary does not sufficiently capture these situations.

The Tsybakov noise model [50] provides a unified framework that significantly extends the Massart noise condition to capture the above scenarios: it prescribes that the label of each example is independently flipped with some probability which is controlled by an adversary, but is not uniformly bounded by a constant less than $1/2$. In particular, it allows the flipping probabilities to be *arbitrarily close to* $1/2$ for a fraction of the examples. Importantly, it makes *no geometric assumptions* about the noise, e.g., that it is only potentially large close to the decision boundary.

Formally, we have the following definition:

Definition 1.1 (PAC Learning with Tsybakov Noise)**.** Let C be a concept class of Boolean-valued functions over $X = \mathbb{R}^d$, $\mathcal{F}$ be a family of distributions on X, $0 < \epsilon < 1$ be the error parameter, and $0 \leq \alpha < 1, A > 0$ be parameters of the noise model.

Let f be an unknown target function in C. A *Tsybakov example oracle*, $\mathrm{EX}^{\mathrm{Tsyb}}(f, \mathcal{F})$, works as follows: Each time $\mathrm{EX}^{\mathrm{Tsyb}}(f, \mathcal{F})$ is invoked, it returns a labeled example $(\mathbf{x}, y)$, such that: (a) $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$, where $\mathcal{D}_{\mathbf{x}}$ is a fixed distribution in $\mathcal{F}$, and (b) $y = f(\mathbf{x})$ with probability $1 - \eta(\mathbf{x})$ and $y = -f(\mathbf{x})$ with probability $\eta(\mathbf{x})$. Here $\eta(\mathbf{x})$ is an *unknown* function that satisfies the (α, A)-Tsybakov noise condition. That is, for any $0 < t \leq 1/2$, $\eta(\mathbf{x})$ satisfies $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\eta(\mathbf{x}) \geq 1/2 - t] \leq A\, t^{\frac{\alpha}{1-\alpha}}$.

Let $\mathcal{D}$ denote the joint distribution on $(\mathbf{x}, y)$ generated by the above oracle. A learning algorithm is given i.i.d. samples from $\mathcal{D}$ and its goal is to output a hypothesis function $h : X \to \{\pm 1\}$ such that with high probability h is ϵ-close to f, i.e., it holds $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[h(\mathbf{x}) \neq f(\mathbf{x})] \leq \epsilon$.

The Tsybakov noise model was proposed in [50], then refined in [57], and subsequently studied in a number of works, see, e.g., [6, 8, 11, 36, 37, 57]. All these prior works address information-theoretic aspects of the model, i.e., do not provide computationally efficient algorithms in high dimensions. The only algorithmic result we are aware of in this model is the prior work by a subset of the authors [29], which gave a *quasi-polynomial* time algorithm for learning homogeneous halfspaces under a family of well-behaved distributions (including log-concave distributions). Obtaining a *polynomial* time algorithm for any *any non-trivial setting* (even under Gaussian Marginals) was a long-standing open problem in learning theory, see, e.g., [2].

It is easy to see that the Tsybakov model becomes more challenging as the parameter α in Definition 1.1 decreases. In particular, it is well-known that $\mathrm{poly}(d, 1/\epsilon^{1/\alpha})$ samples are necessary (and sufficient) to learn halfspaces in this model. That is, an exponential dependence in $1/\alpha$ is information-theoretically required for any algorithm that solves this problem.

We note that the error guarantee of Definition 1.1 is a strong identifiability guarantee for the true function, which is information-theoretically impossible in the agnostic model. In the following remark, we emphasize that even a constant factor approximation to the optimal misclassification error is insufficient for identifiability. This is important as it implies a computational separation between the Tsybakov and agnostic models, even under Gaussian marginals.

Remark 1.2 (Identifiability versus Misclassification Error). *Definition 1.1 requires that the learning algorithm identifies* the true function $f \in C$ *within arbitrary accuracy* ϵ. *A related commonly used loss function is the misclassification error, i.e., the probability* $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]$. *We note that having an efficient algorithm with misclassification error* $\mathrm{OPT} + \epsilon$ *for all* $\epsilon > 0$, *where* $\mathrm{OPT} = \inf_{g\in C} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[g(\mathbf{x}) \neq y]$, *is equivalent to having an efficient algorithm with the guarantee of Definition 1.1. We emphasize however that there is a major qualitative difference between achieving misclassification error of* $\mathrm{OPT} + \epsilon$ *and achieving error* $c \cdot \mathrm{OPT} + \epsilon$, *for a constant* $c > 1$. *The latter guarantee only allows us to approximate* f *within error* $\Omega(\mathrm{OPT})$.

Obtaining error OPT+ϵ in the agnostic model is known to require time $d^{\mathrm{poly}(1/\epsilon)}$ for halfspaces under Gaussian marginals [26, 34, 40]. On the positive side, [5, 17, 25, 30] gave $\mathrm{poly}(d/\epsilon)$ time algorithms for agnostically learning halfspaces under log-concave marginals. These algorithms have error of $O(\mathrm{OPT}) + \epsilon$, which is significantly weaker as explained in Remark 1.2.

1.2 Our Contributions

The existence of a computationally efficient learning algorithm in the presence of Tsybakov noise for any natural concept class and under any distributional assumptions has been a long-standing open problem in learning theory. *In this work, we make significant progress in this direction by essentially resolving the complexity of learning halfspaces in this model.*

In this section, we formally state our contributions. We start by defining the distribution family for which our algorithms succeed.

Definition 1.3 (Well-Behaved Distributions)**.** For $L, R, U > 0$ and $k \in \mathbb{Z}_+$, a distribution $\mathcal{D}_{\mathbf{x}}$ on $\mathbb{R}^d$ is called (k, L, R, U)-well-behaved if for any projection $(\mathcal{D}_{\mathbf{x}})_V$ of $\mathcal{D}_{\mathbf{x}}$ on a k-dimensional subspace V of $\mathbb{R}^d$, the corresponding pdf γ_V on V satisfies the following properties:

(i) $\gamma_V(\mathbf{x}) \geq L$, for all $\mathbf{x} \in V$ with $\|\mathbf{x}\|_2 \leq R$ (anti-anti-concentration), and (ii) $\gamma_V(\mathbf{x}) \leq U$ for all $\mathbf{x} \in V$ (anti-concentration). If, additionally, there exists $\beta \geq 1$ such that, for any $t > 0$ and unit vector $\mathbf{w} \in \mathbb{R}^d$, we have that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle \mathbf{w}, \mathbf{x}\rangle| \geq t] \leq \exp(1 - t/\beta)$ (sub-exponential concentration), we call $\mathcal{D}_\mathbf{x}$ (k, L, R, U, β)-well-behaved.

We focus on the case that the marginal distribution $\mathcal{D}_\mathbf{x}$ on the examples is well-behaved for some values of the relevant parameters. Definition 1.3 specifies the concentration and anti-concentration conditions on the low-dimensional projections of the data distribution that are required for our learning algorithm. Throughout this paper, we will take $k = 3$, i.e., we only require 3-dimensional projections to have such properties.

Interestingly, the class of well-behaved distributions is quite broad. In particular, it is easy to show that the broad class of isotropic log-concave distributions is well-behaved for L, R, U, β being universal constants. Moreover, as Definition 1.3 does not require a specific functional form for the underlying density function, it encompasses a much more general set of distributions.

Since the complexity of our algorithm depends (polynomially) on $1/L, 1/R, U, \beta$, we state here a simplified version of our main result for the case that these parameters are bounded by a universal constant. To simplify the relevant theorem statements, we will sometimes say that a distribution $\mathcal{D}$ of labeled examples in $\mathbb{R}^d \times \{\pm 1\}$ is well-behaved to mean that its marginal distribution $\mathcal{D}_\mathbf{x}$ is well-behaved. We show:

Theorem 1.4 (Learning Tsybakov Halfspaces under Well-Behaved Distributions). *Let $\mathcal{D}$ be a well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)$. There exists an algorithm that draws $N = O_{A,\alpha}(d/\epsilon)^{O(1/\alpha)}$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(N, d)$ time, and computes a vector $\widehat{\mathbf{w}}$ such that, with high probability we have that $\mathrm{err}_{0-1}^{\mathcal{D}_\mathbf{x}}(h_{\widehat{\mathbf{w}}}, f) \leq \epsilon$.*

See Theorem 5.1 for a more detailed statement.

For the class of log-concave distributions, we give a significantly more efficient algorithm:

Theorem 1.5 (Learning Tsybakov Halfspaces under Log-concave Distributions). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)$ and is such that $\mathcal{D}_\mathbf{x}$ is isotropic log-concave. There exists an algorithm that draws $N = \mathrm{poly}(d)\, O(A/\epsilon)^{O(1/\alpha^2)}$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(N, d)$ time, and computes a vector $\widehat{\mathbf{w}}$ such that, with high probability, we have that $\mathrm{err}_{0-1}^{\mathcal{D}_\mathbf{x}}(h_{\widehat{\mathbf{w}}}, f) \leq \epsilon$.*

See Theorem 5.2 for a more detailed statement. Since the sample complexity of the problem is $\mathrm{poly}(d, 1/\epsilon^{1/\alpha})$, the algorithm of Theorem 1.5 is qualitatively close to best possible.

1.3 Overview of Techniques

Here we give an intuitive summary of our techniques in tandem with a comparison to the most relevant prior work. A more detailed technical discussion is provided in the proceeding sections.

Our learning algorithms employ the certificate-based framework of [29]. At a high-level, this framework allows us to efficiently reduce the problem of *finding* a near-optimal halfspace $h_{\widehat{\mathbf{w}}}(\mathbf{x}) = \mathrm{sign}(\langle \widehat{\mathbf{w}}, \mathbf{x}\rangle)$ to the (easier) problem of *certifying* whether a candidate halfspace $h_\mathbf{w}(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}\rangle)$ is "far" from the optimal halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)$. The idea is to use a certificate algorithm (as a black-box) and combine it with an online convex optimization routine. Roughly speaking, starting from an initial guess $\mathbf{w}_0$ for $\mathbf{w}^*$, a judicious combination of these two ingredients allows us to efficiently compute a near-optimal halfspace $\widehat{\mathbf{w}}$, i.e., one that the certifying algorithm cannot reject. We note that a similar approach has been used in [14] for converting non-proper learners to proper learners in the Massart noise model.

With the aforementioned approach as the starting point, the learning problem reduces to that of designing an efficient certifying algorithm. In recent work [29], the authors developed a certifying algorithm for Tsybakov halfspaces based on high-dimensional polynomial regression. This method leads to a certifying algorithm with sample complexity and runtime $d^{\mathrm{polylog}(1/\epsilon)}$, i.e., a quasi-polynomial upper bound. As we will explain in Section 3.1, the [29] approach is inherently limited to quasi-polynomial time and new ideas are needed to obtain a polynomial time algorithm. *The main contribution of this paper is the design of a polynomial-time certificate algorithm for Tsybakov halfspaces under well-behaved distributions.*

The key idea to design a certificate in the Tsybakov noise model is the following simple but crucial observation: If $\mathbf{w}^*$ is the normal vector to true halfspace, then for any non-negative function $T(\mathbf{x})$, it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[T(\mathbf{x}) y \langle \mathbf{w}^*, \mathbf{x}\rangle] \geq 0$. On the other hand, for any $\mathbf{w} \neq \mathbf{w}^*$ there exists a non-negative function $T(\mathbf{x})$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[T(\mathbf{x})\, y \langle \mathbf{w}, \mathbf{x}\rangle] < 0$. In other words, there exists a *reweighting of the space* that makes the expectation of $y \langle \mathbf{w}, \mathbf{x}\rangle$ negative (Fact 3.1). Note that we can always use as $T(\mathbf{x})$ the indicator of the disagreement region between the candidate halfspace $h_\mathbf{w}(\mathbf{x})$ and the optimal halfspace $f(\mathbf{x}) = h_{\mathbf{w}^*}(\mathbf{x})$. Of course, since optimizing over the space of non-negative functions is intractable, we need to restrict our search space to a "simple" parametric family of functions. In [29], squares of low-degree polynomials were used, which led to a quasi-polynomial upper bound.

In this work, we consider certifying functions of the form:

$$T(\mathbf{x}) = \frac{\mathbb{1}\left\{\sigma_1 \leq \langle \mathbf{w}, \mathbf{x}\rangle \leq \sigma_2\,, -t_1 \leq \left\langle \mathbf{v}, \mathrm{proj}_{\mathbf{w}^\perp} \frac{\mathbf{x}}{\langle \mathbf{w}, \mathbf{x}\rangle} \right\rangle \leq -t_2\right\}}{\langle \mathbf{w}, \mathbf{x}\rangle}$$

that are parameterized by a vector $\mathbf{v}$ and scalar thresholds σ_1, σ_2, $t_1, t_2 > 0$. Here $\mathrm{proj}_{\mathbf{w}^\perp}$ denotes the orthogonal projection on the subspace orthogonal to $\mathbf{w}$. It will be important for our approach that functions of this form are specified by $O(d)$ parameters.

Of course, it may not be a priori clear why functions of this form can be used as certifying functions in our setting. The intuition behind choosing functions of this simple form is given in Section 3.1. In particular, in Claim 3.4, we show that for any incorrect guess $\mathbf{w}$ there *exists* a *certifying vector* $\mathbf{v}$ that makes the expectation $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[T(\mathbf{x})\, y \langle \mathbf{w}, \mathbf{x}\rangle]$ negative. In fact, the vector $\mathbf{v} = \mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^* / \left\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\right\|_2 := (\mathbf{w}^*)^{\perp_\mathbf{w}}$ suffices for this purpose.

The key challenge is in finding such a certifying vector $\mathbf{v}$ algorithmically. We note that our algorithm in general does not find $(\mathbf{w}^*)^{\perp_\mathbf{w}}$. But it does find a vector $\mathbf{v}$ with similar behavior, in the sense of making the $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[T(\mathbf{x})\, y \langle \mathbf{w}, \mathbf{x}\rangle]$ sufficiently negative. To achieve this goal, we take a two-step approach: The first step involves computing an initialization vector $\mathbf{v}_0$ that has non-trivial

correlation with $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$. In our second step, we give a perceptron-like update rule that iteratively improves the initial guess until it converges to a certifying vector $\mathbf{v}$. While this algorithm is relatively simple, its correctness relies on a win-win analysis (Lemma 3.12) whose proof is quite elaborate. In more detail, we show that for any *non-certifying* vector $\mathbf{v}$ that is sufficiently correlated with $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$, we can efficiently compute a direction that improves its correlation to $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$. We then argue (Lemma 3.13) that by choosing an appropriate step size this iteration converges to a certifying vector within a small number of steps.

A subtle point is that the aforementioned analysis does not take place in the initial space, where the underlying distribution is well-behaved and the labels are Tsybakov homogeneous halfspaces, but in a transformed space. The transformed space is obtained by restricting our points in a band and then performing an appropriate "perspective" projection on the subspace orthogonal to $\mathbf{w}$ (Section 3.2). Fortunately, we are able to show (Proposition 3.6) that this transformation preserves the structure of the problem: The transformed distribution remains well-behaved (albeit with somewhat worse parameters) and satisfies the Tsybakov noise condition (again with somewhat worse parameters) with respect to a potentially biased halfspace. In fact, this consideration motivated our use of the perspective projection in the definition of $T(\mathbf{x})$.

It remains to argue how to compute an initialization vector $\mathbf{v}_0$ that acts as a warm-start for our algorithm. Naturally, the sample complexity and runtime of our certificate algorithm depend on the quality of the initialization. The simplest way to initialize is by using a random unit vector. With random initialization, we achieve initial correlation roughly $1/\sqrt{d}$, which leads to a certifying algorithm with complexity $(d/\epsilon)^{O(1/\alpha)}$ (Theorem 3.3). This simple initialization suffices to obtain Theorem 1.4 for the general class of well-behaved distributions.

To obtain our faster algorithm for log-concave marginals (Theorem 1.5), we use the exact same approach described above starting from a better initialization. Our algorithm to obtain a better starting vector leverages additional structural properties of log-concave distributions. Our initialization algorithm runs in $\mathrm{poly}(d)$ time (independent of $1/\alpha$) and computes a unit vector whose correlation with $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ is $\Omega(\epsilon^{1/\alpha})$ (Theorem 4.2).

Specifically, our initialization algorithm works as follows:

(1) It starts by conditioning on a random sufficiently narrow band around the current candidate $\mathbf{w}$ and projecting the samples on the subspace $\mathbf{w}^{\perp}$.
(2) It transforms the resulting distribution to ensure that it is isotropic log-concave through rescaling and rejection sampling.
(3) It then computes the degree-2 *Chow parameters* and uses them to construct a low-dimensional subspace V inside which $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ has sufficiently large projection. This subspace V is the span of the degree-1 Chow vector and the large eigenvectors of the degree-2 Chow matrix.
(4) Finally, the algorithm outputs a uniformly random vector in V that can be shown to have the desired correlation with $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$.

The resulting distribution after the initial conditioning in Step 1 is still log-concave and approximately satisfies the Tsybakov noise condition with respect to a near-origin centered halfspace orthogonal to $\mathbf{w}$. However, the distribution may no longer be zero-centered and may contain a tiny amount of non-Tsybakov noise — in the sense that we may end with points $\mathbf{x}$ having $\eta(\mathbf{x}) > 1/2$. As we can control the total non-Tsybakov noise, the latter is not a significant issue. We address the former issue by reweighting the distribution to make it isotropic. We do this by applying rejection sampling with probability $\min(1, \exp(-\langle \mathbf{x}, \mathbf{r}\rangle))$, for some vector $\mathbf{r}$ that we compute via SGD (so that the resulting mean is near-zero) and then rescaling by the inverse covariance matrix.

After the first two steps, our goal is to find any vector with non-trivial correlation $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$, given that the underlying distribution is isotropic log-concave. We show that the labels y must correlate with some degree-2 polynomial in $\left\langle (\mathbf{w}^*)^{\perp_{\mathbf{w}}}, \mathbf{x}\right\rangle$. Our algorithm crucially exploits this property, along with recently established "thin shell" estimates [48] for log-concave distributions, to show that a large part of this correlation is explained by the vector of degree-1 Chow parameters and the top few eigenvectors of the degree-2 Chow matrix. This implies that the subspace V spanned by those vectors contains a non-trivial part of $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$, and thus a random vector from V has non-trivial correlation with $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ with constant probability.

1.4 Related Work

Recent work by a subset of the authors [29] gave the first non-trivial algorithm for learning homogeneous halfspaces with Tsybakov noise under a family of "well-behaved" distributions. The notion of well-behaved distributions in that work is somewhat different than ours, but also contains log-concave distributions. The sample complexity and runtime of the [29] algorithm is $d^{\mathrm{polylog}(1/\epsilon)}$ and the quasi-polynomial upper bound is tight for their techniques.

The Tsybakov noise model lies in between the Massart model [51, 56] and the agnostic model [38, 41]. During the past five years, substantial algorithmic progress has been made on learning with Massart noise in both the distribution-specific setting [3, 4, 28, 61–63] and the distribution-free PAC model [14, 19]. The algorithmic techniques in these prior works are known to inherently fail for the more challenging Tsybakov noise model, and new ideas are needed for this more general setting.

Learning in the agnostic model is known to be computationally hard, even under well-behaved marginals. Specifically, recent work [26, 34] proved Statistical Query lower bounds of $d^{\mathrm{poly}(1/\epsilon)}$ for agnostically learning halfspaces to error $\mathrm{OPT}+\epsilon$ under Gaussian marginals. This lower bound bound is qualitatively matched by the L_1 regression algorithm [40]. A related line of work [5, 17, 25, 30, 45] gave efficient algorithms for agnostically learning halfspaces under log-concave marginals. While these algorithms run in $\mathrm{poly}(d/\epsilon)$ time, they achieve a "semi-agnostic" error guarantee of $O(\mathrm{OPT})+\epsilon$, instead of $\mathrm{OPT}+\epsilon$. As already mentioned in Remark 1.2, this guarantee is significantly weaker and cannot be used to approximate the true function within any desired accuracy.

This work is part of the broader direction of designing robust learning algorithms for a range of statistical models with respect to natural and challenging noise models. A line of work [5, 20–23, 25, 27, 45–47] has given efficient robust learners for a range of

settings in the presence of adversarial corruptions. See [24] for a recent survey on the topic.

1.5 Structure of This Paper

After the required preliminaries in Section 2, in Section 3 we give our certifying algorithm for the class of well-behaved distributions. In Section 4, we give our more efficient certifying algorithm for log-concave distributions. Finally, in Section 5, we review the certificate framework and put everything together to prove our main results.

2 PRELIMINARIES

For $n \in \mathbb{Z}_+$, let $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$. We will 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 $\|\mathbf{x}\|_2 \stackrel{\text{def}}{=} (\sum_{i=1}^d \mathbf{x}_i^2)^{1/2}$ denotes the ℓ_2-norm of $\mathbf{x}$. We will use $\langle \mathbf{x}, \mathbf{y} \rangle$ for the inner product of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\theta(\mathbf{x}, \mathbf{y})$ for the angle between $\mathbf{x}, \mathbf{y}$. We will use $\mathbb{1}_A$ to denote the characteristic function of the set A, i.e., $\mathbb{1}_A(\mathbf{x}) = 1$ if $\mathbf{x} \in A$ and $\mathbb{1}_A(\mathbf{x}) = 0$ if $\mathbf{x} \notin A$.

Let $\mathbf{e}_i$ be the i-th standard basis vector in $\mathbb{R}^d$. For $d \in \mathbb{N}$, let $\mathbb{S}^{d-1} \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$ be the unit sphere. We will denote by $\text{proj}_U(\mathbf{x})$ the projection of $\mathbf{x}$ onto the subspace $U \subset \mathbb{R}^d$. For a subspace $U \subset \mathbb{R}^d$, let $U^\perp$ be the orthogonal complement of U. For a vector $\mathbf{w} \in \mathbb{R}^d$, we use $\mathbf{w}^\perp$ to denote the subspace spanned by vectors orthogonal to $\mathbf{w}$, i.e., $\mathbf{w}^\perp = \{\mathbf{u} \in \mathbb{R}^d : \langle \mathbf{w}, \mathbf{u} \rangle = 0\}$. Finally, we denote by $\mathbf{w}^{\perp_\mathbf{v}}$ the projection of the vector $\mathbf{w}$ on the subspace $\mathbf{v}^\perp$ after normalization, i.e., $\mathbf{w}^{\perp_\mathbf{v}} = \frac{\mathbf{w} - \langle \mathbf{w}, \mathbf{v} \rangle \mathbf{v}}{\|\mathbf{w} - \langle \mathbf{w}, \mathbf{v} \rangle \mathbf{v}\|_2}$.

We use $\mathbf{E}[X]$ for the expectation of the random variable X and $\mathbf{Pr}[\mathcal{E}]$ for the probability of event $\mathcal{E}$.

We study the binary classification setting where labeled examples $(\mathbf{x}, y)$ are drawn i.i.d. from a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \{\pm 1\}$. We denote by $\mathcal{D}_\mathbf{x}$ the marginal of $\mathcal{D}$ on $\mathbf{x}$. The zero-one error between two hypotheses f, h (with respect to $\mathcal{D}_\mathbf{x}$) is $\text{err}_{0-1}^{\mathcal{D}_\mathbf{x}}(f, h) \stackrel{\text{def}}{=} \mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[f(\mathbf{x}) \neq h(\mathbf{x})]$.

3 EFFICIENTLY CERTIFYING NON-OPTIMALITY

In this section, we give an efficient algorithm that can certify whether a candidate weight vector $\mathbf{w}$ defines a halfspace $h_\mathbf{w}(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$ that is far from the optimal halfspace $\text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)$. Before we formally describe and analyze our algorithm, we provide some intuition.

Background: Certifying Non-Optimality. Our approach relies on the following simple but powerful idea, introduced in [29]: If a candidate weight vector $\mathbf{w}$ defines a halfspace $h_\mathbf{w}(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$ that differs from the target halfspace $f(\mathbf{x}) = \text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)$, there exists a *certifying function* of its non-optimality. In more detail, there exists a *reweighting of the space* that makes the expectation of $y \langle \mathbf{w}, \mathbf{x} \rangle$ negative. This intuition is captured in Fact 3.1, stated below. We note that the only assumption required for this to hold is that the underlying distribution on examples assigns positive mass to the symmetric difference of any two distinct halfspaces.

Fact 3.1 (Certifying Function). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that: (a) For any pair of distinct unit vectors $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d$, we have that*

$$\mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[h_\mathbf{v}(\mathbf{x}) \neq h_\mathbf{u}(\mathbf{x})] > 0$$

. (b) $\mathcal{D}$ satisfies the Tsybakov noise condition with optimal classifier $f(\mathbf{x}) = \text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)$. Then we have:

(1) For any $T : \mathbb{R}^d \mapsto \mathbb{R}_+$, we have that $\mathbf{E}_{(\mathbf{x},y) \sim \mathcal{D}}[T(\mathbf{x})\, y \langle \mathbf{w}^*, \mathbf{x} \rangle] \geq 0$.

(2) For any non-zero vector $\mathbf{w} \in \mathbb{R}^d$ such that $\theta(\mathbf{w}, \mathbf{w}^*) > 0$, there exists a function $T : \mathbb{R}^d \mapsto \mathbb{R}_+$ satisfying $\mathbf{E}_{(\mathbf{x},y) \sim \mathcal{D}}[T(\mathbf{x})\, y \langle \mathbf{w}, \mathbf{x} \rangle] < 0$.

Proof. For the first statement, note that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}}[T(\mathbf{x})\, y \left\langle \mathbf{w}^*, \mathbf{x} \right\rangle] = \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[T(\mathbf{x}) | \left\langle \mathbf{w}^*, \mathbf{x} \right\rangle | (1 - \eta(\mathbf{x}))] -$$
$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[T(\mathbf{x}) | \left\langle \mathbf{w}^*, \mathbf{x} \right\rangle | \eta(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[T(\mathbf{x}) | \left\langle \mathbf{w}^*, \mathbf{x} \right\rangle | (1 - 2\eta(\mathbf{x}))] \geq 0\,,$$

where we used the fact that $\eta(\mathbf{x}) \leq 1/2$ and $T(\mathbf{x}) \geq 0$.

For the second statement, let $\mathbf{w} \neq \mathbf{0}$ and $\theta(\mathbf{w}, \mathbf{w}^*) > 0$. By picking as a certifying function T the indicator function of the disagreement region between f and $h_\mathbf{w}$, i.e., $T(\mathbf{x}) \stackrel{\text{def}}{=} \mathbb{1}\{h_\mathbf{w}(\mathbf{x}) \neq f(\mathbf{x})\}$, we have that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}}[T(\mathbf{x})\, y \langle \mathbf{w}, \mathbf{x} \rangle] = - \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[T(\mathbf{x}) | \langle \mathbf{w}, \mathbf{x} \rangle | (1 - 2\eta(\mathbf{x}))]\ .$$

We claim that $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[T(\mathbf{x}) | \langle \mathbf{w}, \mathbf{x} \rangle | (1 - 2\eta(\mathbf{x}))] > 0$, which proves the second statement. To see this, we use our assumption that the symmetric difference between any pair of distinct homogeneous halfspaces has positive probability mass. First, we note that from the Tsybakov condition (for any choice of parameters) we have that $\mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\eta(\mathbf{x}) = 1/2] = 0$. So, it suffices to show that $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[T(\mathbf{x}) | \langle \mathbf{w}, \mathbf{x} \rangle |] > 0$.

Let $\mathbf{w}'$ be a non-zero vector such that the hyperplane $\{\mathbf{x} : \langle \mathbf{w}', \mathbf{x} \rangle = 0\}$ is contained in the disagreement region $\{\mathbf{x} : h_\mathbf{w}(\mathbf{x}) \neq f(\mathbf{x})\}$ and $\theta(\mathbf{w}, \mathbf{w}'), \theta(\mathbf{w}^*, \mathbf{w}') > 0$. This implies that $\{\mathbf{x} : h_\mathbf{w}(\mathbf{x}) \neq f(\mathbf{x})\} \supset \{\mathbf{x} : h_{\mathbf{w}'}(\mathbf{x}) \neq f(\mathbf{x})\}$ and $\mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[h_{\mathbf{w}'}(\mathbf{x}) \neq f(\mathbf{x})] > 0$. Note that $|\langle \mathbf{w}, \mathbf{x} \rangle| > 0$ for all $\mathbf{x}$ with $h_{\mathbf{w}'}(\mathbf{x}) \neq f(\mathbf{x})$. Therefore, we get that

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[T(\mathbf{x}) | \langle \mathbf{w}, \mathbf{x} \rangle |] \geq \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\mathbb{1}\{h_{\mathbf{w}'}(\mathbf{x}) \neq f(\mathbf{x})\} | \langle \mathbf{w}, \mathbf{x} \rangle |] > 0\ .$$

This completes the proof of Fact 3.1. □

Main Result of this Section. Fact 3.1 shows that a certifying function exists. However, in general, finding such a function is information-theoretically and computationally hard. By leveraging our distributional assumptions, we show that a certifying function of a specific simple form exists and can be computed in polynomial time.

For the rest of this section, we work with distributions that are $(3, L, R, \beta)$-well-behaved. These distributions satisfy the same properties as those in Definition 1.3, except the anti-concentration condition. (The anti-concentration condition is only required at the end of our analysis in Section 5 to deduce that small angle between two halfspaces implies small 0-1 error.)

Definition 3.2. For $L, R > 0$, $\beta \geq 1$, and $k \in \mathbb{Z}_+$, a distribution $\mathcal{D}_\mathbf{x}$ on $\mathbb{R}^d$ is called (k, L, R, β)-well-behaved if the following conditions hold: (i) For any projection $(\mathcal{D}_\mathbf{x})_V$ of $\mathcal{D}_\mathbf{x}$ on a k-dimensional subspace V of $\mathbb{R}^d$, the corresponding pdf γ_V on V satisfies $\gamma_V(\mathbf{x}) \geq L$, for all $\mathbf{x} \in V$ with $\|\mathbf{x}\|_2 \leq R$ (anti-anti-concentration). (ii) For any

$t > 0$ and unit vector $\mathbf{w} \in \mathbb{R}^d$, we have that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle \mathbf{w}, \mathbf{x}\rangle| \geq t] \leq \exp(1 - t/\beta)$ (sub-exponential concentration).

Specifically, we have:

Theorem 3.3 (Efficiently Certifying Non-Optimality). *Let $\mathcal{D}$ be a $(3, L, R, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)$. Let $\mathbf{w}$ be a unit vector with $\theta(\mathbf{w}, \mathbf{w}^*) \geq \theta$, where $\theta \in (0, \pi]$. There is an algorithm that, given as input $\mathbf{w}$, θ, and $N = ((A/(LR)) \cdot (d/\theta))^{O(1/\alpha)} \log(1/\delta)$ samples from $\mathcal{D}$, it runs in $\mathrm{poly}(N, d)$ time, and with probability at least $1 - \delta$ returns a certifying function $T_\mathbf{w} : \mathbb{R}^d \mapsto \mathbb{R}_+$ such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T_\mathbf{w}(\mathbf{x})\, y\, \langle \mathbf{w}, \mathbf{x}\rangle] \leq -\frac{1}{\beta}\left(\frac{LR\,\theta}{A\,d}\right)^{O(1/\alpha)} . \tag{1}$$

3.1 Intuition and Roadmap of the Proof

In this subsection, we give an intuitive proof overview of Theorem 3.3 along with pointers to the corresponding subsections where the proof of each component appears. First, we discuss the specific form of the certifying function that we compute. The proof of Fact 3.1 shows that a valid choice for the certifying function would be the characteristic function of the disagreement region between the candidate hypothesis $\mathbf{w}$ and the optimal halfspace $\mathbf{w}^*$, i.e., $T_\mathbf{w}(\mathbf{x}) = \mathbb{1}\{\mathrm{sign}(\langle \mathbf{w}, \mathbf{x}\rangle) \neq \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)\}$. Unfortunately, we do not know $\mathbf{w}^*$ (this is the vector we are trying to approximate!), and therefore it is unclear how to algorithmically use this certifying function.

Our goal is to judiciously define a parameterized family of "simple" certifying functions and optimize over this family to find one that acts similarly to the indicator of the disagreement region. A natural attempt to construct a certifying function for a guess $\mathbf{w}$ would be to focus on a small "band" around the candidate halfspace $\mathbf{w}$. This idea bears some similarity with the technique of "localization", an approach going back to [7], which has previously seen success for the problem of efficiently learning homogeneous halfspaces with Massart noise [3, 4, 28, 62]. Unfortunately, this idea is inherently insufficient to provide us with a certifying function for the following reason: Even an arbitrarily thin band around $\mathbf{w}$ will assign more probability mass on points that do not belong in the disagreement region, and therefore the expectation $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}\{\sigma_1 \leq \langle \mathbf{w}, \mathbf{x}\rangle \leq \sigma_2\} y \langle \mathbf{w}, \mathbf{x}\rangle]$ will be positive. See Figure 1 for an illustration.

Intuitively, we need a way to *boost* the contribution of the disagreement region. One way to achieve this is by constructing a smooth reweighting of the space. In particular, we can look in the direction of the projection of $\mathbf{w}^*$ on the orthogonal complement of $\mathbf{w}$, i.e., the vector

$$(\mathbf{w}^*)^{\perp_\mathbf{w}} = \frac{\mathrm{proj}_{\mathbf{w}^\perp}(\mathbf{w}^*)}{\left\|\mathrm{proj}_{\mathbf{w}^\perp}(\mathbf{w}^*)\right\|_2} ,$$

that lies in the 2-dimensional subspace spanned by $\mathbf{w}$ and $\mathbf{w}^*$; see Figure 1. Notice that the disagreement region is a subset of the points that have negative inner product with $(\mathbf{w}^*)^{\perp_\mathbf{w}}$. Therefore, a candidate reweighting can be obtained by using a polynomial $p(\langle (\mathbf{w}^*)^{\perp_\mathbf{w}}, \mathbf{x}\rangle)$ of moderately large degree that will boost

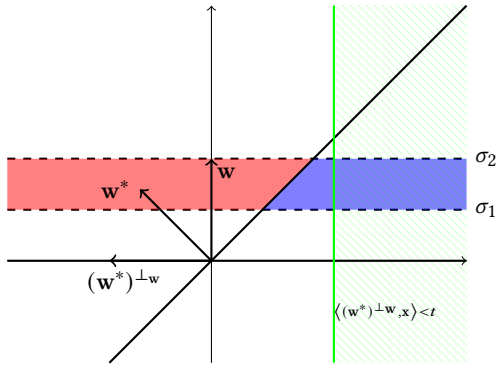


Figure 1: The indicator of a band $\{\mathbf{x} : \sigma_1 \leq \langle \mathbf{w}, \mathbf{x}\rangle \leq \sigma_2\}$ cannot be used as a certificate even when there is no noise and the underlying distribution is the standard Gaussian: the contribution of the positive points (red region) is larger than the contribution of the negative points (blue region). On the other hand, taking the intersection of the band and the halfspace with normal vector $(\mathbf{w}^*)^{\perp_\mathbf{w}}$ and a sufficiently negative threshold $t < 0$ gives us a subset of the disagreement region (intersection of blue and green regions).

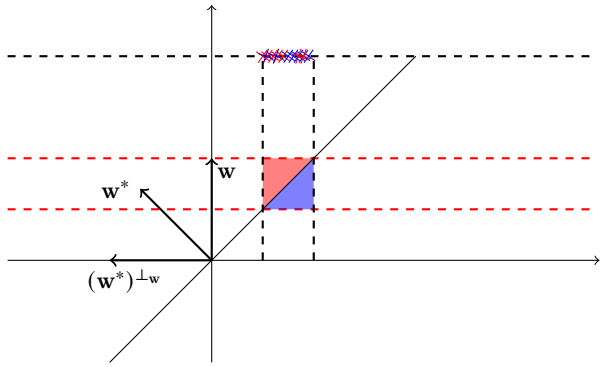
the points that lie in the disagreement region. This was the approach used in the recent work [29]. Since $(\mathbf{w}^*)^{\perp_\mathbf{w}}$ is not known, one needs to formulate a convex program (SDP) over the space of all d-variate polynomials of sufficiently large degree k implying that the corresponding SDP has $d^{\Omega(k)}$ variables. Unfortunately, it is not hard to show that the required degree cannot be smaller than $\Omega(\log(1/\epsilon))$. Therefore, this approach can only give a $d^{\Omega(\log(1/\epsilon))}$, i.e., quasi-polynomial, certificate algorithm.

In this work, we instead use a *hard threshold function* together with a band to isolate (a non-trivial subset of) the disagreement region. Specifically, we consider a function of the form $\mathbb{1}\{\langle (\mathbf{w}^*)^{\perp_\mathbf{w}}, \mathbf{x}\rangle < t\}$ for some scalar threshold t; see Figure 1. Since $(\mathbf{w}^*)^{\perp_\mathbf{w}}$ is unknown, we need to find a certifying vector $\mathbf{v}$ that is perpendicular to $\mathbf{w}$, i.e., $\mathbf{v} \in \mathbf{w}^\perp$ and acts similarly to $(\mathbf{w}^*)^{\perp_\mathbf{w}}$. This leads us to the following **non-convex** optimization problem

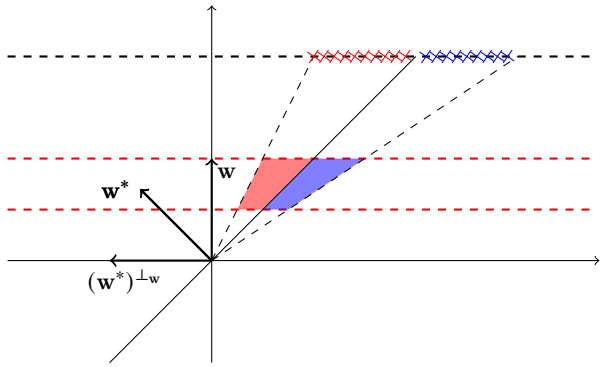
$$\min_{t\in\mathbb{R}, \mathbf{v}\in\mathbf{w}^\perp} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [\mathbb{1}\{\sigma_1 \leq \langle \mathbf{w}, \mathbf{x}\rangle \leq \sigma_2\}\mathbb{1}\{\langle \mathbf{v}, \mathbf{x}\rangle < t\} \langle \mathbf{w}, \mathbf{x}\rangle] .$$

Thus far, we have succeeded in reducing the number of parameters that we want to compute down to $O(d)$, but now we are faced with a non-convex optimization problem. Our main result is an efficient algorithm that computes a *certifying vector* $\mathbf{v}$ and a threshold t that does not necessarily minimize the above non-convex objective, but still suffice to make the corresponding expectation sufficiently negative.

We now describe the main steps we use to compute the certifying vector $\mathbf{v}$. The first obstacle we need to overcome is that, for $\mathbf{v} \in \mathbf{w}^\perp$, the corresponding instance fails to satisfy the Tsybakov noise condition. In particular, when we project the datapoints on $\mathbf{w}^\perp$, the region close to the boundary of the optimal halfspace becomes "fuzzy" even without noise: Points with different labels are mapped to the same point of $\mathbf{w}^\perp$; see Figure 2a. We bypass this difficulty by using a *perspective projection* to map the datapoints onto $\mathbf{w}^\perp$. For



(a) Orthogonal projection.



(b) Perspective projection.

Figure 2: The dotted line on top of the figures corresponds to the subspace $\mathbf{w}^\perp$. When we project the points to $\mathbf{w}^\perp$ orthogonally, we map points with different labels to the same point of $\mathbf{w}^\perp$ and obtain the "fuzzy" region where blue points (classified as negative by $\mathbf{w}^*$) overlap with red points (positive according to $\mathbf{w}^*$). On the other hand, the perspective projection defined in Equation 2 preserves linear separability.

non-zero vectors $\mathbf{w}, \mathbf{x} \in \mathbb{R}^d$, the perspective projection of $\mathbf{x}$ on $\mathbf{w}$ is defined as follows:

$$\pi_{\mathbf{w}}(\mathbf{x}) \stackrel{\text{def}}{=} \text{proj}_{\mathbf{w}^\perp} \frac{\mathbf{x}}{\langle \mathbf{w}, \mathbf{x} \rangle} . \tag{2}$$

Notice that without noise the perspective projection keeps the dataset linearly separable (see Figure 2b), which means that after we perform this projection the label noise of the resulting instance will again satisfy the Tsybakov noise condition. In addition, we show that this transformation will preserve the crucial distributional properties (concentration, anti-anti-concentration) of the underlying marginal distribution $\mathcal{D}_{\mathbf{x}}$. For a detailed discussion and analysis of this data transformation, see Subsection 3.2.

Given this setup, the certificate that our algorithm will compute for a candidate weight vector $\mathbf{w} \in \mathbb{R}^d$ is a function of the form

$$T_{\mathbf{w}}(\mathbf{x}) = \frac{\mathbb{1}\left\{\sigma_1 \le \langle \mathbf{w}, \mathbf{x} \rangle \le \sigma_2 , -t_1 \le \langle \mathbf{v}, \pi_{\mathbf{w}}(\mathbf{x}) \rangle \le -t_2\right\}}{\langle \mathbf{w}, \mathbf{x} \rangle} =: \frac{\psi(\mathbf{x})}{\langle \mathbf{w}, \mathbf{x} \rangle} , \tag{3}$$

for some vector $\mathbf{v} \in \mathbb{R}^d$ and scalars $\sigma_1, \sigma_2, t_1, t_2 > 0$. For an illustration, in Figure 2b we plot the set of the indicator function $\psi(\mathbf{x})$ which is a (high-dimensional) trapezoid.

It is not difficult to verify that by choosing $\mathbf{v} = (\mathbf{w}^*)^{\perp \mathbf{w}}$ and appropriately picking $\sigma_1, \sigma_2, t_1, t_2$, the corresponding certificate function $T_{\mathbf{w}}$ resembles the indicator function of the disagreement region and certifies the *non-optimality* of the candidate halfspace $\mathbf{w}$. In the following claim, we prove that for any non-optimal halfspace there exists a certifying function of the above form.

Claim 3.4. *Let $\mathcal{D}$ be a $(3, L, R, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\mathbf{x}) = \text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)$. Fix any non-zero vector $\mathbf{w}$ such that $\theta(\mathbf{w}, \mathbf{w}^*) > 0$. Then, by setting $\mathbf{v} = (\mathbf{w}^*)^{\perp \mathbf{w}}$ in the definition (3) of $T_{\mathbf{w}}(\mathbf{x})$, there exist $\sigma_1, \sigma_2, t_1, t_2 > 0$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[T_{\mathbf{w}}(\mathbf{x})\, y\, \langle \mathbf{w}, \mathbf{x} \rangle] < 0$.*

We note here that the proof of Claim 3.4 is sketched below for the sake of intuition and is not required for the subsequent analysis.

Proof Sketch. Setting $\mathbf{v} = (\mathbf{w}^*)^{\perp \mathbf{w}}$ in (3), we have

$$\begin{aligned} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T_{\mathbf{w}}(\mathbf{x})\, y\, \langle \mathbf{w}, \mathbf{x} \rangle] &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\psi(\mathbf{x})\, y] \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\psi(\mathbf{x})\,(1 - 2\eta(\mathbf{x}))\,\text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)\right] . \end{aligned}$$

We will show that by appropriate choices of $\sigma_1, \sigma_2, t_1, t_2$ the indicator $\psi(\mathbf{x})$ above corresponds to a subset of the disagreement region $\{\mathbf{x} : \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \neq \text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)\}$. See Figure 3 for an illustration. More precisely, since the distribution satisfies an anti-anti-concentration property, we can choose $\sigma_1, \sigma_2 = \Theta(R)$, so that inside the band $\{\sigma_1 \le \langle \mathbf{w}, \mathbf{x} \rangle \le \sigma_2\}$ there is non-zero probability mass. In particular, by setting $\sigma_1 = \rho R/2$ and $\sigma_2 = \rho R/\sqrt{2}$, for some $\rho \in (0, 1]$, we have that the band has mass roughly $\Omega(\rho R^3)$. For these choices of σ_1 and σ_2, we can pick $t_1 = \Theta(R/\rho)$ and guarantee that the slope of the corresponding line in the two-dimensional subspace is sufficiently small, so that we get a trapezoid whose intersection with the aforementioned horizontal band is large (see Figure 3). It remains to tune the parameter t_2. Since $\theta = \theta(\mathbf{w}, \mathbf{w}^*)$ is known, we may pick $t_2 = \Theta(R \tan\theta/\rho)$ in order to make sure that the trapezoid is a subset of the disagreement region between $\mathbf{w}^*$ and $\mathbf{w}$. □

From the above proof, it is clear that one does not really need to optimize the scalars σ_1, σ_2, t_1. Their values can be chosen according to the parameters of the underlying well-behaved distribution. Our optimization problem will be with respect to the vector $\mathbf{v}$ and the threshold t_2. However, optimizing the expectation of the certifying function $T_{\mathbf{w}}$ of Equation (3) is still a non-convex problem. Given a candidate certifying vector $\mathbf{v}_0$ that has non-trivial correlation with $(\mathbf{w}^*)^{\perp \mathbf{w}}$, our main structural result is a **win-win** statement showing that either there exists a threshold t_2 that, together with $\mathbf{v}_0$, makes the corresponding expectation of $T_{\mathbf{w}}$ sufficiently negative, or a perceptron-like update rule *will improve the correlation* between $(\mathbf{w}^*)^{\perp \mathbf{w}}$ and $\mathbf{w}$. In particular, we show that after roughly $\text{poly}(d/\epsilon)$ updates the correlation between the guess $\mathbf{v}$ and $(\mathbf{w}^*)^{\perp \mathbf{w}}$ will be sufficiently large so that there exists some threshold t_2 that makes $\mathbf{v}$ a certifying vector. Having such a vector $\mathbf{v}$, it is easy to optimize over all possible thresholds and find a value for t_2 that works. For

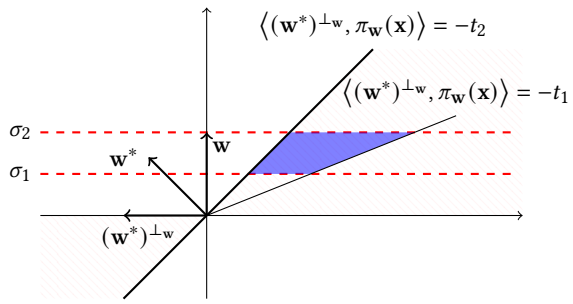


Figure 3: The function $\psi(\mathbf{x})$ for $\mathbf{v} = (\mathbf{w}^*)^{\perp_{\mathbf{w}}} = \frac{\mathrm{proj}_{\mathbf{w}^\perp}(\mathbf{w}^*)}{\|\mathrm{proj}_{\mathbf{w}^\perp}(\mathbf{w}^*)\|_2}$ defined in (3) and appropriate scalars $\sigma_1, \sigma_2, t_1, t_2$ is the indicator of a subset of the disagreement region $\{\mathbf{x} : \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \neq \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)\}$.

the formal statement of this claim and its proof, see Subsection 3.3 and Proposition 3.11.

3.2 Data Transformation

In this subsection, we show that we can simplify the problem of searching for a certifying vector $\mathbf{v}$ in $T_{\mathbf{w}}(\mathbf{x})$ defined in Equation (3) by projecting the samples to an appropriate $(d-1)$-dimensional subspace via the perspective projection (2). The main proposition of this subsection (Proposition 3.6) shows that this operation in some sense preserves the structure of the problem. In more detail, the transformed distribution remains well-behaved and satisfies the Tsybakov noise condition (albeit with somewhat worse parameters).

The transformation we perform is as follows:

(1) We first condition on the band $B = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w} \rangle \in [\sigma_1, \sigma_2]\}$, for some positive parameters σ_1, σ_2.
(2) We then perform the perspective projection on the samples, $\pi_{\mathbf{w}}(\cdot)$, defined in Equation (2).

To facilitate the proceeding formal description, we introduce the following definition.

Definition 3.5 (Transformed Distribution)**.** Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, $B \subseteq \mathbb{R}^d$ and $(\mathbf{x}, y) \sim \mathcal{D}$.

- We use $\mathcal{D}_B$ to denote $\mathcal{D}$ conditioned on $\mathbf{x}$ being in the set B.
- Let $q : \mathbb{R}^d \mapsto \mathbb{R}^d$. We denote by $\mathcal{D}^q$ the distribution of the random variable $(q(\mathbf{x}), y)$.

With the above notation, $\mathcal{D}_B^q$ is the distribution obtained by first conditioning on B and then applying the transformation $q(\cdot)$ to $\mathcal{D}_B$.

With Definition 3.5 in place, the distribution obtained from $\mathcal{D}$ after we condition on the band B is $\mathcal{D}_B$, and the distribution obtained from $\mathcal{D}_B$ after we perform the perspective projection is $\mathcal{D}_B^{\pi_{\mathbf{w}}}$. We can now state the main proposition of this subsection.

Proposition 3.6 (Properties of $\mathcal{D}_B^{\pi_{\mathbf{w}}}$). *Let $\mathcal{D}$ be a $(3, L, R, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)$. Fix any unit vector $\mathbf{w}$ such that $\theta(\mathbf{w}, \mathbf{w}^*) = \theta$, and let $B = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w} \rangle \in [\rho R/2, \rho R/\sqrt{2}]\}$, for some $\rho \in (0, 1]$. Then, for some $c = (LR)^{O(1)}$, the following conditions hold:*

(1) *The distribution $\mathcal{D}_B^{\pi_{\mathbf{w}}}$ on $\mathbb{R}^d \times \{\pm 1\}$ is $\left(2, c\rho^3, \frac{1}{\rho}, \frac{\beta}{c\rho} \log \frac{1}{\rho}\right)$-well-behaved.*
(2) *The distribution $\mathcal{D}_B^{\pi_{\mathbf{w}}}$ satisfies the $\left(\alpha, \frac{A}{c\rho}\right)$-Tsybakov noise condition with optimal classifier $\mathrm{sign}\left(\langle (\mathbf{w}^*)^{\perp_{\mathbf{w}}}, \mathbf{x} \rangle + 1/\tan\theta\right)$.*

The rest of this subsection is devoted to the proof of Proposition 3.6. Before we proceed with the proof, we express the problem of finding a certifying vector $\mathbf{v}$ satisfying (3) in the transformed domain. Indeed, it is not hard to see that after we condition on B and perform the perspective projection $\pi_{\mathbf{w}}$, our goal is to find a vector $\mathbf{v}$ and scalars $t_1, t_2 > 0$ such that

$$\mathop{\mathbf{E}}_{(\mathbf{z}, y) \sim \mathcal{D}_B^{\pi_{\mathbf{w}}}} [\mathbb{1}\{-t_1 \leq \langle \mathbf{v}, \mathbf{z} \rangle \leq -t_2\} y] < 0 \,. \tag{4}$$

More formally, we have the following simple lemma showing that if we find a certifying vector $\mathbf{v}$ and parameters t_1, t_2 in the transformed instance $\mathcal{D}_B^{\pi_{\mathbf{w}}}$ satisfying Equation (4), the same vector and parameters will be a certificate with respect to the initial well-behaved distribution $\mathcal{D}$. The relevant expectation remains negative but is slightly closer to zero.

Lemma 3.7. *Let $\mathcal{D}$ be a $(3, L, R, \beta)$-well-behaved distribution on $\mathbb{R}^d$ and let $B = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w} \rangle \in [\rho R/2, \rho R/\sqrt{2}]\}$, for some $\rho \in (0, 1]$. Let $\mathbf{w} \in \mathbb{R}^d$ be a unit vector and let $\mathbf{v} \in \mathbf{w}^\perp$, $t_1, t_2 > 0$ be such that $\mathbf{E}_{(\mathbf{z}, y) \sim \mathcal{D}_B^{\pi_{\mathbf{w}}}} [\mathbb{1}\{-t_1 \leq \langle \mathbf{v}, \mathbf{z} \rangle \leq -t_2\} y] < -C$, for some $C > 0$. Then we have that $\mathbf{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [T_{\mathbf{w}}(\mathbf{x})\, y \langle \mathbf{w}, \mathbf{x} \rangle] = -\Omega(CLR^3\rho)$.*

Proof of Proposition 3.6. Our goal is to compute a certificate of the form (3). As we already discussed, if we had chosen to simply project the points on the subspace $\mathbf{w}^\perp$, we would have obtained an instance that is not linearly separable — even if the noise rate $\eta(\mathbf{x})$ was identically zero. By first conditioning on the set $B = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w} \rangle \in [\sigma_1, \sigma_2]\}$, where $\sigma_1, \sigma_2 > 0$, and then performing the perspective projection $\pi_{\mathbf{w}}$, we keep the dataset linearly separable (with respect to the noiseless distribution, i.e., for $\eta(\mathbf{x}) = 0$), albeit by a *biased* linear classifier. We have the following lemma.

Lemma 3.8. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that for $(\mathbf{x}, y) \sim \mathcal{D}$ we have that $y = \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)$. Let $\mathbf{w}$ be any unit vector such that $\theta(\mathbf{w}, \mathbf{w}^*) = \theta \in (0, \pi]$. For $(\mathbf{z}, y) \sim \mathcal{D}_B^{\pi_{\mathbf{w}}}$ it holds $y = \mathrm{sign}\left(\langle (\mathbf{w}^*)^{\perp_{\mathbf{w}}}, \mathbf{z} \rangle + \frac{1}{\tan\theta}\right)$, i.e., the transformed distribution is linearly separable by a biased hyperplane.*

Proof. Observe that $\mathbf{w}^* = \lambda_1 (\mathbf{w}^*)^{\perp_{\mathbf{w}}} + \lambda_2 \mathbf{w}$, where $\lambda_1 > 0$. We then have

$$\begin{aligned} \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle) &= \mathrm{sign}\left(\lambda_1 \langle (\mathbf{w}^*)^{\perp_{\mathbf{w}}}, \mathbf{x} \rangle + \lambda_2 \langle \mathbf{w}, \mathbf{x} \rangle\right) \\ &= \mathrm{sign}\left(\langle (\mathbf{w}^*)^{\perp_{\mathbf{w}}}, \pi_{\mathbf{w}}(\mathbf{x}) \rangle + \frac{\lambda_2}{\lambda_1}\right), \end{aligned}$$

where to get the last equality we use the fact that λ_1 and $\langle \mathbf{w}, \mathbf{x} \rangle$ are both positive given that we conditioned on the band B. Observe that if the angle between $\mathbf{w}$ and $\mathbf{w}^*$ is θ, then $\lambda_1 = \sin\theta$ and $\lambda_2 = \cos\theta$. This completes the proof. □

We next show that conditioning on the band B will not make the Tsybakov noise condition substantially worse.

Lemma 3.9. *Let $\mathcal{D}$ be a $(3, L, R, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition*

with respect to an unknown halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)$. *Let* $B = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w}\rangle \in [\rho R/2, \rho R/\sqrt{2}]\}$, *for some* $\rho \in (0,1]$. *Then* $\mathcal{D}_B$ *satisfies the Tsybakov noise condition with parameters* $(\alpha, O(A/(R^3 L\rho)))$ *and optimal linear classifier* $\mathbf{w}^*$.

Proof. We have that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[1-2\eta(\mathbf{x}) > t | \mathbf{x} \in B] \le \mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[1-2\eta(\mathbf{x}) > t]/\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[B]$. From the proof of Lemma 3.7, we have seen that we can use the anti-anti-concentration property of $\mathcal{D}_\mathbf{x}$ to bound $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[B]$ from below. Specifically, we have $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[B] \ge \Omega(LR^3\rho)$. Therefore, $\mathcal{D}_B$ satisfies the Tsybakov noise condition with parameters $(\alpha, O(A/(R^3\rho L))$. □

Finally, we show that the transformation of Equation (2) also preserves the anti-anti-concentration and concentration properties of the marginal distribution $\mathcal{D}_\mathbf{x}$.

Lemma 3.10. *Let* $\mathcal{D}$ *be a* $(3, L, R, \beta)$*-well-behaved distribution. Fix any unit vector* $\mathbf{w}$ *and let* $B = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w}\rangle \in [\rho R/2, \rho R/\sqrt{2}]\}$, *for some* $\rho \in (0,1]$. *Then the transformed distribution* $\mathcal{D}_B^{\pi_\mathbf{w}}$ *is* $(2, \Omega(L\rho^3 R^3), 1/\rho, O(\beta/(R\rho)\log(1/(LR\rho))))$*-well-behaved.*

Proposition 3.6 follows by combining Lemmas 3.8, 3.9, 3.10.

3.3 Efficient Certificate Computation Given Initialization

In this subsection, we give our main algorithm for computing a non-optimality certificate in the transformed instance, i.e., a vector $\mathbf{v}$ and parameters $t_1, t_2 > 0$ satisfying Equation (4). Recall that after the perspective projection transformation of Subsection 3.2, we now have sample access to i.i.d. labeled examples $(\mathbf{x}, y)$ from a well-behaved distribution $\mathcal{D}$ on $\mathbb{R}^d \times \{\pm 1\}$ satisfying the Tsybakov noise condition (albeit with somewhat worse parameters) with the optimal classifier being a non-homogeneous halfspace (see Proposition 3.6.)

Our certificate algorithm in this subsection assumes the existence of an initialization vector, i.e., a vector that has non-trivial correlation with $(\mathbf{w}^*)^{\perp_\mathbf{w}}$. The simplest way to find such a vector is by picking a uniformly random unit vector. A random initialization suffices for the guarantees of this subsection (and in particular for Theorem 3.3). We note that for the family of log-concave distributions, we can leverage additional structure to design a fairly sophisticated initialization algorithm that in turn leads to a faster certificate algorithm (see Section 4).

The main algorithmic result of this section is an efficient algorithm to compute a certifying vector satisfying Equation (4). Note that we are essentially working in $(d-1)$ dimensions, since we have already projected the examples to the subspace $\mathbf{w}^\perp$. As shown in Proposition 3.6, the transformed distribution $\mathcal{D}_B^{\pi_\mathbf{w}}$ is still well-behaved and follows the Tsybakov noise condition, but with somewhat worse parameters than the initial distribution $\mathcal{D}$.

To avoid clutter in the relevant expressions, we overload the notation and use $\mathcal{D}$ instead of $\mathcal{D}_B^{\pi_\mathbf{w}}$ in the rest of this section. Moreover, we use the notation (L, R, β) and (α, A) to denote the well-behaved distribution's parameters and the Tsybakov noise parameters. The actual parameters of $\mathcal{D}_B^{\pi_\mathbf{w}}$ (quantified in Proposition 3.6) are used in the proof of Theorem 3.3. To simplify notation, we will henceforth denote by $\mathbf{v}^*$ the vector $(\mathbf{w}^*)^{\perp_\mathbf{w}}$. We show:

Proposition 3.11. *Let* $\mathcal{D}$ *be a* $(2, L, R, \beta)$*-well-behaved distribution on* $\mathbb{R}^d \times \{\pm 1\}$ *satisfying the* (α, A)*-Tsybakov noise condition with respect to an unknown halfspace* $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{v}^*, \mathbf{x}\rangle + b)$. *Let* $\mathbf{v}_0 \in \mathbb{R}^d$ *be a unit vector such that* $\langle \mathbf{v}_0, \mathbf{v}^*\rangle \ge 4b/R$. *There is an algorithm (Algorithm 1) with the following performance guarantee: Given* $\mathbf{v}_0$ *and* $N = d\,\frac{\beta^2 R^2}{b^2}\left(\frac{A}{RL}\right)^{O(1/\alpha)}\log(1/\delta)$ *samples from* $\mathcal{D}$, *the algorithm runs in* $\mathrm{poly}(N, d)$ *time, and with probability at least* $1-\delta$ *returns a unit vector* $\mathbf{v} \in \mathbb{R}^d$ *and a scalar* $t \in \mathbb{R}_+$ *such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}[-R \le \langle \mathbf{v}, \mathbf{x}\rangle \le -t]\, y] \le -\frac{b}{R\beta}\left(\frac{RL}{A}\right)^{O(1/\alpha)}.$$

Algorithm 1 employs a "perceptron-like" update rule that in polynomially many rounds succeeds in improving the angle between the initial guess $\mathbf{v}_0$ and the target vector $(\mathbf{w}^*)^{\perp_\mathbf{w}} = \mathbf{v}^*$. While the algorithm is relatively simple, its proof of correctness relies on a novel structural result (Lemma 3.12) whose proof is the main technical contribution of this section. Roughly speaking, our structural result establishes the following win-win statement: Given a vector whose correlation with $\mathbf{v}^*$ is non-trivial, either this vector is already a certifying vector (see Item 1 of Lemma 3.12 and Lemma 3.7) or the update step will improve the angle with $\mathbf{v}^*$ (Item 2 of Lemma 3.12).

In more detail, starting with a vector $\mathbf{v}_0$ that has non-trivial correlation with $\mathbf{v}^*$, we consider the following update rule

$$\mathbf{v}^{(t+1)} = \mathbf{v}^{(t)} + \lambda \mathbf{g}, \tag{5}$$

where $\lambda > 0$ is an appropriately chosen step size and

$$\mathbf{g} = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}\{-R \le \langle \mathbf{v}^{(t)}, \mathbf{x}\rangle \le -R/2\}\, y\, \mathrm{proj}_{(\mathbf{v}^{(t)})^\perp}(\mathbf{x})],$$

where $\mathrm{proj}_{(\mathbf{v}^{(t)})^\perp}(\mathbf{x})$ is the projection of $\mathbf{x}$ to the subspace $(\mathbf{v}^{(t)})^\perp$. In Lemma 3.13, we show that if $\mathbf{v}^{(t)}$ is not a certifying vector, i.e., it does not satisfy Item 1 of Lemma 3.13, then there exists an appropriately small step size λ that improves the correlation with $\mathbf{v}^*$ after the update. This is guaranteed by Item 2 of Lemma 3.13, which shows that $\mathbf{g}$ has positive correlation with $(\mathbf{v}^*)^{\perp_\mathbf{v}}$ (the normalized projection of $\mathbf{v}^*$ onto $\mathbf{v}^\perp$), and thus will turn $\mathbf{v}^{(t)}$ towards the direction of $\mathbf{v}^*$ decreasing the angle between them.

We are now ready to state and prove our win-win structural result:

Lemma 3.12 (Win-Win Result). *Let* $\mathcal{D}$ *be a* $(2, L, R, \beta)$*-well-behaved distribution on* $\mathbb{R}^d \times \{\pm 1\}$ *that satisfies the* (α, A)*-Tsybakov noise condition with respect to* $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{v}^*, \mathbf{x}\rangle + b)$, *and* $\mathbf{v} \in \mathbb{R}^d$ *be a unit vector with* $\langle \mathbf{v}, \mathbf{v}^*\rangle \ge 4b/R$. *Consider the band* $B^t = \{\mathbf{x} : -R \le \langle \mathbf{v}, \mathbf{x}\rangle \le -t\}$ *for* $t \in [R/2, R]$ *and define*

$$\mathbf{g} = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}_{B^{R/2}}(\mathbf{x})\, y\, \mathrm{proj}_{\mathbf{v}^\perp}(\mathbf{x})].$$

For some $c = (RL/A)^{O(1/\alpha)}$, *one of the following statements is satisfied:*

1. *There exists* $t_0 \in (R/2, R]$, *such that* $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}_{B^{t_0}}(\mathbf{x})\, y\right] \le -c^2\frac{b}{R\beta}$.
2. *It holds* $\langle \mathbf{g}, \mathbf{v}^*\rangle \ge c^2\frac{\pi b}{4\beta}$.

Moreover, the first condition always holds if $\theta(\mathbf{v}, \mathbf{v}^*) \le b\,c/\beta$.

Algorithm 1 Computing a Certificate Given Initialization

1: **procedure** ComputeCertificate($(L, R, \beta), (A, \alpha), \delta, \mathbf{v}_0, \widehat{\mathcal{D}}$)
2: **Input:** Empirical distribution $\widehat{\mathcal{D}}$ of a $(2, L, R, \beta)$-well-behaved distribution that satisfies the (α, A)-Tsybakov noise condition, initialization vector $\mathbf{v}_0$, confidence probability δ.
3: **Output:** A certifying vector $\mathbf{v}$ and positive scalars t_1, t_2 that satisfy (4).
4: $\mathbf{v}^{(0)} \leftarrow \mathbf{v}_0$
5: $T \leftarrow \mathrm{poly}(1/L, 1/R, A)^{1/\alpha} \cdot \mathrm{poly}(1/b, 1/\beta)$
6: $\lambda \leftarrow \frac{1}{\beta^3}\mathrm{poly}(L, R, 1/A)^{1/\alpha}$; $c \leftarrow \frac{b}{R\beta}\mathrm{poly}(L, R, 1/A)^{1/\alpha}$
7: **for** $t = 1, \ldots, T$ **do**
8: $B^{t'} = \{\mathbf{x} : -R \le \left\langle \mathbf{v}^{(t-1)}, \mathbf{x} \right\rangle \le -t'\}$
9: **if** there exists $t_0 \in (R/2, R]$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[\mathbb{1}_{B^{t_0}}(\mathbf{x})\, y\right] \le -c$
10: **return**$(\mathbf{v}^{(t-1)}, R, t_0)$
11: $\hat{\mathbf{g}}^{(t)} \leftarrow \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[\mathbb{1}_{B^{R/2}}(\mathbf{x})\, y\, \mathrm{proj}_{(\mathbf{v}^{(t-1)})^\perp}(\mathbf{x})\right]$
12: $\mathbf{v}^{(t)} \leftarrow \frac{\mathbf{v}^{(t-1)}+\lambda\hat{\mathbf{g}}^{(t)}}{\|\mathbf{v}^{(t-1)}+\lambda\hat{\mathbf{g}}^{(t)}\|_2}$

In the next lemma, we show that if Item 2 of Lemma 3.12 is satisfied, then an update step decreases the angle between the current vector $\mathbf{v}$ and the optimal vector $\mathbf{v}^*$.

Lemma 3.13 (Correlation Improvement). *For unit vectors $\mathbf{v}^*, \mathbf{v} \in \mathbb{R}^d$, let $\hat{\mathbf{g}} \in \mathbb{R}^d$ such that $\langle \hat{\mathbf{g}}, \mathbf{v}^* \rangle \ge \frac{c}{\beta}$, $\langle \hat{\mathbf{g}}, \mathbf{v} \rangle = 0$, and $\|\hat{\mathbf{g}}\|_2 \le \beta$, with $c > 0$ and $\beta \ge 1$. Then, for $\mathbf{v}' = \frac{\mathbf{v}+\lambda\hat{\mathbf{g}}}{\|\mathbf{v}+\lambda\hat{\mathbf{g}}\|_2}$, with $\lambda = \frac{c}{2\beta^3}$, we have that $\langle \mathbf{v}', \mathbf{v}^* \rangle \ge \langle \mathbf{v}, \mathbf{v}^* \rangle + \lambda^2\beta^2/2$.*

To analyze the sample complexity of Algorithm 1, we require the following simple lemma, which bounds the sample complexity of estimating the update function and testing the current candidate certificate.

Lemma 3.14 (Estimating $\mathbf{g}$). *Let $\mathcal{D}$ be a $(2, L, R, \beta)$-well-behaved distribution. Given $N = O((d\beta^2/\epsilon^2)\log(d/\delta))$ i.i.d samples $(\mathbf{x}^{(i)}, y^{(i)}))$ from $\mathcal{D}$, the estimator $\hat{\mathbf{g}} = \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{B^{R/2}}\left(\mathbf{x}^{(i)}\right) y^{(i)}\mathbf{x}^{(i)}$ satisfies the following with probability at least $1-\delta$:*

- $\|\hat{\mathbf{g}} - \mathbf{g}\|_2 \le \epsilon$, *where* $\mathbf{g} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}_{B^{R/2}}(\mathbf{x})\, y\, \mathbf{x}]$, *and*
- $\|\hat{\mathbf{g}}\|_2 \le e\beta + \epsilon$.

Before we proceed with the proof of Proposition 3.11, we show that we can efficiently check for the certificate in Line 9 of Algorithm 1 with high probability.

Lemma 3.15. *Let $\widehat{\mathcal{D}}_N$ be the empirical distribution obtained from $\mathcal{D}$ with $N = O(\log(1/\delta)/\epsilon^2)$ samples. Then, with probability $1-\delta$, for every $t \in \mathbb{R}_+$, $|\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}_{B^t}(\mathbf{x})\, y\right] - \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}_N}\left[\mathbb{1}_{B^t}(\mathbf{x})\, y\right]| \le \epsilon$.*

We are now ready to prove Proposition 3.11.

Proof of Proposition 3.11. Consider the k-th iteration of Algorithm 1. Let $\mathbf{g}^{(k)} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}_{B_k^{R/2}}(\mathbf{x})y\mathbf{x}]$, where $B_k^{R/2}(\mathbf{x}) = \{\mathbf{x} : -R \le \left\langle \mathbf{x}, \mathbf{v}^{(k)} \right\rangle \le -R/2\}$ and $G := \sqrt{b}(RL/A)^{O(1/\alpha)}$. Moreover, let $\hat{\mathbf{g}}^{(k)} = \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{B_k^{R/2}}\left(\mathbf{x}^{(i)}\right) y^{(i)}\mathbf{x}^{(i)}$ and note that from Lemma 3.14 we have that given $N = O\left(d\beta^2/G^4 \log(1/(LR))\log(dT/\delta)\right)$ samples, for every iteration k, it holds that $\left\|\hat{\mathbf{g}}^{(k)} - \mathbf{g}^{(k)}\right\|_2 \le G^2/(16\beta)$ and $\left\|\hat{\mathbf{g}}^{(k)}\right\|_2 \le e\beta + G^2/(16\beta) \le 3\beta$, with probability $1 - \delta/T$.

We first show that if Condition 1 of Lemma 3.12 is satisfied, then Algorithm 1 terminates at Line 10 returning a certifying vector. The only issue is that we have access to the empirical distribution $\widehat{\mathcal{D}}_N$ instead of $\mathcal{D}$. From Lemma 3.15, we have that the empirical expectation of Line 9 is sufficiently close to the true expectation that appears in Condition 1 of Lemma 3.12, thus it is going to find it.

We now analyze the case when Condition 1 of Lemma 3.12 is not true. From Lemma 3.12, we immediately get that since Condition 1 is not satisfied, Condition 2 is true. Then, using the update rule $\mathbf{v}^{(k+1)} = \frac{\mathbf{v}^{(k)}+\lambda\tilde{\mathbf{g}}^{(k)}}{\|\mathbf{v}^{(k)}+\lambda\tilde{\mathbf{g}}^{(k)}\|_2}$ with $\lambda = G^2/(64\beta^3)$, where $\tilde{\mathbf{g}}^{(k)} = \mathrm{proj}_{(\mathbf{v}^{(k)})^\perp}\hat{\mathbf{g}}^{(k)}$ (here $\tilde{\mathbf{g}}^{(k)}$ is the $\hat{\mathbf{g}}^{(k)}$ with the component on the direction $\mathbf{v}^{(k)}$ removed). Note that this procedure only decreases the norm of $\tilde{\mathbf{g}}$ (by the Pythagorean theorem). Then, from Lemma 3.13, we have $\left\langle \mathbf{v}^{(k+1)}, \mathbf{v}^* \right\rangle \ge \left\langle \mathbf{v}^{(k)}, \mathbf{v}^* \right\rangle + G^4/\beta^4$.

The update rule is repeated for at most $O(\beta^4/G^4)$ iterations. From Lemma 3.12, we have that a certificate exists if the angle with the optimal vector is sufficiently small. Putting everything together, our total sample complexity is $N = \tilde{O}\left(\frac{d\beta^4}{b^2G^4}\right)\log(1/\delta)$. It is also clear that the runtime is $\mathrm{poly}(N, d)$, which completes the proof. □

3.4 Proof of Theorem 3.3

To prove Theorem 3.3, we will use the iterative algorithm developed in Proposition 3.11 initialized with a uniformly random unit vector $\mathbf{v}_0$. It is easy to show that such a random vector will have non-trivial correlation with $\mathbf{v}^*$.

Fact 3.16 (see, e.g., Remark 3.2.5 of [60]). *Let $\mathbf{v}$ be a unit vector in $\mathbb{R}^d$. For a random unit vector $\mathbf{u} \in \mathbb{R}^d$, with constant probability, it holds $|\,\langle \mathbf{v}, \mathbf{u} \rangle\,| = \Omega(1/\sqrt{d})$.*

We now present the proof of Theorem 3.3 putting together the machinery developed in the previous subsections.

Proof of Theorem 3.3. As explained in Section 3.1, we are looking for a certificate function $T_{\mathbf{w}}(\mathbf{x})$ of the form given in Equation (3). As argued in Section 3.2, the search for such a certificate function can be simplified by projecting the samples to a $(d-1)$-dimensional subspace via the perspective projection.

From Proposition 3.6, choosing $\rho = O(\theta/\sqrt{d})$, there is a $c = (LR)^{O(1)}$ such that the resulting distribution $\mathcal{D}_B^{\pi_{\mathbf{w}}}$ is $(2, c\theta/\sqrt{d}, \sqrt{d}/\theta, \beta\sqrt{d}/(c\theta)\log(\sqrt{d}/\theta))$-well-behaved and satisfies the $(\alpha, Ad^{1/2}/(c\theta))$-Tsybakov noise condition.

From Fact 3.16, a random unit vector $\mathbf{v} \in \mathbb{R}^{d-1}$ with constant probability satisfies $\left\langle \mathbf{v}, (\mathbf{w}^*)^{\perp_{\mathbf{w}}} \right\rangle = \Omega(1/\sqrt{d})$. We call this event $\mathcal{E}$.

From Proposition 3.11, conditioning on the event $\mathcal{E}$ and using $\frac{\beta^4}{b^2}\left(\frac{A}{RL}\right)^{O(1/\alpha)}\log(1/\delta)$ samples, with probability $1-\delta$, we get a $(\mathbf{v}', R, t_0)$ such that

$$\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}_B^{\pi_{\mathbf{w}}}}\left[\mathbb{1}[-R \le \left\langle \mathbf{v}', \mathbf{x} \right\rangle \le -t_0]\, y\right] \le -\left(\theta LR/(Ad)\right)^{O(1/\alpha)}/\beta\ .$$

By inverting the transformation (Lemma 3.7), we get that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T_{\mathbf{w}}(\mathbf{x})\,\langle\mathbf{x},\mathbf{w}\rangle\,y] \leq -\left(\theta LR/(Ad)\right)^{O(1/\alpha)}/\beta\,.$$

Overall, we conclude that with constant probability Algorithm 1 returns a valid certificate. Repeating the process $k = O(\log(1/\delta))$ times, we can boost the probability to $1-\delta$. The total number of samples for finding and testing these candidate certificates until we find a correct one with probability at least $1-\delta$ is $N = \left(\frac{dA}{\theta LR}\right)^{O(1/\alpha)}\log(1/\delta)$. It is also clear that the runtime is $\mathrm{poly}(N,d)$, which completes the proof. □

4 MORE EFFICIENT CERTIFICATE FOR LOG-CONCAVE DISTRIBUTIONS

In this section, we present a more efficient certificate algorithm for the important special case of isotropic log-concave distributions. To achieve this, we use Algorithm 1 from the previous section starting from a significantly better initialization vector. To obtain such an initialization, we leverage the structure of log-concave distributions. The main result of this section is the following theorem.

Theorem 4.1 (Certificate for Log-concave Distributions). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$ that satisfies the (α,A)-Tsybakov noise condition with respect to the halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle\mathbf{w}^*,\mathbf{x}\rangle)$ and is such that $\mathcal{D}_{\mathbf{x}}$ is isotropic log-concave. Let $\mathbf{w}$ be a unit vector that satisfies $\theta(\mathbf{w},\mathbf{w}^*)\geq\theta$, where $\theta\in(0,\pi]$. There is an algorithm that, given as input $\mathbf{w}$, θ, and $N = \mathrm{poly}(d)\cdot\left(\frac{A}{\theta}\right)^{O(1/\alpha^2)}\log(1/\delta)$ samples from $\mathcal{D}$, it runs in $\mathrm{poly}(d,N)$ time, and with probability at least $1-\delta$ returns a certifying function $T_{\mathbf{w}}:\mathbb{R}^d\mapsto\mathbb{R}_+$ such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T_{\mathbf{w}}(\mathbf{x})\,y\,\langle\mathbf{w},\mathbf{x}\rangle] \leq -\left(\frac{\theta}{A}\right)^{O(1/\alpha^2)}. \tag{6}$$

In other words, we give an algorithm whose sample complexity and running time as a function of d is a fixed degree polynomial, independent of the noise parameters.

To establish Theorem 4.1, we apply Algorithm 1 starting from a better initialization vector. The main technical contribution of this section is an efficient algorithm to obtain such a vector for log-concave marginals.

Theorem 4.2 (Efficient Initialization for Log-Concave Distributions). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$ that satisfies the (α,A)-Tsybakov noise condition with respect to an unknown halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle\mathbf{w}^*,\mathbf{x}\rangle)$ and is such that $\mathcal{D}_{\mathbf{x}}$ is isotropic log-concave. There exists an algorithm that, given an $\epsilon>0$, a unit vector $\mathbf{w}$ such that $\|\mathbf{w}^*-\mathbf{w}\|_2 = \Theta(\epsilon)$, and $N = \mathrm{poly}(d)\cdot(A/(\alpha\epsilon))^{O(1/\alpha)}$ samples from $\mathcal{D}$, it runs in $\mathrm{poly}(d,N)$ time, and with constant probability returns a unit vector $\mathbf{v}$ such that $\left\langle\mathbf{v},(\mathbf{w}^*)^{\perp_{\mathbf{w}}}\right\rangle\geq(\alpha\epsilon/A)^{O(1/\alpha)}$, where $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ is the component of $\mathbf{w}^*$ perpendicular to $\mathbf{w}$.*

4.1 Intuition and Roadmap of the Proof

Here we sketch the proof of Theorem 4.2 and point to the relevant lemmas in the formal argument. We start with the following definition

Definition 4.3 ((α,β)-isotropic distribution)**.** We say that a distribution $\mathcal{D}$ is (α,β)-isotropic, if for every unit vector $\mathbf{u}\in\mathbb{R}^d$, it holds $|\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[\langle\mathbf{x},\mathbf{u}\rangle]|\leq\alpha$ and $1/\beta\leq\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[\langle\mathbf{x},\mathbf{u}\rangle^2]\leq\beta$.

Given a weight vector $\mathbf{w}$ of unit length, our goal is to find a unit vector $\mathbf{v}$ that has non-trivial correlation with $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$, i.e., such that $\left\langle(\mathbf{w}^*)^{\perp_{\mathbf{w}}},\mathbf{v}\right\rangle$ is roughly $\epsilon^{1/\alpha}$, where $\mathbf{w}^*$ is the optimal halfspace.

Our first step is to condition on a thin band around the current candidate $\mathbf{w}$ (similarly to Section 3, see Figure 1). When the size of the band approaches 0, we get an instance whose separating hyperplane is perpendicular to $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ and has much larger Tsybakov noise rate. After that, we would like (similarly to Section 3) to project the points on the subspace $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$. Instead of having a zero length band, we will instead take a very thin band. We have already seen in Section 3 that we can apply a perspective transformation in order to project the points on $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ and obtain an instance that satisfies the Tsybakov noise condition (with somewhat worse parameters). Unfortunately, for the current setting of log-concave distributions, we cannot use the perspective projection, as it *does not preserve the log-concavity* of the underlying distribution. On the other hand, we know that log-concavity is preserved when we condition on convex sets (such as the thin band we consider here) and when we perform orthogonal projections.

As we have seen (see Figure 2a), an orthogonal projection will create a "fuzzy" region with arbitrary sign. However, we can control the probability of this "fuzzy" region by taking a sufficiently thin random band. In particular, instead of Tsybakov noise, we will end up with the following noise condition: For some small $\xi>0$, with probability $2/3$ the noise $\eta(\mathbf{x})$ is bounded above by $1/2-\xi$, and with probability roughly $\xi^{\Theta(1)}$ we have $\eta(\mathbf{x})>1/2$ (this corresponds to the probability of the "fuzzy" region). For the proof of this statement and detailed discussion on how the random band results in this above noise guarantee.

Lemma 4.4 (Properties of Transformed Instance). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$ that satisfies the (α,A)-Tsybakov noise condition with respect to an unknown halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle\mathbf{w}^*,\mathbf{x}\rangle)$ and is such that $\mathcal{D}_{\mathbf{x}}$ is isotropic log-concave. Fix $\epsilon>0$ and unit vector $\mathbf{w}$ such that $\theta(\mathbf{w},\mathbf{w}^*)=\Theta(\epsilon)$. Let s be a sufficiently small multiple of ϵ. Set $\xi = (\Theta(s/A))^{1/\alpha}$ and $s' = \Theta(\xi^3\,s\,\epsilon)$. Pick x_0 uniformly at random from $[s,2s]$ and define the random band $B_{x_0} = \{\mathbf{x}\in\mathbb{R}^d : \langle\mathbf{x},\mathbf{w}\rangle\in[x_0,x_0+s']\}$.*

Define the distribution $\mathcal{D}^{\perp} = \mathcal{D}^{\mathrm{proj}_{\mathbf{w}^{\perp}}}_{B_{x_0}}$, the classifier $f^{\perp}(\mathbf{x}^{\perp}) = \mathrm{sign}(x_0/\tan\theta + \left\langle\mathbf{x}^{\perp},(\mathbf{w}^*)^{\perp_{\mathbf{w}}}\right\rangle)$, and the noise function

$$\eta^{\perp}(\mathbf{x}^{\perp}) = \mathbf{Pr}_{(\mathbf{z},y)\sim\mathcal{D}^{\perp}}[y\neq f^{\perp}(\mathbf{z})|\mathbf{z}=\mathbf{x}^{\perp}]\,.$$

Then $\mathcal{D}^{\perp}$ is an $(O(1),O(1))$-isotropic log-concave distribution and, with probability at least 99%, satisfies the following noise condition: $\mathbf{Pr}_{\mathbf{x}^{\perp}\sim\mathcal{D}^{\perp}_{\mathbf{x}}}[\eta^{\perp}(\mathbf{x}^{\perp})\leq 1/2-\xi]\geq 2/3$ and $\mathbf{Pr}_{\mathbf{x}^{\perp}\sim\mathcal{D}^{\perp}_{\mathbf{x}}}[\eta^{\perp}(\mathbf{x}^{\perp})\geq 1/2]\leq\xi^3$.

From this point on, we will be working in the subspace $\mathbf{w}^{\perp}$ and assume that the distribution satisfies the aforementioned noise condition. As we have discussed, the marginal distribution on the examples remains log-concave and it is not hard to make its covariance be close to the identity. However, conditioning on the thin slice may result in a distribution with large mean, even though originally the distribution was centered. This is a non-trivial technical issue.

We cannot simply translate the distribution to be origin-centered, as this would result in a potentially very biased optimal halfspace. Our proof crucially relies on the assumption of having a distribution that is *nearly* centered and at the same time for the optimal halfspace to have *small bias*. We overcome this obstacle in Step 1 below.

Lemma 4.5. *Let $\mathcal{D}$ be an isotropic log-concave distribution on $\mathbb{R}^d$. Let $\mathbf{w} \in \mathbb{R}^d$ be a unit vector and let $B = \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{w}, \mathbf{x} \rangle \in [a, b]\}$ for $a, b > 0$ smaller than some universal absolute constant. There exists an algorithm that, given $\gamma > 0$ and $\mathrm{poly}(d/\gamma)$ independent samples from $\mathcal{D}_B^{\mathrm{proj}_{\mathbf{w}^\perp}}$, runs in sample polynomial time, and returns a vector $\mathbf{r}$ such that if $\mathbf{z}$ is obtained from $\mathcal{D}_B^{\mathrm{proj}_{\mathbf{w}^\perp}}$ by rejection sampling, where a sample $\mathbf{x}$ is accepted with probability $\min(1, e^{-\langle \mathbf{r}, \mathbf{x} \rangle})$, then:*

- *A sample is rejected with probability p, where $p \in (0, 1)$ is an absolute constant.*
- *The distribution of $\mathbf{z}$ is $(\gamma, O(1))$-isotropic log-concave.*

Our approach is as follows:

(1) First, we show that there is an efficient rejection sampling procedure that preserves log-concavity and gives us a distribution that is nearly isotropic (see Definition 4.3). For the algorithm and its detailed proof of correctness, see Lemma 4.5.
(2) Then we show the following statement: Under the following assumptions
   (i) the $\mathbf{x}$-marginal is nearly isotropic,
   (ii) the optimal halfspace has sufficiently small bias, and
   (iiii) the noise $\eta(\mathbf{x})$ is bounded away from $1/2$ with constant probability,

   we can compute in polynomial time a vector $\mathbf{v}$ with good correlation to the target $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$. This is established below.

Proposition 4.6. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that $\mathcal{D}_{\mathbf{x}}$ is (α, β)-isotropic log-concave. Let $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{v}^*, \mathbf{x} \rangle - \theta)$ be such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[y \neq f(\mathbf{x})|\mathbf{x}] = \eta(\mathbf{x})$, where for some $\xi > 0$ we have that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\eta(\mathbf{x}) < 1/2 - \xi] \geq 2/3$ and $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\eta(\mathbf{x}) > 1/2] \leq \xi'$, where ξ' is a constant degree polynomial in ξ[1]. Then, as long as $|\alpha| + |\theta|$ is less than a sufficiently small constant multiple of $1/(\log(1/\xi))$, there exists an algorithm with sample complexity and runtime $\mathrm{poly}(d/\xi)$ that with constant probability returns a unit vector $\mathbf{v} \in \mathbb{R}^d$ such that $\langle \mathbf{v}, \mathbf{v}^* \rangle > \mathrm{poly}(\xi)$.*

We start by describing our algorithm to transform the distribution to nearly isotropic position (Step 1 above). We avoid translating the samples by reweighting the distribution using rejection sampling. To achieve this, we find an approximate stationary point of the non-convex objective $F(\mathbf{r}) = \left\|\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{x} \max(1, \exp(-\langle \mathbf{r}, \mathbf{x} \rangle))]\right\|_2^2$. Notice that, since this is a non-convex objective as a function of $\mathbf{r}$, we can only use (projected) SGD to efficiently find a stationary point. In particular, we show that a γ-stationary point $\mathbf{r}$ of $F(\mathbf{r})$ will make the above norm of the expectation roughly $O(\gamma)$. Therefore, in time $\mathrm{poly}(d/\gamma)$, we find a reweighting of the initial distribution whose mean is close to $\mathbf{0}$. Given this point $\mathbf{r}$, we then perform rejection sampling: We draw $\mathbf{x}$ from the initial distribution $\mathcal{D}$ and accept it with probability $\max(1, \exp(-\langle \mathbf{r}, \mathbf{x} \rangle))$, i.e., we "shrink" the distribution along the direction $\mathbf{r}$.

[1]It is not difficult to verify that $\xi' = \Theta(\xi^3)$ suffices.

We now explain how to handle the setting that the distribution is approximately log-concave (Step 2 above). After we make our distribution nearly isotropic, we compute the degree-2 Chow parameters of the distribution, i.e., the vector $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}]$ and the matrix $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{x}\mathbf{x}^\intercal - \mathbf{I})]$. We show that there exists a degree-2 polynomial $p(\langle (\mathbf{w}^*)^{\perp_{\mathbf{w}}}, \mathbf{x} \rangle)$ that correlates non-trivially with the labels y. This means that $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ correlates reasonably with the degree-2 Chow parameters. In particular, $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ has a non-trivial projection on the subspace V spanned by the degree-1 Chow parameters (this is a single vector) and the eigenvectors of the degree-2 Chow matrix with large eigenvalues. Our plan is to return a random unit vector of the subspace V. However, in order for this random vector to have non-trivial correlation with $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$, we also need to show that the dimension of V is not very large.

The last part of our argument shows that V has reasonably small dimension. To prove this, we first show that the dimension of V can be bounded above by the variance of the projection of $\mathcal{D}$ onto V, $\mathcal{D}^{\mathrm{proj}_V}$, $\mathbf{Var}_{\mathbf{x}\sim\mathcal{D}^{\mathrm{proj}_V}}[\|\mathbf{x}\|_2^2]$. Then we make essential use of a recent "thin-shell" result (Lemma 4.7) about log-concave measures that bounds from above $\mathbf{Var}_{\mathbf{x}\sim\mathcal{D}^{\mathrm{proj}_V}}[\|\mathbf{x}\|_2^2]$.

Lemma 4.7 (Corollary 13 of [48]). *Let $\mathcal{D}$ be any isotropic log-concave distribution on $\mathbb{R}^d$. We have that $\mathbf{Var}_{\mathbf{x}\sim\mathcal{D}}[\|\mathbf{x}\|_2^2] \leq d^{3/2}$.*

4.2 Proof of Theorem 4.1

Using Theorem 4.2, we can prove Theorem 4.1. The proof is similar to the proof of Theorem 3.3, but we additionally need to guess how far the current guess $\mathbf{w}$ is from $\mathbf{w}^*$.

Proof of Theorem 4.1. We start by guessing a value $\epsilon = \Omega(\theta)$ such that $\|\mathbf{w} - \mathbf{w}^*\|_2 = \Theta(\epsilon)$. From Proposition 3.6 with $\rho = O(\theta(\alpha\epsilon/A)^{O(1/\alpha)})$, we have that the distribution $\mathcal{D}_B^{\pi_{\mathbf{w}}}$ is then $(2, \Omega(\rho), 1/\rho, O(\log(1/\rho)/\rho))$-well-behaved and also satisfies the $(\alpha, O(A/\rho))$-Tsybakov noise condition, where we used that the values L, R are absolute constants. Using Theorem 4.2, a random unit vector $\mathbf{v} \in \mathbb{R}^d$ with constant probability δ_1 satisfies $\left\langle \mathbf{v}, (\mathbf{w}^*)^{\perp_{\mathbf{w}}} \right\rangle \geq (\alpha\epsilon/A)^{O(1/\alpha)}$. We call this event $\mathcal{E}$. Conditioning on the event $\mathcal{E}$, from Proposition 3.11, using $\frac{\beta^4}{\theta^2}\left(\frac{A}{\theta\alpha}\right)^{O(1/\alpha^2)}\log(1/\delta)$ samples, with probability $1 - \delta$, we get a $(\mathbf{v}', R, t_0)$ such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}_B^{\pi_{\mathbf{w}}}}\left[\mathbb{1}[-R \leq \left\langle \mathbf{v}', \mathbf{x} \right\rangle \leq -t_0]y\right] \leq -(\theta\alpha/A)^{O(1/\alpha^2)}/\beta\,.$$

Using Lemma 3.7, we get that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T_{\mathbf{w}}(\mathbf{x})\langle \mathbf{x}, \mathbf{w} \rangle y] \leq -(\theta\alpha/A)^{O(1/\alpha^2)}/\beta\,.$$

Conditioning on the event $\mathcal{E}^c$, where $\mathcal{E}^c$ is the complement of $\mathcal{E}$, Algorithm 1 either returns a certificate or returns nothing. Thus, by taking $k = O(\log(1/\delta))$ random vectors, we get that the probability that event $\mathcal{E}^c$ happens is at most $(1 - \delta_1)^k \leq e^{-\delta_1 k}$. Thus, by taking $O(\log 1/\delta)$ random vectors and running Algorithm 1 with confidence $\delta/\log(1/\delta)$, we get a certificate with probability $1 - 2\delta$. Moreover, the number of samples needed to construct the empirical distribution is $\left(\frac{A}{\theta\alpha}\right)^{O(1/\alpha^2)}\log(1/\delta)$. Finally, to guess the value of ϵ, it suffices to run the algorithm for the values $\theta, 2\theta, \ldots, 1$ which will increase the complexity by a $\log(1/\theta)$ factor. This completes the proof of Theorem 4.1. □

5 LEARNING A NEAR-OPTIMAL HALFSPACE VIA ONLINE CONVEX OPTIMIZATION

In this section we present a black-box approach that uses our certificate algorithms from the previous sections to learn halfspaces in the presence of Tsybakov noise. In more detail, we provide a generic result showing that one can apply a certificate oracle in a black-box manner combined with online gradient descent to learn the unknown halfspace. We note that an essentially identical approach, with slightly different formalism, was given in [29].

Using the aforementioned approach, we establish the two main algorithmic results of this paper.

Theorem 5.1 (Learning Tsybakov Halfspaces under Well-Behaved Distributions). *Let $\mathcal{D}$ be a $(3, L, R, U, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)$. There exists an algorithm that draws $N = \beta^4 \left(\frac{dUA}{RL\epsilon}\right)^{O(1/\alpha)} \log(1/\delta)$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(N, d)$ time, and computes a vector $\widehat{\mathbf{w}}$ such that, with probability $1-\delta$, we have that $\mathrm{err}_{0-1}^{\mathcal{D}_{\mathbf{x}}}(h_{\widehat{\mathbf{w}}}, f) \leq \epsilon$.*

For the important special case of log-concave distributions on examples, we give a more efficient learning algorithm.

Theorem 5.2 (Learning Tsybakov Halfspaces under Log-concave Distributions). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)$ and is such that $\mathcal{D}_{\mathbf{x}}$ is isotropic log-concave. There exists an algorithm that draws $N = \mathrm{poly}(d) \cdot \left(\frac{A}{\epsilon}\right)^{O(1/\alpha^2)} \log(1/\delta)$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(N, d)$ time, and computes a vector $\widehat{\mathbf{w}}$ such that, with probability $1-\delta$, we have that $\mathrm{err}_{0-1}^{\mathcal{D}_{\mathbf{x}}}(h_{\widehat{\mathbf{w}}}, f) \leq \epsilon$.*

To formally describe the approach of this section, we require the notion of a *certificate oracle*. A certificate oracle is an algorithm that, given a candidate weight vector $\mathbf{w}$ and an accuracy parameter $\rho > 0$, it returns a certifying function $T(\mathbf{x})$. Recall that a certifying function is a non-negative function that satisfies $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[T(\mathbf{x}) y \langle \mathbf{x}, \mathbf{w}\rangle] \leq -\rho$ for some $\rho > 0$. We have already described how to efficiently implement such an oracle in Section 3.

Definition 5.3 (Certificate Oracle)**.** Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)$. For a decreasing function $\rho(\cdot) : \mathbb{R}_+ \mapsto \mathbb{R}_+$, we define $C(\mathbf{w}, \theta, \delta)$ to be the following ρ-certificate oracle: For any unit vector $\mathbf{w}$ and $\theta > 0$, if $\theta(\mathbf{w}, \mathbf{w}^*) \geq \theta$, then a call to $C(\mathbf{w}, \theta, \delta)$, with probability at least $1-\delta$, returns a function $T(\mathbf{x})$, with $\|T\|_\infty \leq 1$ such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T(\mathbf{x}) y \langle \mathbf{x}, \mathbf{w}\rangle] \leq -\rho(\theta),$$

and with probability at most δ returns "FAIL".

Remark 5.4. *We note that the above oracle provides a "one-sided" guarantee in the following sense. When the candidate vector $\mathbf{w}$ satisfies $\theta(\mathbf{w}, \mathbf{w}^*) \geq \theta$, the oracle is required to return a certifying function T with high probability. But it may also return such a function when $\theta(\mathbf{w}, \mathbf{w}^*) \leq \theta$. In other words, the oracle is not required to output "FAIL" with high probability when $\mathbf{w}$ is nearly parallel to $\mathbf{w}^*$. We show that an one-sided oracle of non-optimality suffices for our purposes.*

Remark 5.5. *Using Fact 3.1 the optimal halfspace $\mathbf{w}^*$ satisfies $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[T(\mathbf{x}) y \langle \mathbf{x}, \mathbf{w}^*\rangle] \geq 0$ for any non-negative function T. Therefore, as $\mathbf{w}$ approaches $\mathbf{w}^*$, we have that*

$$\lim_{\theta(\mathbf{w},\mathbf{w}^*)\to 0} \inf_{T:\|T\|_\infty \leq 1} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T(\mathbf{x}) y \langle \mathbf{x}, \mathbf{w}\rangle] = 0,$$

where $\|T\|_\infty$ is the ℓ_∞ norm for functions, i.e., $\|T\|_\infty = \sup_{\mathbf{x}\in\mathbb{R}^d} |T(\mathbf{x})|$. That is, $\lim_{\theta\to 0} \rho(\theta) = 0$ and it is natural that the non-negative function $\rho(\theta)$ is a decreasing function of the (lower bound on the) angle between $\mathbf{w}$ and $\mathbf{w}^*$. Intuitively, the closer $\mathbf{w}$ is to $\mathbf{w}^*$, the harder it is to find a certifying function T that makes $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[T(\mathbf{x}) y \langle \mathbf{x}, \mathbf{w}\rangle]$ sufficiently negative. Moreover, if our goal is to estimate the vector $\mathbf{w}^*$ within angle ϵ, we can always give the oracle this worst-case target angle, i.e., $\theta = \epsilon$. Finally, notice that when the distribution $\mathcal{D}$ is isotropic, we have $\rho(\theta) \leq 1$, as follows from $\|T\|_\infty \leq 1$ and the Cauchy-Schwarz inequality.

Given a certificate oracle, the following result shows we can efficiently approximate the optimal halfspace using projected online gradient descent.

Proposition 5.6 (Certificate-Based Optimization). *Let $\mathcal{D}$ be a $(3, L, R, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)$, and let C be a ρ-certificate oracle. There exists an algorithm that makes at most $T = \frac{1}{\rho^2(\epsilon)} \frac{1}{\alpha} \left(\frac{A}{RL}\right)^{O(1/\alpha)}$ calls to $C(\cdot)$, draws $N = d \frac{T\beta^2}{\rho^2(\epsilon)} \log\left(\frac{dT}{\delta\rho(\epsilon)}\right)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(T, N, d)$, and computes a weight vector $\widehat{\mathbf{w}}$ such that with probability $1-\delta$ we have that $\theta(\widehat{\mathbf{w}}, \mathbf{w}^*) \leq \epsilon$.*

The algorithm establishing Proposition 5.6 is given in pseudocode in Algorithm 2. In the remaining part of this section, we provide a proof sketch of Proposition 5.6.

Proof Sketch. The main idea of the algorithm is to provide a sequence of adaptively chosen convex loss functions to an Online Convex Optimization algorithm, for example Online Gradient Descent (OGD). In more detail, we construct these loss functions using our certificate oracle C. At round t, we call the certificate oracle to obtain a certifying function $T(\mathbf{x})$ and set

$$\ell_t(\mathbf{w}) = -\left\langle \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [(T(\mathbf{x}) + \lambda) y \mathbf{x}], \mathbf{w} \right\rangle,$$

where $\lambda > 0$ acts similarly to a regularizer. The last term $\lambda\langle \mathbf{w}, \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}]\rangle$ prevents the trivial vector $\mathbf{w} = \mathbf{0}$ from being a valid solution (in the sense of one that minimizes regret, see also the full version of the paper)

The crucial property of the above sequence of loss functions is that they are positive and bounded away from 0 when $\mathbf{w}$ is far from $\mathbf{w}^*$. Their value will always be greater than (roughly) $\rho(\epsilon)$, given the guarantee of our certificate oracle from Definition 5.3 for $\theta = \epsilon$ and assuming that the regularizer λ is sufficiently small.

We then provide this convex loss function to the OGD algorithm that updates the guess according to the gradient of $\ell_t(\mathbf{w})$. Our analysis follows from the regret guarantee of OGD. Since we provide convex (and in particular linear) loss functions to OGD, we know

the average regret will converge to 0 as $T \to \infty$ with a convergence rate roughly $O(1/\sqrt{T})$. This means that the oracle can only succeed in returning certifying functions for a bounded number of rounds, since every time the oracle succeeds, OGD suffers loss of at least $\rho(\epsilon)$. Therefore, after roughly $1/\rho(\epsilon)^2$ rounds the regret will be so small that for at least one round the certificate oracle must have failed. Our algorithm then stops and returns the halfspace of that iteration. Even though our certificate is "one-sided", we know that the probability that it failed with $\theta(\mathbf{w}, \mathbf{w}^*)$ being larger than ϵ is very small, which implies that we have indeed found a vector $\mathbf{w}$ very close to $\mathbf{w}^*$. □

Algorithm 2 Learning Halfspaces with Tsybakov Noise using a ρ-certificate oracle C

procedure ALG($\epsilon, \delta, \mathcal{D}, C$) ▷ ϵ: accuracy, δ: confidence
Input: $\mathcal{D}$ is a $(3, L, R, \beta)$-well-behaved distribution that satisfies the (α, A)-Tsybakov noise condition, and C is a ρ-certificate oracle.
Output: A vector $\widehat{\mathbf{w}}$ such that $\mathrm{err}_{0-1}^{\mathcal{D}_{\mathbf{x}}}(h_{\widehat{\mathbf{w}}}, f) \leq \epsilon$ with probability at least $1 - \delta$.
  $\mathbf{w}^{(0)} \leftarrow \mathbf{e}_1$
  $T \leftarrow \frac{1}{\rho(\epsilon)^2 \alpha} \left(\frac{A}{RL}\right)^{O(1/\alpha)}$
  Draw $N = \tilde{O}\left(d \cdot \frac{T\beta^2}{\rho^2(\epsilon)} \log\left(\frac{1}{\delta}\right)\right)$ samples from $\mathcal{D}$ to form the empirical distribution $\widehat{\mathcal{D}}$
  for $t = 1, \ldots, T$ **do**
    $\eta_t \leftarrow 1/(\sqrt{t} + \rho(\epsilon))$
    if $\mathbf{w}^{(t-1)} = \mathbf{0}$ **then**
      Set $\hat{\ell}_t(\mathbf{w}) \leftarrow \left\langle \mathbf{w}, -\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[\frac{\rho(\epsilon)}{2} y\mathbf{x}\right]\right\rangle$
      $\mathbf{w}^{(t)} \leftarrow \Pi_{\mathcal{B}}\left(\mathbf{w}^{(t-1)} - \eta_t \nabla_{\mathbf{w}} \hat{\ell}_t\left(\mathbf{w}^{(t-1)}\right)\right)$
    else
      Ans $\leftarrow C(\mathbf{w}^{(t-1)} / \left\|\mathbf{w}^{(t-1)}\right\|_2, \epsilon, \delta/T)$
      if Ans = FAIL **then**
        return $\mathbf{w}^{(t-1)}$
    $T_{\mathbf{w}^{(t)}}(\mathbf{x}) \leftarrow$ Ans
    Set $\hat{\ell}_t(\mathbf{w}) \leftarrow \left\langle \mathbf{w}, -\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[\left(T_{\mathbf{w}^{(t)}}(\mathbf{x}) + \frac{\rho(\epsilon)}{2}\right) y\mathbf{x}\right]\right\rangle$
    $\mathbf{w}^{(t)} \leftarrow \Pi_{\mathcal{B}}\left(\mathbf{w}^{(t-1)} - \eta_t \nabla_{\mathbf{w}} \hat{\ell}_t\left(\mathbf{w}^{(t-1)}\right)\right)$ ▷
  $\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$

Given Proposition 5.6, it is straightforward to prove our main results. Here we give the proof for the case of log-concave densities and provide a similar argument for well-behaved distributions in the full version of this paper.

Proof of Theorem 5.2. First, we require a ρ-certificate oracle for log-concave distributions. The algorithm of Theorem 4.1 returns a function $T_{\mathbf{w}}$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[T_{\mathbf{w}}(\mathbf{x}) y \langle \mathbf{w}, \mathbf{x}\rangle] \leq -(\theta/A)^{O(1/\alpha^2)}$. From the definition of $T_{\mathbf{w}}$ (i.e., Equation (3)), it is clear that $\|T_{\mathbf{w}}\|_\infty \leq \frac{1}{\min_{\mathbf{x}\in B} |\langle \mathbf{w}, \mathbf{x}\rangle|} \leq \left(\frac{\log A}{\alpha\theta}\right)^{O(1/\alpha)}$, where B is the band from Equation (3). Note that the function $T_{\mathbf{w}}/\|T_{\mathbf{w}}\|_\infty$ satisfies the conditions of the ρ-certificate oracle. Thus, by scaling the output of the algorithm of Theorem 4.1, we obtain a $(\theta\alpha/A)^{O(1/\alpha^2)}$-certificate oracle.

From Proposition 5.6, this gives us an algorithm that returns a vector $\widehat{\mathbf{w}}$ such that $\theta(\widehat{\mathbf{w}}, \mathbf{w}^*) \leq \frac{\epsilon}{\log^2(1/\epsilon)}$ with probability $1 - \delta$. Using the fact that for log-concave distributions $\mathrm{err}_{0-1}^{\mathcal{D}_{\mathbf{x}}}(h_{\widehat{\mathbf{w}}}, f) \leq O\left(\log^2(1/\epsilon)\theta(\widehat{\mathbf{w}}, \mathbf{w}^*)\right) + \epsilon$ the result follows. □

ACKNOWLEDGMENTS

REFERENCES

[1] D. Angluin and P. Laird. 1988. Learning From Noisy Examples. *Mach. Learn.* 2, 4 (1988), 343–370.
[2] P. Awasthi. 2018. Talk at Workshop on Computational Efficiency and High-Dimensional Robust Statistics. TTI, Chicago. http://www.iliasdiakonikolas.org/tti-robust/Awasthi.pdf
[3] P. Awasthi, M. F. Balcan, N. Haghtalab, and R. Urner. 2015. Efficient Learning of Linear Separators under Bounded Noise. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*. 167–190.
[4] P. Awasthi, M. F. Balcan, N. Haghtalab, and H. Zhang. 2016. Learning and 1-bit Compressed Sensing under Asymmetric Noise. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*. 152–192.
[5] P. Awasthi, M. F. Balcan, and P. M. Long. 2017. The Power of Localization for Efficiently Learning Linear Separators with Noise. *J. ACM* 63, 6 (2017), 50:1–50:27.
[6] M.-F. Balcan, A. Z. Broder, and T.Zhang. 2007. Margin Based Active Learning. In *Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007 (Lecture Notes in Computer Science, Vol. 4539)*. Springer, 35–50.
[7] P. L. Bartlett, O. Bousquet, and S. Mendelson. 2005. Local Rademacher complexities. *Ann. Statist.* 33, 4 (08 2005), 1497–1537.
[8] P. L. Bartlett, M. I. Jordan, and J. D. Mcauliffe. 2006. Convexity, Classification, and Risk Bounds. *J. Amer. Statist. Assoc.* 101, 473 (2006), 138–156.
[9] E. Beigman and B. B. Klebanov. 2009. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 280–287.
[10] A. Blum, A. M. Frieze, R. Kannan, and S. Vempala. 1996. A Polynomial-Time Algorithm for Learning Noisy Linear Threshold Functions. In *37th Annual Symposium on Foundations of Computer Science, FOCS '96*. 330–338.
[11] S. Boucheron, O. Bousquet, and G. Lugosi. 2005. Theory of Classification: a Survey of Some Recent Advances. *ESAIM: Probability and Statistics* 9 (2005), 323–375.
[12] T. Bylander. 1994. Learning Linear Threshold Functions in the Presence of Classification Noise *(COLT '94)*. Association for Computing Machinery, New York, NY, USA, 340–347. https://doi.org/10.1145/180139.181176
[13] R. M. Castro and R. D. Nowak. 2008. Minimax bounds for active learning. *IEEE Transactions on Information Theory* 54, 5 (2008), 2339–2353.
[14] S. Chen, F. Koehler, A. Moitra, and M. Yau. 2020. Classification Under Misspecification: Halfspaces, Generalized Linear Models, and Connections to Evolvability. *CoRR* abs/2006.04787 (2020). arXiv:2006.04787 https://arxiv.org/abs/2006.04787
[15] R. S. Chhikara and J. McKeon. 1984. Linear discriminant analysis with misallocation in training samples. *J. Amer. Statist. Assoc.* 79, 388 (1984), 899–906.
[16] I. Dagan, O. Glickman, and B. Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*. Springer, 177–190.
[17] A. Daniely. 2015. A PTAS for Agnostically Learning Halfspaces. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*. 484–502.
[18] A. Daniely. 2016. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual Symposium on Theory of Computing, STOC 2016*. 105–117.
[19] I. Diakonikolas, T. Gouleakis, and C. Tzamos. 2019. Distribution-Independent PAC Learning of Halfspaces with Massart Noise. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 4751–4762.
[20] I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, and Alistair Stewart. 2019. Sever: A Robust Meta-Algorithm for Stochastic Optimization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*. 1596–1606.
[21] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. 2016. Robust Estimators in High Dimensions without the Computational Intractability. In *Proceedings of FOCS'16*. 655–664.
[22] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. 2017. Being Robust (in High Dimensions) Can Be Practical. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*. 999–1008.
[23] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. 2018. Robustly Learning a Gaussian: Getting Optimal Error, Efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018*. 2683–2702.

[24] I. Diakonikolas and D. M. Kane. 2019. Recent Advances in Algorithmic High-Dimensional Robust Statistics. *CoRR* abs/1911.05911 (2019). arXiv:1911.05911 http://arxiv.org/abs/1911.05911

[25] I. Diakonikolas, D. M. Kane, and A. Stewart. 2018. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*. 1061–1073.

[26] I. Diakonikolas, D. M. Kane, and N. Zarifis. 2020. Near-Optimal SQ Lower Bounds for Agnostically Learning Halfspaces and ReLUs under Gaussian Marginals. *CoRR* abs/2006.16200 (2020). arXiv:2006.16200 https://arxiv.org/abs/2006.16200

[27] I. Diakonikolas, W. Kong, and A. Stewart. 2019. Efficient Algorithms and Lower Bounds for Robust Linear Regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*. 2745–2754.

[28] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. 2020. Learning Halfspaces with Massart Noise Under Structured Distributions. In *Conference on Learning Theory, COLT 2020 (Proceedings of Machine Learning Research, Vol. 125)*, Jacob D. Abernethy and Shivani Agarwal (Eds.). PMLR, 1486–1513.

[29] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. 2020. Learning Halfspaces with Tsybakov Noise. *arXiv preprint arXiv:2006.06467* (2020).

[30] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. 2020. Non-Convex SGD Learns Halfspaces with Adversarial Label Noise. In *NeurIPS*.

[31] V. Feldman, P. Gopalan, S. Khot, and A. Ponnuswami. 2006. New Results for Learning Noisy Parities and Halfspaces. In *Proc. FOCS*. 563–576.

[32] B. Frénay and M. Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25, 5 (2013), 845–869.

[33] Y. Freund and R. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* 55, 1 (1997), 119–139.

[34] S. Goel, A. Gollakota, and A. Klivans. 2020. Statistical-Query Lower Bounds via Functional Gradients. *CoRR* abs/2006.15812 (2020). arXiv:2006.15812 https://arxiv.org/abs/2006.15812

[35] V. Guruswami and P. Raghavendra. 2006. Hardness of learning halfspaces with noise. In *Proc. 47th IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE Computer Society, 543–552.

[36] S. Hanneke. 2011. Rates of convergence in active learning. *Ann. Statist.* 39, 1 (02 2011), 333–361.

[37] S. Hanneke and L. Yang. 2015. Minimax analysis of active learning. *J. Mach. Learn. Res.* 16 (2015), 3487–3602.

[38] D. Haussler. 1992. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation* 100 (1992), 78–150.

[39] M. Hopkins, D. M. Kane, S. Lovett, and G. Mahajan. 2020. Noise-tolerant, reliable active classification with comparison queries. In *COLT*.

[40] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. 2008. Agnostically Learning Halfspaces. *SIAM J. Comput.* 37, 6 (2008), 1777–1805.

[41] M. Kearns, R. Schapire, and L. Sellie. 1994. Toward Efficient Agnostic Learning. *Machine Learning* 17, 2/3 (1994), 115–141.

[42] B. B. Klebanov and E. Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics* 35, 4 (2009), 495–503.

[43] B. B. Klebanov and E. Beigman. 2010. Some empirical evidence for annotation noise in a benchmarked dataset. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 438–446.

[44] B. B. Klebanov and E. Beigman. 2010. Some empirical evidence for annotation noise in a benchmarked dataset. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 438–446.

[45] A. Klivans, P. Long, and R. Servedio. 2009. Learning Halfspaces with Malicious Noise. (2009). To appear in *Proc. 17th Internat. Colloq. on Algorithms, Languages and Programming (ICALP)*.

[46] A. R. Klivans, P. K. Kothari, and R. Meka. 2018. Efficient Algorithms for Outlier-Robust Regression. In *Conference On Learning Theory, COLT 2018*. 1420–1430.

[47] K. A. Lai, A. B. Rao, and S. Vempala. 2016. Agnostic Estimation of Mean and Covariance. In *Proceedings of FOCS'16*.

[48] Y. T. Lee and S. S. Vempala. 2017. Eldan's Stochastic Localization and the KLS Hyperplane Conjecture: An Improved Lower Bound for Expansion. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. 998–1007.

[49] W. Maass and G. Turan. 1994. How fast can a threshold gate learn?. In *Computational Learning Theory and Natural Learning Systems*, S. Hanson, G. Drastal, and R. Rivest (Eds.). MIT Press, 381–414.

[50] E. Mammen and A. B. Tsybakov. 1999. Smooth discrimination analysis. *Ann. Statist.* 27, 6 (12 1999), 1808–1829.

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

[52] A. K. Menon, B. Van Rooyen, and N. Natarajan. 2018. Learning from binary labels with instance-dependent noise. *Machine Learning* 107, 8-10 (2018), 1561–1595.

[53] M. Minsky and S. Papert. 1968. *Perceptrons: an introduction to computational geometry*. MIT Press, Cambridge, MA.

[54] A. Novikoff. 1962. On convergence proofs on perceptrons. In *Proceedings of the Symposium on Mathematical Theory of Automata*, Vol. XII. 615–622.

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

[56] R. H. Sloan. 1988. Types of Noise in Data for Concept Learning. In *Proceedings of the First Annual Workshop on Computational Learning Theory* (MIT, Cambridge, Massachusetts, USA) *(COLT '88)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 91–96.

[57] A. Tsybakov. 2004. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics* 32, 1 (2004), 135–166.

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

[59] V. Vapnik. 1998. *Statistical Learning Theory*. Wiley-Interscience, New York.

[60] R. Vershynin. 2018. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press. https://doi.org/10.1017/9781108231596

[61] S. Yan and C. Zhang. 2017. 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*. 1056–1066.

[62] C. Zhang, J. Shen, and P. Awasthi. 2020. Efficient active learning of sparse halfspaces with arbitrary bounded noise. *CoRR* abs/2002.04840 (2020). arXiv:2002.04840

[63] Y. Zhang, P. Liang, and M. Charikar. 2017. A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*. 1980–2022.