Testable Learning of *General* Halfspaces with Adversarial Label Noise

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin Madison

Daniel M. Kane DAKANE@UCSD.EDU
University of California San Diego

Sihan Liu SIL046@UCSD.EDU
University of California San Diego

Nikos Zarifis ZARIFIS@WISC.EDU
University of Wisconsin Madison

Editors: Shipra Agrawal and Aaron Roth

Abstract

We study the task of testable learning of general — not necessarily homogeneous — halfspaces with adversarial label noise with respect to the Gaussian distribution. In the testable learning framework, the goal is to develop a tester-learner such that if the data passes the tester, then one can trust the output of the robust learner on the data. Our main result is the first polynomial time tester-learner for general halfspaces that achieves dimension-independent misclassification error. At the heart of our approach is a new methodology to reduce testable learning of general halfspaces to testable learning of nearly homogeneous halfspaces that may be of broader interest.

Keywords: PAC learning, testable learning, adversarial label noise, linear threshold functions

1. Introduction

A (general) halfspace or Linear Threshold Function (LTF) is any Boolean function $h : \mathbb{R}^d \to \{\pm 1\}$ of the form $h_{\mathbf{w}}(\mathbf{x}) = \mathrm{sign}\,(\mathbf{w} \cdot \mathbf{x} + t)$, where $\mathbf{w} \in \mathbb{R}^d$ is the defining vector, $t \in \mathbb{R}$ is the threshold, and $\mathrm{sign} : \mathbb{R} \to \{\pm 1\}$ is defined as $\mathrm{sign}(t) = 1$ if $t \geq 0$ and $\mathrm{sign}(t) = -1$ otherwise. The family of halfspaces is one of the most basic concept classes in computational learning theory with history dating back to Rosenblatt's perceptron algorithm (Rosenblatt, 1958). While halfspaces are efficiently learnable in Valiant's (realizable) distribution-free PAC model (Valiant, 1984) (i.e., with clean labels), in the agnostic (or adversarial label noise) model (Haussler, 1992; Kearns et al., 1994) even *weak* learning is intractable (Daniely, 2016; Diakonikolas et al., 2022a; Tiegel, 2023), unless one makes assumptions on the distribution of feature vectors.

A long line of work, starting with Kalai et al. (2008), has developed efficient halfspace learners in the distribution-specific agnostic model. Concretely, Kalai et al. (2008) gave an agnostic learner within 0-1 error of $\mathrm{opt} + \epsilon$ — where opt is the 0-1 error of the best-fitting halfspace — with complexity $d^{O(1/\epsilon^2)}$ if the distribution on feature vectors is the standard Gaussian. Unfortunately, if one insists on optimal error of $\mathrm{opt} + \epsilon$, computational limitations arise even in the Gaussian setting. Specifically, the exponential complexity dependence on $1/\epsilon$ has been recently shown to be inherent (Diakonikolas et al., 2020a; Goel et al., 2020; Diakonikolas et al., 2021b, 2023b).

By relaxing the desired accuracy to $f(\mathrm{opt}) + \epsilon$, for an appropriate function $f(t)$ that goes to 0 when $t \to 0$, it is possible to obtain $\mathrm{poly}(d/\epsilon)$ (i.e., *fully* polynomial) time algorithms (Klivans et al., 2009; Awasthi et al., 2017; Daniely, 2015; Diakonikolas et al., 2018, 2020b, 2022b). Specifically,

for the class of *homogeneous* halfspaces — corresponding to the case that the threshold $t = 0$ or equivalently that the separating hyperplane goes through the origin — Awasthi et al. (2017) first gave a $\mathrm{poly}(d/\epsilon)$ time algorithm with error $C\mathrm{opt} + \epsilon$, for some universal constant $C > 1$, that succeeds under the standard Gaussian (and, more generally, isotropic logconcave distributions). Interestingly, the case of *general* halfspaces turns out to be more challenging: with the exception of Diakonikolas et al. (2018, 2022b), all prior approaches inherently fail for non-homogeneous halspaces. We will return to this point in the subsequent discussion.

The model of *testable learning* (Rubinfeld and Vasilyan, 2023) was defined to alleviate the following conceptual limitation of distribution-specific agnostic learning: if the assumptions on the marginal distribution on examples (e.g., Gaussianity or log-concavity) are not satisfied, a vanilla agnostic learner provides no guarantees. Ideally, one would like to guarantee the following desiderata: (a) if the learner accepts, then we can trust its output, and (b) it is unlikely that the learner rejects if the data satisfies the distributional assumptions. This has led to the following definition:

Definition 1 (Testable Learning with Adversarial Label Noise (Rubinfeld and Vasilyan, 2023)) *Fix $\epsilon, \tau \in (0, 1]$ and let $f : [0, 1] \mapsto \mathbb{R}_+$. A tester-learner $\mathcal{A}$ (approximately)* testably learns *a concept class $\mathcal{C}$ with respect to the distribution $D_{\mathbf{x}}$ on $\mathbb{R}^d$ with N samples, and failure probability τ if the following holds. For any distribution D on $\mathbb{R}^d \times \{\pm 1\}$, the tester-learner $\mathcal{A}$ draws a set S of N i.i.d. samples from D. In the end, it either rejects S or accepts S and produces a hypothesis $h : \mathbb{R}^d \mapsto \{\pm 1\}$. Moreover, the following conditions must be met:*

- *(Completeness) If D truly has marginal $D_{\mathbf{x}}$, $\mathcal{A}$ accepts with probability at least $1 - \tau$.*
- *(Soundness) The probability that $\mathcal{A}$ accepts and outputs a hypothesis h for which $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] > f(\mathrm{opt}) + \epsilon$, where $\mathrm{opt} := \min_{g\in\mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[g(\mathbf{x}) \neq y]$ is at most τ.*

The probability in the above statements is over the randomness of the sample S and the internal randomness of the tester-learner $\mathcal{A}$.

Since the introduction of the model, algorithmic aspects of testable learning have been investigated in a number of works (Rubinfeld and Vasilyan, 2023; Gollakota et al., 2023a; Diakonikolas et al., 2023a; Gollakota et al., 2023b,c). The earlier works (Rubinfeld and Vasilyan, 2023; Gollakota et al., 2023a) studied the agnostic setting where $f(t) = t$ (corresponding to error $\mathrm{opt} + \epsilon$). These works developed general moment-matching methodology that gives testable learners for general halfspaces (and other concept classes) with runtime $d^{\mathrm{poly}(1/\epsilon)}$. Since testable learning with error guarantees $\mathrm{opt}+\epsilon$ is at least as hard as the standard agnostic setting with the same error guarantees, the resulting algorithms necessarily incur exponential dependence in $1/\epsilon$. Subsequent work (Diakonikolas et al., 2023a; Gollakota et al., 2023b,c) developed testable learners with weaker error guarantees that run in fully-polynomial time. Specifically, these algorithms achieve error $O(\mathrm{opt})+\epsilon$ under the Gaussian distribution, and more recently, for a subclass of log-concave distributions. *Importantly, all of the existing fully-polynomial time algorithms work only for* homogeneous *halfspaces and inherently fail for general halfspaces.*

The distinction between homogeneous and general halfspaces might seem inconsequential at first glance. Indeed, one can trivially reduce a general halfspace to a homogeneous one by adding an extra constant coordinate to each datapoint. While this is a valid reduction for distribution-free learning, it alters the marginal distribution on the feature vectors. Therefore, it does not generally work in the distribution-specific setting we study here. Beyond this, as we will elaborate in the

proceeding section, the techniques underlying previous testable learning algorithms are insufficient to obtain any non-trivial error guarantee for general halfspaces.

Motivated by this gap in our understanding, in this work we ask whether it is possible to obtain a fully-polynomial time testable learner for general halfspaces with dimension-independent error. Specifically, we study the following question:

Is there a $\mathrm{poly}(d/\epsilon)$ *time tester-learner for* general *halfspaces with error* $f(\mathrm{opt}) + \epsilon$*?*

As our main result, formally stated below, we provide an affirmative answer to this question.

Theorem 2 (Testable Learning General Halfspaces under Gaussian Marginals) *Let* $\epsilon, \tau \in (0,1)$ *and* $\mathcal{C}$ *be the class of general halfspaces on* $\mathbb{R}^d$*. There exists a tester-learner for* $\mathcal{C}$ *with respect to* $\mathcal{N}(\mathbf{0}, \mathbf{I})$ *up to* 0-1 *error* $\widetilde{O}\left(\sqrt{\mathrm{opt}}\right) + \epsilon$*, where* opt *is the* 0-1 *error of the best fitting function in* $\mathcal{C}$*, that fails with probability at most* τ*. Furthermore, the algorithm draws* $N = \mathrm{poly}(d, 1/\epsilon)\log(1/\tau)$ *samples, and runs in time* $\mathrm{poly}(N)$*.*

We reiterate that this is the first polynomial-time testable learner for *general* halfspaces that achieves non-trivial (i.e., dimension-independent) error guarantee. As already mentioned, prior works either focus on *homogeneous* halfspaces or incur super-polynomial runtime.

We leave open the questions of obtaining quantitatively better error guarantee (with $O(\mathrm{opt}) + \epsilon$ as the ideal goal) and extending to broader classes of distributions beyond the Gaussian. Either potential improvement stumbles upon non-trivial obstacles; see Remark 13 and 14.

1.1. Some natural attempts and why they fail

One may wonder whether previously known testable learners for homogeneous halfspaces can be applied (directly or with minor modifications) in our setting to achieve non-trivial learning error, e.g., opt^c for some $c \in (0,1)$. We will demonstrate below that known methods fail to achieve testable learning for general halfspaces with even (sufficiently small) constant error.

Difficulty of Estimating Chow Parameters For learning general halfspaces under the Gaussian distribution with adversarial label noise, the only prior works that achieve dimension-independent error are Diakonikolas et al. (2018) and Diakonikolas et al. (2022b) (in the non-testable setting). In both of these works, the algorithms are required at some point to estimate the Chow-parameters, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim D}[y\mathbf{x}]$. The intuition behind this choice is that for any halfspace $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)$, we have that $\mathbf{E}_{(\mathbf{x},y)\sim D}[h(\mathbf{x})\mathbf{x}] = \mathbf{w}G(t)$, where $G(t)$ is the pdf of the standard normal. In the non-testable regime, this expectation is easily computable with $\mathrm{poly}(d/\epsilon)$ samples and runtime. Unfortunately, in the testable regime, one needs to first certify that the marginal distribution $D_{\mathbf{x}}$ is similar to a Gaussian in the sense that $\mathbf{E}_{\mathbf{x}\sim D_{\mathbf{x}}}[h(\mathbf{x})\mathbf{x}] \sim \mathbf{w}G(t)$ for any halfspace h. To the best of our knowledge, the only way to certify this property is to certify that at least $\Omega(1/G(t))$ moments of the distribution $D_{\mathbf{x}}$ match those of the standard normal (see Gollakota et al. (2023a)), which requires at least $d^{\Omega(1/G(t))}$ samples and runtime. In other words, this approach would lead to exponential runtime if $1/G(t)$ is on the order of $\mathrm{poly}(1/\epsilon)$.

Adapting Testable Learners for Homogeneous Halfspaces For testable learning of homogeneous halfspaces, two techniques have been used in prior work. The first one uses an adaptive localization method (Diakonikolas et al., 2023a), and the second is to construct non-convex SGD tester-learners (Gollakota et al., 2023b,c). At a high-level, both of these techniques localize in a band *close to*

the origin in order to maximize the error conditioned on the band of the current hypothesis and the optimal one. Unfortunately, for general halfspaces, one cannot localize to maximize the error, *unless* the angle between the current hypothesis (weight vector) and the optimal one is sufficiently small. This holds because if the angle is large, then no band can contain large enough error, except bands that are very far away (which due to the noise model can be arbitrarily noisy). The only known method to obtain a good enough initialization point is to rely on Chow-parameters — but as we discussed above, this is hard in the testable regime.

1.2. Overview of Techniques

Reduction to Nearly Homogeneous Halfspaces The overall strategy of our algorithmic approach is to efficiently reduce the testable learning of general halfspaces to the testable learning of "nearly" homogeneous halfspaces, i.e., halfspaces with thresholds of size ϵ [1]. Assuming such an efficient reduction, our main theorem then follows via a careful analysis showing that (essentially) the main algorithm of the prior work by Diakonikolas et al. (2023a) (on testable learning of homogeneous halfspaces) can testably learn halfspaces with small biases. The key notion in implementing such a reduction is what we refer to as *Good Localization Centers* (Definition 3). In short, a good localization center is a point that is close to the decision boundary of the target halfspace and at the same time not too far from the origin. As an example, the intersection between the line along the defining vector of the target halfspace and the separating hyperplane of the halfspace will be a good localization center.

Then, leveraging the technique developed in Diakonikolas et al. (2018), one can localize to regions around the good localization center via rejection sampling. After such a localization procedure, using the fact that the good localization center is almost on the target halfspace, one can show that a general halfspace will become nearly homogeneous. Moreover, since the center is also required to be not too far from the origin, one can further demonstrate that a decent fraction of the samples can still survive the rejection sampling, ensuring that the proceeding invocation of the homogeneous halfspace tester-learner is sample and computationally efficient. Our main technical contribution is an efficient testable learner that computes a list of candidates containing at least one good localization center. This implies that the learner has inherently learned a good constant approximation to the defining vector of the unknown halfspace, which is the major obstacle against applying prior methods (developed for homogeneous halfspaces) in the general halfspace setting (see Section 1.1 for more details).

Finding a Good Localization Center We now describe a testable learning procedure that either produces a good localization center or reports that the underlying distribution is not Gaussian. We can assume that the optimal halfspace is $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} - t^*)$ for some $t^* > 0$. As mentioned in the preceding discussion, the ideal localization center would be the intersection between the line along the defining vector $\mathbf{v}^*$ and the separating hyperplane of the target halfspace.

Naively, one may try to estimate the Chow parameters $\tilde{\mathbf{v}} := \mathbf{E}_{(\mathbf{x},y)\sim D}[y\mathbf{x}]$, and pick some $\mathbf{w}$ along the direction of $\tilde{\mathbf{v}}$ — as the angle between $\tilde{\mathbf{v}}$ and $\mathbf{v}^*$ ought to be small when the underlying $\mathbf{x}$-marginal is Gaussian. Yet, as we discussed in Section 1.1, this cannot be done efficiently in the testable regime. For this reason, we instead calculate the mean of the points that are positively

1. Our main technical contribution is to show that we can efficiently construct a list of $\tilde{O}(1/\epsilon)$ candidate learning instances one of which is guaranteed to correspond to a halfspace with a threshold of size at most ϵ; and reduce testable learning of the unknown general halfspace to testable learning of these nearly homogeneous halfspaces.

labeled (formally, the tail points; see Definition 10), i.e., $\bar{\boldsymbol{\mu}}_+ = \mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{x} \mid y = 1]$. In the absence of corrupted labels, this gives us a point located in the positive side of the halfspace, i.e., $h(\bar{\boldsymbol{\mu}}_+) = 1$, as $\mathbf{v}^* \cdot \bar{\boldsymbol{\mu}}_+ = \mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{v}^* \cdot \mathbf{x} \mid h(\mathbf{x}) = 1] \geq t^*$. In other words, there exists a $\lambda \in (0, 1)$ so that $\lambda\bar{\boldsymbol{\mu}}_+ \cdot \mathbf{v}^* = t^*$, and if we center our distribution at the point $\mathbf{w} = \lambda\bar{\boldsymbol{\mu}}_+$ via rejection sampling (see Definition 4), then the optimal halfspace becomes exactly homogeneous (Figure 1).

There are two obstacles that we need to circumvent in order for such an approach to work: (i) We do not have access to the true labels and, by conditioning on the positive labels, the noise can make the corrupted mean $\boldsymbol{\mu}_+ = \mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{x} \mid y = 1]$ appear on the negative side, i.e., $h(\boldsymbol{\mu}_+) = -1$; and (ii) the point $\lambda\boldsymbol{\mu}_+$ can be very far from the origin, i.e., $\|\lambda\boldsymbol{\mu}_+\|_2 \gtrsim \sqrt{\log(1/\mathrm{opt})}$, and applying rejection sampling to re-center the distribution at this point can result in a corruption level that is far worse than the current one. Fortunately, we can ensure that these situations cannot happen by certifying the following two properties of the empirical distribution over samples: (i) the distribution has bounded covariance, and (ii) the cumulative density function of the distribution projected along the direction of $\boldsymbol{\mu}_+$ is sufficiently close to that of the standard Gaussian.

First, if we assume that (a) the mass of the positive points is at least $B \geq \sqrt{\mathrm{opt}}\mathrm{poly}(\log(1/\mathrm{opt}))$, and (b) the sample covariance has spectral norm bounded from above by 2, then we can ensure that $\|\boldsymbol{\mu}_+ - \bar{\boldsymbol{\mu}}_+\|_2 \leq 1/\mathrm{poly}(\log(1/\mathrm{opt}))$ via the following simple calculation: for any unit vector $\mathbf{u}$, $|\mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{u} \cdot \mathbf{x}(\mathbb{1}\{h(\mathbf{x}) = 1\} - \mathbb{1}\{y = 1\})]| \leq \mathbf{E}_{(\mathbf{x},y)\sim D}[(\mathbb{1}\{h(\mathbf{x}) = 1\} - \mathbb{1}\{y = 1\})^2]^{1/2}\,\mathbf{E}_{(\mathbf{x},y)\sim D}[(\mathbf{u} \cdot \mathbf{x})^2]^{1/2} \leq \sqrt{2\mathrm{opt}}$; hence, the error $\|\boldsymbol{\mu}_+ - \bar{\boldsymbol{\mu}}_+\|_2$ is at most $\sqrt{2\mathrm{opt}}/B \leq 1/\mathrm{poly}(\log(1/\mathrm{opt}))$. Note that assumption (b) can be efficiently verified and assumption (a) on $B \geq \sqrt{\mathrm{opt}}\mathrm{poly}(\log(1/\mathrm{opt}))$ is without loss of generality — otherwise, the constant hypothesis, i.e., $h(\mathbf{x}) = -1$ for all $\mathbf{x}$, would get $\widetilde{O}(\sqrt{\mathrm{opt}})$ error [2]. This suffices to show that the intersection point $\mathbf{w}$ is close to $\boldsymbol{\mu}_+$. This sketch is formalized in Lemma 11.

It remains to show that $\|\boldsymbol{\mu}_+\|_2$ is not very large. To this end, we certify that the empirical distribution projected along $\boldsymbol{\mu}_+$ has its cumulative density function close to that of the standard Gaussian. If so, $\|\boldsymbol{\mu}_+\|_2$ can then be bounded from below by $\Phi^{-1}(M)$, where M denotes the mass of the samples inside the region $\{\mathbf{x} \in S : \boldsymbol{\mu} \cdot \mathbf{x} > \|\boldsymbol{\mu}\|_2^2\}$. Conditioned on that the "Gaussian closeness" CDF test passes, we show that the distribution of the tail points must be approximately "centered" around $\boldsymbol{\mu}$ in the $\boldsymbol{\mu}$-direction. Thus, $\{\mathbf{x} \in S : \boldsymbol{\mu} \cdot \mathbf{x} > \|\boldsymbol{\mu}\|_2^2\}$ must contain a large fraction of the tail points, allowing us to derive a lower bound on the mass of the set M, from which the upper bound on $\|\boldsymbol{\mu}_+\|_2$ follows. See Lemma 12 for the details of the argument.

Conditioned on the relevant tests all passing, we can construct a list of candidate localization centers along the direction of $\boldsymbol{\mu}_+$ whose ℓ_2-norms form an ϵ-cover of the segment $[0, O(\sqrt{\log(1/B)})]$. It is not hard to see that such a list is guaranteed to contain one localization center as good as $\mathbf{w}$. We then perform the reduction to near-homogeneous halfspaces with each of the candidate localization centers in our list. Finally, we simply choose the best halfspace thus obtained, by examining the empirical errors of the list of the halfspaces learned. This concludes the overview of our approach.

1.3. Preliminaries

We use small boldface characters for vectors and capital bold characters for matrices. We use $[d]$ to denote the set $\{1, 2, \ldots, d\}$. For a vector $\mathbf{x} \in \mathbb{R}^d$ and $i \in [d]$, $\mathbf{x}_i$ denotes the i-th coordinate of $\mathbf{x}$,

2. In the end, we can simply compare the halfspace learned via localization with the constant hypothesis in terms of their empirical errors to select the best.

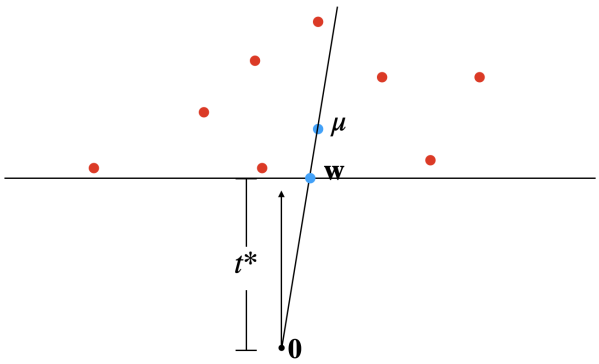

Figure 1: When there is no noise, the tail points are at distance at least t^* from the origin, since they all lie on the side of the hyperplane not containing the origin. Consequently, the line along their mean $\boldsymbol{\mu}$ must first intersect the separating hyperplane of the halfspace (crossing $\mathbf{w}$) and then cross the mean vector $\boldsymbol{\mu}$ of the tail points, regardless of the underlying marginal distribution. If we re-center the distribution at $\mathbf{w}$, the halfspace will then become exactly homogeneous.

and $\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^d \mathbf{x}_i^2}$ the ℓ_2 norm of $\mathbf{x}$. We use $\mathbf{x} \cdot \mathbf{y} := \sum_{i=1}^n \mathbf{x}_i \mathbf{y}_i$ as the inner product between them. We use $\mathbb{1}\{E\}$ to denote the indicator function of some event E.

We use $\mathbf{E}_{\mathbf{x}\sim D}[\mathbf{x}]$ for the expectation of the random variable $\mathbf{x}$ according to the distribution D and $\mathbf{Pr}[E]$ for the probability of event E. For simplicity of notation, we may omit the distribution when it is clear from the context. For $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, we denote by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For $(\mathbf{x}, y) \in \mathcal{X}$ distributed according to D, we denote $D_{\mathbf{x}}$ to be the marginal distribution of $\mathbf{x}$. Let $f : \mathbb{R}^d \mapsto \{\pm 1\}$ be a boolean function and D a distribution over $\mathbb{R}^d$. The (degree-1) Chow parameter vector of f with respect to D is defined as $\mathbf{E}_{\mathbf{x}\sim D}[f(\mathbf{x})\mathbf{x}]$. For a halfspace $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + t)$, we say that $\mathbf{v}$ is the defining vector of h and t is the threshold of h.

We denote by $\Phi(t)$ the cumulative density function (CDF) of the one-dimensional standard Gaussian, i.e., $\Phi(t) = \mathbf{Pr}_{x\sim\mathcal{N}(0,1)}[x > t]$, and by $G(t)$ its probability density function, i.e., $G(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$. An elementary inequality we use is that $\Phi(t) \leq \frac{1}{t\sqrt{2\pi}} \exp(-t^2/2) = G(t)/t$. We say that two one-dimensional distributions X, Y are ϵ-close in CDF (or Kolmogorov) distance if they satisfy the inequality $\sup_{t\in R} |\mathbf{Pr}[X > t] - \mathbf{Pr}[Y > t]| \leq \epsilon$.

The asymptotic notation $\tilde{O}$ (resp. $\tilde{\Omega}$) suppresses logarithmic factors in its argument, i.e., $\tilde{O}(f(n)) = O(f(n)\log^c f(n))$ and $\tilde{\Omega}(f(n)) = \Omega(f(n)/\log^c f(n))$, where $c > 0$ is a universal constant. We write $a \lesssim b$ (resp. $a \gtrsim b$) to denote that $\alpha \leq cb$ for a sufficiently small absolute constant $c > 0$ (resp. for a sufficiently large absolute constant $c > 0$).

2. Testable Learning of General Halfspaces to Error $\widetilde{O}\left(\sqrt{\mathrm{opt}}\right)$

In this section, we establish our main result (Theorem 2). Our main strategy is to efficiently reduce testable learning of general halfspaces to testable learning of *nearly* homogeneous halfspaces, i.e., halfspaces whose thresholds have small magnitude. The main technical tool enabling such a reduction is the notion of a *good localization center*, defined below.

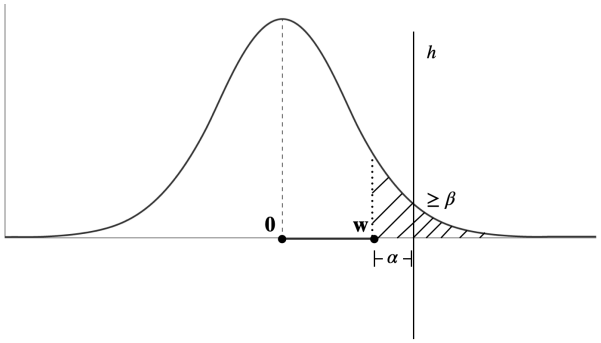

Figure 2: The figure illustrates a good localization center $\mathbf{w}$ with respect to the halfspace h. $\mathbf{w}$ is α-far from the halfspace, and $\Phi(\|\mathbf{w}\|_2)$ still captures enough mass.

Definition 3 (Good Localization Center) *Let $\alpha > 0$ and $\beta \in (0,1)$. Given a general halfspace $h(\mathbf{x})$, we say that a point $\mathbf{w} \in \mathbb{R}^d$ is an* (α, β)-good localization center with respect to *h, if (i) the distance from the separating hyperplane of h to $\mathbf{w}$ is bounded by α, and (ii) $\Phi(\|\mathbf{w}\|_2) \geq \beta$, where Φ is the cdf of $\mathcal{N}(0,1)$.*

In Section 2.1, we demonstrate how a good localization center can be leveraged to perform such a reduction. In the process, we also develop additional required technical machinery, including soft-localization (Lemma 5) and a Wedge-Bound (Lemma 8). In Section 2.2, we describe an algorithm (Algorithm 1) that testably constructs a list of candidate points containing at least one good localization center; this is the main technical contribution of this work. In Section 2.3, we put everything together to complete the proof of Theorem 2.

2.1. Good Localization Center and its Properties

Let $\mathbf{w}$ be an (α, β)-good localization center with respect to the halfspace h. If $\mathbf{w}$ is sufficiently close to the separating hyperplane of the halfspace and β is bounded from below, we show that $\mathbf{w}$ can be leveraged to reduce testable learning of general halfspaces to testable learning halfspaces that are *nearly homogeneous*. Specifically, Lemma 5 demonstrates that if we "localize" the $\mathbf{x}$-marginal around the center $\mathbf{w}$ using rejection sampling (see Definition 4), the halfspace will be transformed into a new one whose threshold is bounded above by $O(\alpha\sqrt{\log(1/\beta)})$.

Definition 4 *For $\mathbf{w} \in \mathbb{R}^d$ and $\sigma \in (0,1)$, the $(\mathbf{w}, \sigma)$-rejection procedure is as follows: given $\mathbf{x} \in \mathbb{R}^d$, the procedure accepts it with probability* $\exp\left(-(\sigma^{-2}-1)(\mathbf{w}/\|\mathbf{w}\|_2 \cdot \mathbf{x} + \|\mathbf{w}\|_2/(1-\sigma^2))^2/2\right)$, *and rejects it otherwise.*

At a high level, if one performs $(\mathbf{w}, \sigma)$-rejection sampling on a Gaussian distribution, the resulting distribution will be $G := \mathcal{N}(\mathbf{w}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the matrix with eigenvalue σ^2 in the $\mathbf{w}$-direction, and eigenvalues 1 in all orthogonal directions. This follows from the fact that the formula of the acceptance probability in Definition 4 is precisely the ratio between the probability density functions of G and the standard Gaussian. Let S_h be the separating hyperplane of h. Assume that $\mathbf{w}$ is α-close to S_h. This immediately implies that S_h will be at most α-far from the center of the new distribution G. If we transform the space to make G isotropic again, h will then consequently get transformed into a new halfspace with small threshold. This is formally shown in the following lemma, whose proof can be found in Appendix B.

Lemma 5 (Localization With A Good Center) *Let* $\mathbf{w}$ *be an* (α, β)*-localization center with respect to* $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t^*)$ *and* $\sigma := \min(\|w\|_2^{-1}, \sqrt{1/2})$*. The following hold:*

1. *The* $(\mathbf{w}, \sigma)$*-rejection sampling procedure (Definition 4) applied on the standard Gaussian, results in a distribution* $G := \mathcal{N}(\mathbf{w}, \mathbf{\Sigma})$ *with* $\mathbf{\Sigma} = \mathbf{I} - (1 - \sigma^2)\mathbf{w}\mathbf{w}^T / \|\mathbf{w}\|_2^2$*. Moreover, the acceptance probability is* $\Omega(\beta)$*, and* $\sigma \geq \Omega(1/\sqrt{\log(1/\beta)})$*.*

2. *Applying the transformation* $(\mathbf{x}, y) \mapsto (\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \mathbf{w}), y)$ *to the distribution of* $(\mathbf{x}, h(\mathbf{x}))$*, where* $\mathbf{x} \sim G$*, leads to the distribution* $(\mathbf{x}, h'(\mathbf{x}))$*, where* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$*, and* $h'(\mathbf{x}) = \mathrm{sign}(\mathbf{\Sigma}^{1/2}\mathbf{v}^* \cdot \mathbf{x} + \mathbf{v}^* \cdot \mathbf{w} + t^*)$*. Moreover, the resulting halfspace* h' *has its threshold size bounded from above by* $|\mathbf{v}^* \cdot \mathbf{w} + t^*| / \|\mathbf{\Sigma}^{1/2}\mathbf{v}^*\|_2 = O(\alpha\sqrt{\log(1/\beta)})$*.*

Given a good localization center, Lemma 5 allows us to transform the original halfspace into one with a small threshold. We show that such a nearly homogeneous halfspace can be testably learned effectively in the proposition below.

Proposition 6 (Testable Learning of Nearly Homogeneous Halfspaces) *Let* $\epsilon, \tau \in (0, 1)$*. Let* D *be a distribution over* $\mathbb{R}^d \times \{\pm 1\}$ *and* h *be a halfspace that achieves error* opt *under* D *with* $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t^*)$*, where* $\mathbf{v}^* \in \mathbb{R}^d$ *is some unit vector, and* $|t| < \epsilon$*. Then, given* $N = \mathrm{poly}(d/\epsilon)\log(1/\tau)$ *many i.i.d. samples from* D*, there exists an efficient tester-learner that runs in time* $\mathrm{poly}(N)$*, and either reports* $D_{\mathbf{x}}$ *is not Gaussian or outputs a vector* $\mathbf{v}$*. Moreover, with probability at least* $1 - \tau$*, we have (i) if the algorithm reports* $D_{\mathbf{x}}$ *is not Gaussian, the report is correct, and (ii) if the algorithm returns a vector* $\mathbf{v}$*,* $\mathbf{v}$ *satisfies that* $\|\mathbf{v} - \mathbf{v}^*\|_2 \leq O(\mathrm{opt}) + \epsilon$*.*

This proposition is established by a careful analysis of the testable learner for homogeneous halfspaces by Diakonikolas et al. (2023a). The main idea is that whenever the threshold of the optimal halfspace is bounded by some absolute constant, the Chow parameters will be a constant multiplicative factor of the defining vector. We can testably learn the Chow parameters up to constant error, which then gives a constant approximation to the defining vector of the halfspace. After that, the algorithm proceeds in an iterative manner. In particular, it performs localization via rejection sampling using the current estimate of the defining vector. The localization steps may blow up the threshold t^* of the optimal halfspace by at most a factor of $1/\epsilon$. Since t^* is initially bounded from above by ϵ, the threshold stays bounded from above by some absolute constant. This ensures that the Chow-parameters vector stays within a constant multiplicative factor of the defining vector, thereby allowing us to iteratively approximate the defining vector by testably learning the Chow parameters to constant error. The detailed proof can be found in Appendix A.

Since Lemma 5 gives the form of the transformation on the defining vector exactly, after learning the defining vector of the nearly homogeneous halfspace, we can revert the transformation to get back the defining vector of the original halfspace. While the argument for bounding the error of the vector after reverting the transformation is a standard piece in most localization based learning procedures (see, e.g., Diakonikolas et al. (2021a)), one technical issue in our scenario is that such an argument usually requires the angle between the localization direction $\mathbf{w}$ and the defining vector $\mathbf{v}^*$ to be bounded away from $\pi/2$ by some constant. On the other hand, we may need to apply the lemma in cases where $\mathbf{w}$ and $\mathbf{v}^*$ are almost orthogonal. To deal with this issue, we require the following lemma which carefully bounds how the error grows as the angle between $\mathbf{w}$ and $\mathbf{v}^*$ increases (see Appendix B for the proof).

Lemma 7 (Transformation Error) *Let $\sigma, \delta \in (0,1)$, $\beta \in (0, \sqrt{2})$, $\mathbf{w}, \mathbf{v}, \mathbf{v}^* \in \mathbb{R}^d$ be unit vectors, and $\mathbf{\Sigma} = \mathbf{I} - (1-\sigma^2)\mathbf{w}\mathbf{w}^T$. Furthermore, assume that $\|\mathbf{v}^* - \mathbf{w}\|_2 < \beta$, $\|\mathbf{v} - \mathbf{\Sigma}^{1/2}\mathbf{v}^*/\|\mathbf{\Sigma}^{1/2}\mathbf{v}^*\|_2\|_2 < \delta$. Then it holds $\|\mathbf{\Sigma}^{1/2}\mathbf{v}/\left\|\mathbf{\Sigma}^{1/2}\mathbf{v}\right\|_2 - \mathbf{v}^*\| \leq O(\delta)(\sigma + \beta)\left(\sigma^{-1}\frac{\beta}{1-\beta^2/2} + 1\right)$.*

After learning the defining vector, we search for the threshold of the halfspace in a brute-force manner. In particular, we construct a list of candidate halfspaces with $\mathrm{poly}(1/\epsilon)$ many possible thresholds, and then select the one with the smallest empirical error after drawing sufficiently many samples. The list is guaranteed to contain some halfspace $\tilde{h}$ whose defining vector and threshold are both close to those of the optimal halfspace, and such a search procedure based on the empirical error is guaranteed to yield some halfspace whose overall error is no worse than the best among the list (up to a constant factor). The last remaining piece is a procedure to certify that closeness to the optimal halfspace in parameter distance implies small 0-1 error. This is achieved by running a Wedge-Bound algorithm, very similar to the one from Diakonikolas et al. (2023a).

Though the algorithm is essentially the same to the one in Diakonikolas et al. (2023a), we need to generalize its analysis to handle general halfspaces. The formal guarantee is specified in the following lemma (see Algorithm 2 for the detailed pseudocode and Appendix B for the proof).

Lemma 8 (Wedge Bound for General Halfspaces) *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. Let $\mathbf{v}, \mathbf{v}^*$ be two unit vectors in $\mathbb{R}^d$ satisfying $\|\mathbf{v}^* - \mathbf{v}\|_2 < \delta$ and t, t^* be two positive numbers in $\mathbb{R}^d$ satisfying $0 < t - t^* < \eta$. Assume that D passes the tests in Algorithm 2 with tolerance η along the direction $\mathbf{v}$. Denote $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + t)$ and $h^*(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t^*)$. Then it holds $\mathbf{Pr}_{\mathbf{x} \sim D_\mathbf{x}}[h(\mathbf{x}) \neq h^*(\mathbf{x})] \leq O(\delta + \eta)$.*

We remark that though the lemma above suffices for the purpose of learning general halfspaces up to error $\widetilde{O}\left(\sqrt{\mathrm{opt}}\right)$, it is not as efficient as its counterpart for homogeneous halfspaces. In particular, for a general halfspace with threshold t, if the distribution is indeed Gaussian one should expect some $\Phi(t)$ dependence in the equation, but there is no such dependency in the lemma above. This is one of the main bottlenecks of our approach in achieving error $O(\mathrm{opt})$ instead of $\widetilde{O}(\sqrt{\mathrm{opt}})$.

2.2. Finding a Good Localization Center

In this section, we present an efficient algorithm for finding a good localization center given that the distribution satisfies certain certifiable properties. Our main algorithmic ingredient returns a small list of points, at least one of which is a good localization center.

Proposition 9 *Let $\epsilon \in (0,1)$, D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$, and $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t^*)$ be a halfspace with 0-1 error at most opt with respect to D. Define the parameter $B := \min\left(\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y = +1], \mathbf{Pr}_{(\mathbf{x},y)\sim D}[y = -1]\right)$. Then Algorithm 1 draws $\mathrm{poly}(d/\epsilon)$ i.i.d. samples from D, and either reports that the distribution is not Gaussian or returns a list L containing at most $2\log(1/B)/\epsilon^2$ many points. Moreover, the following hold with high constant probability:*

1. *If the algorithm reports anything, the report is correct.*
2. *If we have that $\max(\sqrt{\epsilon}, \sqrt{\mathrm{opt}}\log^2(1/\mathrm{opt})) < B$, then there exists $\mathbf{w} \in L$ such that $\mathbf{w}$ is an $(\epsilon^2, C\,B/\log(1/B))$-good localization center, where $C > 0$ is a universal constant.*

Note that in Proposition 9 we assume that the mass of the points with the "minority" label is at least $B \gtrsim \max\left(\sqrt{\mathrm{opt}}\log^2(1/\mathrm{opt}), \sqrt{\epsilon}\right)$ in the completeness case, i.e., the case when the algorithm

does not reject. This would be an issue (and thereby a bottleneck) if the goal were to achieve a learning error of $O(\mathrm{opt})$. However, we claim that this is a minor assumption when our target error is $\widetilde{O}(\sqrt{\mathrm{opt}})$. Indeed, if the assumption on B does not hold, then there exists a constant halfspace ($h(\mathbf{x}) \equiv 1$ or $h(\mathbf{x}) \equiv -1$) so that $\mathbf{Pr}[h(\mathbf{x}) \neq y] \leq B = \widetilde{O}(\sqrt{\mathrm{opt}})$. Since our algorithm in the end returns the halfspace with the minimum error over testing samples among all the candidate halfspaces found, we can easily meet the error guarantee if we include the constant halfspaces in our hypothesis list.

As discussed in Section 1.2, when the halfspace has a non-trivial threshold, in order to obtain directional information about its defining vector, we need to focus on the *tail points*.

Definition 10 (Tail point) *Let $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t^*)$ be a halfspace. If $t^* > 0$ (resp. $t^* < 0$) we say that all the points with label -1 (resp. $+1$) after corruption are the tail points.*

Note that we can assume without loss of generality that the identity of the tail points is known to us. This is because we can run the same procedure twice — the first time treating the points with label $+1$ as the tail points and the second time treating the points with label -1 as the tail points, and in the end combine the candidate localization centers obtained.

The algorithm starts by computing the mean vector of the tail points, which we denote by $\boldsymbol{\mu}$. Then, the algorithm (1) tests that the first and second moments of the empirical distribution are close to those of $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and (2) tests that the marginal distribution projected along the direction of the mean vector $\boldsymbol{\mu}$ is sufficiently close to the Gaussian distribution. Conditioned on the event that the tests pass, the algorithm outputs a list of points that cover the segment from $\mathbf{0}$ to $O(\sqrt{\log(1/B)})\,\boldsymbol{\mu}$. The detailed pseudocode is given in Algorithm 1.

We now proceed to show the list of points returned by the algorithm with high constant probability contains a good localization center. Specifically, we argue that the intersection between the line along the mean vector $\boldsymbol{\mu}$ and the separating hyperplane of the optimal halfspace h [3] is a $(0, \Omega(B/\log(1/B)))$-good localization center, where B is the mass of the points with the minority label. Since the list is guaranteed to contain some point close to the intersection, it follows that it contains at least one $(\epsilon, \Omega(B/\log(1/B)))$-good localization center. To show that the intersection point is a good localization center, we first argue it cannot be much further from the origin than the mean vector $\boldsymbol{\mu}$. Second, we argue that the mean vector $\boldsymbol{\mu}$ itself cannot be too far from the origin. We now give the formal statement of the first property.

Lemma 11 (Distance Between Intersection And Mean) *Let $B \in (0,1)$, D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$, and $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t^*)$ be a halfspace that achieves the optimal error $\mathrm{opt} \in (0,1)$ with respect to D. Moreover, assume that $t^* > 10$ and $B \gtrsim \sqrt{\mathrm{opt}}\log^2(1/\mathrm{opt})$. Let S be a set of $N = \mathrm{poly}(d)/B^2$ i.i.d. samples drawn from D, $\boldsymbol{\mu}$ be the mean vector of the tail points among the samples with respect to h (see Definition 10), and $\mathbf{v}$ be the point of intersection between the separating hyperplane of h and the line along $\boldsymbol{\mu}$. Assume that the empirical distribution over S has its covariance matrix bounded from above by $2\mathbf{I}$, and the fraction of tail points among the samples is at least B. Then it holds that $\|\mathbf{v}\|_2 \leq \|\boldsymbol{\mu}\|_2 + O(1/\sqrt{\log(1/B)})$.*

We now give some high-level ideas of the proof. When there are no corruptions, we have that all the tail points are on a different side of the hyperplane than the origin. As illustrated in Figure 1, it is

3. We say that $\mathbf{x}$ is the point of intersection of the separating hyperplane of h and the line along $\mathbf{u}$ if $\mathbf{v} \cdot \mathbf{x} + t = 0$ and $\mathbf{x} = \lambda \mathbf{u}$ for some $\lambda \in \mathbb{R}$. When the halfspace has non-zero threshold, there will be exactly 1 intersection point when the line is not orthogonal to the defining vector of the halfspace, and 0 intersection point otherwise.

easy to see that the line from the origin along $\boldsymbol{\mu}$ must first intersect with the separating hyperplane of the halfspace, and then pass through the mean vector. This immediately gives us $\|\mathbf{v}\|_2 \leq \|\boldsymbol{\mu}\|_2$, where $\mathbf{v}$ is the intersection point. To deal with the noise, we will use the fact that Algorithm 1 certifies that (i) the mass of the outliers is only a small fraction of the mass of the tail points, and (ii) the sample covariance matrix has spectral norm bounded from above by some constant. The above turns out to be sufficient to certify that the outliers cannot move the mean by more than $O(1/\log(1/B))$ in ℓ_2 distance. Lemma 11 then follows via a careful geometric argument. The detailed proof can be found in Appendix C.

Next we argue that the distance from the mean vector $\boldsymbol{\mu}$ of the tail point to the origin can be bounded from above such that $\Phi(\|\boldsymbol{\mu}\|_2) \geq \Omega(B/\log(1/B))$ conditioned on that the tests in Algorithm 1 regarding the empirical distribution projected along the direction of $\boldsymbol{\mu}$ pass. Intuitively, this is due to the fact that a decent fraction of the tail points have to lie on the further side of $\boldsymbol{\mu}$, i.e., the set $\{\mathbf{x} : \mathbf{x} \cdot \boldsymbol{\mu} \geq \|\boldsymbol{\mu}\|_2^2\}$. If this were not the case, in order to balance the contributions from the points on the other side of the mean, the tail points would have to locate on the extreme end of the Gaussian tail, which would result in a violation to the CDF test along the direction of $\boldsymbol{\mu}$. The formal proof can be found in Appendix C.

Lemma 12 *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$, and $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t^*)$ be a halfspace that achieves the optimal error* opt *with respect to D, where $t^* > 10$. Let S be a set of $N = \mathrm{poly}(d)/\epsilon^2$ i.i.d. samples from D, $\boldsymbol{\mu}$ be the mean vector of the tail points among the samples with respect to h (see Definition 10). Suppose Lines 6b, and 6c from Algorithm 1 pass. Then it holds that $\Phi(\|\boldsymbol{\mu}\|_2) \geq \Omega(B/\log(1/B))$.*

One may have noticed that Lemma 11 and Lemma 12 require $t^* > 10$. When the assumption does not hold, we claim that there is a straightforward way to find a good localization center. In particular, we can now just testably learn the Chow parameters of the halfspace. Since now the Chow parameters are of size $G(t) = \Theta(1)$, it suffices that we can testably learn it up to constant accuracy, which can be readily achieved by Lemma 2.3 from Diakonikolas et al. (2023a). The formal statement and its proof can be found in Lemma 19 in Appendix C.

We are now ready to conclude the proof of Proposition 9.

Proof [Proof of Proposition 9] We defer the proof of the soundness of the algorithm to Lemmas 18 and 19. We now argue that the list returned by the algorithm will contain a good localization center with high constant probability. First, assume that $t^* \leq 10$. The existence of a good localization center follows from Lemma 19. Next we consider the case $t^* > 10$. Let $\boldsymbol{\mu}$ be the mean vector of the tail points, and $\mathbf{v}$ be the intersection of the line along $\boldsymbol{\mu}$ and the separating hyperplane of the optimal halfspace h. The argument of $\Phi(\|\mathbf{v}\|_2) \geq \Omega(B/\log(1/B))$ relies on the following two claims, which follow from Lemma 12 and Lemma 11 respectively: (a) $\Phi(\|\boldsymbol{\mu}_+\|_2) \geq \Omega(B/\log(1/B))$, and (b) $\|\mathbf{v}\|_2 \leq \|\boldsymbol{\mu}_+\|_2 + O(1/\sqrt{\log(1/B)})$. Assuming the above claims, we immediately have that

$$\Phi(\|\mathbf{v}\|_2) \geq \Phi\left(\|\boldsymbol{\mu}_+\|_2 + O\left(1/\sqrt{\log(1/B)}\right)\right) \geq \Omega(1)\,\Phi(\|\boldsymbol{\mu}_+\|_2) \geq \Omega(B/\log(1/B))\,,$$

where in the first inequality we use (b), in the second inequality we use Claim 20 and the fact that $\|\boldsymbol{\mu}_+\|_2 \leq O\left(\sqrt{\log(1/B)}\right)$, which is implied by (a), and in the last inequality we again use (a). This shows that $\mathbf{v}$ is a $(0, \Omega(B/\sqrt{\log(1/B)}))$-good localization center. It is easy to see that the output list contains some point that is ϵ^2-close to $\mathbf{v}$, and closer to the origin than $\mathbf{v}$. Hence, it follows

that the list contains some $\left(\epsilon^2, \Omega(B/\sqrt{\log(1/B)})\right)$-good localization center. This concludes the proof of Proposition 9. ■

Input: Sample access to a distribution D over $\mathbb{R}^d \times \{\pm 1\}$; tolerance parameter ϵ.
Output: Reject or a list of vectors containing a good localization center.

1. Set $N = \mathrm{poly}(d/\epsilon)$. Draw N i.i.d. samples, and denote the set of samples as S.
2. Return an empty list if either the fraction of points labeled with $+1$ or -1 is less than $\epsilon/2$.
3. Verify that $\mathbf{E}_{(\mathbf{x},y)\sim S}\left[\mathbf{x}\mathbf{x}^T\right] \preccurlyeq 2\mathbf{I}.$, and that $\left\|\mathbf{E}_{(\mathbf{x},y)\sim S}\left[\mathbf{x}\right]\right\|_2 < \epsilon$.
4. Compute $\boldsymbol{\mu}_+ = \mathbf{E}_{(\mathbf{x},y)\sim S}[\mathbf{x}|y=+1]$, $\boldsymbol{\mu}_- = \mathbf{E}_{(\mathbf{x},y)\sim S}[\mathbf{x}|y=-1]$.
5. Initialize an empty list L.
6. For $\boldsymbol{\mu} \in \{\boldsymbol{\mu}_+, \boldsymbol{\mu}_-\}$, do the following:
   (a) For each $(\mathbf{x}, y) \in S$, project it along the direction of $\boldsymbol{\mu}$ to obtain the new pair $(x', y) \in \mathbb{R} \times \{\pm 1\}$ where $x' = \boldsymbol{\mu} \cdot \mathbf{x}/\left\|\boldsymbol{\mu}\right\|_2$. Denote the resulting set as H.
   (b) Verify that the empirical distribution over H and $\mathcal{N}(0,1)$ are ϵ-close in CDF distance.
   (c) Verify that removing at most ϵ-fraction of points from H changes the mean by at most $O(\epsilon\sqrt{\log(1/\epsilon)})$.
   (d) Add $\left\{\mathbf{v}^{(i)} := i\epsilon^2\boldsymbol{\mu}/\left\|\boldsymbol{\mu}\right\|_2\right\}_i$, where $i \in \{0\} \cup \lceil \frac{\|\boldsymbol{\mu}\|_2 + 1/\log(1/B)}{\epsilon^2/\log(1/B)} \rceil$, to the list L.
7. Run the algorithm from Lemma 19, and add the localization centers found into L. Return L.

Algorithm 1: Find-Localization-Center

2.3. Putting things together

We describe our algorithm and its analysis at a high level in this section. We first invoke the routine Find-Localization-Center from Proposition 9 to obtain a list of candidate centers, which is guaranteed to contain some $\left(\epsilon^2, \widetilde{\Omega}(B)\right)$-localization center if the routine does not reject. For each candidate center $\mathbf{w}$, we use $(\mathbf{w}, \sigma := \min(\|\mathbf{w}\|_2^{-1}, \sqrt{2}))$-rejection sampling (Definition 4) to re-center the marginal distribution at $\mathbf{w}$, and then transform the marginal back into a standard Gaussian. If $\mathbf{w}$ is indeed a good localization center, by Lemma 5, the halfspace becomes $h'(\mathbf{x}) = \mathrm{sign}\left(\boldsymbol{\Sigma}^{1/2}\mathbf{v}^* \cdot \mathbf{x}/\|\boldsymbol{\Sigma}^{1/2}\mathbf{v}^*\|_2 + t'\right)$ after the transformation, where $\boldsymbol{\Sigma} = \mathbf{I} - (1-\sigma^2)\mathbf{w}\mathbf{w}^\top/\left\|\mathbf{w}\right\|_2^2$, and t' is some threshold with size at most ϵ. Moreover, the fraction of points that survive the rejection sampling is at least $\widetilde{\Omega}(B)$, which ensures that the fraction of outliers among the surviving points is at most $\widetilde{O}(\mathrm{opt}/B) = \widetilde{O}(\sqrt{\mathrm{opt}})$ under the assumption $B \gtrsim \sqrt{\mathrm{opt}}$ (otherwise, we can always output a constant halfspace in the end). We then run Nearly-Homogeneous-Halfspace-Testable-Learner, which gives us an $\widetilde{O}(\mathrm{opt}/B)$-approximation to $\boldsymbol{\Sigma}^{1/2}\mathbf{v}^*/\|\boldsymbol{\Sigma}^{1/2}\mathbf{v}^*\|_2$. By Lemma 7, revert-

ing the transformation $\mathbf{\Sigma}^{1/2}$ gives a vector $\hat{\mathbf{v}}$, which will be an $\widetilde{O}(\mathrm{opt}/B)$-approximation to $\mathbf{v}^*$ [4]. Since we search for the threshold in a brute-force manner, it is guaranteed that we will have some halfspace $\hat{h}$ whose parameters are all $\widetilde{O}(\sqrt{\mathrm{opt}})$-close to the optimal halfspace in our final hypothesis list. Conditioned on that the Wedge-Bound algorithm passes, Lemma 8 then guarantees that $\hat{h}$ will have learning error at most $\widetilde{O}(\sqrt{\mathrm{opt}})$. In the end, we simply draw some fresh samples, and output the halfspace with the smallest testing error. The detailed proof and pseudocode can be found in Appendix D.

3. Additional Remarks Regarding Theorem 2

We provide some additional remarks regarding our main theorem, addressing some of its limitations.

Remark 13 It is natural to ask whether our algorithmic result can be generalized to hold for other structured distributions, e.g., for a large subclass of isotropic log-concave distributions. Such a generalization turns out to be possible for *homogeneous halfspaces*; see Gollakota et al. (2023c). It is important to note, however, that obtaining polynomial-time agnostic learners for general halfspaces under non-Gaussian distributions is a challenging task — even without the testable requirement. In particular, the only known efficient agnostic learners for general halfspaces that achieve dimension-independent error are the ones from Diakonikolas et al. (2018, 2022b), which both work only under the Gaussian distribution.

Remark 14 Theorem 2 achieves error of $\widetilde{O}(\sqrt{\mathrm{opt}}) + \epsilon$. On the other hand, error of $O(\mathrm{opt}) + \epsilon$ can be achieved in the same setting for homogeneous halfspaces. We discuss here a technical barrier of our approach preventing further improvements. A critical ingredient in our algorithm is a procedure which robustly estimates the mean of samples sharing the same label (in a testable way), in order to extract directional information about the weight vector of the optimal halfspace. Our current method, relying on certifying bounded second moments of the samples, ceases to work when the amount of corruption approaches $\sqrt{B}$, where B is the probability mass of the points having that label. It is plausible that certifying boundedness of higher moments is an avenue for progress here.

4. We also need to show that the angle between $\mathbf{w}$ and $\mathbf{v}^*$ is not too large. The details of this argument can be found in Appendix D.

Acknowledgments

Ilias Diakonikolas acknowledges support under NSF Medium Award CCF-2107079 and a DARPA Learning with Less Labels (LwLL) grant. Daniel M. Kane acknowledges support under NSF Award CCF-1553288 (CAREER) and NSF Medium Award CCF-2107079. Sihan Liu acknowledges support under NSF Award CCF-1553288 (CAREER) and NSF Medium Award CCF-2107079. Nikos Zarifis acknowledges support under NSF Medium Award CCF-2107079, and a DARPA Learning with Less Labels (LwLL) grant.

References

P. Awasthi, M. F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. *J. ACM*, 63(6):50:1–50:27, 2017.

A. Daniely. A PTAS for agnostically learning halfspaces. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pages 484–502, 2015.

A. Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual Symposium on Theory of Computing, STOC 2016*, pages 105–117, 2016.

I. Diakonikolas and D. M. Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.

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

I. Diakonikolas, D. M. Kane, and N. Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under Gaussian marginals. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020a.

I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Non-convex SGD learns halfspaces with adversarial label noise. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020b.

I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Agnostic proper learning of halfspaces under gaussian marginals. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021a.

I. Diakonikolas, D. M. Kane, T. Pittas, and N. Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the SQ model. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021b.

I. Diakonikolas, D. M. Kane, P. Manurangsi, and L. Ren. Cryptographic hardness of learning halfspaces with massart noise. In *Advances in Neural Information Processing Systems*, 2022a.

I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning general halfspaces with adversarial label noise via online gradient descent. In *International Conference on Machine Learning*, pages 5118–5141. PMLR, 2022b.

I. Diakonikolas, D. M Kane, V. Kontonis, S. Liu, and N. Zarifis. Efficient testable learning of halfspaces with adversarial label noise. In *Advances in Neural Information Processing Systems*, 2023a.

I. Diakonikolas, D. M. Kane, and L. Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals. In *ICML*, 2023b.

S. Goel, A. Gollakota, and A. R. Klivans. Statistical-query lower bounds via functional gradients. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

A. Gollakota, A. Klivans, and P. Kothari. A moment-matching approach to testable learning and a new characterization of rademacher complexity. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1657–1670, 2023a.

A. Gollakota, A. R. Klivans, K. Stavropoulos, and A. Vasilyan. An efficient tester-learner for halfspaces. *arXiv*, 2023b. Conference version to appear in ICLR'24.

A. Gollakota, A. R Klivans, K. Stavropoulos, and A. Vasilyan. Tester-learners for halfspaces: Universal algorithms. In *Advances in Neural Information Processing Systems*, 2023c.

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

A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008. Special issue for FOCS 2005.

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

A. Klivans, P. Long, and R. Servedio. Learning Halfspaces with Malicious Noise. *Journal of Machine Learning Research*, 10:2715–2740, 2009.

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

R. Rubinfeld and A. Vasilyan. Testing distributional assumptions of learning algorithms. In *STOC*, 2023.

S. Tiegel. Hardness of agnostically learning halfspaces from worst-case lattice problems. In *COLT*, 2023.

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

Appendices

Appendix A. Testable Learner for Nearly Homogeneous Halfspaces

We restate and show the following:

Proposition 15 *Let $\epsilon \in (0,1)$. Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ and h be a halfspace that achieves error opt under D with $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t)$ where $|t| < \epsilon$. Then, given i.i.d. sample access to D, there exists an efficient tester-learner which either reports $D_{\mathbf{x}}$ is not Gaussian or outputs a vector $\mathbf{v}$ satisfying $\|\mathbf{v} - \mathbf{v}^*\|_2 \leq O(\mathrm{opt}) + \epsilon$.*

Proof The algorithm of Proposition 6 is identical to the algorithm of Diakonikolas et al. (2023a), but we need to argue that the small offset does not make the algorithm to fail. If the distribution was Gaussian, then trivially the $\mathbf{Pr}[\mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t) \neq \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x})] = O(\epsilon)$, therefore, we could just assume that there exists a homogeneous halfspace with error $\mathrm{opt} + O(\epsilon)$. Unfortunately, in our setting we cannot do that. The reason is that, we do not have any guarantee that all the bands of the form $\{|\mathbf{u} \cdot \mathbf{x}| \leq \epsilon\}$ have mass of order ϵ for all the vectors $\mathbf{u}$.

The whole approach of Diakonikolas et al. (2023a) boils down into calculating the $\mathbf{E}[y\mathbf{x}]$ up to a good accuracy. This is done in Proposition 2.1 of this work. We need to show that if the optimal halfspace is of the form $\mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t)$ with $|t| \leq 10$, then we can robustly estimate $\mathbf{E}[y\mathbf{x}]$. We first need to show that small $|t|$ does not change the $\mathbf{E}[y\mathbf{x}]$ by a lot. We use the following fact

Fact 16 (see, e.g., Lemma 4.3 of Diakonikolas et al. (2018)) *Let $\mathbf{v}$ be a unit vector and $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + t)$ be the corresponding halfspace. If $\mathbf{x}$ is drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, then we have that $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[h(\mathbf{x})\mathbf{x}] = 2G(t)\,\mathbf{v}$, where $G(\cdot)$ is the pdf of the standard Gaussian.*

Note that from Fact 16, we can see that if $|t'| \leq 10$ then $G(t') \geq C$, for some sufficiently small absolute constant $C > 0$. Therefore, the proof of Proposition 2.1 remains the same as it only changes by a constant factor the calculation of $\mathbf{E}[y\mathbf{x}]$.

Furthermore, note that in each iteration, the algorithm of Diakonikolas et al. (2023a) changes the covariance matrix as $\mathbf{\Sigma} = \mathbf{I} - (1 - \sigma^2)\mathbf{w}\mathbf{w}^\top$ for some $\sigma \geq \Omega(\epsilon)$ via rejection sampling. This means that norm of the vector $\mathbf{v}^*$ after the application of of the operator $\mathbf{\Sigma}^{1/2}$ will be at least σ, hence after normalization we have that $\mathrm{sign}(\mathbf{\Sigma}^{1/2}\mathbf{v}^* \cdot \mathbf{x} + t) = \mathrm{sign}((\mathbf{\Sigma}^{1/2}\mathbf{v}^* \cdot \mathbf{x} + t)/\|\mathbf{\Sigma}^{1/2}\mathbf{v}^*\|_2)$, which means that $|t|$ will be increased to at most $|t|/\sigma$. Moreover, note that $|t| \leq \epsilon$ and $\sigma = \Omega(\epsilon)$, hence $|t'| = |t|/\sigma \leq O(1)$ and we can choose the constants so that $|t'| \leq 10$. Hence, the error $\|\mathbf{v} - \mathbf{v}^*\|_2$ steadily shrinks by a constant factor in each iteration until we have $\|\mathbf{v} - \mathbf{v}^*\|_2 = \Theta(\max(\epsilon, \mathrm{opt}))$. This completes the proof. ■

Appendix B. Omitted Proofs for General to Near-Homogeneous Reduction

B.1. Proof of Properties of Good Localization Center (Lemma 5)

Proof By the definition of an (α, β)-good localization center, $\mathbf{w}$ satisfies $\Phi(\|\mathbf{w}\|_2) \geq \beta$. Note that $\Phi(\|\mathbf{w}\|_2)$ can be bounded from above by $\frac{1}{\sqrt{2\pi}} \frac{1}{\|\mathbf{w}\|_2} \exp(-\|\mathbf{w}\|_2^2/2)$. It then follows that

$$\exp(-\|\mathbf{w}\|_2^2/2) \geq C \ \|\mathbf{w}\|_2 \, \beta.$$

From here one can conclude that $\|\mathbf{w}\|_2 \leq O\left(\sqrt{\log(1/\beta)}\right)$.

The form of the rejection sampling distribution conditioned on acceptance follows from Diakonikolas et al. (2018). To analyze the acceptance probability, we first assume $\|\mathbf{w}\|_2 > \sqrt{2}$. If we perform $(\mathbf{w}, 1/\|\mathbf{w}\|_2)$-rejection sampling on the standard Gaussian distribution, the acceptance probability will be

$$\begin{aligned}\|\mathbf{w}\|_2^{-1} \exp\left(-\frac{\|\mathbf{w}\|_2^2}{2\,(1-\|\mathbf{w}\|_2^{-2})}\right) &= \|\mathbf{w}\|_2^{-1} \exp\left(-\frac{\|\mathbf{w}\|_2^2}{2}\left(1+\frac{\|\mathbf{w}\|_2^{-2}}{(1-\|\mathbf{w}\|_2^{-2})}\right)\right)\\ &= \|\mathbf{w}\|_2^{-1} \exp\left(-\frac{\|\mathbf{w}\|_2^2}{2}\right)\exp\left(-\frac{1}{2(1-\|\mathbf{w}\|_2^{-2})}\right).\end{aligned}$$

Recall that in this case we assume $\|\mathbf{w}\|_2$ is within the range $[\sqrt{2},\infty)$. Hence, $\|\mathbf{w}\|_2^{-2}$ is within the range $(0,1/2]$. As a result, $\exp\left(-\frac{1}{2(1-\|\mathbf{w}\|_2^{-2})}\right)$ is bounded from above and below by constants. Hence, we can bound from below the overall acceptance probability by

$$\frac{1}{\|\mathbf{w}\|_2}\exp\left(-\|\mathbf{w}\|_2^2/\left(2\,(1-\|\mathbf{w}\|_2^{-2})\right)\right) \geq C\,\Phi(\|\mathbf{w}\|) \geq C\beta.$$

Now, suppose $\|\mathbf{w}\|_2 < \sqrt{2}$. We now perform rejection sampling with parameter $\sigma = \sqrt{1/2}$. The acceptance probability is then $2\,\exp\left(-\|\mathbf{w}\|_2^2\right) \geq \Omega(1)$. This completes the argument of Property (1).

After rejection sampling, the standard Gaussian becomes $\mathcal{N}(\mathbf{w},\mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{I} - (1-\sigma^2)\mathbf{w}\mathbf{w}^T/\|\mathbf{w}\|_2^2$ for $\sigma = \min(\sqrt{1/2}, \|\mathbf{w}\|_2^{-1}) > \Omega\left(1/\sqrt{\log(1/\beta)}\right)$. Recall the original halfspace is given by $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t)$. After we transform the space to make $\mathcal{N}(\mathbf{w},\mathbf{\Sigma})$ isotropic, the halfspace then becomes

$$h'(\mathbf{x}) = \mathrm{sign}(\mathbf{\Sigma}^{1/2}\mathbf{v}^*\cdot\mathbf{x} + \mathbf{v}^*\cdot\mathbf{w} + t^*).$$

Now we analyze the ℓ_2 norm of the new defining vector and offset separately. For the new defining vector, we decompose $\mathbf{v}^*$ into its component in $\mathbf{w}$ and the orthogonal part, i.e. $\mathbf{v}^* = a\mathbf{w}/\|\mathbf{w}\|_2 + b\mathbf{u}$ where $a^2+b^2=1$. Then, it is not hard to see that $\mathbf{\Sigma}^{1/2}\mathbf{v} = a\mathbf{w}/\|\mathbf{w}\|_2 + \sigma\ b\mathbf{u}$. Hence, the ℓ_2 norm is at least $\max(a,\sigma b) \geq \sigma/\sqrt{2} > \Omega(1/\log(1/\beta))$. For the threshold, we note that by the definition of the localization center, $\mathbf{w}$ is at most α far from the halfspace h. Therefore, it holds $|\mathbf{v}^*\cdot\mathbf{w}+t^*| \leq \alpha$. Combining our analysis then gives the new halfspace h' is at most $O\left(\alpha\sigma^{-1}\right) = O\left(\alpha\sqrt{\log(1/\beta)}\right)$ far from the origin. This concludes the proof of Property (2), and also Lemma 5. ■

B.2. Proof of Transformation Error Bound (Lemma 7)

Proof For convenience, we define $\lambda = \left\|\mathbf{\Sigma}^{1/2}\mathbf{v}^*\right\|_2$. We can decompose $\mathbf{v}^*$ as $\mathbf{v}^* = a\ \mathbf{w} + b\ \mathbf{y}$ for some unit vector $\mathbf{y}$ orthogonal to $\mathbf{w}$ and positive coefficients a, b satisfying $a^2+b^2=1$. Notice that we have $(1-a)^2 + b^2 = \|\mathbf{w}-\mathbf{v}^*\|_2 \leq \beta^2$. This implies that b is bounded from above by β. Consequently, since $a^2+b^2=1$, we have

$$a \geq 1-\beta^2/2. \tag{1}$$

Note that the operator $\mathbf{\Sigma}^{1/2}$ scales any vector along the direction of $\mathbf{w}$ by a factor of σ while leaving any vector orthogonal to $\mathbf{w}$ unchanged. Therefore, we have $\mathbf{\Sigma}^{1/2}\mathbf{v}^*$, which further implies that $\lambda = \left\|\mathbf{\Sigma}^{1/2}\mathbf{v}^*\right\|_2 \leq a\sigma + b$.

Since the ℓ_2 distance between $\mathbf{v}$ and $\lambda^{-1}\mathbf{\Sigma}^{1/2}\mathbf{v}^*$ is at most δ, we can write

$$\mathbf{v} = \lambda^{-1}\mathbf{\Sigma}^{1/2}\mathbf{v}^* + c\,\mathbf{w} + d\,\mathbf{z}.$$

for some unit vector $\mathbf{z}$ orthogonal to $\mathbf{w}$ and positive coefficients $c, d < \delta$.
Multiplying both sides by $\mathbf{\Sigma}^{-1/2}$ then gives

$$\mathbf{\Sigma}^{-1/2}\mathbf{v} = \lambda^{-1}\mathbf{v}^* + c\,\sigma^{-1}\,\mathbf{w} + d\,\mathbf{z}.$$

We can substitute the decomposition of $\mathbf{v}^*$ into the equation. After some algebra, we get that

$$\begin{aligned}\mathbf{\Sigma}^{-1/2}\mathbf{v} &= \left(\lambda^{-1} + c\,\sigma^{-1}\,a^{-1}\right)\,a\,\mathbf{w} + \left(\lambda^{-1} + c\,\sigma^{-1}\,a^{-1}\right)\,b\,\mathbf{y} - c\,\sigma^{-1}\,a^{-1}b\mathbf{y} + d\,\mathbf{z}.\\ &= \left(\lambda^{-1} + c\,\sigma^{-1}\,a^{-1}\right)\mathbf{v}^* - c\,\sigma^{-1}\,a^{-1}b\,\mathbf{y} + d\,\mathbf{z}.\end{aligned}$$

Now denote $\xi := \lambda^{-1} + c\,\sigma^{-1}a^{-1}$. Note that since $\lambda < a\,\sigma + \beta$, we have $\xi^{-1} < a\,\sigma + \beta$ as well. We can then divide both sides by ξ and merge the terms involving $\mathbf{y}, \mathbf{z}$, which gives us

$$\xi^{-1}\mathbf{\Sigma}^{-1/2}\mathbf{v} = \mathbf{v}^* + \rho\mathbf{u}\,,$$

where $\mathbf{u}$ is a unit vector and ρ is a positive number bounded by

$$\xi^{-1}\left(c\sigma^{-1}a^{-1}b + d\right) \leq (a\sigma+\beta)\left(c\sigma^{-1}a^{-1}b + d\right) \leq (a\sigma+\beta)\,\left(\delta\sigma^{-1}\frac{\beta}{1-\beta^2/2} + \delta\right).$$

Note that the right hand side is exactly the bound we want for $\left\|\mathbf{\Sigma}^{1/2}\mathbf{v}/\left\|\mathbf{\Sigma}^{1/2}\mathbf{v}\right\|_2 - \mathbf{v}^*\right\|_2$ up to some constant. If the right hand side is $\Omega(1)$, the conclusion is trivial since the distance between two unit vectors is always upper bounded by a constant. Otherwise, since $\xi^{-1}\mathbf{\Sigma}^{-1/2}\mathbf{v}$ is ρ-close to $\mathbf{v}^*$ in ℓ_2 distance and $\rho = o(1)$, it follows that the vector, after being normalized to have norm 1, is still $O(\rho)$ closed to $\mathbf{v}^*$ in ℓ_2 distance. Our result then follows from the observation that

$$\xi^{-1}\mathbf{\Sigma}^{-1/2}\mathbf{v}/\left\|\xi^{-1}\mathbf{\Sigma}^{-1/2}\mathbf{v}\right\|_2 = \mathbf{\Sigma}^{-1/2}\mathbf{v}/\left\|\mathbf{\Sigma}^{-1/2}\mathbf{v}\right\|_2.$$

This concludes the proof of Lemma 7. ■

B.3. Proof of General Wedge Bound (Lemma 8)

We provide the pseudocode of the Wedge-Bound algorithm below for completeness.

Input: Sample access to a distribution $D_{\mathbf{x}}$ over $\mathbb{R}^d$; tolerance parameter $\eta > 0$; unit vector $\mathbf{v} \in \mathbb{R}^d$; failure probability $\tau \in (0,1)$.
Output: Certifies the (conditional) moments of D approximately match with $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

1. Set $B = \lceil \sqrt{\log(1/\eta)}/\eta \rceil$.
2. Let $\widetilde{D}$ be the empirical distribution obtained by drawing $\mathrm{poly}(d, 1/\eta)\log(1/\tau)$ samples from $D_{\mathbf{x}}$.
3. For integers $-B-1 \le i \le B$, define E_i to be the event that $\{\mathbf{v}\cdot\mathbf{x} \in [i\eta, (i+1)\eta]\}$ and E_{B+1} to be the event that $\{|\mathbf{v}\cdot\mathbf{x}| \ge \sqrt{\log(1/\eta)}\}$.
4. Verify that $\sum_{i=-B-1}^{B+1} \left|\mathbf{Pr}_{\mathcal{N}(\mathbf{0},\mathbf{I})}[E_i] - \mathbf{Pr}_{\widetilde{D}}[E_i]\right| \le \eta$.
5. Let S_i the distribution of $\widetilde{D}$ conditioned on E_i and $S_i^{\perp}$ be S_i projected on the subspace orthogonal to $\mathbf{v}$.
6. For each i, verify that $S_i^{\perp}$ has bounded covariance, i.e., check that $\mathbf{E}_{\mathbf{x}\sim S_i^{\perp}}[\mathbf{x}\mathbf{x}^{\top}] \preccurlyeq 2\mathbf{I}$.

Algorithm 2: Wedge-Bound

Proof [Proof of Lemma 8]

Claim 17 *Suppose D is a distribution passing the Wedge Bound Test with tolerance η along the direction $\mathbf{v}$. Let a, b be two positive number satisfying $a^2+b^2=1$ and $b<\delta$. Let $\mathbf{u}$ be a unit vector orthogonal to $\mathbf{v}$. Then it holds*

$$\mathbf{Pr}[-b\mathbf{u}\cdot\mathbf{x} > a\mathbf{v}\cdot\mathbf{x}+t, a\mathbf{v}\cdot\mathbf{x}+t>0]\,,\ \mathbf{Pr}[b\mathbf{u}\cdot\mathbf{x} > -a\mathbf{v}\cdot\mathbf{x}-t, -a\mathbf{v}\cdot\mathbf{x}-t>0] \le O(\delta+\eta).$$

Proof For convenience, we define $x := -\mathbf{u}\cdot\mathbf{x}$, and $y := \mathbf{v}\cdot\mathbf{x}$. We have

$$\begin{aligned}
&\mathbf{Pr}[-bx > ay+t, ay+t>0] \\
&\le \mathbf{Pr}[0 < ay+t<\delta] + \sum_{i=1}^{\infty}\mathbf{Pr}[-bx > i\delta, i\delta < ay+t < (i+1)\delta] \\
&\le O(\delta+\eta) + O(\delta+\eta)\sum_{i=1}^{\infty}\frac{1}{i^2} \le O(\delta+\eta)\,,
\end{aligned}$$

where the first inequality uses the law of total probability, the second inequality follows from that we verify that the cumulative density function of y is η-close to that of the standard Gaussian in Kolmogorov distance, and that the distribution of x conditioned on $i\delta < ay+t < (i+1)\delta$ has its second moment bounded from above by some constant, the last inequality follows from that the series $\sum_{i=1}^{\infty}\frac{1}{i^2}$ is converging. The argument for bounding the term $\mathbf{Pr}[b\mathbf{u}\cdot\mathbf{x} > -a\mathbf{v}\cdot\mathbf{x}-t, -a\mathbf{v}\cdot\mathbf{x}-t>0]$ is symmetric, and we hence omit it here. This concludes the proof of Claim 17. ■

We first decompose $\mathbf{v}^*$ as $\mathbf{v}^* = a\,\mathbf{v}\cdot\mathbf{x} + b\,\mathbf{u}\cdot\mathbf{x}$ where $\mathbf{u}$ is a unit vector orthogonal to $\mathbf{v}$ and a, b are two positive coefficients satisfying $a^2+b^2=1$ and $b<\delta$. Essentially, there are two cases where

$h(\mathbf{x})$ and $h^*(\mathbf{x})$ could differ. In the first case, we have $a\ \mathbf{v}\cdot\mathbf{x} + b\ \mathbf{u}\cdot\mathbf{x} + t^* < 0$ and $\mathbf{v}\cdot\mathbf{x} + t > 0$. The first inequality can be rewritten as $-b\ \mathbf{u}\cdot\mathbf{x} > a\ \mathbf{v}\cdot\mathbf{x} + t^*$ Using Claim 17, we have

$$\Pr_{\mathbf{x}\sim D}\left[-b\ \mathbf{u}\cdot\mathbf{x} > a\ \mathbf{v}\cdot\mathbf{x} + t^*\,,\ a\ \mathbf{v}\cdot\mathbf{x} + t^* \geq 0\right] \leq O(\delta+\eta).$$

On the other hand, since we must also have $\mathbf{v}\cdot\mathbf{x} + t > 0$, the probability that $\mathbf{v}\cdot\mathbf{x} + t > 0$ and $a\ \mathbf{v}\cdot\mathbf{x} + t^* < 0$ are both true cannot be too large. In particular, if both of them are true, we must have $t^*/a < -\mathbf{v}\cdot\mathbf{x} < t$. Since $|t - t^*| < \eta$, and a is of order $1 \pm O(\delta)$. It follows this happens with probability at most $O(\eta+\delta)$. Altogether, we then have

$$\begin{aligned}
&\Pr_{\mathbf{x}\sim D}\left[a\ \mathbf{v}\cdot\mathbf{x} + b\ \mathbf{u}\cdot\mathbf{x} + t^* < 0\,,\ \mathbf{v}\cdot\mathbf{x} + t > 0\right] \\
&\leq \Pr_{\mathbf{x}\sim D}\left[-b\ \mathbf{u}\cdot\mathbf{x} > a\ \mathbf{v}\cdot\mathbf{x} + t^*\,,\ a\ \mathbf{v}\cdot\mathbf{x} + t^* \geq 0\right] + \Pr_{\mathbf{x}\sim D}\left[\mathbf{v}\cdot\mathbf{x} + t > 0\,,\ a\ \mathbf{v}\cdot\mathbf{x} + t^* < 0\right] \\
&\leq O(\delta+\eta).
\end{aligned}$$

In the second case, we have $a\ \mathbf{v}\cdot\mathbf{x} + b\ \mathbf{u}\cdot\mathbf{x} + t^* > 0$ and $\mathbf{v}\cdot\mathbf{x} + t < 0$. Similarly, we can rewrite the first inequality as $b\ \mathbf{u}\cdot\mathbf{x} > -a\ \mathbf{v}\cdot\mathbf{x} - t^*$. Now, since $-\mathbf{v}\cdot\mathbf{x} - t > 0$ and $t > t^*$, we must have $-a\ \mathbf{v}\cdot\mathbf{x} - t^* > 0$. Hence, we can just use Claim 17 to conclude that this happens with probability at most $O(\eta+\delta)$. Combining the bound on the probability mass in the two cases then concludes the proof of Lemma 8. ■

Appendix C. Omitted Proofs for Good Localization Center Search

C.1. Soundness of the Tests in Algorithm 1

In this subsection, we show that Algorithm 1 is sound, i.e., it rarely falsely rejects $D_{\mathbf{x}}$ when we have $D_{\mathbf{x}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Lemma 18 (Soundness of Tests) *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. Assume $D_{\mathbf{x}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then Algorithm 1 does not reject on Lines 3, 3, 6b and 6c with high constant probability.*

Proof We first show that if $D_{\mathbf{x}}$ is indeed $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the algorithm does not report violation of the distributional assumptions with high constant probability. Lines 3 and 3 verify that the first and second moments are close to the ones of the standard normal; it follows from the fact that if $D_{\mathbf{x}}$ is the standard Gaussian, then by drawing $N \gtrsim d^2/\epsilon^2$ i.i.d. samples from $D_{\mathbf{x}}$, the first and the second empirical moments should be close to those of $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with high constant probability. For Line 6b, we remark that if $D_{\mathbf{x}}$ is standard Gaussian, then H is empirical distribution of $\mathcal{N}(0, 1)$ and N i.i.d. samples from $D_{\mathbf{x}}$ are enough to test it with high constant probability. To argue the soundness of the test we argue that the uniform distribution over $\{\mathbf{v}\cdot\mathbf{x} \mid \mathbf{x} \in S\}$ is close to $\mathcal{N}(0, 1)$ in Kolmogorov distance for every $\mathbf{v}$. We note that the quantity can be expressed as

$$\sup_{\mathbf{v}\in\mathbb{R}^d, b\in\mathbb{R}} \left| \Pr_{\mathbf{x}\sim S}\left[\mathbf{v}\cdot\mathbf{x} \geq b\right] - \Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\mathbf{v}\cdot\mathbf{x} \geq b\right] \right|.$$

Hence, by the VC-inequality, with high constant probability, the above is bounded from above by ϵ when we take $N \gtrsim d/\epsilon^2$ many samples. Conditioned on the event that Line 6b does not reject, we verify that the test passes Line 6c with high constant probability. For Line 6c, we note that if one

takes $N \gtrsim d/\epsilon^2$ many samples from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the empirical distribution will satisfy the following stability property (see Diakonikolas and Kane (2023) e.g.)

$$\left| \frac{1}{|S'|} \sum_{\mathbf{x} \in S'} \boldsymbol{\mu}_+ \cdot \mathbf{x} \right| \leq O\left(\epsilon \sqrt{\log(1/\epsilon)}\right), \tag{2}$$

for every $S' \subseteq S$ with $|S'| \epsilon (1-\epsilon) |S|$. On the other hand, since Line 3 passes, we immediately have that $\mathbf{E}_{x \sim H}[x] \leq \epsilon$ and combining it with Equation (2), it follows that the algorithm does not reject with high constant probability. This concludes the proof of Lemma 18. ■

C.2. Proof of Lemma 11

Proof [Proof of Lemma 11] Suppose the optimal halfspace is given by $h(\mathbf{x}) = \operatorname{sign}(\mathbf{v}^* \cdot \mathbf{x} + t^*)$. We first argue that

$$\mathbf{v}^* \cdot \left(\mathbf{v} - \boldsymbol{\mu}_+\right) \leq O(1/\log(1/B)). \tag{3}$$

Since $\mathbf{v}$ is defined to be the intersection between the halfspace and some line passing the origin, we always have $\mathbf{v} \cdot \mathbf{v}^* = |t^*|$. Let $\bar{\boldsymbol{\mu}}_+$ be the empirical mean over the points with label $+1$ when there are no outliers, i.e., all points are labeled by $h(\mathbf{x})$. Then we have $\mathbf{v}^* \cdot \bar{\boldsymbol{\mu}}_+ \geq |t^*| = \mathbf{v} \cdot \mathbf{v}^*$. Thus, it suffices to show that

$$\mathbf{v}^* \cdot \left(\bar{\boldsymbol{\mu}}_+ - \boldsymbol{\mu}_+\right) \leq \frac{1}{\log(1/B)}. \tag{4}$$

Let $\bar{B}$ be the fraction of samples within S such that $h(\mathbf{x}) = +1$, and $\tilde{B}$ be the fraction of samples within S such that $y = +1$. We have $\left|\bar{B} - \tilde{B}\right| \leq \mathrm{opt}$ by definition since the adversarial can at most flip the labels of opt fraction of points. Moreover, conditioned on that the algorithm does not declare reject in Line 2, S then must contain at least $B/2$-fraction of points with label $+1$, which implies that $\bar{B}, \tilde{B} \geq \Omega(B)$. In order to show Equation (4), we define the following "mis-normalized" empirical mean $\hat{\boldsymbol{\mu}}_+$

$$\hat{\boldsymbol{\mu}}_+ := \frac{\tilde{B}}{\bar{B}} \boldsymbol{\mu}_+ = \frac{1}{\bar{B}|S|} \left(\sum_{(\mathbf{x}, y) \in S} \mathbb{1}\{y = 1\} \mathbf{x} \right).$$

We note that $\hat{\boldsymbol{\mu}}_+$ and $\boldsymbol{\mu}_+$ are close to each other:

$$\boldsymbol{\mu}_+ - \hat{\boldsymbol{\mu}}_+ = \left(1 - \frac{\tilde{B}}{\bar{B}}\right) \frac{1}{\tilde{B}|S|} \left(\sum_{(\mathbf{x}, y) \in S} \mathbb{1}\{y = 1\} \mathbf{x} \right) = O(\mathrm{opt}/B) \boldsymbol{\mu}_+ ,$$

where the last equality follows from that $\left|\tilde{B} - \hat{B}\right| \leq \mathrm{opt}$ and $\tilde{B} > \Omega(B)$. Since Lemma 12 implies that $\left\|\boldsymbol{\mu}_+\right\|_2 \leq O\left(\sqrt{\log(1/\mathrm{opt})}\right)$ and $B \gtrsim \sqrt{\mathrm{opt}} \log(1/\mathrm{opt})$, it then follows that

$$\mathbf{v}^* \cdot (\hat{\boldsymbol{\mu}}_+ - \boldsymbol{\mu}_+) \leq \tilde{O}\left(\sqrt{\mathrm{opt}}\right). \tag{5}$$

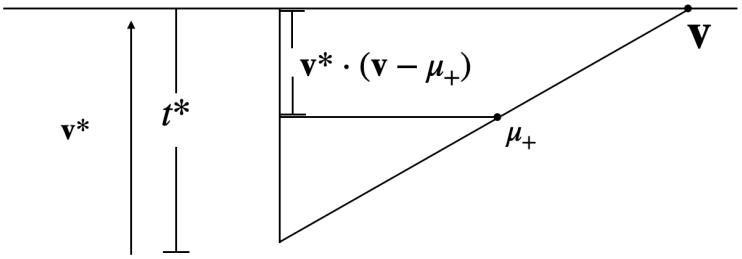

Figure 3: Bound $\|\mathbf{v}\|_2 - \|\boldsymbol{\mu}_+\|_2$ via $\mathbf{v}^* \cdot (\mathbf{v} - \boldsymbol{\mu}_+)$.

We then bound from above the distance between $\mathbf{v}^* \cdot \hat{\boldsymbol{\mu}}_+$ and $\mathbf{v}^* \cdot \bar{\boldsymbol{\mu}}_+$. We have that

$$\begin{aligned}
\mathbf{v}^* \cdot \hat{\boldsymbol{\mu}}_+ - \mathbf{v}^* \cdot \bar{\boldsymbol{\mu}}_+ &= \frac{1}{\bar{B}|S|} \sum_{(\mathbf{x},y)\in S} \left(\mathbb{1}\{h(\mathbf{x}) = +1\} - \mathbb{1}\{y = +1\}\right) \mathbf{v}^* \cdot \mathbf{x} \\
&\le \frac{1}{\bar{B}|S|} \sqrt{\sum_{(\mathbf{x},y)\in S} \left(\mathbb{1}\{h(\mathbf{x}) = +1\} - \mathbb{1}\{y = +1\}\right)^2 \mathbf{v}^{*\top} \left(\sum_{(\mathbf{x},y)\in S} \mathbf{x}\mathbf{x}^\top\right) \mathbf{v}^*} \\
&\le \frac{1}{\bar{B}} O\left(\sqrt{\mathrm{opt}}\right) < O\left(\frac{1}{\log(1/B)}\right), \qquad (6)
\end{aligned}$$

where in the first inequality we use Cauchy's inequality, in the second inequality we use the fact that there are at most opt fraction of points with $h(\mathbf{x}) \neq y$, and that the empirical second moments of S is bounded by $2\mathbf{I}$, and in the last inequality we use $\tilde{B} \gtrsim \sqrt{\mathrm{opt}}\log(1/\mathrm{opt})$. Combining Equation (5) and Equation (6) then concludes the proof of Equation (4).

With Equation (4) in our hand, we proceed to finish the proof. Note that $\mathbf{v}$ is the intersection point between the halfspace and the line $\boldsymbol{\mu}_+$. That means that for some $\lambda \in \mathbb{R}$, we have $\mathbf{v} = \lambda\boldsymbol{\mu}_+$ and hence using that $\mathbf{v} \cdot \mathbf{v}^* = -t^* = |t^*|$, we have that $\boldsymbol{\mu}_+ \cdot \mathbf{v}^* = |t^*|/\lambda$. Furthermore, note that $\mathbf{v}^* \cdot \bar{\boldsymbol{\mu}}_+ \ge |t^*|$, hence $\lambda \in (0, 1)$ for the noiseless $\bar{\boldsymbol{\mu}}_+$ and from Equation (4) one can reduce that λ is $(1 + o(1))$ away from the optimal's one. Note that $\mathbf{v}^* \cdot (\mathbf{v} - \boldsymbol{\mu}_+) = |t^*|(\lambda - 1)/\lambda$, hence

$$\frac{\|\mathbf{v}\|_2 - \|\boldsymbol{\mu}_+\|_2}{\|\mathbf{v}\|_2} = \frac{\lambda\|\boldsymbol{\mu}_+\|_2 - \|\boldsymbol{\mu}_+\|_2}{\lambda\|\boldsymbol{\mu}_+\|_2} = \frac{\mathbf{v}^* \cdot (\mathbf{v} - \boldsymbol{\mu}_+)}{t^*} \le O(1/\log(1/B)),$$

where the first equality is illustrated in Figure 3, and the last inequality is follows from Equation (3) and $t^* > 10$.

Therefore, it follows that

$$\|\mathbf{v}\|_2 \le \|\boldsymbol{\mu}_+\|_2 (1 + O(1/\log(1/B))).$$

By Lemma 12, we must have $\|\boldsymbol{\mu}_+\|_2 \le O(\sqrt{\log(1/B)})$. We therefore have

$$\|\mathbf{v}\|_2 \le \|\boldsymbol{\mu}_+\|_2 + O(\sqrt{\log(1/B)}/\log(1/B)) \le \|\boldsymbol{\mu}_+\|_2 + O(1/\sqrt{\log(1/B)}).$$

This concludes the proof of Lemma 11 ■

C.3. Localization Center Search for Halfspaces with Constant Thresholds

Lemma 19 *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ and $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t^*)$ be a halfspace that achieves* opt *error with respect to D. Then, there exists an algorithm that takes $N := \mathrm{poly}(d/\epsilon)$ many samples, and runs in time $\mathrm{poly}(N)$ and with high constant probability, either*

1. *reports violation of the distribution assumption, in which case the report is correct.*
2. *returns a list of at most $O(1/\epsilon^2)$ points. In addition, if it holds that $t^* \leq 10$, the list contains at least one $(\epsilon^2, \Theta(1))$-good localization center, where $C > 0$ is a universal constant.*

Proof Let S be the empirical distribution over $N := \mathrm{poly}(d/\epsilon)$ samples. Let $\eta > 0$ be some sufficiently small constant. The algorithm first verifies that the degree up to $k = \Theta(\log(1/\eta)/\eta^2)$ moments of the empirical distribution S match with those of $\mathcal{N}(\mathbf{0}, \mathbf{I})$ up to an additive error of $\Delta = \frac{1}{kd^k}\left(\frac{1}{C\sqrt{k}}\right)^{k+1}$. Then it computes the Chow parameter $\mathbf{w} := \mathbf{E}_{(\mathbf{x},y)\sim S}[y\mathbf{x}]$. Lastly, it returns the list of points $i\ \epsilon^2\ \mathbf{w}/\left\|\mathbf{w}\right\|_2$ for $i \in [10/\epsilon^2]$.

The soundness of the test follows from standard concentration inequalities of low degree moments of $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We hence omit the proof for conciseness.

We now argue that, with high constant probability, when the algorithm does not reject, we find a good localization center. Let $\mathbf{v}$ be the intersection between the line along $\mathbf{w}$ and the halfspace. We will argue that $\left\|\mathbf{v}\right\|_2$ is bounded from above by some constant, from which the existence of a good localization center in the output list follows. Since we assume that $t^* \leq 10$, it suffices to show that the angle between $\mathbf{w}$ and $\mathbf{v}^*$ is bounded from above by some sufficiently small constant.

Using Lemma 2.3 of Diakonikolas et al. (2023a), we have that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[h(\mathbf{x})\mathbf{x}] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h(\mathbf{x})\mathbf{x}] \leq O\left(\sqrt{\eta}\right). \tag{7}$$

From Lemma 4.3 of Diakonikolas et al. (2018), it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h(\mathbf{x})\mathbf{x}] = G(t^*)\ \mathbf{v}^*, \tag{8}$$

where $G(t^*)$ is the probability density function of the one dimensional standard normal. Finally, let $\mathbf{u}$ be an arbitrary unit vector in $\mathbb{R}^d$. Since we verify that the eigenvalues of the empirical covariance of the samples is bounded from above by 2, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[y\mathbf{x}\cdot\mathbf{u}] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[h(\mathbf{x})\mathbf{x}\cdot\mathbf{u}] \leq \sqrt{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}\left[\left(h(\mathbf{x})-y\right)^2\right]\ \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[\mathbf{u}^T\mathbf{x}\mathbf{x}^T\mathbf{u}]} \leq O\left(\sqrt{\mathrm{opt}}\right), \tag{9}$$

where the first inequality follows from Cauchy's inequality, and the second inequality follows from that h has error at most opt and S has its covariance bounded by $2\mathbf{I}$. Let $\mathbf{w}$ be the Chow parameter estimated. Combining Equations (7) to (9), we have that $\angle(\mathbf{w}, \mathbf{v}^*) \leq O(\sqrt{\eta})$, which can be made sufficiently small if we choose η to be some sufficiently small constant. This then concludes the proof of Lemma 19. ■

C.4. Proof of Lemma 12

Before presenting the analysis, we first present some useful bounds related to the standard Gaussian CDF function $\Phi(t) := \mathbf{Pr}_{x\sim\mathcal{N}(0,1)}[x > t]$.

Claim 20 *Let $x > 10$, and $b \le O(x^{-1})$. Then it holds $\Phi(x+b) \ge \Omega(1)\Phi(x)$.*

Proof We have

$$\begin{aligned}
\Phi(x+b) &= \frac{1}{\sqrt{2\pi}} \int_{t=x+b}^{\infty} \exp(-t^2/2) \\
&= \frac{1}{\sqrt{2\pi}} \int_{t=x}^{\infty} \exp(-(t+b))^2/2) \\
&\ge \frac{1}{\sqrt{2\pi}} \int_{t=x}^{10\,x} \exp(-(t+b))^2/2) \\
&\ge \frac{1}{\sqrt{2\pi}} \int_{t=x}^{10\,x} \exp(-t^2/2 - O(1)) \\
&= \Omega(1)\ (\Phi(x) - \Phi(10\,x)) \ge \Omega(1)\ \Phi(x)\,,
\end{aligned}$$

where the first inequality follows from the positivity of the $\exp(\cdot)$ function, and the second inequality follows from that $tb, b^2 = O(1)$ since $b \le O(1/x) \le O(1/t)$. ■

We now give the proof of Lemma 12.

Proof [Proof of Lemma 12] Recall that S is a set of N i.i.d. samples from D where $N = \mathrm{poly}(d, 1/\epsilon)$, $\boldsymbol{\mu}_+$ is the empirical mean of the samples in S with label $+1$, H is the resulting set after projecting the points in S along the direction of $\boldsymbol{\mu}_+$, and $\tilde{B}$ is the fraction of points in H (or S) with label $+1$. For convenience, we define $u := \left\|\boldsymbol{\mu}_+\right\|_2$. As a slight abuse of notation, we also denote by H the empirical distribution over the set H.

In our analysis, it is easier to bound $\Phi(\|\mathbf{u}_+\|_2)$ in terms of $\tilde{B}$ thatn to bound it in terms of B. Nonetheless, we remark that B and $\tilde{B}$ are approximately the same. In particular, using standard concentration inequalities, we have that $\tilde{B} := \mathbf{Pr}_{(\mathbf{x},y)\sim H}[y = 1]$ should satisfy $\left|\tilde{B} - B\right| \le \epsilon/10$. By our assumption that $\epsilon < B^2$, we have that that $\tilde{B} > B/2$, which implies that $\tilde{B}/\log(1/\tilde{B}) > (B/2)/\log(2/B) = \Omega(B/\log(1/B))$, since $x/\log(1/x)$ is monotonically increasing for $x \in [0, 1)$. Hence, it suffices to show that

$$\Phi(\left\|\boldsymbol{\mu}_+\right\|_2) \ge \Omega\left(\tilde{B}/\log(1/\tilde{B})\right). \tag{10}$$

We further introduce the following sets to facilitate the discussion. We divide H into two sets M, P, where M is the set of points with label -1, and P is the set of points with label $+1$. Among P, we further divide into P_L and P_R, where $P_L := \{x \in P \text{ such that } x \le u\}$, and $P_R := \{x \in P \text{ such that } x > u\}$. We proceed to examine two cases separately.

Case I In the first case, suppose more than half of the points from P are within the range $[u - 1/\sqrt{\log(1/\tilde{B})}, \infty]$. Since $|P|/|H| = \tilde{B}$, we have

$$\mathop{\mathbf{Pr}}_{x\sim H}\left[x > u - 1/\sqrt{\log(1/\tilde{B})}\right] \ge \mathop{\mathbf{Pr}}_{x\sim H}\left[x \in P \text{ and } x > u - 1/\sqrt{\log(1/\tilde{B})}\right] \ge \tilde{B}/2.$$

Recall that in Line 6b we explicitly verify that H is close to $\mathcal{N}(0,1)$ in Kolmogorov distance. It follows that

$$\Phi(u - 1/\sqrt{\log(1/\tilde{B})}) \geq \mathop{\mathbf{Pr}}_{x\sim H}\left[x > u - 1/\sqrt{\log(1/\tilde{B})})\right] - \epsilon \geq \Omega(\tilde{B}).$$

This allows us to bound $u - 1/\sqrt{\log(1/\tilde{B})}$ from above by $O\left(\sqrt{\log(1/\tilde{B})}\right)$. Therefore, applying Claim 20 gives that $\Phi(u) \geq \Omega(1)\Phi\left(u - 1/\sqrt{\log(1/\tilde{B})}\right) \geq \Omega(\tilde{B})$, which implies Equation (10).

Case II In the second case, we have more than half of the points from P lying in the range $[-\infty, u - 1/\sqrt{\log(1/\tilde{B})}]$. In this case, we can bound from below the contribution from the points in P_L to the mean of P by

(11)

$$\frac{1}{|P|}\sum_{x\in P_L}(u-x) \geq \frac{1}{|P|}\sum_{x\in P_L}\mathbb{1}\left\{x < u - \frac{1}{\sqrt{\log(1/\tilde{B})}}\right\}\frac{1}{\sqrt{\log(1/\tilde{B})}} \geq \frac{1}{2\sqrt{\log(1/\tilde{B})}}.$$

This then implies that the points in P_R need to satisfy that

$$\sum_{x\in P_R} x \geq \sum_{x\in P_R}(x-u) = |H|\,\frac{|P|}{|H|}\,\frac{1}{|P|}\sum_{x\in P_L}(u-x) \geq |H|\frac{\tilde{B}}{2\sqrt{\log(1/\tilde{B})}}\,, \tag{12}$$

where the first inequality holds because $u := \left\|\boldsymbol{\mu}_+\right\|_2 \geq 0$, and the equality holds since $\sum_{x\in P_R}(x-u) = \sum_{x\in P_L}(u-x)$, and the last inequality follows from the definition $\tilde{B} := |P|/|H|$ and Equation (11). For the sake of contradiction, assume that $\mathbf{Pr}_{x\sim H}[x \in P_R] \leq c\,\tilde{B}/\log(1/\tilde{B})$ for some sufficiently small constant $c > 0$. Then, conditioned on the event that Line 6c passes the test; if we remove all points from P_R, the mean deviates by at most $O\left(c\,\frac{\tilde{B}}{\log(1/B)}\sqrt{\log\left(\frac{\sqrt{\log(1/B)}}{\tilde{B}}\right)}\right) = O\left(c\,\tilde{B}/\sqrt{\log(1/B)}\right)$. In particular, the mean of the remaining samples must satisfy

$$\left|\sum_{x\in H\setminus P_R} x\right| \leq O\left(c\,\tilde{B}/\sqrt{\log(1/B)}\right)\,|H\setminus P_R| \leq O\left(c\,\tilde{B}/\sqrt{\log(1/B)}\right)\,|H|. \tag{13}$$

Conditioned on the event that Line 2 passes, we have that

$$\left|\sum_{x\in H} x\right| = \epsilon\,|H|. \tag{14}$$

Note that $\sum_{x\in P_R} x = \sum_{x\in H} x + \left(\sum_{x\in H\setminus} -x\right)$. Thus, by using Equations (13) and (14), and the triangle inequality, we have that

$$\left|\sum_{x\in P_R} x\right| \leq O\left(c\frac{\tilde{B}}{\sqrt{\log(1/\tilde{B})}} + \epsilon\right)|H|\,. \tag{15}$$

Recall that $\epsilon \lesssim \tilde{B}$ by our assumption. Thus, by choosing c to be a sufficiently small constant, Equation (15) contradicts Equation (12). Therefore, we have that $\mathbf{Pr}_{x\sim H}[x > u] = \mathbf{Pr}_{x\sim H}[x \in P_R] > \Omega\left(\tilde{B}/\log(1/\tilde{B})\right)$. With similar arguments as in Case I, conditioned on the event that Line 6b passes, we then have that

$$\Phi(\left\|\boldsymbol{\mu}_+\right\|_2) \geq \mathop{\mathbf{Pr}}_{\mathbf{x}\in H}\left[x > \left\|\boldsymbol{\mu}_+\right\|_2\right] - \epsilon \geq \Omega\left(\tilde{B}/\log(1/\tilde{B})\right) .$$

Therefore Equation (10) holds for Case II, and this concludes the proof of Lemma 12. ■

Appendix D. Proof of Main Theorem (Theorem 2)

Input: Sample access to a distribution D over $\mathbb{R}^d \times \{\pm 1\}$; tolerance ϵ; failure probability τ.
Output: Reject or output a halfspace h, with error at most $\tilde{O}(\sqrt{\mathrm{opt}} + \epsilon)$.

1. Run Find-Localization-Center (Algorithm 1) to find a set of localization centers Localization-Centers $:= \{\mathbf{v}^{(i)}\}_{i=1}^{\log(1/B)/\epsilon^2}$
2. Initialize an empty set S for storing candidate halfspaces.
3. For $\mathbf{w} \in$ Localization-Centers
   (a) Set $\sigma := \min\left(\|\mathbf{w}\|_2^{-1}, \sqrt{1/2}\right)$.
   (b) Define G as the distribution over $\mathbb{R}^d \times \{\pm 1\}$ obtained by performing $(\mathbf{w}, \sigma)$-rejection sampling on the $\mathbf{x}$-marginal of D.
   (c) Define D' as the distribution obtained by applying the transformation $h(\mathbf{x}) = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \mathbf{w})$ on the $\mathbf{x}$-marginal of G, where $\boldsymbol{\Sigma} := \mathbf{I} - (1-\sigma^2)\mathbf{w}\mathbf{w}^T/\|\mathbf{w}\|_2^2$.
   (d) Run Almost-Homogeneous-Testable-Learn from Proposition 6 with sample access to D', and denote the learned halfspace vector as $\mathbf{v}$.
   (e) Run Wedge-Bound (Algorithm 2) with sample access to D, and the unit vector $\boldsymbol{\Sigma}^{-1/2}\mathbf{v}$.
   (f) For $i = -\lceil\log(1/\epsilon)/\epsilon\rceil, \cdots, \lceil\log(1/\epsilon)/\epsilon\rceil$, add the halfspace $h(\mathbf{x}) = \mathrm{sign}\left(\boldsymbol{\Sigma}^{-1/2}\mathbf{v} + i\epsilon\right)$ to S.
4. Compute the empirical errors of the halfspaces within S using $\mathrm{poly}(d/\epsilon)$ i.i.d. samples from D.
5. Return the halfspace with the smallest empirical error.

Algorithm 3: Testable-Learn-General-Halfspace

Proof [Proof of Main Theorem] Our algorithm produces a list S of halfspaces and picks the one with the smallest empirical error. It suffices to show that at least one of the halfspaces contained in the list S achieves error at most $\tilde{O}(\sqrt{\mathrm{opt}} + \epsilon)$. Let B be the mass of the minority label. Our list always contains a hypothesis that outputs either the label $+1$ or -1 for any point, therefore if

$B < \max(\sqrt{\epsilon}, \sqrt{\mathrm{opt}})$ then one of these two halfspaces obtains error at most $\sqrt{\mathrm{opt}} + \sqrt{\epsilon}$. Hence, we can assume $B > \max(\sqrt{\epsilon}, \sqrt{\mathrm{opt}})$, for the rest of the analysis of the algorithm.

Let $h(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t^*)$ be the optimal halfspace. If $t^* > 10$, we can apply Proposition 9 to obtain a list of candidate localization centers with the guarantee that at least one of them is an $(\epsilon^2, \Omega(B/\log(1/B)))$-good localization center. If $t^* < 10$, we directly estimate the Chow vector (enhanced with moment matching) which, up to normalization, gives us a constant approximation to the defining vector $\mathbf{v}^*$ of the optimal halfspace (see Lemma 19). Then, we guess the intersection between the Chow vector and the halfspace h up to accuracy ϵ^2. This gives us another list containing at least one point that is $(\epsilon^2, \Theta(1))$-good localization center. The list L in both cases will have size at most $\mathrm{poly}(1/\epsilon)$.

For each $\mathbf{w} \in L$, we apply $(\mathbf{w}, \sigma)$ rejection sampling with $\sigma = \min(\sqrt{1/2}, \|\mathbf{w}\|_2^{-1})$. The resulting distribution is then $\mathcal{N}(\mathbf{w}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{I} - (1-\sigma^2)\mathbf{w}\mathbf{w}^T/\|\mathbf{w}\|_2^2$. Next we apply the appropriate transformation to make the $\mathbf{x}$-marginal isotropic, and call this distribution $D'_{\mathbf{w}}$. By assumption, there exists at least one $\mathbf{w} \in L$ that is $(\epsilon^2, \Omega(B/\log(1/B)))$-good localization center. Let $\mathbf{w}$ be an $(\epsilon^2, \Omega(B/\log(1/B)))$-good localization center. From Lemma 5, after the transformation, an optimal halfspace of $D'_{\mathbf{w}}$ is a halfspace $\tilde{h}$ with offset $O(1)\ \epsilon^2 \log(1/B) \le \epsilon$. The acceptance probability of a sample in the distribution $D'_{\mathbf{w}}$ is at least $\Omega(B/\log(1/B))$. It follows that the error of the halfspace under the new distribution is at most $O(\mathrm{opt}\ \log(1/B)/B) < \tilde{O}(\sqrt{\mathrm{opt}})$.

Let $\tilde{h}(\mathbf{x}) = \mathrm{sign}(\tilde{\mathbf{v}} \cdot \mathbf{x} + \tilde{t})$ be the halfspace after applying the transformation that makes the new distribution isotropic. Using Proposition 6, we can then learn a vector $\mathbf{v}$ satisfying

$$\|\mathbf{v} - \tilde{\mathbf{v}}\|_2 \le \tilde{O}(\sqrt{\mathrm{opt}} + \epsilon).$$

Notice that by Lemma 5, $\tilde{\mathbf{v}}$ is parallel to the vector $\mathbf{\Sigma}^{1/2}\mathbf{v}^*$. We argue the estimation $\mathbf{v}$ after reverting the transformation $\mathbf{\Sigma}^{1/2}$ gives us a good approximation of $\mathbf{v}^*$. To do so, we need to apply Lemma 7. A crucial requirement of Lemma 7 is that the angle between the localization direction $\mathbf{w}$ and $\mathbf{v}^*$ must not be too large. To show this, we need to consider two cases depending on whether we have $|t^*| < 10$.

In the case that $|t^*| < 10$, the angle between $\mathbf{w}$ and $\mathbf{v}^*$ will be bounded by a small constant (since the direction of $\mathbf{w}$ is obtained through Chow parameter estimation in this case).

The harder case is when $|t^*| > 10$. In this case, we write $\mathbf{w} = b\mathbf{v}^* + a\mathbf{z}$ where $\mathbf{z}$ is orthogonal to $\mathbf{v}^*$. Since the distance from $\mathbf{w}$ to the halfspace is at most ϵ^2, the component of $\mathbf{w}$ along $\mathbf{v}^*$ must be at least $(10 - \epsilon^2)$, i.e., $b > (10 - \epsilon^2)$. On the other hand, since $\mathbf{w}$ satisfies $\Phi(\|\mathbf{w}\|_2) > \Omega(B/\log(1/B))$, it follows that $\|\mathbf{w}\|_2 \le O(\sqrt{\log(1/B)})$, which further implies that $a < O(\sqrt{\log(1/B)})$. Combining the two observations we have that $a/b \le O(\sqrt{\log(1/B)})$. Let $x = b/a$, we have that

$$\begin{aligned}\|\mathbf{v}^* - \mathbf{w}/\|\mathbf{w}\|_2\|_2^2 &= \left(1 - \frac{1}{\sqrt{1+x^2}}\right)^2 + \left(\frac{x}{\sqrt{1+x^2}}\right)^2 = 1 + \frac{1}{1+x^2} - \frac{2}{\sqrt{1+x^2}} + \frac{x^2}{1+x^2} \\ &= 2 - \frac{2}{\sqrt{1+x^2}} \le 2 - \frac{2}{O(\sqrt{\log(1/B)})}.\end{aligned}$$

For this particular localization center $\mathbf{w}$, we also have $\sigma \ge \Omega(1/\sqrt{\log(1/B)}) \ge \Omega(1/\sqrt{\log(1/\mathrm{opt})})$ by Lemma 5.

We can therefore apply Lemma 7 with $\beta = \sqrt{2 - \frac{2}{O(\sqrt{\log(1/B)})}}$, $\delta = \tilde{O}(\sqrt{\mathrm{opt}}) + \epsilon$ and $\sigma = \Omega(1/\sqrt{\log(1/\mathrm{opt})})$. Note that

$$\frac{\beta}{1-\beta^2/2} \leq O(\sqrt{\log(1/B)}).$$

This implies that, after reverting the transformation, we obtain a vector $\mathbf{v}$ satisfying that

$$\begin{aligned}\|\mathbf{v}-\mathbf{v}^*\|_2 &\leq O(1)\ (\sigma+\beta)\ \left(\delta\sigma^{-1}\frac{\beta}{1-\beta^2/2}+\delta\right)\\ &\leq O(\delta)\ \left(\sqrt{\log(1/\mathrm{opt})}\ \sqrt{\log(1/\mathrm{opt})}\right) \leq \tilde{O}(\sqrt{\mathrm{opt}}+\epsilon).\end{aligned}$$

We then guess the offset of the optimal halfspace up to accuracy ϵ. Denote the best guess as $\bar{t}$. Applying Lemma 8, we have that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+\bar{t}) \neq \mathrm{sign}(\mathbf{v}^*\cdot\mathbf{x}+t^*)] \leq O(1)\ (\|\mathbf{v}-\mathbf{v}^*\|_2 + |\bar{t}-t^*|) \leq \tilde{O}(\sqrt{\mathrm{opt}}+\epsilon).$$

The above concludes the proof of Theorem 2 for constant failure probability τ. In order to boost the success probability, we will run the algorithm $10\ \log(1/\tau)$ many times. If the algorithm outputs reject for more than half of the time, we output reject. Otherwise, we collect the halfspaces learned, draw $O\left(d\log\log(1/\tau)/\epsilon^2\right)$ many i.i.d. samples, and output the halfspace that achieves the smallest empirical error. If D truly has marginal $D_{\mathbf{x}}$, each invocation of the algorithm outputs accepts with high constant probability. Since we only output reject when the algorithm outputs reject in more than half of the trials, it follows that we reject with probability at most τ. If we do not output reject, we must have collected at least $5\log(1/\tau)$ many halfspaces, and each halfspace achieves the required learning error with high constant probability. So there exists a halfspace in the learned list with error $\widetilde{O}(\sqrt{\mathrm{opt}})+\epsilon$ with probability at least $1-\tau/2$. Besides, with probability at least $1-\tau/2$, we can estimate the error of these halfspaces up to accuracy ϵ. Conditioned on the above two events, which hold simultaneously with probability at least $1-\tau$ by the union bound, it then follows that the final halfspace picked will have error at most $\widetilde{O}(\sqrt{\mathrm{opt}})+2\epsilon$. This concludes the proof of Theorem 2. ■