Learning Intersections of Two Margin Halfspaces under Factorizable Distributions

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

Mingchen Ma MINGCHEN@CS.WISC.EDU
University of Wisconsin-Madison

Lisheng Ren LREN29@WISC.EDU
University of Wisconsin-Madison

Christos Tzamos CTZAMOS@GMAIL.COM
University of Athens and Archimedes AI

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Learning intersections of halfspaces is a central problem in Computational Learning Theory. Even for just two halfspaces, it remains a major open question whether learning is possible in polynomial time with respect to the margin γ of the data points and their dimensionality d. The best-known algorithms run in quasi-polynomial time $d^{O(\log 1/\gamma)}$, and it has been shown that this complexity is unavoidable for any algorithm relying solely on correlational statistical queries (CSQ).

In this work, we introduce a novel algorithm that provably circumvents the CSQ hardness barrier. Our approach applies to a broad class of distributions satisfying a natural, previously studied, factorizability assumption. Factorizable distributions lie between the distribution-specific and distribution-free settings, and significantly extend previously known tractable cases. For these distributions, we show that CSQ-based methods still require quasipolynomial time even for weak learning. Our main result is a learning algorithm for intersections of two margin halfspaces under factorizable distributions that achieves $\mathrm{poly}(d, 1/\gamma)$ time by leveraging more general statistical queries (SQ). As a corollary, we establish a strong separation between CSQ and SQ for this fundamental PAC learning problem. Our main result is grounded in a rigorous analysis utilizing a novel duality framework that characterizes the moment tensor structure induced by the marginal distributions. Building on these structural insights, our learning algorithm combines a refined variant of Jennrich's Algorithm with PCA over random projections of the moment tensor, along with a gradient-descent-based non-convex optimization framework.

Keywords: Intersections of Halfspaces, Efficient Learning Algorithms, Statistical Query Learning

1. Introduction

A halfspace $h = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) : \mathbb{R}^d \to \{\pm 1\}$ is a Boolean function defined by its weight vector $\mathbf{u}^* \in \mathbb{R}^d$ and threshold $t_1 \in \mathbb{R}$. Halfspace learning is one of the oldest and most fundamental problems in Machine Learning (Rosenblatt, 1958; Block, 1962). While learning a single halfspace is well-understood, learning intersections of halfspaces is significantly more challenging. Even for intersections of two halfspaces, polynomial-time algorithms are known only under strong distributional assumptions about the datapoints $\mathbf{x}$, which are e.g., assumed to be drawn from a Gaussian or log-concave distribution (Blum and Kannan, 1997; Vempala, 2010a,b). Beyond these assumptions, little is known about the problem's complexity. Prior work (Klivans and Sherstov, 2007, 2009;

Daniely and Shalev-Shwartz, 2016; Tiegel, 2024) establish hardness results for learning intersections of $\omega_d(1)$ halfspaces. It is a central open question in computational learning theory whether an intersection of even two halfspaces can be efficiently learned in the distribution-free setting.

To design efficient learning algorithms in the more challenging distribution-free setting, a popular approach is to assume that the underlying distribution has a margin with respect to the target hypothesis; see, e.g., (Arriaga and Vempala, 2006; Klivans and Servedio, 2004a) (see Definition 1). Under the γ-margin assumption, it is well-known that the Perceptron algorithm properly learns a single halfspace in time $\tilde{O}(d/(\gamma^2\epsilon))$; see, e.g., Cristianini (2000). Unfortunately, a similar result does not hold for learning an intersection of two halfspaces. Even under a margin assumption, it is computationally hard to output an intersection of any constant number of halfspaces with an error better than $1/2$ (Khot and Saket, 2008). The best known algorithm (Klivans and Servedio, 2004a) in this setting, developed over 20 years ago, runs in time $d^{O(\log(1/\gamma))}$ and outputs a polynomial threshold function with degree $O(\log(1/\gamma))$. Such a learning algorithm not only has super-polynomial time complexity, but also needs super-polynomial time to evaluate its hypothesis on a single example. An important open question, posed in Klivans and Servedio (2004b), is whether a $\mathrm{poly}(d, 1/\gamma, 1/\epsilon)$ time learning algorithm exists under only a γ-margin assumption.

Specifically, the algorithm developed by Klivans and Servedio (2008) closely relates to the notion of Correlation Statistical Queries (CSQ), which are queries of the form $\mathbf{E}_{(x,y)\sim D}[yq(x)]$, where q is an arbitrary bounded function. The main idea of this type of algorithm relies on the fact that the target hypothesis can be represented as a high-degree polynomial threshold function, and thus one can make CSQ queries—independent of the marginal distribution—to obtain a weak hypothesis; a strong hypothesis can then be obtained via boosting. Algorithms of this type usually do not leverage useful structural properties of the underlying learning problem, and are hard to adapt to obtain more efficient algorithms. In fact, even for weak learning, $d^{\Omega(\log(1/\gamma))}$ complexity is the best one can hope for via a CSQ algorithm. This suggests that, to make progress toward a polynomial time algorithm, a new algorithmic framework is needed. In particular, one needs to design *instance-dependent* statistical queries by learning information about the marginal distribution D_X. In this work, we introduce a novel algorithm that provably circumvents the CSQ-hardness barrier under the γ-margin assumption. Our algorithm runs in fully-polynomial time for a broad class of distributions satisfying a factorizability assumption. We now formally define the problem we study in this paper.

Definition 1 (Learning Intersections of Margin Halfspaces Under Factorizable Distributions) *Let $V \subseteq \mathbb{R}^d$ be an unknown two-dimensional subspace and $W = V^\perp \subseteq \mathbb{R}^d$ be the orthogonal complement of V. Let $h^*(\mathbf{x}) = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2) : \mathbb{R}^d \to \{\pm 1\}$, where $\mathbf{u}^*, \mathbf{v}^* \in V \cap S^{d-1}$ be the target directions and $t_1, t_2 \in \mathbb{R}$ are the thresholds of the defining halfspaces. Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ satisfying the following:*

1. *The distribution D is consistent with an instance of learning intersections of two halfspaces h^*, i.e., for $(\mathbf{x}, y) \sim D$, $y = h^*(\mathbf{x})$ holds almost surely.*
2. *The distribution D satisfies the γ-margin assumption, i.e., for $\mathbf{x} \sim D_X$, it holds $\|\mathbf{x}\|_2 \leq 1$ and $|\mathbf{u}^* \cdot \mathbf{x}_V + t_1| \geq \gamma$, $|\mathbf{v}^* \cdot \mathbf{x}_V + t_2| \geq \gamma$ holds almost surely. Here $\mathbf{x}_V$ is the projection of $\mathbf{x}$ on V.*
3. *We say that D is factorizable if $D_X = D_V \times D_W$, where D_V is the marginal distribution of D_X over D_V and D_W is the marginal distribution of D_X over W.*

Given parameters $\epsilon, \delta \in (0, 1)$, a learning algorithm $\mathcal{A}$ draws a set $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ of m examples i.i.d. from D and outputs a hypothesis $\hat{h} : \mathbb{R}^d \to \{\pm 1\}$ such that with probability at least $1 - \delta$, $\mathrm{err}(\hat{h}) := \mathbf{Pr}_{(\mathbf{x},y)\sim D}\left(\hat{h}(\mathbf{x}) \neq y\right) \leq \epsilon$.

Throughout this paper, we will use D_X for the marginal distribution of D on the feature space, D^+ to denote the marginal distribution of D_X on positive examples, and D^- to denote the marginal distribution of D_X on negative examples.

Discussion Factorizable distributions lie between the distribution-specific and the distribution-free settings. Specifically, an efficient learning algorithm for such distributions would significantly extend previously known tractable settings, such as under the Gaussian or uniform distribution over the unit sphere, as no assumptions are made over D_V and D_W. Factorizable distributions are not new in this learning context. The original motivation of studying them can be traced back at least to Blum (1994) in the context of learning k-juntas under the uniform distribution on the hypercube, and to (Klivans et al., 2008; Vempala, 2010a) for learning convex concepts under the Gaussian distribution. In both of these settings, the target hypothesis only depends on the projection of the points on some unknown low-dimensional subspace, the marginal distributions are factorizable and satisfy additional strong assumptions. With this motivation, Vempala and Xiao (2011) first formally proposed the setting of learning k-subspace juntas (functions that only depend on the projection on a k-dimensional subspace) under factorizable distributions. The original observation of Vempala and Xiao (2011) was that if there are k directions in V along which the moments of D_V are different from those of a standard Gaussian, then one can information-theoretically recover the subspace V. However, as we will discuss in detail in Section 3, this approach incurs an exponential dependence on d and the accuracy parameter. To obtain computationally efficient algorithms, Vempala and Xiao (2011) additionally assume that D_W is the standard Gaussian. Under this assumption, if D_V satisfies the aforementioned moment conditions and H satisfies some additional robustness assumptions, they gave an algorithm that approximately recovers the relevant subspace and then learns over a low-dimensional space. In contrast, our work focuses on the original setting where no additional assumptions are made over D_V, D_W for the basic case of intersections of two large-margin halfspaces. In this context, we establish novel structural results for intersections of two halfspaces, and design a fully-polynomial time learning algorithm. We summarize our results below.

Our Contribution and Technical Overview We start by establishing a quasi-polynomial $d^{\Omega(\log(1/\gamma))}$ Correlational Statistical Query (CSQ) lower bound (Theorem 2) for our learning task (Definition 1). Our CSQ lower bound shows that, unlike learning under the Gaussian distribution where there exists a fully-polynomial CSQ algorithm, learning in the factorizable setting is more challenging and CSQ algorithms even fail to efficiently weakly learn. Furthermore, the CSQ lower bound we obtain matches the running time of the algorithm developed by Klivans and Servedio (2008) for learning intersections of two γ-margin halfspaces (even without the factorizable assumption). This suggests that a new algorithmic framework is required to obtain more efficient algorithms. Due to space limitations, the proof of this lower bound is deferred to Appendix G.

Theorem 2 (CSQ Lower Bound) *Let $\gamma > 0$, $q, d \in \mathbb{N}$, $\tau \in (0,1)$ and $d' = \min(d, 1/\gamma^2)$. Any CSQ algorithm that learns intersections of two halfspaces with γ-margin in d dimensions under factorizable distributions to error $1/2 - \max(d'^{-\Omega(\log(1/\gamma))}, 2^{-d'^{\Omega(1)}})$ requires q queries of tolerance at most τ, where $q/\tau^2 \geq \min(d'^{\Omega(\log(1/\gamma))}, 2^{d'^{\Omega(1)}})$.*

Our main algorithmic result bypasses the CSQ-hardness by giving a polynomial-time algorithm for learning any intersection of two halfspaces with γ-margin, as long as D_X is factorizable. This implies the first strong separation between CSQ and SQ algorithms for *weak* (realizable) PAC learning of a natural concept class. We refer the reader to the related work for more detailed discussion.

Theorem 3 (Main Result) *There is an algorithm that for any distribution D over $\mathbb{B}^d(1) \times \{\pm 1\}$ satisfying the conditions of Definition 1, for any $0 < \epsilon, \delta < 1$, has the following guarantees: it draws $n = \mathrm{poly}(d, 1/\gamma, 1/\epsilon, \log(1/\delta))$ labeled examples from D, runs in $\mathrm{poly}(n, d)$ time, and outputs a hypothesis $\hat{h} : \mathbb{B}^d(1) \to \{\pm 1\}$ such that with probability at least $1 - \delta$, $\mathrm{err}(\hat{h}) \leq \epsilon$.*

The full proof of Theorem 3 and the main learning algorithm are deferred to Appendix F. In the rest of this section, we give a detailed outline of the techniques involved in proving this result.

Intuitively, the CSQ lower bound of Theorem 2 implies that without learning some information about the marginal distribution D_X, it is impossible to efficiently find a CSQ, q, such that $\left|\mathbf{E}_{(\mathbf{x},y)} yq(\mathbf{x})\right| > \mathrm{poly}(\gamma/d)$; as otherwise, one could output $\mathrm{sign}(q(\mathbf{x}) - t), t \sim [-1, 1]$, as a weak hypothesis with $1/2 - \mathrm{poly}(\gamma/d)$ error. This suggests that a plausible approach is to first learn (some information about) the marginal distribution D_X, and use it to design *instance-dependent* statistical queries. The construction of these queries hinges on the following observation. If we are given a direction $\mathbf{w}$ that is $\mathrm{poly}(\gamma)$-close to V, then restricted over bands $B_i := \{\mathbf{x} \mid \mathbf{w} \cdot \mathbf{x} \in [i\gamma, (i+1)\gamma]\}$, the labels y are consistent with a degree-2 polynomial threshold function. This in turn implies that a CSQ of the form $q(\mathbf{x}) = \mathbb{1}(\mathbf{x} \in B_i)p(\mathbf{x})$, for some degree-2 polynomial p, can be used to give a weak hypothesis with $\mathrm{poly}(\gamma)$ advantage. Once we have a weak hypothesis, we can run a standard boosting algorithm to get a strong hypothesis. With this goal, the question is how to efficiently find such a direction $\mathbf{w}$. To achieve this, we start with an easier case, where $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}^{\otimes m}\|_F$ is large. Since D_X is factorizable, we show in Theorem 14 that any local maximum or local minimum of the objective function $f(\mathbf{u}) = (\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})(\mathbf{u} \cdot \mathbf{x})^m$ must be in V. Though finding exact locally optimal solutions for f is computationally intractable, we show that an approximate solution, obtained by running a standard gradient-descent method, suffices for our purposes.

The more challenging case is when the underlying instance is indeed CSQ-hard. In this case, a low-degree polynomial function q will also satisfy $\left|\mathbf{E}_{(\mathbf{x},y)\sim D} yq(\mathbf{x})\right| \approx 0$, which implies that $\mathbf{E}_{\mathbf{x}\sim D^+} \mathbf{x}^{\otimes m} \approx \mathbf{E}_{\mathbf{x}\sim D^-} \mathbf{x}^{\otimes m}$ for small $m \in \mathbb{Z}_+$. Leveraging the fact that the label y only depends on the subspace V, we establish the following novel structural property for D_V. Our key structural result can be summarized as follows.

Theorem 4 (Informal statement of Theorem 6) *For any distribution D satisfying the conditions of Definition 1, if for $m \in [3]$, $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}^{\otimes m}\|_F \leq \mathrm{poly}(\gamma)$ and $\|\mathbf{E}_{\mathbf{x}\sim D_X} \mathbf{x}\|_F \leq \mathrm{poly}(\gamma)$, then $\left\|\mathbf{E}_{\mathbf{x}\sim D_V} \mathbf{x}_V^{\otimes 3}\right\|_F \geq \mathrm{poly}(\gamma)$.*

That is, for any distribution D consistent with our learning task, if the first three moments of D^+, D^- are nearly matching and the mean of D_X is close to 0, then the third moment tensor of D_V must significantly deviate from 0. The proof of this structural result is rather technical, because the only condition we assumed about D_V is the margin assumption. To prove this result, we develop a novel technique which we term *one-sided polynomial approximation.* Roughly speaking, if we are able to appropriately construct a polynomial function $p(\mathbf{x})$ such that for some $z \in \{\pm 1\}$, $\mathrm{sign}(p(\mathbf{x})) = z, \forall \mathbf{x}, h^*(\mathbf{x}) = z$, then such a polynomial can be used as a certificate to show that distributions satisfying certain moment conditions do not exist. One-sided polynomial approximation is a powerful tool for proving moment properties of distributions. In Section 2, we will carefully design these polynomials to prove useful results for the marginal distribution D_V.

The next step is to efficiently find a direction $\mathbf{w}$ close to V, by leveraging our structural result. Our initial attempt was inspired by the observation of Vempala and Xiao (2011) on generalized independent component analysis: for $m \geq 3$ if D_V has m-th moment different from that of a standard

Gaussian, but the first $m-1$ moments are the same as those of a Gaussian, then a locally optimal solution to $f(\mathbf{u}) = \mathbf{E}_{\mathbf{x}\sim D_X}(\mathbf{u}\cdot\mathbf{x})^m$ must be either in V or W. Unfortunately, such an idea does not lead to an efficient algorithm for the following reason. Since we make no distributional assumptions over D_W, D_W could also be factorized as $\Omega(d)$ many distributions that are isomorphic to D_V. This implies that information-theoretically we are only able to find a list of $O(d)$ unit vectors such that one of them is close to V. The natural approach is to optimize f in order to find one direction, and optimize the same function over the orthogonal subspace to find the next directions. As we explain in Section 3, such an approach has exponential sample complexity (as also demonstrated in Vempala and Xiao (2011)). Interestingly, we show that by carefully adapting such an idea, we are able to obtain an efficient SQ algorithm, which in turn leads to $\mathrm{poly}(d/\gamma)$ sample complexity. Still, it is unclear how to get a computationally efficient learning algorithm. To overcome this difficulty, we develop a completely different approach inspired by the tensor-decomposition literature. Though the third moment tensor T of D_X is in general very high rank and does not admit a low-rank decomposition, we know that the third-moment tensor T_V of D_V significantly deviates from 0. This implies that if we take the product of T with a Gaussian random vector $\mathbf{v}$, then with good probability one of the eigenvalues of $T_V\cdot\mathbf{v}$ must be significantly different from the other eigenvalues of $T\cdot\mathbf{v}$. By the factorizable assumption, an eigenvector of $T\cdot\mathbf{v}$ must be close to V. Thus, even for the hard instance mentioned above, we are able to efficiently find a direction close to V. We give Algorithm 1, a sketch of our approach for learning intersections of two halfspaces, and give the full algorithm in Appendix F.

Algorithm 1 LEARNINGINTERSECTIONS (Sketch)

1: Find a list $\mathcal{O}$ of unit vectors as follows: if D^+ and D^- have nearly matching first three moments, run the algorithm in Section 3.1; otherwise, run the algorithm in Section 3.2.
2: For each $\mathbf{w}\in\mathcal{O}$, run Adaboost with the algorithm in Section 4 to get a hypothesis $h_{\mathbf{w}}$.
3: Return the best hypothesis $h\in\{h_{\mathbf{w}}\mid\mathbf{w}\in\mathcal{O}\}$ via a standard hypothesis testing approach.

1.1. Related Work

Learning Intersections of Halfspaces Learning intersections of halfspaces is one of the central problems in learning theory. Despite a long line of work studying this problem algorithmically (Baum, 1990; Long and Warmuth, 1994; Kwek and Pitt, 1996; Blum and Kannan, 1997; Klivans et al., 2004; Klivans and Servedio, 2008; Klivans et al., 2008, 2009; Vempala, 2010a,b), relatively little is known about its complexity. In the distribution-specific setting, Blum and Kannan (1997) first showed that under the uniform distribution over the unit sphere (or under the Gaussian distribution), for any fixed constant k, one can learn an intersection of k halfspaces in polynomial time via PCA. Vempala (2010b) later extend this result to isotropic log-concave distributions via a random sampling method. For discrete distributions, Klivans et al. (2004), developed Fourier-based algorithms for learning an intersection of any constant number of halfspaces under the uniform distribution over the Boolean hypercube via Fourier analysis (albeit with complexity exponential in the inverse of the accuracy parameter ϵ). Many subsequent works in the distribution-specific setting developed algorithms with improved complexity under the Gaussian or uniform distribution over the hypercube. It remains an open question whether, under these strong distributional assumptions, an intersection of k halfspaces can be learned in fully-polynomial time. For the special case of two halfspaces, (Baum, 1990; Klivans et al., 2009) developed polynomial-time algorithms for learning an intersection of

two homogeneous halfspaces under mean zero symmetric/isotropic log-concave distributions. We emphasize that the distributional assumptions of these two works are fairly strong and the underlying algorithms do not work if the defining halfspaces do not go through the origin.

In the distribution-free setting, much less is understood. (Klivans and Sherstov, 2007, 2009; Daniely and Shalev-Shwartz, 2016; Tiegel, 2024) showed that if the number of halfspaces $k = \omega_d(1)$, then distribution-free learning is hard. Tiegel (2024) recently gave an SQ lower bound of $d^{\Omega(k)}$ for distribution-free learning intersections of k halfspaces. However, no super-polynomial SQ lower bound (or any other representation-independent hardness result) is known for learning intersections of a constant number of halfspaces.

Independent Component Analysis and Its Generalization The learning setting where the marginal distribution is factorizable is closely related to the work of Vempala and Xiao (2011) on an unsupervised learning setting, known as generalized independent component analysis. Independent component analysis (Jutten and Herault, 1991), originally considered the following problem: given examples $y \in \mathbb{R}^d$ generated by $y = A\mathbf{x}$, where $A \in \mathbb{R}^{d\times d}$ is an unknown matrix and $\mathbf{x} \in \mathbb{R}^d$ is a random vector such that $\mathbf{x}_i, i \in [d]$ are independent, recover the underlying direction $\mathbf{x}_1, \ldots, \mathbf{x}_d$. ICA is a natural generalization of PCA, another task that identifies the source components given only their linear combinations. PCA can be viewed as the task of finding vectors on the unit sphere that are local optima of the second moment of the observed data. Such an approach fails when eigenvalues repeat and ICA bypasses the difficulty by considering finding a local optimum of functions related to higher-order moments of observed data. It is known that the original ICA problem can be efficiently solved via second-order method (Frieze et al., 1996; Arora et al., 2012). Generalized ICA (Vempala and Xiao, 2011) instead aims to recover the distribution D_V given examples generated from a factorizable distribution $D_X = D_V \times D_W$. Such a problem can also be viewed as a generalization of Non-Gaussian Component Analysis(NGCA) (Tan and Vershynin, 2018; Goyal and Shetty, 2019). However, unlike the original ICA problem, the generalized ICA suffers issues of numerical instability, and no fully polynomial-time algorithm is known for the problem so far.

SQ Model and CSQ Model The CSQ model (Bshouty and Feldman, 2002) is a subset of the SQ model (Kearns, 1998), where the oracle access is of a special form (see Appendix A). In particular, any SQ query function $q_{\text{sq}} : X \times \{0, 1\} \to [-1, 1]$ can always be decomposed to $q_{\text{sq}}(\mathbf{x}, y) = q_1(\mathbf{x}) + q_2(\mathbf{x}, y)$, where $q_1(\mathbf{x})$ is a query function independent of the label y, and $q_2(\mathbf{x}, y)$ is a CSQ query. An intuitive interpretation is that, compared with the SQ model, the CSQ model loses exactly the power to make label-independent queries about the distribution, i.e., the power to ask queries about the marginal distribution of $\mathbf{x}$. In the context of learning Boolean-valued functions, the two models are known to be equivalent in the distribution-specific setting (i.e., when the marginal distribution on feature vectors is known to the learner) (Bshouty and Feldman, 2002). However, they are not in general equivalent in the distribution-free PAC model. In the realizable PAC setting, there are known natural separations between the CSQ and SQ models. Notably, Feldman (2011) showed that Boolean halfspaces are not efficiently learnable up to an arbitrarily small accuracy via CSQ algorithms (even though they are efficiently SQ learnable). Intuitively, such a separation exists because even though CSQ algorithms can be used to learn a weak hypothesis, without using stronger SQs, we cannot implement boosting algorithms to get a strong hypothesis. Our results for learning intersections of two halfspaces exhibit a new separation between CSQ and SQ models in the realizable PAC learning setting, in terms of weak learning. That is to say, under the assumption of factorizable distributions, it is hard to efficiently find a hypothesis with error $1/2 - o(1)$ using CSQ

algorithms; but this is possible via efficient SQ algorithms. This shows the necessity of using SQ queries for PAC learning is not only due to boosting, but also due to the hardness of finding a weak hypothesis.

Organization In Section 2, we present our main structural result for learning intersections of two halfspaces. In Section 3, we make use of the structural results developed in Section 2 to design computationally efficient algorithms that find a direction close to V. In Section 4, we show how to use the direction we find in Section 3 to obtain an efficient weak learner.

2. Structural Result for Distribution-Free Learning Intersections of Two Halfspaces

By Theorem 2, we know that without looking at the marginal distribution D_X, it is impossible to find a Correlational Statistical Query (CSQ) that can detect the correlation between D_X and D_y; thus, even performing weak learning efficiently via CSQs is impossible. To bypass this inherent limitation of CSQ algorithms, we need to design "instance-dependent" CSQs by first looking at the marginal D_X. Motivated by this intuition, we provide a novel structural result for learning intersections of two halfspaces, which we will make essential use of in Section 3 to design instance-dependent statistical queries. To start with, we present the following (α, m)-moment matching condition that will be heavily discussed throughout the paper.

Definition 5 ((α, m)-moment matching condition) *Let $\alpha \geq 0$, $m \in \mathbb{Z}_+$, and D be a distribution of $(\mathbf{x}, y)$ over $\mathbb{R}^d \times \{\pm 1\}$. We say that D satisfies the (α, m)-moment matching condition if $\left\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\,\mathbf{x}^{\otimes m}\right\|_F \leq \alpha$.*

To simplify the notation, in this section we make the following assumption about the subspace V. We assume that V, the subspace spanned by $\mathbf{u}^*, \mathbf{v}^*$, is exactly equal to $\mathbf{span}\{\mathbf{e}_1, \mathbf{e}_2\}$. Such an assumption can be made without loss of generality, as applying a rotation transformation will not affect the γ-margin assumption. Since the label of an example $\mathbf{x}$ only depends on its projection $\mathbf{x}_V$ onto V, to simplify notation in this section, we restrict attention to the dimensions of V and consider examples $\mathbf{x} \in \mathbb{R}^2$ drawn from D^+ (or D^-). Our main structural result, Theorem 6, shows that if D^+, D^- have nearly matched their first three moments, and $\mathbf{E}_{\mathbf{x}\sim D_X}(\mathbf{x})$ is close to 0, then the third-moment tensors of D^+, D^- must significantly deviate from 0.

Theorem 6 *Let D be a distribution over $\mathbb{B}^2(1) \times \{\pm 1\}$ that is consistent with an instance of learning intersections of two halfspaces with γ-margin. Let $c > 0$ be* any *suitably large constant. Suppose that D satisfies the (γ^c, m)-moment matching condition for $m \in [3]$, and $\|\mathbf{E}_{\mathbf{x}\sim D^+}\,\mathbf{x}\|_F \leq \gamma^c$. Then $\left\|\mathbf{E}_{\mathbf{x}\sim D^+}\,\mathbf{x}^{\otimes 3}\right\|_F, \left\|\mathbf{E}_{\mathbf{x}\sim D^-}\,\mathbf{x}^{\otimes 3}\right\|_F = \Omega(\gamma^{15})$.*

Here we present some high-level intuition for the proof of Theorem 6. The formal proof is given in Appendix B.7. To prove Theorem 6, we establish two technical lemmas. The first lemma, Lemma 8, shows that the conditions in Theorem 6 imply that the smallest eigenvalue of the covariance matrix of both D^+ and D^- will be at least γ^c. Thus, by properly rescaling D (by at most a $\mathrm{poly}(\gamma)$ factor), we can create a new distribution D' such that both $(D')^+$ and $(D')^-$ have isotropic covariance matrices and the distribution D' still satisfies a γ'-margin condition with respect to an intersection of two halfspaces, where $\gamma' = \gamma^{c'}$. Now assuming, for the purpose of contradiction, that $\left\|\mathbf{E}_{\mathbf{x}\sim D^+}\,\mathbf{x}^{\otimes 3}\right\|_F, \left\|\mathbf{E}_{\mathbf{x}\sim D^-}\,\mathbf{x}^{\otimes 3}\right\|_F < \mathrm{poly}(\gamma)$, we must also have $\left\|\mathbf{E}_{\mathbf{x}\sim D'^+}\,\mathbf{x}^{\otimes 3}\right\|_F, \left\|\mathbf{E}_{\mathbf{x}\sim D'^-}\,\mathbf{x}^{\otimes 3}\right\|_F < \mathrm{poly}(\gamma)$ on the new distribution D' as well. This assumption

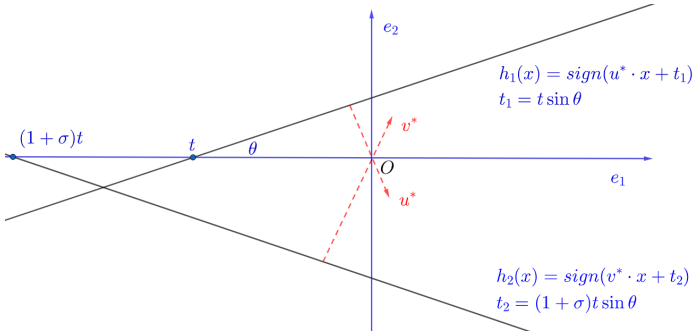

Figure 1: Geometry of Intersection of Two Halfspaces under Assumption 1.

implies that both $(D')^+$ and $(D')^-$ must have their first and third moment tensors roughly 0. Our second lemma, Lemma 9, shows that the scale of the covariance matrices of D'^+ and D'^- must be significantly different, which contradicts the previous assumption that the first three moments of D'^+ and D'^- are nearly matched. The proofs of these technical lemmas rely on a novel technique, which we term *polynomial one-sided approximation*, that leverages weak duality between distributions with specific moments and polynomial certificates. We summarize the property of one-sided polynomial approximation in the following theorem and defer its proof to Appendix B.1.

Theorem 7 (Polynomial One-sided Approximation) *For any $d, m \in \mathbb{N}$, $C \subseteq \mathbb{R}^d$, $T_i \in (\mathbb{R}^d)^{\otimes i}$ for $i \in [m]$, and $\tau \in \mathbb{R}_{\geq 0}$, at most one of the following conditions can be satisfied:*
a) There is a distribution D supported on C such that $\| \mathbf{E}_{\mathbf{x}\sim D}(\mathbf{x}^{\otimes i}) - T_i\|_F \leq \tau$ for any $i \in [m]$;
b) There is a degree-m polynomial $p : \mathbb{R}^d \to \mathbb{R}$, defined as $p(\mathbf{x}) = \sum_{i=1}^m A_i \cdot \mathbf{x}^{\otimes i}$, where $p(\mathbf{x}) \geq 0$ for any $\mathbf{x} \in C$, $\sum_{i=1}^m \|A_i\|_F \leq 1$ and $\sum_{i=1}^m A_i \cdot T_i < -\tau$.
We call such a polynomial a one-sided approximation for C w.r.t. to moments T_i and tolerance τ.

Given Theorem 7, in order to certify the non-existence of certain distributions on C with specific moments, it suffices to construct a one-sided approximating polynomial, as stated in Theorem 7. However, a direct application of Theorem 7 may not be intuitive. In the rest of the section, we explain how to use this idea to prove our two technical lemmas. To give a clearer intuition, it is convenient to parameterize the instance of learning intersections of two halfspaces as in Assumption 1 (see Figure 1 for a geometric illustration). We defer further discussion to Appendix B.2 showing that such a parameterization can be made without loss of generality.

Assumption 1 *Given an intersection of two halfspaces $h^* = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ and a distribution D over $\mathbb{B}^2(1) \times \{\pm 1\}$ satisfying the γ-margin condition w.r.t. h^*, we parameterize h^* by an angle $\theta \in (0, \pi/2)$, and thresholds $t \geq 0, \sigma \geq 0$, where $\mathbf{u}^* = \sin\theta \mathbf{e}_1 - \cos\theta \mathbf{e}_2, t_1 = t \sin\theta$ and $\mathbf{v}^* = \sin\theta \mathbf{e}_1 + \cos\theta \mathbf{e}_2, t_2 = (1+\sigma)t \sin\theta$ such that $\|\mathbf{E}_{\mathbf{x}\sim D^+} \mathbf{x}\| \leq \gamma^c$, $t_1, t_2 \geq \gamma$, $|t_1|, |t_2| \leq 1$.*

Moment-matched Distributions Cannot Have Ill-Conditioned Covariance Matrices Our first technical lemma (Lemma 8) shows that if D^+, D^- have nearly matched their first two moments, then the covariance matrices of D^+, D^- cannot have small eigenvalues.

Lemma 8 *Let D be a distribution over $\mathbb{B}^2(1) \times \{\pm 1\}$ that is consistent with an instance of learning intersections of two halfspaces with γ-margin, where γ is smaller than some sufficiently small*

constant. Let $c > 0$ be any suitably large constant. Suppose that D satisfies the (γ^c, m)-moment matching condition for $m \in [2]$ and $\|\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}\|_F \le \gamma^c, \|\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}\|_F \le \gamma^c$. Then $\left\|(\Sigma^+)^{-1}\right\|_2 \le O(1/\gamma^4)$ and $\left\|(\Sigma^-)^{-1}\right\|_2 \le O(1/\gamma^4)$, where $\Sigma^+ := \mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}\mathbf{x}^T$ and $\Sigma^- := \mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}\mathbf{x}^T$.

The proof strategy of Lemma 8 is to construct a suitable polynomial function as a certificate, as stated in Theorem 7. We will use the polynomial function $f(\mathbf{x}) = (\mathbf{u}^* \cdot \mathbf{x} + t_1)(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ as a certificate. Intuitively, if the first two moments of D^+, D^- are nearly matched, then $\mathbf{E}_{\mathbf{x}\sim D^+} f(x) \approx \mathbf{E}_{\mathbf{x}\sim D^-} f(\mathbf{x})$. Furthermore, by the γ-margin assumption, for every positive example $\mathbf{x}$, $f(\mathbf{x}) \ge \gamma^2$ and for every negative example $\mathbf{x}$ with $h_1(\mathbf{x})h_2(\mathbf{x}) = -1$, $f(\mathbf{x}) \le -\gamma^2$. This implies that if the probability $\mathbf{Pr}_{\mathbf{x}\sim D^-}(h_1(\mathbf{x}) = h_2(\mathbf{x}))$ is small, then $\mathbf{E}_{\mathbf{x}\sim D^+} f(\mathbf{x}) - \mathbf{E}_{\mathbf{x}\sim D^-} f(\mathbf{x}) \ge \Omega(\gamma^2)$, which gives a contradiction. Geometrically, if $\mathbf{Pr}_{\mathbf{x}\sim D^-}(h_1(\mathbf{x}) = h_2(\mathbf{x}))$ is large, then examples in this region must have a significant contribution to Σ^- along every direction $\mathbf{v}$ to make $\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}$ close to 0, which contradicts the fact that Σ^- has a direction with small variance.

Moment-matched Distributions Must Have Large Third Moments By Lemma 8, it is safe to assume that the marginal distribution D_X has an isotropic covariance matrix. Our main structural result, Lemma 9, shows that if D^+, D^- have nearly matched first three moments and their covariance matrices are nearly isotropic, then their third moments must significantly deviate from 0. We defer the proof of Lemma 9 to Appendix B.6.

Lemma 9 *Let D be a distribution over $\mathbb{B}^2(1) \times \{\pm 1\}$ that is consistent with an instance of learning intersections of two halfspaces with γ-margin. Let $c > 0$ be any suitably large constant. Suppose*

1. $\|\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}\|_F \le \gamma^c, \|\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}\|_F \le \gamma^c$.
2. $\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}\mathbf{x}^T = \alpha^2 I + \Delta_+, \mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}\mathbf{x}^T = \alpha^2 I + \Delta_-$, *where $\Delta_+, \Delta_- \in \mathbb{R}^{2\times 2}$ are symmetric matrices such that $\|\Delta_+\|_F \le \gamma^c, \|\Delta_-\|_F \le \gamma^c$ and $\alpha^2 > 0$.*
3. $\left\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\,\mathbf{x}^{\otimes 3}\right\|_F \le \gamma^c$.

Then we have $\left\|\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}^{\otimes 3}\right\|_F \ge \Omega(\gamma^2), \left\|\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}^{\otimes 3}\right\|_F \ge \Omega(\gamma^2)$.

The intuition behind Lemma 9 relies on Fact 1 and Fact 2 below that characterize the covariance matrix of *any* pair of distributions D^+, D^- with zero mean and zero third moment tensor.

Fact 1 *Let $h^* = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ be the target hypothesis of an instance of the problem of learning intersections of two halfspaces, and D be a distribution that is consistent with h^*. Under Assumption 1, if $\mathbf{E}_{\mathbf{x}\sim D^+}(\mathbf{x}) = 0, \mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}^{\otimes 3} = 0$ and for every $\mathbf{v} \in S^{d-1} \cap V$, it holds $\mathbf{E}_{\mathbf{x}\sim D^+}(\mathbf{v}\cdot\mathbf{x})^2 = \alpha^2$, then $\alpha^2 \le t^2 \sin^2\theta$.*

Fact 2 *Let $h^* = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ be the target hypothesis of an instance of the problem of learning intersections of two halfspaces, and D be a distribution that is consistent with h^*. Under Assumption 1, if $\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x} = 0, \mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}^{\otimes 3} = 0$ and for every $\mathbf{v} \in S^{d-1} \cap V$, it holds $\mathbf{E}_{\mathbf{x}\sim D^-}(\mathbf{v}\cdot\mathbf{x})^2 = \beta^2$, then $\beta^2 \ge (1+\sigma)t^2 \tan^2\theta$.*

Here we give an overview of the proof techniques behind Fact 1 and Fact 2, and defer the full proofs to Appendix B.5. We take Fact 1 as an example and show how the certificate one-sided approximating polynomial used in Theorem 7 is derived. For every $\mathbf{x}$ that is labeled positive by h^*, denote by $p(\mathbf{x})$ the variable of the density of a distribution D^+ over $\mathbb{R}^2$. Notice that any distribution

D^+ that satisfies the statement of Fact 1 gives a feasible solution to the following LP (1). Thus, to upper bound the variance of D^+, it is equivalent to upper bound the optimal value of LP (1).

$$\begin{aligned} \max\ \alpha^2 \quad \text{s.t.}\ & \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x} = 0, \quad \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}\mathbf{x}^T = \alpha^2 I, \\ & \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}^{\otimes 3} = 0, \quad \sum_{\mathbf{x}} p(\mathbf{x}) = 1, \quad p(\mathbf{x}) \geq 0, \forall \mathbf{x} \in \mathbf{supp}(D^+) \end{aligned} \tag{1}$$

To upper bound the optimal value of LP (1), we use LP duality theory (Bertsimas and Tsitsiklis, 1997; Shapiro, 2001). The dual linear program to LP (1) is defined by LP (2), whose variable is defined over the coefficients of $f(\mathbf{x})$, a degree-3 polynomial over $\mathbb{R}^2$, and the objective function is given by its constant term, namely

$$\min\ a_0 \quad \text{s.t.}\ \forall \mathbf{x} \in \mathbf{supp}(D^+),\ f(\mathbf{x}) \geq 0, \quad a_{11} + a_{22} = -1. \tag{2}$$

Here, a_{11}, a_{22} are the coefficients of f with respect to monomials $\mathbf{x}_1^2, \mathbf{x}_2^2$. Thus, to give tight bounds for α^2, β^2, the key technical difficulty is to design a pair of polynomials that are feasible to the dual LPs with nearly optimal objective values. The polynomials we used here are given by Lemma 10 and Lemma 11 (illustrated in Figure 1), the proofs of which are deferred to Appendix B.4.

Lemma 10 *Let $h^* = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ be the target hypothesis of an instance of the problem of learning intersections of two halfspaces and D be a distribution that is consistent with h^*. Under Assumption 1, the polynomial $f^*(\mathbf{x}) = \frac{1}{t \sin\theta}(\mathbf{u}^* \cdot \mathbf{x} - t\sin\theta)^2(\mathbf{u}^* \cdot \mathbf{x} + t\sin\theta)$ satisfies $f^*(\mathbf{x}) \geq 0, \forall \mathbf{x} \in \boldsymbol{supp}(D^+)$.*

Lemma 11 *Let $h^* = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ be the target hypothesis of an instance of the problem of learning intersections of two halfspaces and D be a distribution that is consistent with h^*. Under Assumption 1, the polynomial $f^*(\mathbf{x}) = a_0 + a_1\mathbf{x}_1 + a_2\mathbf{x}_2 - \mathbf{x}_2^2$, where $a_0 = (1+\sigma)\tan^2\theta t^2$, $a_1 = (2+\sigma)\tan^2\theta t$ and $a_2 = -\sigma\tan\theta t$ satisfies $f^*(\mathbf{x}) \leq 0, \forall \mathbf{x} \in \boldsymbol{supp}(D^-)$.*

Given Fact 1 and Fact 2, under the γ-margin assumption, we know that if D^+, D^- have zero means, zero third moments, and isotropic covariance matrices, then the variances of D^+, D^- must differ by at least γ^2. In other words, if D^+, D^- have zero means and matched second, third moments, then their third moments must differ from 0 by $\Omega(\gamma^c)$. However, in general, we are not able to guarantee that the moments of D^+, D^- satisfy the condition of Fact 1 and Fact 2 exactly. Thus, we need Lemma 9, a robust version of the above argument. Importantly, the polynomials we construct in Lemma 10 and Lemma 11 are stable enough and can still be used in our proof even if the moment conditions are perturbed. Thus, using these one-sided approximating polynomials as certificates, we are able to prove Lemma 9.

3. Relevant Direction Extraction for Intersections of Two Halfspaces

In Section 2, we developed structural results for the marginal distribution D_V of any instance of learning intersections of two halfspaces with a margin assumption. In this section, we will show that with these structural results, we are able to efficiently find a unit vector $\mathbf{w}$ that is close to V for *any* factorizable distribution D consistent with an instance of learning intersections of two halfspaces. As we will show later, with such a unit vector $\mathbf{w}$, we are able to design non-smoothed statistical queries that can be used for weak learning. Recall by Theorem 2 that a necessary condition that makes CSQ

algorithms not work is that low-degree moments of D^+, D^- are nearly the same. So, in Section 3.1, we will present algorithms that work under this "hard" condition, while in Section 3.2, we will give an algorithm under the easier condition where the low-degree moments of D^+, D^- are mismatched.

3.1. Relevant Direction Extraction with Matched Moments

When D^+, D^- have nearly the same low-degree moments, the moments of both of D^+, D^- look like those of D_X. As we mentioned in Section 2, in this case, the third moment of D_X must deviate from 0 significantly. In this step, we will make use of this property to perform certain unsupervised learning tasks over D_X to find some $\mathbf{w} \in S^{d-1}$ close to V.

An Efficient SQ Algorithm for Relevant Direction Extraction with Matched Moments Our initial attempt was inspired by independent component analysis (ICA) (Frieze et al., 1996; Arora et al., 2012) and its generalization (Vempala and Xiao, 2011). The generalized ICA studies the following problem. Consider a distribution D_X over $\mathbb{R}^d$, where there is a k-dimensional subspace V such that D_V and D_W ($D_{V^\perp}$) are independent. Given sample access to D_X, ICA is asked to recover the subspaces V and W. The core idea of generalized ICA is to solve some non-convex optimization task based on higher moments of D_X. Vempala and Xiao (2011) observed that if D_X has the same first $m-1$ moments as the standard Gaussian, for $m \geq 3$, but has a different mth moment, then any local maximum (minimum) $\mathbf{u}^*$ of $f^*(\mathbf{u}^*)$ over S^{d-1} with $f^*(\mathbf{u}^*) > \gamma_m$ ($f^*(\mathbf{u}^*) < \gamma_m$) must be either in V or W. Here, $f^*(\mathbf{u}) = \mathbf{E}_{\mathbf{x}\sim D_X}(\mathbf{u}\cdot\mathbf{x})^m$ and γ_m is the m-th moment of the standard normal distribution. In particular, if the distribution D_W is a standard Gaussian, then any $\mathbf{u}^*$ obtained above must be in V, which gives a reasonable method that finds one direction in V.

However, for two general distributions D_V, D_W, this observation does not immediately give a method to find a direction $\mathbf{u} \in V$, because we are not able to guarantee whether the local optimum of $f^*(\mathbf{u})$ is in V or W. In fact, for the problem of learning intersections of two halfspaces, V only has dimension 2; but it is possible that D_W is also a factorizable distribution that can be factorized into $\Omega(d)$ distributions, each of which is isomorphic to D_V. Thus, information-theoretically, finding a list of $O(d)$ directions such that one of them is close to V is the best one can achieve. To do this, the direct attempt is to find the first local optimum $\mathbf{u}^{(1)}$, look at the subspace $(\mathbf{u}^{(1)})^\perp$, find $\mathbf{u}^{(2)}$, the next local optimum of f^* within $(\mathbf{u}^{(1)})^\perp$, and perform this procedure recursively d times. This can be done because every time we make a projection, the resulted distribution is still factorizable. Unfortunately, such a direct approach cannot be turned into an efficient algorithm, and no fully polynomial time algorithm for generalized ICA is known so far. This is because, due to the sampling error and optimization error for optimizing f^*, we are only able to find an approximate solution for $\mathbf{u}^{(1)}$, which is not in V or W exactly. Thus, the local optimum of f^* restricted at $(\mathbf{u}^{(1)})^\perp$ is not guaranteed to be a local optimum of f^* over S^{d-1}. Such an error can accumulate exponentially fast with respect to the order in the output list (as demonstrated in Vempala and Xiao (2011)). Since we are not able to guarantee which $\mathbf{u}$ in the list is close to V, in the worst case, our target direction could be the last few discovered directions in the list. To guarantee that these vectors are still close to V, the first several solutions must be found with error $\gamma^{-\Omega(d)}$. Our first result in this section shows that, although such a framework cannot give us a computationally efficient algorithm, we can modify it to get an SQ-efficient algorithm that outputs a list of $\mathrm{poly}(d/\gamma)$ unit vectors such that at least one of them is $\mathrm{poly}(\gamma/d)$-close to V. In other words, we show that extracting one relevant direction can be done in a sample-efficient manner. Formally, we establish the following theorem (we defer the algorithm and the proof to Appendix C.1).

Theorem 12 *There is a Statistical Query learning algorithm $\mathcal{A}$ such that for $c > 0$, a suitably large constant, and for an instance of learning intersections of two γ-margin halfspaces under factorizable distributions, if the input distribution D satisfies the (γ^c, m)-moment matching condition for $m \in [3]$, the algorithm makes $\mathrm{poly}(d/\gamma)$ many statistical queries, each of which has tolerance $\mathrm{poly}(\gamma/d)$, and outputs a direction $\mathbf{w} \in \mathbb{R}^d$ such that with probability at least $\mathrm{poly}(\gamma/d)$, $\|\mathbf{w}_W\|_2 \leq \mathrm{poly}(\gamma/d)$.*

Computationally Efficient Algorithm for Relevant Direction Extraction with Matched Moments Given the above discussion, finding a relevant direction via a direct non-convex optimization method is technically challenging. In summary, since there is no structural assumption over D_W, the function $f(\mathbf{u}) = \mathbf{E}_{\mathbf{x}\sim D_X}(\mathbf{u}\cdot\mathbf{x})^3$ could have too many locally optimal solutions and some of them (including the ones that are close to V) are hard to find; this makes the error accumulate fast when sequentially finding each local optimum. Thus, to avoid such an issue of error accumulation and get a computationally efficient algorithm, one hope is to find directions in V and W simultaneously. Following this idea, we give a fully-polynomial time algorithm that solves this task using techniques from the tensor decomposition literature. Formally, we have the following theorem.

Theorem 13 *There is a learning algorithm $\mathcal{A}$ such that for every c, a suitably large constant, and any instance of learning intersections of two γ-margin halfspaces under factorizable distributions, if the input distribution D satisfies the (γ^c, m)-moment matching condition for $m \in [3]$, $\mathcal{A}$ runs in $\mathrm{poly}(d, 1/\gamma)$ time and outputs a list of d unit vectors $\mathcal{O}$ such that at least one direction $\mathbf{w} \in \mathcal{O}$ satisfies $\|\mathbf{w}_W\|_2 \leq \mathrm{poly}(\gamma)$ with probability $\Omega(\gamma/d)$.*

Tensor decomposition techniques usually deal with problems of the following type. Given a tensor $T \in \mathbb{R}^{d\times d\times d}$ of the form $T = \sum_{i=1}^k (\mathbf{v}^{(i)})^{\otimes 3}$, recover $\mathbf{v}^{(i)}, i \in [k]$ for some small k. A number of prior works address this problem from a computational point of view. Unfortunately, for the moment tensor of a general distribution, k can be large and it can be challenging to compute the decomposition. This also happens for our problem. However, our goal is to find a direction $\mathbf{w} \in V$, instead of doing a complete tensor decomposition. Assuming that D_X has zero mean, then we can write $T^* := \mathbf{E}_{\mathbf{x}\sim D}\,\mathbf{x}^{\otimes 3} = \mathbf{E}_{\mathbf{x}\sim D_V}\,\mathbf{x}_V^{\otimes 3} + \mathbf{E}_{\mathbf{x}\sim D_W}\,\mathbf{x}_W^{\otimes 3}$. Notice that for every $\mathbf{v} \in \mathbb{R}^d$, we have $M = T^*\cdot\mathbf{v} = \mathbf{E}_{\mathbf{x}\sim D_V}\,\mathbf{x}_V\mathbf{x}_V^T(\mathbf{x}_V\cdot\mathbf{v}) + \mathbf{E}_{x\sim D_W}\,\mathbf{x}_W\mathbf{x}_W^T(\mathbf{x}_W\cdot\mathbf{v}) = M_V + M_W$, where $M_V = \mathbf{E}_{\mathbf{x}\sim D_V}\,\mathbf{x}_V\mathbf{x}_V^T(\mathbf{x}_V\cdot\mathbf{v})$ and $M_W = \mathbf{E}_{\mathbf{x}\sim D_W}\,\mathbf{x}_W\mathbf{x}_W^T(\mathbf{x}_W\cdot\mathbf{v})$. Since $V \perp W$, every eigenvector $\mathbf{w}$ of M_V must also be an eigenvector of M. Thus, as long as M_W does not have a common eigenvalue as M_V, we are able to find one direction $\mathbf{w} \in V$ using eigendecomposition algorithms. On the other hand, if the eigenvalues of M_V are the same as (or close to) the eigenvalues of M_W, vectors that have heavy components in both V and W can also be eigenvectors of M, which makes finding $\mathbf{w} \in V$ hard. To overcome this difficulty, we choose $\mathbf{v} \sim N(0, \frac{1}{d}I)$. Such a choice makes the eigenvalues of M_V and M_W are independent. Importantly, by Theorem 6, T_V significantly deviates from 0. Thus, if we write $\mathbf{v}_V = \alpha^2\mathbf{v}_V^0$, where $\mathbf{v}_V^0$ is the direction of $\mathbf{v}_V$, then with constant probability M_V/α^2 has at least one eigenvalue σ_1 with magnitude at least γ^c. On the other hand, the corresponding eigenvalue $\alpha^2\sigma_1$ of M_V is a random variable that satisfies an anti-concentration property. In the proof of Theorem 13, we will show that this anti-concentration property can make $\alpha^2\sigma_1$ far away from any eigenvalue of M_W with a non-trivial probability. Thus, as long as we estimate the moment tensor of T^* up to $\mathrm{poly}(\gamma/d)$ accuracy, we are able to find a direction $\mathbf{w}$ close to V with a non-trivial probability. We defer the algorithm and the proof of Theorem 13 to Appendix C.2.

3.2. Relevant Direction Extraction with Mismatched Moments

In Section 3.1, we described an efficient algorithm that outputs a direction $\mathbf{w}$ that is close to V when D^+ and D^- have nearly matched low-degree moments. In this section, we focus on a different regime, where the low-degree moments of D^+, D^- are not matched. Recall that in Definition 5, we use the (α, m)-moment matching condition to measure the level of mismatch of the low-degree moments of D^+, D^-. This characterizes the difficulty of using polynomials to detect the correlation between the labels and the unlabeled examples. If for small m, the (α, m) moment matching condition always does not hold, then one can use the polynomial regression method to output a degree-m Polynomial Threshold Function (PTF) with $\mathrm{poly}(\alpha)$ advantage. However, this does not imply that we are able to run a boosting algorithm with PTFs. Indeed, this would require the moment-matching condition to not hold throughout the process. However, when the distribution D is factorizable, instead of boosting using polynomials, our strategy will be to extract a direction $\mathbf{u}$ in the relevant subspace V by solving a carefully defined non-convex optimization problem. The main result we obtain is summarized in Theorem 14. We defer the full proof of Theorem 14 and the corresponding algorithm to Appendix D.

Theorem 14 *There is an algorithm $\mathcal{A}$ (Algorithm 4) such that for any instance of learning intersections of two halfspaces under factorizable distributions, if the distribution D does not satisfy the (α, m)-moment matching condition and D satisfies the $(\alpha^2 d^{-c}/2^m, t)$-moment matching condition for any $t \leq m-1, m \leq 3$ and a sufficiently large universal constant c, then $\mathcal{A}$ draws $\mathrm{poly}(d, 1/\alpha)$ i.i.d. samples from D, runs in time $\mathrm{poly}(d, 1/\alpha)$, and outputs a unit vector $\mathbf{u} \in S^{d-1}$ such that $\|\mathbf{u}_W\| = O(\alpha)$ with probability $2/3$.*

In the rest of the section, we provide an overview of the proof of Theorem 14. Suppose that $\mathbf{E}_{\mathbf{x}\sim D^+}(\mathbf{x}_V^{\otimes t})$ and $\mathbf{E}_{\mathbf{x}\sim D^-}(\mathbf{x}_V^{\otimes t})$ are different, for some $t = m \in \mathbb{N}$. Then the tensor $T = (\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})(\mathbf{x}_V^{\otimes m})$ is nonzero and $T \cdot \mathbf{u}^{\otimes m} = 0$ for all points $\mathbf{u} \in W$. Therefore, the function $f(\mathbf{u}) = \mathbf{u}^{\otimes m} \cdot T$ obtains a local maximum/minimum only if $\mathbf{u}^{\otimes m} \in V^{\otimes m}$, which implies that $\mathbf{u} \in V$. Unfortunately, we are not able to solve such an optimization problem exactly. As we will show in the proof, an approximate solution (that can be efficiently found via a gradient-descent method) suffices for our purposes. Moreover, there is still a technical challenge to implement this approach. Since V is unknown to us, it is impossible for us to estimate T. However, if $\mathbf{E}_{\mathbf{x}\sim D^+}(\mathbf{x}^{\otimes t})$ and $\mathbf{E}_{\mathbf{x}\sim D^-}(\mathbf{x}^{\otimes t})$ are the same for all $t \leq m-1$, then $(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})(\mathbf{x}^{\otimes m}) = T$, for which we can efficiently estimate with samples. We will show that if we take m to be the smallest index such that D does not satisfy the (α, m)-moment matching condition, but satisfies the $(\alpha^2 d^{-c}/2^m, t)$-moment matching condition, then we are able to take $\mathrm{poly}(d^m, 1/\alpha)$ examples from D^+, D^- to estimate T, so that any approximate solution to the estimated function gives a direction close to V. In particular, to use this approach to learn an intersection of two halfspaces, we only need $m \leq 3$.

4. Localization with the Relevant Direction and Learning Intersections of Halfspaces

In the previous sections, we have shown that for every factorizable distribution D that is consistent with an instance of learning intersections of two halfspaces with a margin, we are able to efficiently find one direction $\mathbf{w}$ that is close to the relevant subspace V. Based on these results, a natural attempt is to find the next relevant direction so that we can approximately find the relevant subspace V; and do a brute-force search over all intersections of two halfspaces over V. However, the structural

result we obtained in Section 2 only allows us to find one direction. Furthermore, since we make no distributional assumptions over D_V, to make the brute-force search method succeed, a small mismatch between V and the approximate subspace we found could lead to a large error. On the other hand, instead of trying to recover the relevant subspace, we make the following observation.

Lemma 15 *Let D be a joint distribution of $(\mathbf{x}, y)$ on $\mathbb{B}^d(1) \times \{\pm 1\}$ that is consistent with an intersection of halfspaces with γ-margin and $\mathbf{w} \in S^{d-1}$ such that $\|\mathbf{w}_V\|_2 \leq c\gamma$, for some small constant c, where V is the relevant subspace of the intersection of halspaces. Then for any band $B_t := \{\mathbf{x} \in \mathbb{B}^d(1) \mid \mathbf{x} \cdot \mathbf{w} \in [t, t + c\gamma]\}$ where $t \in \mathbb{R}$ and c is a sufficiently small constant, the distribution of $(\mathbf{x}, y)$ conditioned on $\mathbf{x} \in B_t$ is consistent with an instance of learning a degree-2 polynomial threshold function with $\Omega(\gamma^2)$-margin.*

Lemma 15 states that for any instance of learning an intersection of two halfspaces with a margin, if we cut the space into bands $B_i, i = 1, 2, \ldots$, along a direction $\mathbf{u}$ that is close to V, then the distribution in each band is consistent with a degree-2 PTF. This means that some correlational statistical query of the form $q(\mathbf{x}) = p(\mathbf{x})\mathbb{1}(\mathbf{x} \in B_i)$ can be used to detect the correlation between D_X and D_Y, and allows us to efficiently find a weak hypothesis h with $\mathrm{poly}(\gamma)$-advantage. We state the algorithmic result for the weak learning algorithm in Theorem 16.

Theorem 16 *There is an algorithm $\mathcal{A}$ such that for every instance of learning intersections of two halfspaces with γ-margin, given $\mathbf{w} \in S^{d-1}$ such that $\|\mathbf{w}_W\|_2 \leq c\gamma$ where c is a sufficiently small constant, $\mathcal{A}$ draws $\mathrm{poly}(d, 1/\gamma)$ examples from D, runs in $\mathrm{poly}(d, 1/\gamma)$ time, and outputs a hypothesis $h : \mathbb{B}^d(1) \to \{\pm 1\}$ such that with probability at least $2/3$, $\mathrm{err}(h) \leq 1/2 - \Omega(\gamma)$.*

We emphasize that Theorem 16 holds *without the assumption that D is factorizable*. This immediately implies that we are able to get an efficient strong learning algorithm via boosting algorithms (Schapire and Freund, 2013). Due to space limitations, we defer the proofs in this section to Appendix E.

5. Conclusion

The question of whether the intersection of two halfspaces with a margin can be learned in fully polynomial time is a central problem in Computational Learning Theory that has been open for over two decades. Our work makes progress on this problem by bypassing the previously known limitations through a novel algorithmic framework, yielding new techniques and structural insights. While our approach does not resolve the problem in full generality, the case of factorizable distributions that we consider is a fairly challenging setting and we expect that some of the ideas introduced here will be useful even beyond our factorization assumptions. The key and most difficult step in learning intersections of halfspaces is finding statistical queries that enable weak learning. Our approach to this is to identify a direction that is close to the relevant subspace. Most known learning lower-bounds involving statistical learning algorithms are based on hiding a subspace among irrelevant directions. Our results show that one can efficiently address these cases using SQ algorithms and establish that common SQ lower-bound constructions are not applicable to our setting. Thus, even if an SQ lower bound exists, it would require a novel construction with non-factorizable distributions. Furthermore, our algorithmic result establishes a strong separation between CSQ and SQ algorithms for *weakly* realizable PAC learning. While it is known that SQ is needed for efficient strong realizable PAC learning, our work gives the first natural setting where SQ is even necessary for efficient weakly learning. Our learning framework builds on such a separation, and we expect understanding such a separation may lead to faster algorithms for learning other hypothesis classes.

References

N. Alon. Problems and results in extremal combinatorics–i. *Discret. Math.*, 273(1-3):31–53, 2003.

S. Arora, R. Ge, A. Moitra, and S. Sachdeva. Provable ica with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. *Advances in Neural Information Processing Systems*, 25, 2012.

R. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine learning*, 63:161–182, 2006.

E. B. Baum. A polynomial time algorithm that learns two hidden unit nets. *Neural Computation*, 2 (4):510–522, 1990.

D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

H. Block. The perceptron: A model for brain functioning. i. *Reviews of Modern Physics*, 34(1):123, 1962.

A. Blum. Relevant examples and relevant features: Thoughts from computational learning theory. In *AAAI Fall Symposium on 'Relevance*, volume 5, page 1, 1994.

A. Blum and R. Kannan. Learning an intersection of a constant number of halfspaces over a uniform distribution. *Journal of Computer and System Sciences*, 54(2):371–380, 1997.

N. Bshouty and V. Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2:359–395, 2002.

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

A. Daniely and S. Shalev-Shwartz. Complexity theoretic limitations on learning dnf's. In *Conference on Learning Theory*, pages 815–830. PMLR, 2016.

I. S Dhillon, B. N Parlett, and C. Vömel. The design and implementation of the mrrr algorithm. *ACM Transactions on Mathematical Software (TOMS)*, 32(4):533–560, 2006.

I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pages 73–84, 2017.

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*, 2021.

I. Diakonikolas, D. M. Kane, L. Ren, and Y. Sun. SQ lower bounds for non-gaussian component analysis with weaker assumptions. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2023.

V. Feldman. Distribution-independent evolvability of linear threshold functions. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 253–272. JMLR Workshop and Conference Proceedings, 2011.

V. Feldman. Statistical query learning. In *Encyclopedia of Algorithms*, pages 2090–2095. 2016.

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *J. ACM*, 64(2):8:1–8:37, 2017a.

V. Feldman, C. Guzman, and S. S. Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In Philip N. Klein, editor, *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017*, pages 1265–1277. SIAM, 2017b.

A. Frieze, M. Jerrum, and R. Kannan. Learning linear transformations. In *Proceedings of 37th Conference on Foundations of Computer Science*, pages 359–368. IEEE, 1996.

N. Goyal and A. Shetty. Non-gaussian component analysis using entropy methods. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 840–851, 2019.

C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10, 1991.

M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, 1998.

S. Khot and R. Saket. On hardness of learning intersection of two halfspaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 345–354, 2008.

A. Klivans and R. Servedio. Learning intersections of halfspaces with a margin. *Journal of Computer and System Sciences*, 74(1):35–48, 2008.

A. Klivans, R. O'Donnell, and R. Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences*, 68(4):808–840, 2004.

A. Klivans, R. O'Donnell, and R. Servedio. Learning geometric concepts via gaussian surface area. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 541–550. IEEE, 2008.

A. R Klivans and R. A Servedio. Learning intersections of halfspaces with a margin. In *Learning Theory: 17th Annual Conference on Learning Theory, COLT 2004, Banff, Canada, July 1-4, 2004. Proceedings 17*, pages 348–362. Springer, 2004a.

A. R Klivans and R. A Servedio. Perceptron-like performance for intersections of halfspaces. In *International Conference on Computational Learning Theory*, pages 639–640. Springer, 2004b.

A. R Klivans and A. A Sherstov. Unconditional lower bounds for learning intersections of halfspaces. *Machine Learning*, 69(2):97–114, 2007.

A. R Klivans and A. A Sherstov. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009.

A. R. Klivans, P. M Long, and A. K Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 588–600. Springer, 2009.

S. Kwek and L. Pitt. Pac learning intersections of halfspaces with membership queries. In *Proceedings of the ninth annual conference on Computational learning theory*, pages 244–254, 1996.

P. M Long and M. K Warmuth. Composite geometric concepts and polynomial predictability. *Information and Computation*, 113(2):230–252, 1994.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

R. E. Schapire and Y. Freund. Boosting: Foundations and algorithms. *Kybernetes*, 42(1):164–166, 2013.

A. Shapiro. *On Duality Theory of Conic Linear Problems*, pages 135–165. Springer US, Boston, MA, 2001. ISBN 978-1-4757-3403-4. doi: 10.1007/978-1-4757-3403-4_7. URL https://doi.org/10.1007/978-1-4757-3403-4_7.

A. Sherstov. The intersection of two halfspaces has high threshold degree. In *Proc. 50th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2009.

Y. S. Tan and R. Vershynin. Polynomial time and sample complexity for non-gaussian component analysis: Spectral methods. In *Conference On Learning Theory*, pages 498–534. PMLR, 2018.

S. Tiegel. Improved hardness results for learning intersections of halfspaces. *arXiv preprint arXiv:2402.15995*, 2024.

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

S. S. Vempala. Learning convex concepts from gaussian distributions with pca. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 124–130. IEEE, 2010a.

S. S Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *Journal of the ACM (JACM)*, 57(6):1–14, 2010b.

S. S. Vempala and Y. Xiao. Structure from local optima: Learning subspace juntas via higher order pca. *arXiv preprint arXiv:1108.3329*, 2011.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Appendix

Structure of Appendix We give an overview of the structure of the appendix. In Appendix A, we provide a complete list of notations and preliminaries. In Appendix B, we provide missing proofs and discussions in Section 2. In Appendix C, we provide omitted proofs in Section 3. In Appendix E, we provide omitted proofs in Section 4. In Appendix F, we give a complete description of our main algorithm and provide the proof of Theorem 3. In Appendix D, we give a complete proof of Theorem 2, the CSQ lower bound for learning intersections of two halfspaces under factorizable distributions.

Appendix A. Preliminaries and Notations

In this section, we present a complete list of notations, preliminaries and related background on the statistical learning model.

Basic Notations In this paper, we use small boldface characters for vectors and use capital lightface characters for subspaces, matrices and tensors. For $n \in \mathbb{Z}_+$, we denote by $[n] := \{1, \ldots, n\}$. For $\mathbf{x} \in \mathbb{R}^d$, and $i \in [d]$, we use $\mathbf{x}_i$ to denote the i-coordinate of $\mathbf{x}$. For $i \in [d]$, we denote by $\mathbf{e}_i$ the i-th standard basis of $\mathbb{R}^d$. Let $V \subseteq \mathbb{R}^d$ be a subspace, we denote by $\mathbf{x}_V := \mathrm{proj}_V(\mathbf{x})$, the projection of $\mathbf{x}$ onto the subspace V and denote by $V^\perp$ the orthogonal complement of V. For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, we use $\mathbf{u} \cdot \mathbf{v}$ to denote the inner product of $\mathbf{u}$ and $\mathbf{v}$ and we use $\|\mathbf{u}\|_2$ to denote the ℓ_2 norm of $\mathbf{u}$. We use $S^{d-1} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$ to denote the d-dimensional unit sphere and $B^d(r)$ the d-dimensional ball with radius r.

For any distribution D, we use $\mathbf{E}_{\mathbf{x}\sim D}(\mathbf{x})$ to denote the expectation of D. Let D be a distribution of $(\mathbf{x}, y)$ over $\mathbb{R}^d \times \{\pm 1\}$. We use D_X to denote the marginal distribution of D over $\mathbb{R}^d$. For any subspace $V \subseteq \mathbb{R}^d$, we use D_V to denote the marginal distribution of D for $\mathbf{x}_V$. For $z \in \{\pm 1\}$, we use D_V^z to denote the marginal distribution of D over $\mathbf{x}_V$ condition on $y = z$ and use D^z to denote the marginal distribution D_X over $\mathbf{x}$ condition on $y = z$ For a distribution D_X over $\mathbb{R}^d$, we say D_X has an isotropic covariance matrix if there is some $\alpha \geq 0$ such that $\mathbf{E}_{\mathbf{x}\sim D_X} \mathbf{x}\mathbf{x}^\intercal = \alpha I$, where α is called the scale of $\mathbf{E}_{\mathbf{x}\sim D_X} \mathbf{x}\mathbf{x}^\intercal$.

For tensors, we will consider a k-tensor to be an element in $(\mathbb{R}^n)^{\otimes k} \cong \mathbb{R}^{n^k}$. A symmetric tensor is a tensor that is invariant under a permutation of its vector arguments. We use $\|T\|_F$ to denote the Frobinius norm of T. We will use $T_{i_1,\ldots,i_k}$ to denote the coordinate of a k-tensor T indexed by the k-tuple $(i_1, \ldots, i_k)$. For a tensor $T \in (\mathbb{R}^d)^{\otimes m}$ and $\pi : [m] \to [m]$ be a permutation of indices, we use $\pi(T)$ to denote the tensor permuted by π defined as $\pi(T)_{(i_1),\cdots,(i_m)} = T_{\pi(i_1),\cdots,\pi(i_m)}$. We define $\mathrm{sym}(T)$ as $\frac{1}{m!}\sum_{\pi\in\Pi} \pi(T)$, where Π is the set of all possible permutations of $[m]$. By abuse of notation, we will sometimes treat a tensor $T \in (\mathbb{R}^d)^{\otimes m}$ as a linear mapping, i.e., for $\mathbf{v} \in \mathbb{R}^d$, we use $T \cdot \mathbf{v}$ to denote applying the linear mapping $T : \mathbb{R}^d \to (\mathbb{R}^d)^{\otimes m-1}$ specified by T on $\mathbf{v}$. For a vector $\mathbf{v} \in \mathbb{R}^n$, we denote by $\mathbf{v}^{\otimes k}$ to be a vector (linear object) in $\mathbb{R}^{n^k}$. For a matrix $M \in \mathbb{R}^{n\times m}$, we denote by $\|M\|_2, \|M\|_F$ to be the operator norm and Frobenius norm respectively.

We present the following fact that will be useful in the analysis of our algorithms.

Fact 3 *Let $T \in (\mathbb{R}^d)^{\otimes m}$ and $\|T\|_F = 1$ be a symmetric tensor for $m \leq 3$, then $\max_{\mathbf{u}\in S^{d-1}} \mathbf{u}^{\otimes m} \cdot T \geq 1/\mathrm{poly}(d)$.*

Proof [Proof of Fact 3]

The statement trivially holds for $m = 1, 2$. Therefore, we only need to consider the case $m = 3$. We first show that any symmetric $T \in (\mathbb{R}^d)^{\otimes 3}$ can be written as $T = \sum_{i=1}^N \alpha_i \mathbf{u}_i^{\otimes 3}$, where $N = \mathrm{poly}(d)$, $\sum_{i=1}^N |\alpha_i| = \mathrm{poly}(d)$ and each $\mathbf{u}_i$ is a unit vector. Suppose we have shown that this is true, then it is easy to see that $1 = T \cdot T = \sum_{i=1}^N \alpha_i {\mathbf{u}_i}^{\otimes 3} \cdot T \leq \mathrm{poly}(d) \max_i({\mathbf{u}_i}^{\otimes 3} \cdot T)$, therefore at least one $\mathbf{u}_i$ satisfies ${\mathbf{u}_i}^{\otimes 3} \cdot T = 1/\mathrm{poly}(d)$.

Therefore, we just need to show that the statement above about decomposition is true. Since $\mathrm{sym}(\mathbf{v}_i \otimes \mathbf{v}_j \otimes \mathbf{v}_k)$ where $\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k$ are standard basis vectors span the space of symmetric tensors. Therefore, it suffices for us to show that the statement holds true for any $T = \mathrm{sym}(\mathbf{v}_i \otimes \mathbf{v}_j \otimes \mathbf{v}_k)$. Notice that for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$

$$
\begin{aligned}
&(\mathbf{u} + \mathbf{v})^{\otimes 3} - \mathbf{u}^{\otimes 3} - \mathbf{v}^{\otimes 3} \\
=&\mathbf{u} \otimes \mathbf{v}^{\otimes 2} + \mathbf{v} \otimes \mathbf{u} \otimes \mathbf{v} + \mathbf{v}^{\otimes 2} \otimes \mathbf{u} \\
&+ \mathbf{v} \otimes \mathbf{u}^{\otimes 2} + \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{u} + \mathbf{u}^{\otimes 2} \otimes \mathbf{v} \\
=&3\mathrm{sym}(\mathbf{u} \otimes \mathbf{v}^{\otimes 2}) + 3\mathrm{sym}(\mathbf{v} \otimes \mathbf{u}^{\otimes 2}) \ ,
\end{aligned}
$$

and

$$
\begin{aligned}
&(\mathbf{u} - \mathbf{v})^{\otimes 3} - \mathbf{u}^{\otimes 3} + \mathbf{v}^{\otimes 3} \\
=&\mathbf{u} \otimes \mathbf{v}^{\otimes 2} + \mathbf{v} \otimes \mathbf{u} \otimes \mathbf{v} + \mathbf{v}^{\otimes 2} \otimes \mathbf{u} \\
&- (\mathbf{v} \otimes \mathbf{u}^{\otimes 2} + \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{u} + \mathbf{u}^{\otimes 2} \otimes \mathbf{v}) \\
=&3\mathrm{sym}(\mathbf{u} \otimes \mathbf{v}^{\otimes 2}) - 3\mathrm{sym}(\mathbf{v} \otimes \mathbf{u}^{\otimes 2}) \ .
\end{aligned}
$$

Taking the difference of the above two equations shows that the decomposition statement is true for any $T = \mathrm{sym}(\mathbf{v} \otimes \mathbf{u}^{\otimes 2})$. Therefore, we just need to decompose $\mathrm{sym}(\mathbf{v}_i \otimes \mathbf{v}_j \otimes \mathbf{v}_k)$ as linear combination of tensors of the form $\mathbf{v}^{\otimes 3}$ and $\mathrm{sym}(\mathbf{v} \otimes \mathbf{u}^{\otimes 2})$. Then notice that since sym is a linear operator

$$
\begin{aligned}
&\mathrm{sym}(\mathbf{v}_i \otimes (\mathbf{v}_j + \mathbf{v}_k)^{\otimes 2}) - \mathrm{sym}(\mathbf{v}_i \otimes \mathbf{v}_j^{\otimes 2}) - \mathrm{sym}(\mathbf{v}_i \otimes \mathbf{v}_k^{\otimes 2}) \\
=&\mathrm{sym}(\mathbf{v}_i \otimes (\mathbf{v}_j + \mathbf{v}_k)^{\otimes 2} - \mathbf{v}_i \otimes \mathbf{v}_j^{\otimes 2} - \mathbf{v}_i \otimes \mathbf{v}_k^{\otimes 2}) \\
=&\mathrm{sym}(\mathbf{v}_i \otimes \mathbf{v}_j \otimes \mathbf{v}_k + \mathbf{v}_i \otimes \mathbf{v}_k \otimes \mathbf{v}_j) \\
=&\mathrm{sym}(\mathbf{v}_i \otimes \mathbf{v}_j \otimes \mathbf{v}_k) \ .
\end{aligned}
$$

This completes the proof. ■

Background on Statistical Query Model SQ algorithms are a class of algorithms that are allowed to query expectations of bounded functions on the underlying distribution through an (SQ) oracle rather than directly access samples. The model was introduced by Kearns (1998) as a natural restriction of the PAC model (Valiant, 1984) in the context of learning Boolean functions. Since then, the SQ model has been extensively studied in a range of settings, including unsupervised learning (Feldman, 2016). The class of SQ algorithms is broad and captures a range of known algorithmic techniques in machine learning including spectral techniques, moment and tensor methods, local search (e.g., EM), and many others (see, e.g., Feldman et al. (2017a,b) and references therein).

Definition 17 (SQ Model) *Let D be a distribution on X. A* statistical query *is a bounded function $q : X \to [-1, 1]$. We define $\mathrm{STAT}(q, \tau)$ to be the oracle that given any such query q, outputs a value v such that $|v - \mathbf{E}_{\mathbf{x}\sim D}[q(\mathbf{x})]| \leq \tau$, where $\tau > 0$ is the* tolerance *parameter of the query. A* statistical query (SQ) algorithm *is an algorithm whose objective is to learn some information about an unknown distribution D by making adaptive calls to the corresponding $\mathrm{STAT}(q, \tau)$ oracle.*

Basics of Correlational Statistical Query(CSQ) Model In particular, given D is a distribution on $X \times \{-1, 1\}$, we can define the Correlational Statistical Query (CSQ) model as a further restriction of the SQ model.

Definition 18 (CSQ Model) *Let D be a distribution on $X \times \{-1, 1\}$. A* correlational statistical query *is a bounded function $q : X \times \{-1, 1\} \to [-1, 1]$. We define $\mathrm{CSTAT}(\tau)$ to be the oracle that given any such query q, outputs a value $v \in [-1, 1]$ such that $|v - \mathbf{E}_{(\mathbf{x},y)\sim D}[yq(\mathbf{x})]| \leq \tau$, where $\tau > 0$ is the* tolerance *parameter of the query. A* statistical query (SQ) algorithm *is an algorithm whose objective is to learn some information about an unknown distribution D by making adaptive calls to the corresponding $\mathrm{STAT}(q, \tau)$ oracle.*

Definition 19 (Function Representation of Distribution for CSQ) *Let D be a joint distribution of $(\mathbf{x}, y)$ supported on $\mathbb{R}^d \times \{\pm 1\}$ where D^+ and D^- has probability density functions $P_{D^+}, P_{D^-} : \mathbb{R}^d \to \mathbb{R}_+$. Let D_0 be a distribution on $\mathbb{R}^d$ with density function $P_{D_0} : \mathbb{R}^d \to \mathbb{R}+$ where the support of D contains the support of D^+ and D^-. Then, the function representation of D for CSQ w.r.t. D_0 is defined as a function $f_{D,D_0} : \mathbb{R}^d \to \mathbb{R}$ such that $f_{D,D_0}(\mathbf{x}) = (P_{D^+}(\mathbf{x}) - P_{D^-}(\mathbf{x}))/P_{D_0}(\mathbf{x})$.*

Definition 20 (Pairwise Correlation) *For functions $f, g : \mathbb{R}^d \mapsto \mathbb{R}_+$, we defined the correlation between f and g under the distribution D_0 to be the expectation $\mathbf{E}_{\mathbf{x}\sim D_0}[f(\mathbf{x})g(\mathbf{x})]$.*

Definition 21 *We say that a set of functions F mapping $\mathbb{R}^d \to \mathbb{R}$ is (γ, β)-correlated relative to a distribution D_0 if for any $f_i, f_j \in F$, the correlation $\mathbf{E}_{\mathbf{x}\sim D_0}[f_i(\mathbf{x})f_j(\mathbf{x})] \leq \gamma$ for all $i \neq j$ and $\mathbf{E}_{\mathbf{x}\sim D_0}[f_i(\mathbf{x})f_j(\mathbf{x})] \leq \beta$ for $i = j$.*

Definition 22 (Decision Problem over Distributions) *Let D be a fixed distribution and $\mathcal{D}$ be a distribution family. We denote by $\mathcal{B}(\mathcal{D}, D)$ the decision problem in which the input distribution D' is promised to satisfy either (a) $D' = D$ or (b) $D' \in \mathcal{D}$, and the goal is to distinguish the two cases with high probability.*

Definition 23 (Correlational Statistical Query Dimension) *For $\beta, \gamma > 0$, a decision problem $\mathcal{B}(\mathcal{D}, D)$, where D is a fixed distribution and $\mathcal{D}$ is a family of distribution both over $X \times \{\pm 1\}$, and $f_{D,D_0} \equiv 0$. Let s be the maximum integer such that there exists a finite set of distributions $\mathcal{D}_D \subseteq \mathcal{D}$ such that $\{f_{D,D_0} \mid D \in \mathcal{D}_D\}$ is (γ, β)-correlated relative to D_0 and $|\mathcal{D}_D| \geq s$. The Correlational Statistical Query dimension with pairwise correlations (γ, β) of $\mathcal{B}$ is defined to be s, and denoted by $s = \mathrm{CD}(\mathcal{B}, \gamma, \beta)$.*

Lemma 24 *Let $\mathcal{B}(\mathcal{D}, D)$ be a decision problem, where D is the reference distribution and $\mathcal{D}$ is a class of distribution. For $\gamma, \beta > 0$, let $s = \mathrm{CD}(\mathcal{B}, \gamma, \beta)$. For any $\gamma' > 0$, any CSQ algorithm for $\mathcal{B}$ requires queries of tolerance at most $\sqrt{\gamma + \gamma'}$ or makes at least $s\gamma'/(\beta - \gamma)$ queries.*

Appendix B. Omitted Proofs from Section 2

In this section, we present missing details in Section 2.

B.1. Proof of Theorem 7

In this section, we give the proof of Theorem 7. For convenience, we restate Theorem 7 below.

Theorem 25 (restatement of Theorem 7) *For any $d, m \in \mathbb{N}$, $C \subseteq \mathbb{R}^d$, $T_i \in \mathrm{sym}((\mathbb{R}^d)^{\otimes i})$ for $i \in [m]$ and $\tau \in \mathbb{R}_{\geq 0}$, at most one of the following conditions can be satisfied:*

a) there exists a distribution D supported on C such that $\|\mathbf{E}_{\mathbf{x}\sim D}(\mathbf{x}^{\otimes i}) - T_i\|_F \leq \tau$ for any $i \in [m]$;

b) there exists a degree-m polynomial $p : \mathbb{R}^d \to \mathbb{R}$ defined as $p(\mathbf{x}) = \sum_{i=1}^m A_i \cdot \mathbf{x}^{\otimes i}$ where $p(\mathbf{x}) \geq 0$ for any $\mathbf{x} \in C$, $\sum_{i=1}^m \|A_i\|_F \leq 1$ and $\sum_{i=1}^m A_i \cdot T_i < -\tau$.

We call such a polynomial above a one-sided approximation polynomial for C w.r.t. to moments information T_i and tolerance τ.

Proof [Proof of Theorem 25] We prove Theorem 25 by contradiction. Suppose that the two conditions in Theorem 25 are satisfied simultaneously. Since for every $\mathbf{x} \in C$, $p(x) \geq 0$, we know that $\mathbf{E}_{\mathbf{x}\sim D} p(\mathbf{x}) \geq 0$. On the other hand, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim D} p(\mathbf{x}) &= \sum_{i=1}^m A_i \cdot \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathbf{x}^{\otimes i} = \sum_{i=1}^m (A_i \cdot T_i) + \sum_{i=1}^m A_i \cdot (\mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathbf{x}^{\otimes i} - T_i) \\ &< -\tau + \sum_{i=1}^m A_i \cdot (\mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathbf{x}^{\otimes i} - T_i) \leq 0.\end{aligned}$$

This gives a contradiction. ■

B.2. Discussion on Structural Assumptions made in Section 2

In this section, we explain Assumption 1 as well as other structural assumptions made in Section 2 can be made without loss of generality.

We first argue that we can assume V, the subspace spanned by $\mathbf{u}^*, \mathbf{v}^*$ is exactly equal to $\mathbf{span}\{\mathbf{e}_1, \mathbf{e}_2\}$. If $\mathbf{u}^*, \mathbf{v}^*$ are parallel to each other, then the problem degenerates to the problem of learning a single halfspace or learning a degree-2 polynomial threshold function. Furthermore, since any rotation matrix U will not change the inner product between two points in $\mathbb{R}^d$, if $V \neq \mathbf{span}\{\mathbf{e}_1, \mathbf{e}_2\}$, we can apply a rotation matrix U that maps $\mathbf{u}^*, \mathbf{v}^*$ to $\mathbf{span}\{\mathbf{e}_1, \mathbf{e}_2\}$, and every example $U^\intercal \mathbf{x}$ still satisfies γ-margin assumption and has the same label as $\mathbf{x}$. Based on this, in the rest of the section, we argue that Assumption 1 can be made without loss of generality.

Assumption 2 (restatement of Assumption 1) *Given an intersection of two halfspaces $h^* = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ and a distribution D over $\mathbb{B}^2(1) \times \{\pm 1\}$ that consistent with h^* with the γ-margin condition, we parameterize h^* by $\theta \in (0, \pi/2), t \geq 0, \sigma \geq 0$ where $\mathbf{u}^* = \sin\theta \mathbf{e}_1 - \cos\theta \mathbf{e}_2, t_1 = t\sin\theta$ and $\mathbf{v}^* = \sin\theta \mathbf{e}_1 + \cos\theta \mathbf{e}_2, t_2 = (1+\sigma)t\sin\theta$. Furthermore, we assume $\|\mathbf{E}_{\mathbf{x}\sim D^+} \mathbf{x}\|_2 \leq \gamma^c$, $t_1, t_2 \geq \gamma$, $|t_1|, |t_2| \leq 1$, where c is some large constant.*

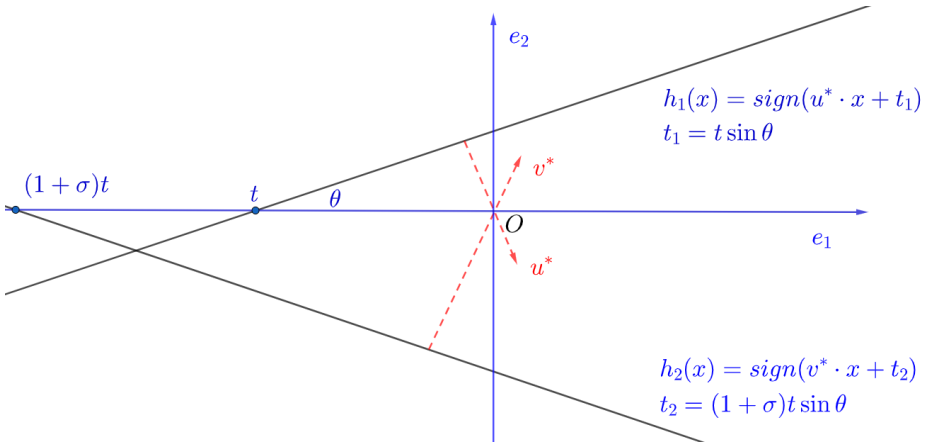

Figure 2: Geometrical Illustration of Assumption 1. Two halfspaces $h_1 = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1)$ and $h_2 = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ are colored in black. Red dashed lines represent the directions of weight vectors $\mathbf{u}^*, \mathbf{v}^*$.

First, we argue that we can without loss of generality assume $\|\mathbf{E}_{\mathbf{x}\sim D^+} \mathbf{x}\|_2 \leq \gamma^c$, for any large constant c. Assuming $\|\mathbf{E}_{\mathbf{x}\sim D^+} \mathbf{x}\|_2 \geq \gamma^c$ instead, by drawing $\mathrm{poly}(d/\gamma)$ positive examples from D^+_X, we are able to estimate some $\hat{\mathbf{x}} \in V$ such that $\|\hat{\mathbf{x}} - \mathbf{E}_{\mathbf{x}\sim D^+} \mathbf{x}\|_2 \leq \gamma^c/2$. Since each example $\mathbf{x}$ has $\|\mathbf{x}\|_2 \leq 1$, we know that $\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq 2$, thus by rescaling, $(\mathbf{x} - \hat{\mathbf{x}})/2$ satisfies $\gamma/2$-margin assumption and the resulting positive example has mean $\gamma^c/2$-close to the origin.

Consider the target halfspace $h^* = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$. We furthermore argue we can without loss of generality make the following two assumptions on h^*

1. Under the assumption $\|\mathbf{E}_{\mathbf{x}\sim D^+} \mathbf{x}\|_2 \leq \gamma^c$, $t_1, t_2 \geq 0$. This is because if $t_1 \leq 0$ (without loss of generality), then every positive example $\mathbf{x}$ satisfies $\mathbf{u}^* \cdot \mathbf{x} \geq \gamma$, which implies $\|\mathbf{E}_{\mathbf{x}\sim D^+} \mathbf{x}\|_2 \geq \gamma$.

2. $|t_1|, |t_2| \leq 1$. If this is not the case, then the problem degenerates to learning a single halfspace and can be solved trivially.

Given the above assumptions, it will be convenient for us to parameterize $h^* = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ to be described by $t \in [0, \infty)$, $\theta \in [0, \pi/2)$ and $\sigma \in [0, \infty)$ where we have $\mathbf{u}^* = \sin\theta \mathbf{e}_1 - \cos\theta \mathbf{e}_2$, $t_1 = t \sin\theta$ and $\mathbf{v}^* = \sin\theta \mathbf{e}_1 + \cos\theta \mathbf{e}_2$, $t_2 = (1+\sigma)t \sin\theta$ (as illustrated in Figure 2). This will be convenient for later calculations.

Notice that for every $\mathbf{u}^*, \mathbf{v}^*$ such that $\theta(\mathbf{u}^*, \mathbf{v}^*) = \pi - 2\theta$, there is a rotation matrix U such that $U\mathbf{u}^* = \sin\theta \mathbf{e}_1 - \cos\theta \mathbf{e}_2, \mathbf{v}^* = \sin\theta \mathbf{e}_1 + \cos\theta \mathbf{e}_2$. Since rotation matrix U maintains the γ-margin assumption, parameterizing h^* in such a way does not lose the generality.

B.3. Proof of Lemma 8

In this section, we present the proof of Lemma 8. For convenience, we restate Lemma 8 as follows.

Lemma 26 (restatement of Lemma 8) *Let D be a distribution over $\mathbb{B}^2(1) \times \{\pm 1\}$ that is consistent with an instance of learning intersections of two halfspaces with γ-margin assumption, where γ is smaller than some sufficiently small constant. Let $c > 0$ be any suitable large constant. Suppose*

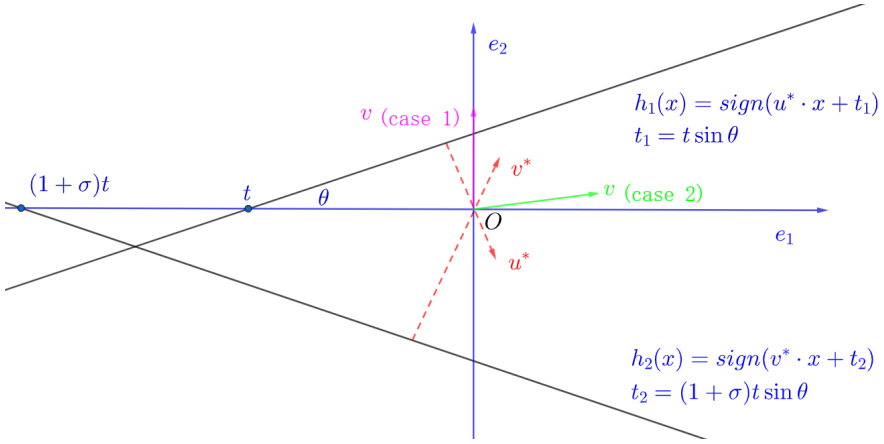

Figure 3: Geometrical illustration for the proof of Lemma 8. The vector colored in purple corresponds to case 1 in the proof and the vector colored in green corresponds to case 2 in the proof.

1. $\|\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}\|_F \le \gamma^c$, $\|\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}\|_F \le \gamma^c$, $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}\|_F \le \gamma^c$
2. $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}\mathbf{x}^\intercal\|_F \le \gamma^c$,

then $\left\|(\Sigma^+)^{-1}\right\|_2 = O(1/\gamma^4)$ *and* $\left\|(\Sigma^-)^{-1}\right\|_2 = O(1/\gamma^4)$, *where* $\Sigma^+ := \mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}\mathbf{x}^\intercal$ *and* $\Sigma^- := \mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}\mathbf{x}^\intercal$.

We first give some high-level intuition behind Lemma 26. For the purpose of contradiction, we assume that $\left\|(\Sigma^+)^{-1}\right\|_2 > \Omega(1/\gamma^c)$ or $\left\|(\Sigma^-)^{-1}\right\|_2 > \Omega(1/\gamma^c)$. Therefore, there must be a unit vector $\mathbf{v}$ such that $\mathbf{v}^\intercal\Sigma^+\mathbf{v} \le O(\gamma^c)$ or $\mathbf{v}^\intercal\Sigma^+\mathbf{v} \le O(\gamma^c)$. However, since Σ^+ and Σ^- are close to each other in Frobinous norm, it must be that $\mathbf{v}^\intercal\Sigma^+\mathbf{v} \le O(\gamma^c)$ and $\mathbf{v}^\intercal\Sigma^-\mathbf{v} \le O(\gamma^c)$. Roughly speaking, this means most of the samples are inside a thin band along the direction of $\mathbf{v}^\perp$, i.e., inside the band region $B := \{\mathbf{x} \in \mathbb{R}^2 \mid |\mathbf{x}\cdot\mathbf{v}| \le \gamma/2\}$. Let $f^*(\mathbf{x}) = \mathrm{sign}(\mathbf{u}^*\cdot x + t_1) \wedge \mathrm{sign}(\mathbf{v}^*\cdot x + t_2)$ be the true concept function, and let A^+, A^- be the region of $\{\mathbf{x} \in \mathbb{R}^2 \mid f^*(\mathbf{x}) = 1\}$ and $\{\mathbf{x} \in \mathbb{R}^2 \mid f^*(\mathbf{x}) = -1\}$ respectively. Notice that there are two cases for this band region B(see Figure 3 for illustration): either

1. $\mathrm{sign}(\mathbf{u}^*\cdot\mathbf{v}^\perp) = \mathrm{sign}(\mathbf{v}^*\cdot\mathbf{v}^\perp)$. In this case, we show that the first moments of D^+, D^- are not close along the direction of $v^\perp$.
2. $\mathrm{sign}(\mathbf{u}^*\cdot\mathbf{v}^\perp) \ne \mathrm{sign}(\mathbf{v}^*\cdot\mathbf{v}^\perp)$. In this case, we show that the moment information must differ by giving a degree-2 polynomial p that $\mathbf{E}_{\mathbf{x}\sim D^+}[p(\mathbf{x})]$ and $\mathbf{E}_{\mathbf{x}\sim D^-}[p(\mathbf{x})]$ differs from each other, where we choose this $p(\mathbf{x}) := (\mathbf{u}^*\cdot x + t_1)(\mathbf{v}^*\cdot x + t_2)$.

In both cases, this contradicts the assumption that D is a moment-matching distribution. We give the formal proof of Lemma 26 below.

Proof [Proof of Lemma 26] To prove the statement, it suffices for us to show that there exists a universal constant $c' > 0$, given

1. $\|\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}\|_F \le c'\gamma^4/100$, $\|\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}\|_F \le c'\gamma^4/100$ and
2. $\|\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}\mathbf{x}^\intercal - \mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}\mathbf{x}^\intercal\|_F \le c'\gamma^4/100$,

then for any $\mathbf{v} \in S^1$, $\mathbf{v}^\intercal \Sigma^- \mathbf{v} \geq c'\gamma^4$. Suppose we can prove the above. Then given c is a sufficiently large constant and γ is at most a sufficiently small constant in Lemma 26, the assumption of the above statement must be satisfied. Therefore, we must have $\mathbf{v}^\intercal \Sigma^- \mathbf{v} \geq c'\gamma^4$ and $\mathbf{v}^\intercal \Sigma^- \mathbf{v} = \mathbf{v}^\intercal \Sigma^+ \mathbf{v} - \mathbf{v}^\intercal(\Sigma^- - \Sigma^+)\mathbf{v} \geq c'\gamma^4/2$, which implies that $\left\|(\Sigma^+)^{-1}\right\|_2 = O(1/\gamma^4)$ and $\left\|(\Sigma^-)^{-1}\right\|_2 = O(1/\gamma^4)$.

We will prove $\mathbf{v}^\intercal \Sigma^- \mathbf{v} \geq c'\gamma^4$ for two cases. Let $\mathbf{v}^\perp$ be the unique unit vector up to negation that $\mathbf{v}^\perp \cdot \mathbf{v} = 0$. We consider the cases that:

1. $\mathrm{sign}(\mathbf{u}^* \cdot \mathbf{v}^\perp) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{v}^\perp)$ or (either $\mathbf{u}^* \cdot \mathbf{v}^\perp$ or $\mathbf{v}^* \cdot \mathbf{v}^\perp = 0$), and
2. $\mathrm{sign}(\mathbf{u}^* \cdot \mathbf{v}^\perp) \neq \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{v}^\perp)$.

For the case $\mathrm{sign}(\mathbf{u}^* \cdot \mathbf{v}^\perp) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{v}^\perp)$, let $\mathbf{v}^\perp$ be the unit direction that $\mathbf{v}^\perp \cdot \mathbf{v} = 0$, $\mathbf{u}^* \cdot \mathbf{v}^\perp \geq 0$ and $\mathbf{v}^* \cdot \mathbf{v}^\perp \geq 0$. Take $c' > 0$ to be a sufficiently small constant and assume for the purpose of contradiction that $\mathbf{v}^\intercal \Sigma^- \mathbf{v} \leq c'\gamma^3$. Let $B := \{\mathbf{x} \in \mathbb{B}^2(1) \mid |\mathbf{x} \cdot \mathbf{v}| \leq \gamma/2\}$. Then notice that by Markov's inequality,

$$\Pr_{\mathbf{x} \sim D^-}[\mathbf{x} \notin B] \leq \mathop{\mathbf{E}}_{\mathbf{x} \sim D^-}[(\mathbf{x} \cdot \mathbf{v})^2]/(\gamma^2/4) \leq \mathbf{v}^\intercal \Sigma^- \mathbf{v}/(\gamma^2/4) \leq c\gamma \,,$$

where c is a sufficiently small constant. Now, we show that for any $\mathbf{x} \in B$ such that $\mathbf{x} \cdot \mathbf{u}^* + t_1 \leq -\gamma$, $\mathbf{x} \cdot \mathbf{v}^\perp \leq -\gamma/2$. First notice that

$$\mathbf{x} \cdot \mathbf{u}^* = \mathbf{x} \cdot \mathrm{proj}_{\mathbf{v}} \mathbf{u}^* + \mathbf{x} \cdot \mathrm{proj}_{\mathbf{v}^\perp} \mathbf{u}^* = (\mathbf{x} \cdot \mathbf{v})(\mathbf{v} \cdot \mathbf{u}^*) + (\mathbf{x} \cdot \mathbf{v}^\perp)(\mathbf{v}^\perp \cdot \mathbf{u}^*) \,. \tag{3}$$

Suppose $\mathbf{u}^* \cdot \mathbf{v}^\perp = 0$, then we immediately get $\mathbf{x} \cdot \mathbf{u}^* + t_1 \geq -|\mathbf{x} \cdot \mathbf{v}| \geq -\gamma/2$. Therefore, no such $\mathbf{x}$ exists, and we can assume without loss of generality that $\mathbf{u}^* \cdot \mathbf{v}^\perp > 0$. Furthermore, notice that for any $\mathbf{x} \in B$ and $\mathbf{x} \cdot \mathbf{v}^* + t_1 \leq -\gamma$, given Equation (3), we get

$$\mathbf{x} \cdot \mathbf{v}^\perp = \frac{\mathbf{x} \cdot \mathbf{u}^* - (\mathbf{x} \cdot \mathbf{v})(\mathbf{v} \cdot \mathbf{u}^*)}{\mathbf{v}^\perp \cdot \mathbf{u}^*} \leq \frac{-\gamma - (\mathbf{x} \cdot \mathbf{v})(\mathbf{v} \cdot \mathbf{u}^*)}{\mathbf{v}^\perp \cdot \mathbf{u}^*} \leq -\gamma - (\mathbf{x} \cdot \mathbf{v})(\mathbf{v} \cdot \mathbf{u}^*) \leq -\gamma/2 \,,$$

where the third from the last inequality follows from $\mathbf{x} \cdot \mathbf{u}^* \leq -\gamma - t_1 \leq -\gamma$ and the last inequality follows from $\mathbf{x} \in B$ and $|\mathbf{v} \cdot \mathbf{u}^*| \leq 1$. Similarly, we also have that for any $\mathbf{x} \in B$ and $\mathbf{x} \cdot \mathbf{v}^* + t_2 \leq -\gamma$, $\mathbf{x} \cdot \mathbf{v}^\perp \leq -\gamma/2$. Combining the above two and the γ-margin condition gives that, for any $\mathbf{x} \in B$ and $\mathbf{x} \in \mathbf{supp}(D^-)$, $\mathbf{x} \cdot \mathbf{v}^\perp \leq -\gamma/2$. Then we get

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x} \sim D^-}[\mathbf{v}^\perp \cdot \mathbf{x}] &= \mathop{\mathbf{E}}_{\mathbf{x} \sim D^-}[\mathbf{v}^\perp \cdot \mathbf{x} | \mathbf{x} \in B] \Pr_{\mathbf{x} \sim D^-}[\mathbf{x} \in B] + \mathop{\mathbf{E}}_{\mathbf{x} \sim D^-}[\mathbf{v}^\perp \cdot \mathbf{x} | \mathbf{x} \notin B] \Pr_{\mathbf{x} \sim D^-}[\mathbf{x} \notin B] \\
&\leq (-\gamma/2)(1 - c\gamma) + c\gamma \\
&\leq -\gamma/4 + \gamma/8 \leq -\gamma/8 \,,
\end{aligned}$$

where that last inequality follows from that c is a sufficiently small constant. This contradicts that $\|\mathbf{E}_{\mathbf{x} \sim D^-} \mathbf{x}\|_F \leq c'\gamma^4/100$ in the assumption. Therefore, we must have $\mathbf{v}^\intercal \Sigma^- \mathbf{v} \geq c'\gamma^3 \geq c'\gamma^4$. This proves the statement for Case 1.

For the case that $\mathrm{sign}(\mathbf{u}^* \cdot \mathbf{v}^\perp) \neq \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{v}^\perp)$, let $B := \{\mathbf{x} \in \mathbb{B}^2(1) \mid |\mathbf{v} \cdot \mathbf{x}| \leq \gamma/2\}$. Let the c' in the statement be a sufficiently small constant. Suppose we can prove that $\mathbf{Pr}_{\mathbf{x} \sim D^-}[\mathbf{x} \notin B] = \Omega(\gamma^2)$, then we are done, since from the definition of B, we immediately get

$$\mathbf{v}^\intercal \Sigma^- \mathbf{v} \geq \Pr_{\mathbf{x} \sim D^-}[(\mathbf{x} \cdot \mathbf{v})^2 \mid \mathbf{x} \notin B] \Pr_{\mathbf{x} \sim D^-}[\mathbf{x} \notin B] = \Omega(\gamma^4) \,.$$

To show that $\mathbf{Pr}_{\mathbf{x}\sim D^-}[\mathbf{x} \notin B] = \Omega(\gamma^2)$, we consider the degree-2 polynomial $p(\mathbf{x}) := (\mathbf{u}^* \cdot \mathbf{x} + t_1)(\mathbf{v}^* \cdot \mathbf{x} + t_2)$. By Assumption 1, we know that $t_2 \leq 1, t_1 \leq 1$. Therefore,

$$\begin{aligned}
&\left| \mathop{\mathbf{E}}_{\mathbf{x}\sim D^+} p(\mathbf{x}) - \mathop{\mathbf{E}}_{\mathbf{x}\sim D^-} p(\mathbf{x}) \right| \\
=&\left| (\mathbf{u}^*)^\intercal (\Sigma^+ - \Sigma^-)\mathbf{v}^* + (t_2\mathbf{u}^* + t_1\mathbf{v}^*) \cdot (\mathop{\mathbf{E}}_{x\sim D^+} x_V - \mathop{\mathbf{E}}_{x\sim D^-} x_V) \right| \\
=&\left| \|(\mathbf{u}^*)^\intercal \mathbf{v}^*\|_F \left\|\Sigma^+ - \Sigma^-\right\|_F + \|t_2\mathbf{u}^* + t_1\mathbf{v}^*\|_2 \left\| \mathop{\mathbf{E}}_{x\sim D^+} x_V - \mathop{\mathbf{E}}_{x\sim D^-} x_V \right\|_2 \right| \leq c\gamma^4 \,,
\end{aligned}$$

where c is a sufficiently small constant. From the γ-margin assumption, we have $\mathbf{E}_{\mathbf{x}\sim D^+} p(\mathbf{x}) \geq \gamma^2$. Therefore, we must have $\mathbf{E}_{\mathbf{x}\sim D^-} p(\mathbf{x}) \geq \gamma^2/2$. We show that in order to satisfy $\mathbf{E}_{\mathbf{x}\sim D^-} p(\mathbf{x}) \geq \gamma^2/2$, we must have $\mathbf{Pr}_{\mathbf{x}\sim D^-}[\mathbf{x} \notin B] = \Omega(\gamma^2)$. Notice that for any $\mathbf{x} \in \mathbf{supp}(D^-)$ and $\mathbf{x} \in B$, we must have either $\mathbf{u}^* \cdot \mathbf{x} + t_1 \leq -\gamma$ or $\mathbf{v}^* \cdot \mathbf{x} + t_2 \leq -\gamma$. Suppose that $\mathbf{u}^* \cdot \mathbf{x} + t_1 \leq -\gamma$, then we have $\mathbf{u}^* \cdot \mathbf{x} \leq -\gamma - t_1 \leq -\gamma$ ($t_1 \geq 0$ from Assumption 1). Combining the above with $\mathbf{u}^* \cdot \mathbf{x} = \mathrm{proj}_{\mathbf{v}}\mathbf{u}^* \cdot \mathbf{x} + \mathrm{proj}_{\mathbf{v}^\perp}\mathbf{u}^* \cdot \mathbf{x}$ and $|\mathrm{proj}_{\mathbf{v}}\mathbf{u}^* \cdot \mathbf{x}| \leq |\mathbf{v} \cdot \mathbf{x}| \leq \gamma/2$, we get $\mathrm{proj}_{\mathbf{v}^\perp}\mathbf{u}^* \cdot \mathbf{x} \leq 0$. Notice that $\mathrm{proj}_{\mathbf{v}^\perp}\mathbf{u}^* \cdot \mathbf{x} = (\mathbf{u}^* \cdot \mathbf{v}^\perp)(\mathbf{v}^\perp \cdot \mathbf{x}) \leq 0$ and we assumed $\mathrm{sign}(\mathbf{v}^\perp \cdot \mathbf{u}^*) \neq \mathrm{sign}(\mathbf{v}^\perp \cdot \mathbf{v}^*)$, then $\mathrm{proj}_{\mathbf{v}^\perp}\mathbf{v}^* \cdot \mathbf{x} = (\mathbf{v}^* \cdot \mathbf{v}^\perp)(\mathbf{v}^\perp \cdot \mathbf{x}) \geq 0$. Plug it into the equation below, we get

$$\mathbf{v}^*\cdot\mathbf{x}+t_2 = \mathrm{proj}_{\mathbf{v}}\mathbf{v}^*\cdot\mathbf{x}+\mathrm{proj}_{\mathbf{v}^\perp}\mathbf{v}^*\cdot\mathbf{x}+t_2 \geq (\mathbf{v}\cdot\mathbf{v}^*)(\mathbf{v}\cdot\mathbf{x})+(\mathbf{v}^*\cdot\mathbf{v}^\perp)(\mathbf{v}^\perp\cdot\mathbf{x})+t_2 \geq -\gamma/2+\gamma \geq \gamma/2 \,,$$

where the second from the last inequality comes from $\mathbf{x} \in B$ and $t_2 \geq \gamma$ in Assumption 1. Similarly, we can also show that for any $\mathbf{x} \in \mathbf{supp}(D^-)$ and $\mathbf{x} \in B$, if $\mathbf{u}^*\cdot\mathbf{x}+t_1 \leq -\gamma$, then $\mathbf{v}^*\cdot\mathbf{x}+t_2 \geq \gamma/2$. Combining the two cases, we get that for any $\mathbf{x} \in \mathbf{supp}(D^-)$ and $\mathbf{x} \in B$,

$$p(\mathbf{x}) = (\mathbf{u}^* \cdot \mathbf{x} + t_1)(\mathbf{v}^* \cdot \mathbf{x} + t_2) \leq -\gamma^2/2 \,.$$

Therefore, we get

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}[p(\mathbf{x})] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}[p(\mathbf{x}) \mid \mathbf{x} \in R] \mathop{\mathbf{Pr}}_{\mathbf{x}\sim D^-}[\mathbf{x} \in R] + \mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}[p(\mathbf{x}) \mid \mathbf{x} \notin R] \mathop{\mathbf{Pr}}_{\mathbf{x}\sim D^-}[\mathbf{x} \notin R] \\
&\leq -\gamma^2/2 \mathop{\mathbf{Pr}}_{\mathbf{x}\sim D^-}[\mathbf{x} \in R] + \mathop{\mathbf{Pr}}_{\mathbf{x}\sim D^-}[\mathbf{x} \notin R] \,.
\end{aligned}$$

Combining the above with that $\mathbf{E}_{\mathbf{x}\sim D^-} p(\mathbf{x}) \geq \gamma^2/2$, we get $\mathbf{Pr}_{\mathbf{x}\sim D^-}[\mathbf{x} \notin R] \geq \gamma^2/3$. This completes the proof.

■

B.4. Proof of Lemma 10 and Lemma 11

In this section, we present the proof of Lemma 10 and Lemma 11. For convenience, we restate the lemmas as follows.

Lemma 27 (restatement of Lemma 10) *Let $h^* = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ be the target hypothesis of an instance of the problem of learning intersections of two halfspaces and D be a distribution that is consistent with h^* Under Assumption 1, the following polynomial*

$$f^*(\mathbf{x}) = \frac{1}{t\sin\theta}(\mathbf{u}^* \cdot \mathbf{x} - t\sin\theta)^2(\mathbf{u}^* \cdot \mathbf{x} + t\sin\theta)$$

satisfies $f^*(\mathbf{x}) \geq 0, \forall \mathbf{x} \in \boldsymbol{supp}(D^+)$.

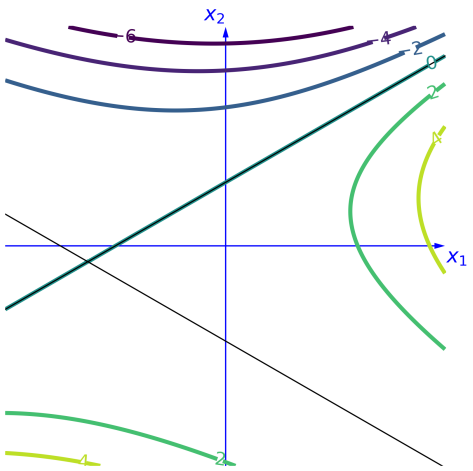

Figure 4: Illustration for Lemma 10. The target intersection of two halfspace h^* is plotted in black. Colored lines represent the contours of the polynomial f^*. $f^*(\mathbf{x}) > 0$ for every example $\mathbf{x}$ labeled positive by h^*.

For a clear intuition, we plot the contour of the polynomial constructed in Lemma 10 in Figure 4.

Proof [Proof of Lemma 27] For every positive example $\mathbf{x}$, we have $\mathbf{u}^* \cdot x + t_1 = \mathbf{u}^* \cdot x + t \sin\theta \geq 0$. Thus, $\forall x \in \textbf{supp}(D^+)$,

$$f^*(\mathbf{x}) = \frac{1}{t\sin\theta}(\mathbf{u}^* \cdot \mathbf{x} - t\sin\theta)^2(\mathbf{u}^* \cdot \mathbf{x} + t\sin\theta) \geq 0.$$

■

Lemma 28 (restatement of Lemma 11) *Let $h^* = \text{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \text{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ be the target hypothesis of an instance of the problem of learning intersections of two halfspaces and D be a distribution that is consistent with h^* Under Assumption 1, the following polynomial*

$$\begin{aligned} f^*(\mathbf{x}) &= a_0 + a_1\mathbf{x}_1 + a_2\mathbf{x}_2 - \mathbf{x}_2^2 \\ a_0 &= (1+\sigma)\tan^2\theta t^2 \\ a_1 &= (2+\sigma)\tan^2\theta t \\ a_2 &= -\sigma\tan\theta t \end{aligned}$$

satisfies $f^*(\mathbf{x}) \leq 0, \forall \mathbf{x} \in \textbf{supp}(D^-)$.

For a clear intuition, we plot the contour of the polynomial constructed in Lemma 11 in Figure 4.

Proof [Proof of Lemma 28]

To show for every $\mathbf{x} \in \mathbb{R}^d$ such that $h^*(\mathbf{x}) = -1$, $f^*(\mathbf{x}) \leq 0$., we partition the region of negative examples $N := \{\mathbf{x} \mid h^*(\mathbf{x}) = -1\}$ into regions $N_1 := \{\mathbf{x} \in V \mid \mathbf{x}_2 \geq -\sigma t\tan\theta/2, \mathbf{u}^* \cdot \mathbf{x} + t_1 \leq 0\}$ and $N_2 := \{\mathbf{x} \in V \mid \mathbf{x}_2 \leq -\sigma t\tan\theta/2, \mathbf{v}^* \cdot \mathbf{x} + t_2 \leq 0\}$, and show that in each region $f^*(\mathbf{x}) \leq 0$.

We first consider the region $N_1 := \{\mathbf{x} \in V \mid \mathbf{x}_2 \geq -\sigma t \tan\theta/2, \mathbf{u}^* \cdot \mathbf{x} + t_1 \leq 0\}$. To start with, we focus on examples that are on the boundary of N_1. Let $\mathbf{x}$ be any example on the decision boundary $\{\mathbf{x} \mid \mathbf{u}^* \cdot \mathbf{x} + t_1 = 0\}$. By Assumption 1, we know that $\mathbf{x}$ satisfies $\mathbf{x}_2 = \tan\theta(\mathbf{x}_1 + t)$. So,

$$\begin{aligned} f^*(\mathbf{x}) &= -\tan^2\theta\mathbf{x}_1^2 + (a_1 + a_2\tan\theta - 2t\tan^2\theta)\mathbf{x}_1 + (a_0 + a_2\tan\theta t - \tan^2\theta t^2) \\ &= -\tan^2\theta\mathbf{x}_1^2 \leq 0. \end{aligned} \tag{4}$$

Thus, $f^*(\mathbf{x}) \leq 0$ for every $\mathbf{x}$ that satisfies $\mathbf{u}^* \cdot \mathbf{x} + t_1 = 0$. Based on (4), we show that $f^*(\mathbf{x}) \leq 0$ holds for every example $\mathbf{x}$ with $\mathbf{x}_2 = -\sigma t\tan\theta/2, \mathbf{x}_1 \leq -(1+\sigma/2)\tan\theta t$.

For every fixed $\mathbf{x}_2$, the partial derivative of $f^*(\mathbf{x})$ with respect to $\mathbf{x}_1$ is

$$\frac{\partial f^*(\mathbf{x})}{\partial \mathbf{x}_1} = a_1 = (2+\sigma)\tan^2\theta t > 0. \tag{5}$$

By (4), we know that the point $\mathbf{x}' := (-(1+\sigma/2)\tan\theta t, -\sigma t\tan\theta/2)$, the only vertex of the region N_1 satisfies $f^*(\mathbf{x}') \leq 0$. This implies that $f^*(\mathbf{x}) \leq 0$ holds for every example $\mathbf{x}$ with $\mathbf{x}_2 = -\sigma t\tan\theta/2, \mathbf{x}_1 \leq -(1+\sigma/2)\tan\theta t$.

So far, we have shown that $f^*(\mathbf{x}) \leq 0$ for every example $\mathbf{x}$ on the boundary of N_1. We next show that $f^*(\mathbf{x}) \leq 0$ holds for every example $\mathbf{x}$ in the interior of N_1. Fix any $\mathbf{x}_1$, the partial derivative of $f^*(\mathbf{x})$ with respect to $\mathbf{x}_2$ is

$$\frac{\partial f^*(\mathbf{x})}{\partial \mathbf{x}_2} = -2\mathbf{x}_2 + a_2 = -2\mathbf{x}_2 - \sigma\tan\theta t \leq \sigma\tan\theta t - \sigma\tan\theta t = 0, \tag{6}$$

when $\mathbf{x}_2 \geq -\sigma t\tan\theta/2$. Since $f^*(\mathbf{x}) \leq 0$ holds for every $\mathbf{x}$ on the boundary of N_1, (6) implies that $f^*(\mathbf{x}) \leq 0$ for every $\mathbf{x}$ in the interior of N_1.

In the rest of the proof, we show that $f^*(\mathbf{x}) \leq 0$ for every $\mathbf{x} \in N_2 = \{\mathbf{x} \in V \mid \mathbf{x}_2 \leq -\sigma t\tan\theta/2, \mathbf{v}^* \cdot \mathbf{x} + t_2 \leq 0\}$. Let $\mathbf{x}$ be any example on the decision boundary $\{\mathbf{x} \mid \mathbf{v}^* \cdot \mathbf{x} + t_2 = 0\}$. By Assumption 1 of $\mathbf{u}^*$ and t_1, we know that $\mathbf{x}$ satisfies $\mathbf{x}_2 = -\tan\theta(\mathbf{x}_1 + (1+\sigma)t)$.

$$\begin{aligned} f^*(\mathbf{x}) &= -\tan^2\theta\mathbf{x}_1^2 + (a_1 - a_2\tan\theta - 2(1+\sigma)t\tan^2\theta)\mathbf{x}_1 \\ &\quad + (a_0 - (1+\sigma)\tan\theta t a_2 - (1+\sigma)^2\tan^2\theta t^2) \\ &= -\tan^2\theta\mathbf{x}_1^2 \leq 0. \end{aligned}$$

Thus, $f^*(\mathbf{x}) \leq 0$ for every $\mathbf{x}$ that satisfies $\mathbf{v}^* \cdot \mathbf{x} + t_2 = 0$. Recall that $f^*(\mathbf{x}) \leq 0$ holds for every example $\mathbf{x}$ with $\mathbf{x}_2 = -\sigma t\tan\theta/2, \mathbf{x}_1 \leq -(1+\sigma/2)\tan\theta t$. Thus, $f^*(\mathbf{x}) \leq 0$ for every example $\mathbf{x}$ on the boundary of N_2. We next show that $f^*(\mathbf{x}) \leq 0$ holds for every example $\mathbf{x}$ in the interior of N_2. Fix any $\mathbf{x}_1$, the partial derivative of $f^*(\mathbf{x})$ with respect to $\mathbf{x}_2$ is

$$\frac{\partial f^*(\mathbf{x})}{\partial \mathbf{x}_2} = -2\mathbf{x}_2 + a_2 = -2\mathbf{x}_2 - \sigma\tan\theta t \geq \sigma\tan\theta t - \sigma\tan\theta t = 0, \tag{7}$$

when $\mathbf{x}_2 \leq -\sigma t\tan\theta/2$. Since $f^*(\mathbf{x}) \leq 0$ holds for every $\mathbf{x}$ on the boundary of N_2, (7) implies that $f^*(\mathbf{x}) \leq 0$ for every $\mathbf{x}$ in the interior of N_2. ■

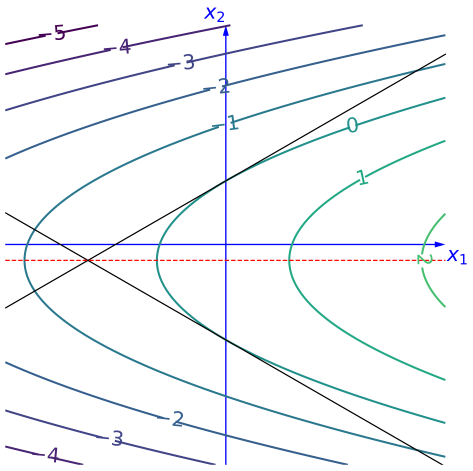

Figure 5: Illustration for Lemma 11. The target intersection of two halfspace h^* is plotted in black. h^* is symmetric according to the red dashed line $\mathbf{x}_2 = -\sigma t \tan\theta/2$. The red dashed line partitions the region of negative examples into two regions $N_1 := \{\mathbf{x} \in V \mid \mathbf{x}_2 \geq -\sigma t \tan\theta/2, \mathbf{u}^* \cdot \mathbf{x} + t_1 \leq 0\}$ and $N_2 := \{\mathbf{x} \in V \mid \mathbf{x}_2 \leq -\sigma t \tan\theta/2, \mathbf{v}^* \cdot \mathbf{x} + t_2 \leq 0\}$. Colored lines represent the contours of the polynomial f^*. $f^*(\mathbf{x}) < 0$ for every example $\mathbf{x}$ labeled negative by h^*.

B.5. Proof of Fact 1 and Fact 2

Fact 4 (restatement of Fact 1) *Let $h^* = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$ be the target hypothesis of an instance of the problem of learning intersections of two halfspaces and D be a distribution that is consistent with h^* Under Assumption 1, if $\mathbf{E}_{\mathbf{x}\sim D^+}(\mathbf{x}) = 0, \mathbf{E}_{\mathbf{x}\sim D^+} \mathbf{x}^{\otimes 3} = 0$ and for every $\mathbf{v} \in S^{d-1} \cap V, \mathbf{E}_{\mathbf{x}\sim D^+}(\mathbf{v} \cdot \mathbf{x})^2 = \alpha^2$, then $\alpha^2 \leq t^2 \sin^2\theta$.*

Proof [Proof of Fact 1]

For every $\mathbf{x}$ that is labeled positive by h^*, denote by $p(\mathbf{x})$ the variable of the density of a distribution D^+ over $\mathbb{R}^d$. Notice that any distribution D^+ that satisfies the statement of Fact 1, gives a feasible solution to the following LP (8). Thus, to upper bound the variance of D^+, it is equivalent to upper bound the optimal value of LP (8).

$$\begin{aligned} &\max \alpha^2 \\ &\text{s.t.} \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x} = 0 \\ &\quad \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}\mathbf{x}^\intercal = \alpha^2 I \\ &\quad \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}^{\otimes 3} = 0 \\ &\quad \sum_{\mathbf{x}} p(\mathbf{x}) = 1 \\ &\quad p(\mathbf{x}) \geq 0 \quad \forall \mathbf{x} \end{aligned} \tag{8}$$

To do this, we use an LP duality argument to derive a tight upper bound for the optimal value of LP (8).

Write $\mathbf{x} \in V$ as $\mathbf{x} = \mathbf{x}_1 e_1 + \mathbf{x}_2 e_2$ and $\mathbf{x}_0 = 1 \in \mathbb{R}$ for the simplicity of notation. Let $f(\mathbf{x}) = \sum_{i,j,k=0}^{2} a_{ijk}\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k$ be a degree-3 polynomial defined over $V = \mathbf{span}\{e_1, e_2\}$. The coefficient of $f(\mathbf{x})$ for the monomial $\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k$ is denoted by a_{ijk}. The dual linear program to LP (1) is defined by LP (2), whose variable is defined over the coefficients of $f(\mathbf{x})$.

$$\begin{aligned} &\min\ a_0 \\ &\text{s.t. } f(\mathbf{x}) \geq 0, \quad \forall \mathbf{x} \in \mathbf{supp}(D^+) \\ &\quad a_{11} + a_{22} = -1 \end{aligned} \tag{9}$$

Here, a_{11}, a_{22} are coefficients of polynomial f with respect to monomials $\mathbf{x}_1^2, \mathbf{x}_2^2$.

Every feasible solution to (9) defines a degree-3 polynomial $f(\mathbf{x})$ such that $f(\mathbf{x}) \geq 0$ for every example $\mathbf{x}$ that is labeled positive by h^*. In particular, by LP duality theory (Bertsimas and Tsitsiklis, 1997; Shapiro, 2001), the constant term a_0 of any feasible polynomial $f(\mathbf{x})$ to LP (9) gives an upper bound for the optimal value α^2 to LP (8). We explicitly construct the following polynomial feasible to LP (9), with a small constant term.

$$f^*(\mathbf{x}) = \frac{1}{t\sin\theta}(\mathbf{u}^* \cdot \mathbf{x} - t\sin\theta)^2(\mathbf{u}^* \cdot \mathbf{x} + t\sin\theta)$$

Notice that the constant term $a_0 = f^*(0) = t^2\sin^2\theta$, which means if $f^*(\mathbf{x})$ gives a feasible solution to LP (8), then $\alpha^2 \leq f^*(0) = t^2\sin^2\theta$. So, in the rest of the proof, we show that $f^*(\mathbf{x})$ gives a feasible solution to LP (9). By Lemma 10, we know that for every $\mathbf{x}$ such that $h^*(\mathbf{x}) = +1$, $f^*(\mathbf{x}) \geq 0$.

On the other hand, we show that the sum of coefficients of $f^*(\mathbf{x})$ for monomials $\mathbf{x}_1^2, \mathbf{x}_2^2$ is equal to -1. Notice that

$$f^*(\mathbf{x}) = \frac{1}{t\sin\theta}\left((\mathbf{u}^* \cdot \mathbf{x})^3 - t\sin\theta(\mathbf{u}^* \cdot \mathbf{x})^2 - (t\sin\theta)^2(\mathbf{u}^* \cdot \mathbf{x}) + (t\sin\theta)^3\right),$$

where the sum of coefficients of $f^*(\mathbf{x})$ for monomials $\mathbf{x}_1^2, \mathbf{x}_2^2$ is $-(\mathbf{u}_1^*)^2 - (\mathbf{u}_2^*)^2 = -\|\mathbf{u}^*\|^2 = -1$. This proves $f^*(\mathbf{x})$ gives a feasible solution to (9).

■

Fact 5 (restatement of Fact 2) *Let $h^* = \mathrm{sign}(\mathbf{u}^*\cdot\mathbf{x}+t_1)\wedge\mathrm{sign}(\mathbf{v}^*\cdot\mathbf{x}+t_2)$ be the target hypothesis of an instance of the problem of learning intersections of two halfspaces and D be a distribution that is consistent with h^* Under Assumption 1, if $\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x} = 0, \mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}^{\otimes 3} = 0$ and for every $\mathbf{v} \in S^{d-1} \cap V, \mathbf{E}_{\mathbf{x}\sim D^-}(\mathbf{v}\cdot\mathbf{x})^2 = \beta^2$, then $\beta^2 \geq (1+\sigma)t^2\tan^2\theta$.*

Proof [Proof of Fact 2] For every $\mathbf{x}$ that is labeled negative by h^*, denote by $p(\mathbf{x})$ the variable of the density of a distribution D^- over $\mathbb{R}^d$. Notice that any distribution D^- that satisfies the statement of Fact 2, gives a feasible solution to the following LP (10). Thus, to lower bound the variance of D^-, it is equivalent to upper bound the optimal value of LP (10).

$$\begin{aligned}
&\min \beta^2 \\
&\text{s.t.} \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x} = 0 \\
&\quad \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}\mathbf{x}^\intercal = \beta^2 I \\
&\quad \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}^{\otimes 3} = 0 \\
&\quad \sum_{\mathbf{x}} p(\mathbf{x}) = 1 \\
&\quad p(\mathbf{x}) \geq 0 \quad \forall \mathbf{x}
\end{aligned} \tag{10}$$

To do this, we use an LP duality argument to derive a tight lower bound for the optimal value of LP (10).

Write $\mathbf{x} \in V$ as $\mathbf{x} = \mathbf{x}_1 e_1 + \mathbf{x}_2 e_2$ and $\mathbf{x}_0 = 1 \in \mathbb{R}$ for the simplicity of notation. Let $f(\mathbf{x}) = \sum_{i,j,k=0}^{2} a_{ijk}\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k$ be a degree-3 polynomial defined over $V = \mathbf{span}\{e_1, e_2\}$. The coefficient of $f(x)$ for the monomial $\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k$ is denoted by a_{ijk}. The dual linear program to LP (10) is defined by LP (11), whose variable is defined over the coefficients of $f(x)$.

$$\begin{aligned}
&\max a_0 \\
&\text{s.t. } f(\mathbf{x}) \leq 0, \quad \forall \mathbf{x} \in \mathbf{supp}(D^-) \\
&\quad a_{11} + a_{22} = -1
\end{aligned} \tag{11}$$

Here, a_{11}, a_{22} are coefficients of polynomial f with respect to monomial x_1^2, x_2^2. Every feasible solution to (11) defines a degree-3 polynomial $f(\mathbf{x})$ such that $f(\mathbf{x}) \leq 0$ for every example $\mathbf{x}$ that is labeled negative by h^*. In particular, by LP duality theory (Bertsimas and Tsitsiklis, 1997), the constant term a_0 of any feasible polynomial $f(\mathbf{x})$ to LP (11) gives a lower bound for the optimal value β^2 to LP (10). We explicitly construct the following polynomial feasible to LP (11), with a large constant term.

$$\begin{aligned}
f^*(\mathbf{x}) &= a_0 + a_1\mathbf{x}_1 + a_2\mathbf{x}_2 - \mathbf{x}_2^2 \\
a_0 &= (1+\sigma)\tan^2\theta t^2 \\
a_1 &= (2+\sigma)\tan^2\theta t \\
a_2 &= -\sigma\tan\theta t
\end{aligned}$$

Notice that the sum of coefficients of $f^*(\mathbf{x})$ for monomials $\mathbf{x}_1^2, \mathbf{x}_2^2$ is equal to -1 and by Lemma 11, $f(\mathbf{x}) \leq 0, \quad \forall \mathbf{x} \in \mathbf{supp}(D^-)$. As the constant term a_0 of $f^*(\mathbf{x})$ is equal to $(1+\sigma)\tan^2\theta t^2$, this concludes the proof of Fact 2.

■

B.6. Proof of Lemma 9

In this section, we give the Proof of Lemma 9. For convenience, we restate Lemma 9 as follows.

Lemma 29 (restatement of Lemma 9) *Let D be a distribution over $\mathbb{B}^2(1)\times\{\pm 1\}$ that is consistent with an instance of learning intersections of two halfspaces with γ-margin assumption. Let $c > 0$ be any suitably large constant. Suppose*

1. $\|\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}\|_F \le \gamma^c$, $\|\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}\|_F \le \gamma^c$ *and* $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}\|_F$.
2. $\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}\mathbf{x}^\intercal = \alpha^2 I + \Delta_+, \mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}\mathbf{x}^\intercal = \alpha^2 I + \Delta_-$, *where* $\Delta_+, \Delta_- \in \mathbb{R}^{2\times 2}$ *are symmetric matrices such that* $\|\Delta_+\|_F \le \gamma^c, \|\Delta_-\|_F \le \gamma^c$ *and* $\alpha^2 > 0$.
3. $\left\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}^{\otimes 3}\right\|_F \le \gamma^c$

then $\left\|\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}^{\otimes 3}\right\|_F \ge \Omega(\gamma^2)$, $\left\|\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}^{\otimes 3}\right\|_F \ge \Omega(\gamma^2)$.

Proof [Proof of Lemma 29] We prove Lemma 9 by contradiction. Assuming $\left\|\mathbf{E}_{\mathbf{x}_V\sim D^+}\mathbf{x}_V^{\otimes 3}\right\|_F \le O(\gamma^2)$ or $\left\|\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}_V^{\otimes 3}\right\|_F \le O(\gamma^2)$ holds. We show there is no α^2 that can be used to fulfill the second condition in the statement of Lemma 9.

For every $\mathbf{x}$ that is labeled positive by h^*, denote by $p(\mathbf{x})$ the variable of the density of a distribution D^+ over $\mathbb{R}^d$. Under margin assumption, every positive example $\mathbf{x}$ satisfies $\mathbf{u}^*\cdot\mathbf{x}+t_1 \ge \gamma$ and $\mathbf{v}^*\cdot\mathbf{x}+t_2 \ge \gamma$. Denote by $S_\gamma^+ := \{\mathbf{x}\in\mathbb{R}^2 \mid \mathbf{u}^*\cdot\mathbf{x}+t_1 \ge \gamma \text{ and } \mathbf{v}^*\cdot\mathbf{x}+t_2 \ge \gamma\}$. Let $b = \gamma^c \ge 0$ be a small positive number that represents the level of perturbation for the LP (8). Notice that Frobenius norm is always an upper bound of infinity norm. Thus, any distribution D^+ that satisfies the statement of Lemma 9 under the γ-margin assumption, gives a feasible solution to the following LP (12).

$$
\begin{aligned}
&\max \alpha^2 \\
&\text{s.t. } -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1 \le b \\
&\quad -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_2 \le b \\
&\quad -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1^2 - \alpha^2 \le b \\
&\quad -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_2^2 - \alpha^2 \le b \\
&\quad -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1\mathbf{x}_2 \le b \\
&\quad -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1^3 \le b \\
&\quad -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1^2\mathbf{x}_2 \le b \\
&\quad -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1\mathbf{x}_2^2 \le b \\
&\quad -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_2^3 \le b \\
&\quad \sum_{\mathbf{x}} p(\mathbf{x}) = 1 \\
&\quad p(\mathbf{x}) \ge 0 \quad \forall \mathbf{x}\in S_\gamma^+
\end{aligned}
\tag{12}
$$

Write $\mathbf{x} \in V$ as $\mathbf{x} = \mathbf{x}_1 e_1 + \mathbf{x}_2 e_2$ and $\mathbf{x}_0 = 1$ for the simplicity of notation. Let $f(\mathbf{x}) = \sum_{i,j,k=0}^{2} a_{ijk}\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k$ be a degree-3 polynomial defined over $V = \mathbf{span}\{e_1, e_2\}$. The coefficient of $f(\mathbf{x})$ for the monomial $\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k$ is denoted by a_{ijk}. The dual linear program to LP (12) is defined by LP (13), whose variable is defined over the coefficients of $f(\mathbf{x})$.

$$\begin{aligned} &\min\ a_0 + b(a_1 + a_2 + a_{11} + a_{22} + a_{111} + a_{112} + a_{122} + a_{222}) \\ &\text{s.t. } f(\mathbf{x}) \le 0, \quad \forall \mathbf{x} \in S_\gamma^+ \\ &\quad a_{11} + a_{22} \le -1 \end{aligned} \tag{13}$$

We construct an upper bound for the optimal value α^2 by constructing a feasible solution to (13) with a small objective value. Consider the following polynomial

$$f^*(\mathbf{x}) = \frac{1}{(t\sin\theta - \gamma/2)}(\mathbf{u}^* \cdot \mathbf{x} - (t\sin\theta - \gamma/2))^2(\mathbf{u}^* \cdot \mathbf{x} + (t\sin\theta - \gamma/2))$$

Under margin assumption, every positive example $\mathbf{x}$ satisfies $\mathbf{u}^* \cdot \mathbf{x} + t_1 \ge \gamma$ and $\mathbf{v}^* \cdot \mathbf{x} + t_2 \ge \gamma$. That is to say, D^+ is also consistent with an intersection of halfspaces $h'(\mathbf{x}) = \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1 - \gamma/2) \wedge \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2 - \gamma/2)$. Thus, $f^*(\mathbf{x}) \ge 0$ for every positive example by Lemma 10.

On the other hand,

$$f^*(\mathbf{x}) = \frac{1}{(t\sin\theta - \gamma/2)}\left((\mathbf{u}^* \cdot \mathbf{x})^3 - (t\sin\theta - \gamma/2)(\mathbf{u}^* \cdot \mathbf{x})^2 - (t\sin\theta - \gamma/2)^2(\mathbf{u}^* \cdot \mathbf{x}) + (t\sin\theta - \gamma/2)^3\right),$$

where the sum of coefficients of $f^*(\mathbf{x})$ for monomials $\mathbf{x}_1^2, \mathbf{x}_2^2$ is $-(\mathbf{u}_1^*)^2 - (\mathbf{u}_2^*)^2 = -\|\mathbf{u}^*\|^2 = -1$. This proves $f^*(\mathbf{x})$ gives a feasible solution to (13). Notice that $t\sin\theta$, the distance between the origin and halfspace h_1^* is at least γ, otherwise, the origin is not labeled positive by h^*. On the other hand, $t\sin\theta \le 1$, because otherwise, no example in $\mathbb{B}(1)$ is labeled negative and D^- is not well-defined under the γ-margin assumption. This implies the objective value corresponding to the solution to LP (12) is

$$\begin{aligned} obj(f^*) &:= (t\sin\theta - \gamma/2)^2 + \frac{b}{t\sin\theta - \gamma/2}((\mathbf{u}_1^*)^3 + (\mathbf{u}_1^*)^2\mathbf{u}_2^* + \mathbf{u}_1^*(\mathbf{u}_2^*)^2 + (\mathbf{u}_2^*)^3) - b(t\sin\theta - \gamma/2)(\mathbf{u}_1^* + \mathbf{u}_2^*) \\ &\le t^2\sin^2\theta - \gamma/4 + O(b/\gamma) \\ &\le t^2\sin^2\theta - \gamma/8, \end{aligned}$$

when $b \le O(\gamma^2)$. This implies that $\alpha^2 \le t^2\sin^2\theta - \Omega(\gamma)$

On the other hand, we derive a lower bound for α^2. For every $\mathbf{x}$ that is labeled negative by h^*, denote by $p(\mathbf{x})$ the variable of the density of a distribution D^- over $\mathbb{R}^d$. Denote by $S_\gamma^- := \{\mathbf{x} \in \mathbb{R}^2 \mid \mathbf{u}^* \cdot \mathbf{x} + t_1 \le -\gamma \text{ or } \mathbf{v}^* \cdot \mathbf{x} + t_2 \le -\gamma\}$. Under γ-margin assumption, any example $\mathbf{x}$ with $h^*(\mathbf{x}) = -1$ satisfies $\mathbf{x} \in S_\gamma^-$. Let $b = \gamma^c \ge 0$ be a small positive number that represents the level of perturbation for the LP (10). Notice that Frobenius norm is always an upper bound of infinity norm. Thus, any distribution D^- that satisfies the statement of Lemma 9 under the γ-margin assumption. Thus, to lower bound the variance of D^-, it is equivalent to lower bound the optimal value of LP (14).

$$
\begin{aligned}
\min\ & \alpha^2 \\
\text{s.t.}\ & -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1 \le b \\
& -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_2 \le b \\
& -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1^2 - \alpha^2 \le b \\
& -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_2^2 - \alpha^2 \le b \\
& -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1\mathbf{x}_2 \le b \\
& -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1^3 \le b \\
& -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1^2\mathbf{x}_2 \le b \\
& -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_1\mathbf{x}_2^2 \le b \\
& -b \le \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}_2^3 \le b \\
& \sum_{\mathbf{x}} p(\mathbf{x}) = 1 \\
& p(\mathbf{x}) \ge 0 \quad \forall \mathbf{x} \in S_\gamma^-
\end{aligned}
\tag{14}
$$

To do this, we use an LP duality argument to derive a tight lower bound for the optimal value of LP (10).

Write $\mathbf{x} \in V$ as $\mathbf{x} = \mathbf{x}_1 e_1 + \mathbf{x}_2 e_2$ and $\mathbf{x}_0 = 1 \in \mathbb{R}$ for the simplicity of notation. Let $f(\mathbf{x}) = \sum_{i,j,k=0}^{2} a_{ijk}\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k$ be a degree-3 polynomial defined over $V = \mathbf{span}\{e_1, e_2\}$. The coefficient of $f(\mathbf{x})$ for the monomial $\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k$ is denoted by a_{ijk}. The dual linear program to LP (10) is defined by LP (11), whose variable is defined over the coefficients of $f(\mathbf{x})$.

$$
\begin{aligned}
\max\ & a_0 - b(a_1 + a_2 + a_{111} + a_{112} + a_{122} + a_{222}) \\
\text{s.t.}\ & f(\mathbf{x}) \le 0, \quad \forall \mathbf{x} \in S_\gamma^- \\
& a_{11} + a_{22} \ge -1
\end{aligned}
\tag{15}
$$

We construct a lower bound for the optimal value α^2 by constructing a feasible solution to (14) with a large objective value. Consider the following polynomial

$$
\begin{aligned}
f^*(\mathbf{x}) &= a_0 + a_1\mathbf{x}_1 + a_2\mathbf{x}_2 - \mathbf{x}_2^2 \\
a_0 &= (1+\sigma)\tan^2\theta t^2 \\
a_1 &= (2+\sigma)\tan^2\theta t \\
a_2 &= -\sigma\tan\theta t
\end{aligned}
$$

By Lemma 11, we know that $f^*(\mathbf{x})$ gives a feasible solution to (15). In the rest of the proof, we show the feasible solution corresponds to $f^*(\mathbf{x})$ has a large objective value. When $b = \gamma^c < O(\gamma^2)$, the objective value is

$$\begin{aligned} obj(f^*) &:= (1+\sigma)\tan^2\theta t^2 - b((2+\sigma)\tan^2\theta t - \sigma\tan\theta t) \\ &\geq (1+\sigma)\tan^2\theta t^2 - b(2+\sigma)\tan^2\theta t = \tan^2\theta t^2(1+\sigma - \frac{b(2+\sigma)}{t}) \\ &\geq \tan^2\theta t^2(1+\sigma - \frac{b(2+\sigma)}{t\sin\theta}) \geq \tan^2\theta t^2(1+\sigma - \frac{b(2+\sigma)}{\gamma}) \\ &= \tan^2\theta t^2(1-O(\gamma)) + \sigma\tan^2\theta t^2(1-O(\gamma)) \geq \tan^2\theta t^2(1-O(\gamma)). \end{aligned}$$

Here, we use the fact that $t\sin\theta > \gamma$. We consider two cases. In the first case, $\cos^2\theta \leq 1 - O(\gamma)$. In this case, we have $obj(f^*) \geq \sin^2\theta t^2$. In the second case, $\cos^2\theta \geq 1 - O(\gamma)$. In this case, we have

$$obj(f^*) \geq \tan^2\theta t^2 - O(\gamma)\sin^2\theta t^2 \geq \sin^2\theta t^2 - O(\gamma) \geq \sin^2\theta t^2 - \gamma/16.$$

Thus, we conclude $\alpha^2 \geq \sin^2\theta t^2 - \gamma/16$.

To conclude the proof of Lemma 9, we without loss of generality to assume $\left\|E_{\mathbf{x}\sim D^-}\mathbf{x}^{\otimes 3}\right\|_F \leq O(\gamma^c)$. Since $E_{\mathbf{x}\sim D^+}\mathbf{x}^{\otimes 3}$ is close to $E_{\mathbf{x}\sim D^-}\mathbf{x}^{\otimes 3}$, we know that D^+ gives a feasible solution to LP (12) with $\alpha^2 \leq t^2\sin^2\theta - \Omega(\gamma)$ and D^- gives a feasible solution to LP (12) with $\alpha^2 \geq \sin^2\theta t^2 - \gamma/16$, which gives a contradiction.

■

B.7. Proof of Theorem 6

In this section, present the full proof of Theorem 6. For convenience, we restate Theorem 6 as follows.

Theorem 30 *(restatement of Theorem 6)*
Let D be a distribution over $\mathbb{B}^2(1) \times \{\pm 1\}$ that is consistent with an instance of learning intersections of two halfspaces with γ-margin assumption. Let $c > 0$ be any *suitably large constant. Suppose*

1. $\|\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}\|_F, \|\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}\|_F \leq \gamma^c$.
2. $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^+})\mathbf{x}\mathbf{x}^\intercal\|_F \leq \gamma^c$
3. $\left\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}^{\otimes 3}\right\|_F \leq \gamma^c$,

then $\left\|\mathbf{E}_{\mathbf{x}\sim D^+}\mathbf{x}^{\otimes 3}\right\|_F, \left\|\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}^{\otimes 3}\right\|_F = \Omega(\gamma^{15})$.

Proof [Proof of Theorem 30] For every example $\mathbf{x} \in V$, we consider the following linear transformation of $\mathbf{x}$.

$$\tilde{\mathbf{x}} := \left\|(\Sigma^+)^{-1/2}\right\|_2^{-1}(\Sigma^+)^{-1/2}\mathbf{x},$$

Notice that

$$\|\tilde{\mathbf{x}}\|_2 = \left\|(\Sigma^+)^{-1/2}\right\|_2^{-1} \left\|(\Sigma^+)^{-1/2}\mathbf{x}\right\|_2 \leq \left\|(\Sigma^+)^{-1/2}\right\|_2^{-1} \left\|(\Sigma^+)^{-1/2}\right\|_2 \|\mathbf{x}\|_2 \leq 1$$

Denote by $y(\tilde{\mathbf{x}}) := h^*(\mathbf{x})$, we will show that $\tilde{\mathbf{x}}$ is labeled by another intersections of two halfspaces $\tilde{h}(\tilde{\mathbf{x}}) = \tilde{h}_1(\tilde{\mathbf{x}}) \wedge \tilde{h}_2(\tilde{\mathbf{x}})$ with $\mathrm{poly}(\gamma)$-margin assumption. Consider the first ground truth halfspace $h_1^* := \mathrm{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1)$. We have

$$\begin{aligned} h_1^*(\mathbf{x}) &= \mathrm{sign}\left(\mathbf{u}^* \cdot \mathbf{x} + t_1\right) = \mathrm{sign}\left((\Sigma^+)^{1/2}\mathbf{u}^* \cdot (\Sigma^+)^{-1/2}\mathbf{x} + t_1\right) \\ &= \mathrm{sign}\left(\left\|(\Sigma^+)^{-1/2}\right\|_2 (\Sigma^+)^{1/2}\mathbf{u}^* \cdot \left\|(\Sigma^+)^{-1/2}\right\|_2^{-1} (\Sigma^+)^{-1/2}\mathbf{x} + t_1\right) \\ &= \mathrm{sign}\left(\mathbf{u}' \cdot \tilde{\mathbf{x}} + t_1\right) = \mathrm{sign}\left(\tilde{\mathbf{u}} \cdot \tilde{\mathbf{x}} + t_1 / \left\|\mathbf{u}'\right\|_2\right) = \tilde{h}_1(\tilde{\mathbf{x}}). \end{aligned}$$

Here $\mathbf{u}' := \left\|(\Sigma^+)^{-1/2}\right\|_2 \Sigma^{1/2}\mathbf{u}^*$ and $\tilde{\mathbf{u}} = \mathbf{u}' / \|\mathbf{u}'\|_2$. Since $\left\|(\Sigma^+)^{1/2}\right\|_2 \leq 1$ and by Lemma 8, $\left\|(\Sigma^+)^{-1/2}\right\|_2 \leq \gamma^{-2}$, we know that $\|\mathbf{u}'\|_2 \leq \gamma^{-2}$. Since $\mathbf{x}$ satisfies γ-margin assumption with respect to h^* and $\|\mathbf{u}'\|_2 \leq \gamma^{-2}$, we know that

$$\left|\tilde{\mathbf{u}} \cdot \tilde{\mathbf{x}} + t_1 / \left\|\mathbf{u}'\right\|_2\right| = \left|\mathbf{u}' \cdot \tilde{\mathbf{x}} + t_1\right| / \left\|\mathbf{u}'\right\|_2 = \left|\mathbf{u}^* \cdot \mathbf{x} + t_1\right| / \left\|\mathbf{u}'\right\|_2 \geq \gamma / \left\|\mathbf{u}'\right\|_2 = \gamma^3.$$

Similarly, for the second ground truth halfspace $h_2^* := \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2)$, we have

$$\begin{aligned} h_2^*(\mathbf{x}) &= \mathrm{sign}\left(\mathbf{v}^* \cdot \mathbf{x} + t_2\right) = \mathrm{sign}\left((\Sigma^+)^{1/2}\mathbf{v}^* \cdot (\Sigma^+)^{-1/2}\mathbf{x} + t_2\right) \\ &= \mathrm{sign}\left(\left\|(\Sigma^+)^{-1/2}\right\|_2 (\Sigma^+)^{1/2}\mathbf{v}^* \cdot \left\|(\Sigma^+)^{-1/2}\right\|_2^{-1} (\Sigma^+)^{-1/2}\mathbf{x} + t_2\right) \\ &= \mathrm{sign}\left(\mathbf{v}' \cdot \tilde{\mathbf{x}} + t_2\right) = \mathrm{sign}\left(\tilde{\mathbf{v}} \cdot \tilde{\mathbf{x}} + t_2 / \left\|\mathbf{v}'\right\|_2\right) = \tilde{h}_2(\tilde{\mathbf{x}}). \end{aligned}$$

Here $\mathbf{v}' := \left\|(\Sigma^+)^{-1/2}\right\|_2 \Sigma^{1/2}\mathbf{v}^*$ and $\tilde{\mathbf{v}} = \mathbf{v}' / \|\mathbf{v}'\|_2$. Since $\left\|(\Sigma^+)^{1/2}\right\|_2 \leq 1$ and by Lemma 8, $\left\|(\Sigma^+)^{-1/2}\right\|_2 \leq \gamma^{-2}$, we know that $\|\mathbf{v}'\|_2 \leq \gamma^{-2}$. Since $\mathbf{x}$ satisfies γ-margin assumption with respect to h^* and $\|\mathbf{v}'\|_2 \leq \gamma^{-2}$, we know that

$$\left|\tilde{\mathbf{v}} \cdot \tilde{\mathbf{x}} + t_1 / \left\|\mathbf{v}'\right\|_2\right| = \left|\mathbf{v}' \cdot \tilde{\mathbf{x}} + t_2\right| / \left\|\mathbf{v}'\right\|_2 = \left|\mathbf{v}^* \cdot \mathbf{x} + t_1\right| / \left\|\mathbf{v}'\right\| \geq \gamma / \left\|\mathbf{v}'\right\| = \gamma^3.$$

This implies that $\tilde{\mathbf{x}}$ is labeled by an intersections of two halfspaces $\tilde{h}(\tilde{\mathbf{x}}) = \tilde{h}_1(\tilde{\mathbf{x}}) \wedge \tilde{h}_2(\tilde{\mathbf{x}})$ with γ^3-margin assumption.

Next, we show that the marginal distribution of $\tilde{\mathbf{x}}$ satisfies the conditions in the statement of Lemma 9. Recall that the linear transformation $\tilde{\mathbf{x}}$ preserves the labels of $\mathbf{x}$. Consider the distributions of $\tilde{\mathbf{x}}$ with positive labels. We have

$$\begin{aligned} \left\|\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+} \tilde{\mathbf{x}}\right\|_2 &= \mathop{\mathbf{E}}_{\mathbf{x}\sim D^+} \left\|\left\|(\Sigma^+)^{-1/2}\right\|_2^{-1} (\Sigma^+)^{-1/2}\mathbf{x}\right\| = \left\|(\Sigma^+)^{-1/2}\right\|_2^{-1} \left\|\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+} (\Sigma^+)^{-1/2}\mathbf{x}\right\|_2 \\ &\leq \left\|(\Sigma^+)^{-1/2}\right\|_2^{-1} \left\|(\Sigma^+)^{-1/2}\right\|_2 \left\|\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+} \mathbf{x}\right\|_2 = \left\|\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+} \mathbf{x}\right\|_2 \leq \gamma^c. \end{aligned}$$

Similarly, consider the distributions of $\tilde{\mathbf{x}}$ with positive labels. We have

$$\left\|\mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}\tilde{\mathbf{x}}\right\|_2 = \mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}\left\|\left\|(\Sigma^+)^{-1/2}\right\|_2^{-1}(\Sigma^+)^{-1/2}\mathbf{x}\right\|_2 = \left\|(\Sigma^+)^{-1/2}\right\|_2^{-1}\left\|\mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}(\Sigma^+)^{-1/2}\mathbf{x}\right\|_2$$
$$\leq \left\|(\Sigma^-)^{-1/2}\right\|_2^{-1}\left\|(\Sigma^+)^{-1/2}\right\|_2\left\|\mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}\mathbf{x}\right\|_2 = \left\|\mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}\mathbf{x}\right\|_2 \leq \gamma^c.$$

Thus, $\tilde{\mathbf{x}}$ satisfies the first condition of Lemma 9. We next show that $\tilde{\mathbf{x}}$ satisfies the second condition of Lemma 9. The covariance matrix of the positive $\tilde{\mathbf{x}}$ is

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal = \left\|(\Sigma^+)^{-1}\right\|_2^{-1}(\Sigma^+)^{-1/2}\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}\mathbf{x}\mathbf{x}^\intercal(\Sigma^+)^{-1/2} = \left\|(\Sigma^+)_2^{-1}\right\|^{-1}I.$$

On the other hand, the covariance of the negative $\tilde{\mathbf{x}}$ is

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal = \left\|(\Sigma^+)^{-1}\right\|_2^{-1}(\Sigma^+)^{-1/2}\mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}\mathbf{x}\mathbf{x}^\intercal(\Sigma^+)^{-1/2} = \left\|(\Sigma^+)^{-1}\right\|_2^{-1}(\Sigma^+)^{-1/2}\Sigma^-(\Sigma^+)^{-1/2}$$
$$= \left\|(\Sigma^+)^{-1}\right\|_2^{-1}(\Sigma^+)^{-1/2}\Sigma^+(\Sigma^+)^{-1/2} - \left\|(\Sigma^+)^{-1}\right\|_2^{-1}(\Sigma^+)^{-1/2}(\Sigma^+-\Sigma^-)(\Sigma^+)^{-1/2}$$
$$= \left\|(\Sigma^+)^{-1}\right\|_2^{-1}I - \left\|(\Sigma^+)^{-1}\right\|_2^{-1}(\Sigma^+)^{-1/2}(\Sigma^+-\Sigma^-)(\Sigma^+)^{-1/2}.$$

Since $\|(\Sigma^+-\Sigma^-)\|_F \leq \gamma^c$, we know that

$$\left\|(\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}-\mathop{\mathbf{E}}_{\mathbf{x}\sim D^-})\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal\right\|_F = \left\|\left\|(\Sigma^+)^{-1}\right\|_2^{-1}(\Sigma^+)^{-1/2}(\Sigma^+-\Sigma^-)(\Sigma^+)^{-1/2}\right\|_F \leq \left\|(\Sigma^+-\Sigma^-)\right\|_F \leq \gamma^c.$$

Thus, the marginal distribution of $\tilde{\mathbf{x}}$ satisfies the conditions in the statement of Lemma 9.

Finally, we show the third condition of Lemma 9 holds. We have

$$\left\|(\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}-\mathop{\mathbf{E}}_{\mathbf{x}\sim D^-})\tilde{\mathbf{x}}^{\otimes 3}\right\|_F = \left\|\left\|(\Sigma^+)^{-1/2}\right\|_2^{-3}((\Sigma^+)^{-1/2})^{\otimes 3}\left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}\mathbf{x}^{\otimes 3}-\mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}\mathbf{x}^{\otimes 3}\right)\right\|_F$$
$$\leq O(\left\|\left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}\mathbf{x}^{\otimes 3}-\mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}\mathbf{x}^{\otimes 3}\right)\right\|_F) \leq \gamma^c.$$

Thus, the third condition in the statement of Lemma 9 holds for $\tilde{\mathbf{x}}$. By Lemma 9, $\left\|\mathbf{E}_{\mathbf{x}\sim D^+}\tilde{\mathbf{x}}^{\otimes 3}\right\|_F \geq \Omega(\gamma^6)$, $\left\|\mathbf{E}_{\mathbf{x}\sim D^-}\tilde{\mathbf{x}}^{\otimes 3}\right\|_F \geq \Omega(\gamma^6)$. By Lemma 8, we know that $\left\|(\Sigma^+)^{-1/2}\right\|_F \leq O(\gamma^2)$. Thus, $\left\|\mathbf{E}_{\mathbf{x}_V\sim D^+}\mathbf{x}_V^{\otimes 3}\right\|_F \geq \Omega(\gamma^{15})$, $\left\|\mathbf{E}_{\mathbf{x}\sim D^-}\mathbf{x}_V^{\otimes 3}\right\|_F \geq \Omega(\gamma^{15})$.

■

Appendix C. Omitted Proofs from Section 3

C.1. Proof of Theorem 12

In this section, we present the proof of Theorem 12. The algorithm we will analyze is Algorithm 2. For convenience, we restate Theorem 12 below.

Theorem 31 (restatement of Theorem 12) *There is a learning algorithm $\mathcal{A}$ such that for every c, a suitably large constant and any instance of learning intersections of two halfspaces under factorizable distribution with γ-margin assumption if the input distribution D satisfies*

Algorithm 2 SQ-DIRECTION FINDING (SQ-efficient algorithm for finding relevant direction with matched moments)

1: **Input:** $\gamma \in (0,1)$ and i.i.d. sample access to a distribution D on $\mathbb{B}^d(1) \times \{\pm 1\}$ that is an instance of learning intersections of halfspaces under product distribution. Suppose that D satisfies the conditions in the statement of Theorem 12.

2: **Output:** $\mathcal{O}$, a list of $\mathrm{poly}(d)$, $\mathbf{w} \in S^{d-1}$ such that at least one of $\mathbf{w} \in \mathcal{O}$ satisfies $\|\mathrm{proj}_{V^\perp}\mathbf{w}\| \leq \mathrm{poly}(\gamma)$.

3: Take S_1, a set of $m_1 = \mathrm{poly}(d/\gamma)$ i.i.d. samples from D_X to estimate $\mu := \mathbf{E}_{\mathbf{x}\sim D_X}\mathbf{x}$ with

$$\hat{\mu} := \frac{1}{m_1}\sum_{\mathbf{x}\in S_1}\mathbf{x}$$

up to $\mathrm{poly}(\gamma/d)$ error

4: Take S, a set of $N = \mathrm{poly}(d/\gamma)$ i.i.d. samples from D_X and estimate

$$\hat{T} = \frac{1}{N}\sum_{\mathbf{x}\in S}(\mathbf{x}-\hat{\mu})^{\otimes 3}$$

5: Define $\hat{f} : S^{d-1} \to \mathbb{R}$ as $\hat{f}(\mathbf{u}) := \hat{T}\cdot\mathbf{u}^{\otimes 3} = \hat{\mathbf{E}}_{\mathbf{x}\sim S}\left((\mathbf{x}-\hat{\mu})\cdot\mathbf{u}\right)^3$.

6: $\mathcal{O} := \emptyset$

7: **for** $t = 1, \ldots, T = \mathrm{poly}(d)$ **do**

8: Find a $(\gamma/d)^{c'}$-approximate solution $\mathbf{u}_t$ to $\hat{f}$ such that for every $\mathbf{u} \in \mathcal{O}$, $|\mathbf{u}_t \cdot \mathbf{u}| \leq \mathrm{poly}(\gamma/d)$, where $c' > 0$ is a large constant.

9: $\mathcal{O} \leftarrow \mathcal{O} \cup \{\mathbf{u}_t\}$. (If no such $\mathbf{u}_t$ is found, return $\mathcal{O}$)

10: **return** $\mathcal{O}$

1. $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}\|_F \leq \gamma^c$.
2. $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}\mathbf{x}^\intercal\|_F \leq \gamma^c$
3. $\left\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}^{\otimes 3}\right\|_F \leq \gamma^c$

then it makes $\mathrm{poly}(d, 1/\gamma)$ *many statistical queries, where each query has tolerance at most* $\mathrm{poly}(1/d, \gamma)$ *and outputs a list of* $\mathrm{poly}(d/\gamma)$ *directions* $\mathbf{w} \in \mathbb{R}^d$ *such that one* $\mathbf{w}$ *in the list satisfies* $\|\mathbf{w}_W\|_2 \leq \mathrm{poly}(\gamma/d)$.

Notice that Algorithm 2 only uses $\mathrm{poly}(d/\gamma)$ i.i.d. unlabeled examples drawn from D_X to estimate the mean and third-moment tensor of D_X up to $\mathrm{poly}(\gamma/d)$ accuracy, thus these steps can be implemented via $\mathrm{poly}(d)$ SQs, each of which has tolerance $\mathrm{poly}(\gamma/d)$. For detailed background about the SQ model, we refer the reader to Section A.

Here, we give an overview of the proof of Theorem 12. To simplify the notation, we define η-approximate solution for a polynomial function defined over the unit sphere.

Definition 32 (η-approximate solution) *Let $f : S^{d-1} \to \mathbb{R}$ be a polynomial function. For $\eta > 0$ and $\mathbf{u} \in S^{d-1}$, we say a point $\mathbf{u} \in S^{d-1}$ is a η-approximate solution if*

1. $f(\mathbf{u}) \geq \eta_0 = \Omega(\eta)$

2. $\|\mathrm{proj}_{\mathbf{u}^\perp} \nabla f(\mathbf{u})\| \le \eta_1 = O(\eta^2)$

3. $\max_{\mathbf{z} \in S^{(d-1)} \cap \mathbf{u}^\perp} \mathbf{z}^\intercal \nabla^2 f(\mathbf{u}) \mathbf{z} \le \eta_2 = O(\eta^2)$

Recall the observation obtained by Vempala and Xiao (2011) summarized as Fact 6.

Fact 6 (Lemma 4 in Vempala and Xiao (2011)) *Let $D_X = D_V \times D_W$ be factorizable distribution over $\mathbb{R}^d$ such that D_X has the same $m-1 (m \ge 3)$ moments as Gaussian but has a different mth moment. Then any local maximum(minimum) $\mathbf{u}^*$ of $f^*(\mathbf{u}^*)$ over S^{d-1} with $f^*(\mathbf{u}^*) > \gamma_m (f^*(\mathbf{u}^*) < \gamma_m)$ must be either in V or W. Here $f^*(\mathbf{u}) = \mathbf{E}_{\mathbf{x} \sim D} (\mathbf{u} \cdot \mathbf{x})^m$ and γ_m is the m-th moment of a standard 1-dimensional Gaussian distribution.*

The first ingredient of the proof is to show a robust version of Fact 6 for the polynomial $f^*(\mathbf{u}) = T^* \cdot \mathbf{u}^{\otimes 3} = \mathbf{E}_{\mathbf{x} \sim D_X} (\mathbf{x} \cdot \mathbf{u})^3$. That is to say, given $\mathbf{E}_{\mathbf{x} \sim D_X}(\mathbf{x})$ is close to 0, any $\mathrm{poly}(\gamma/d)$ approximate solution to f^* must be $\mathrm{poly}(\gamma/d)$-close to V or W. In particular, by Theorem 6, the moment tensor of D_V must have norm at least γ^c for some constant c, which implies that some point in V must be a $\mathrm{poly}(\gamma/d)$-approximate solution to f^*. On the other hand, since a polynomial function is completely described by its moment tensor. By estimating each entry of the moment tensor up to error $\mathrm{poly}(\gamma/d)$ error, the approximate objective function $\hat{f}$ has a similar structure as the one of f^* and optimizing $\hat{f}$ is enough to give us an approximate solution to f^*.

The second ingredient is the key difference from the prior work (Frieze et al., 1996; Vempala and Xiao, 2011). After obtaining the first approximate solution $\mathbf{u}_1$, instead of restricting $\hat{f}$ over the subspace $(\mathbf{u}_1)^\perp$, we will look at a small band $B = \{\mathbf{x} \in S^{d-1} \mid |\mathbf{u}_1 \cdot x| \le \mathrm{poly}(\gamma/d)\}$. As long as no $\mathbf{u}$, a $\mathrm{poly}(\gamma/d)$ approximate solution close to V is added to the list $\mathcal{O}$, we must be able to add another point to $\mathcal{O}$. By Alon (2003), if each pair of $\mathbf{u}, \mathbf{u}' \in \mathcal{O}$ satisfies $|\mathbf{u} \cdot \mathbf{u}'| \le \mathrm{poly}(\gamma/d)$, the size of $\mathcal{O}$ is at most $\mathrm{poly}(d/\gamma)$. This implies that Algorithm 2 can terminate. Now we are able to present the full proof.

Proof [Proof of Theorem 31] To start with, we consider the case where we have the exact access to the moment tensor of D_X. Without loss of generality, we assume $\|\mu := \mathbf{E}_{\mathbf{x} \sim D_X} \mathbf{x}\|_2 \le o((\gamma/d)^{2c_1})$ for some large constant $c_1 > c$, because otherwise, we can estimate μ up to error $(\gamma/d)^{2c_1}$ and shift D_X to μ and rescale each shifted example by a factor of 2 to make $\|\mathbf{x}\|_2 \le 1$. This will only decrease the margin γ by a factor of at most 2. Let $T^* := \mathbf{E}_{\mathbf{x} \sim D_X} \mathbf{x}^{\otimes 3}$ be the 3rd moment tensor of D_X and $f^*(\mathbf{u}) = T^* \cdot \mathbf{u}^{\otimes 3} = \mathbf{E}_{\mathbf{x} \sim D_X} (\mathbf{x} \cdot \mathbf{u})^3$.

Consider any $\mathbf{u} \in S^{(d-1)}$. We write $\mathbf{u} = s\mathbf{u}_V^0 + t\mathbf{u}_W^0$, where $s \ge 0, t \ge 0, s^2 + t^2 = 1$ and $\mathbf{u}_V^0 := \mathbf{u}_V / \|\mathbf{u}_V\|, \mathbf{u}_W^0 := \mathbf{u}_W / \|\mathbf{u}_W\|$. We show that if $\mathbf{u}$ is an $\eta := (\gamma/d)^{c_1}$ approximate solution to f^*, then $\|\mathbf{u}_W\| \le \mathrm{poly}(\gamma/d)$ or $\|\mathbf{u}_W\| \le \mathrm{poly}(\gamma/d)$. Observe that

$$\begin{aligned} f^*(\mathbf{u}) &= \mathop{\mathbf{E}}_{\mathbf{x} \sim D_X} (\mathbf{x} \cdot \mathbf{u})^3 = \mathop{\mathbf{E}}_{\mathbf{x} \sim D_X} (\mathbf{x}_V \cdot \mathbf{u}_V + \mathbf{x}_W \cdot \mathbf{u}_W)^3 \\ &= \mathop{\mathbf{E}}_{\mathbf{x} \sim D_X} (\mathbf{x}_V \cdot \mathbf{u}_V)^3 + (\mathbf{x}_W \cdot \mathbf{u}_W)^3 + 3(\mathbf{x}_V \cdot \mathbf{u}_V)(\mathbf{x}_W \cdot \mathbf{u}_W)^2 + 3(\mathbf{x}_V \cdot \mathbf{u}_V)^2 (\mathbf{x}_W \cdot \mathbf{u}_W) \\ &= g^*(\mathbf{u}) + h^*(\mathbf{u}). \end{aligned}$$

Here, $g^*(\mathbf{u}) = \mathbf{E}_{\mathbf{x} \sim D_X} (\mathbf{x}_V \cdot \mathbf{u}_V)^3 + (\mathbf{x}_W \cdot \mathbf{u}_W)^3$ and $h^*(\mathbf{u}) = \mathbf{E}_{\mathbf{x} \sim D_X} 3(\mathbf{x}_V \cdot \mathbf{u}_V)(\mathbf{x}_W \cdot \mathbf{u}_W)^2 + 3(\mathbf{x}_V \cdot \mathbf{u}_V)^2 (\mathbf{x}_W \cdot \mathbf{u}_W)$.

In particular, since $\|\mu\| \le (\gamma/d)^{2c_1}$ and D is factorizable, we know that h^* is a degree-3 polynomial that can be characterized with a tensor with Frobenius norm at most $o((\gamma/d)^{c_1})$.

Write

$$g^*(\mathbf{u}) = \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X}(\mathbf{x}_V \cdot \mathbf{u}_V)^3 + (\mathbf{x}_W \cdot \mathbf{u}_W)^3 = s^3 \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X}(\mathbf{x}_V \cdot \mathbf{u}_V^0)^3 + t^3 \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X}(\mathbf{x}_W \cdot \mathbf{u}_W^0)^3.$$

For simplicity, we define $a_V := \mathbf{E}_{\mathbf{x}\sim D_X}(\mathbf{x}_V \cdot \mathbf{u}_V^0)^3, a_W := \mathbf{E}_{\mathbf{x}\sim D_X}(\mathbf{x}_W \cdot \mathbf{u}_W^0)^3$. By the symmetry of g^*, without loss of generality, we assume $a_V \geq 0$.

We consider two cases. In the first case, we assume $a_W \leq 0$.

Since $f^*(\mathbf{u}) \geq \eta_0$ and $h^*(\mathbf{u}) \leq \eta$, we know that $a_V \geq \eta_0/2$ and $s^3 \geq \eta_0/2$. Consider point $\mathbf{z} := t\mathbf{u}_V^0 - s\mathbf{u}_W^0 \in S^{(d-1)} \cap \mathbf{u}^\perp$. Since $\nabla f^*(\mathbf{u}) = \nabla g^*(\mathbf{u}) + \nabla h^*(\mathbf{u})$ and

$$\nabla g^*(\mathbf{u}) = 3 \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X}\left((\mathbf{x}_V \cdot \mathbf{u}_V)^2 \mathbf{x}_V + (\mathbf{x}_W \cdot \mathbf{u}_W)^2 \mathbf{x}_W\right),$$

we have

$$\nabla f^*(\mathbf{u}) \cdot \mathbf{z} = \nabla g^*(\mathbf{u}) \cdot \mathbf{z} + \nabla h^*(\mathbf{u}) \cdot \mathbf{z} = 3(ts^2 a_V - st^2 a_W) + \nabla h^*(\mathbf{u}) \cdot \mathbf{z} \geq 3ts^2 a_V - o((\gamma/d)^{c_1}).$$

Since $\|\mathrm{proj}_{\mathbf{u}^\perp} \nabla f^*(\mathbf{u})\| \leq \eta_1$, we know that

$$3ts^2 a_V \leq \nabla f^*(\mathbf{u}) \cdot \mathbf{z} + o((\gamma/d)^{c_1}) \leq O(\eta_1).$$

Because $a_V \geq \gamma_0/2$ and $s^3 \geq \gamma_0/2$, this implies $s \geq 1 - \mathrm{poly}(\gamma/d)$. Thus, $\|u_W\| \leq \mathrm{poly}(\gamma/d)$.

In the second case, we have $a_W \geq 0$ and $a_V \geq 0$. By symmetry, without loss of generality, we assume $a_V \geq a_W \geq 0$. We will show that if $s \geq (\gamma/d)^{c_1}$ and $t \geq (\gamma/d)^{c_1}$, and $\|\mathrm{proj}_{\mathbf{u}^\perp} \nabla f^*(\mathbf{u})\| \leq \eta_1$, then $\max_{\mathbf{z}\in S^{(d-1)}\cap \mathbf{u}^\perp} \mathbf{z}^\intercal \nabla^2 f^*(\mathbf{u}) \mathbf{z}$ must be sufficiently large, which gives a contradiction.

Observe that

$$\nabla^2 g^*(\mathbf{u}) = 6 \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X}\left((\mathbf{x}_V \cdot \mathbf{u}_V)\mathbf{x}_V \mathbf{x}_V^\intercal + (\mathbf{x}_W \cdot \mathbf{u}_W)\mathbf{x}_W \mathbf{x}_W^\intercal\right).$$

Recall that $\mathbf{z} := t\mathbf{u}_V^0 - s\mathbf{u}_W^0 \in S^{(d-1)} \cap \mathbf{u}^\perp$. We have

$$\begin{aligned}
\mathbf{z}^\intercal \nabla^2 f^*(\mathbf{u}) \mathbf{z} &= \mathbf{z}^\intercal \nabla^2 g^*(\mathbf{u}) \mathbf{z} + \mathbf{z}^\intercal \nabla^2 h^*(\mathbf{u}) \mathbf{z} = 6st^2 a_V + 6s^2 t a_W + \mathbf{z}^\intercal \nabla^2 h^*(\mathbf{u}) \mathbf{z} \\
&= 6s^3(t^2/s^2) a_V + 6t^3(s^2/t^2) a_W + \mathbf{z}^\intercal \nabla^2 h^*(\mathbf{u}) \mathbf{z} \\
&\geq (\gamma/d)^{c_1} \eta_0/2 - o((\gamma/d)^{2c_1}),
\end{aligned}$$

which gives a contradiction to the fact that $\mathbf{u}$ is a $(\gamma/d)^{c_1}$-approximate solution. Thus, we have $t \leq \mathrm{poly}(\gamma/d)$ and $s \geq 1 - \mathrm{poly}(\gamma/d)$.

So far, we have shown that if any point $\mathbf{u} \in S^{(d-1)}$ that is a η-approximate solution to f^* must be $\mathrm{poly}(\gamma/d)$-close to either V or W. We next show that there must be a point that is close to V and is a η-approximate solution to f^*. Since D_X satisfies the statement of Theorem 6, we know that $\left\|\mathbf{E}_{\mathbf{x}\sim D_X^+} \mathbf{x}_V^{\otimes 3}\right\| \geq \Omega(\gamma^{15}), \left\|\mathbf{E}_{\mathbf{x}\sim D_X^-} \mathbf{x}_V^{\otimes 3}\right\| \geq \Omega(\gamma^{15})$. This implies that $\left\|\mathbf{E}_{\mathbf{x}\sim D_V} \mathbf{x}^{\otimes 3}\right\| \geq \Omega(\gamma^c)$, which implies that $\max_{\mathbf{u}\in S^{(d-1)}\cap V} f^*(\mathbf{u}) \geq \Omega(\gamma^c)$. Here c is a constant smaller than c_1.

Let $\mathbf{u}^* \in S^{(d-1)} \cap V$ such that $g^*(\mathbf{u}^*) = \max_{\mathbf{u}\in S^{(d-1)}\cap V} g^*(\mathbf{u})$. Since $\mathbf{u}^*$ is a maximal solution of $g^*(\mathbf{u})$ restricted at V, we know that $\left\|\mathrm{proj}_{(\mathbf{u}^*)^\perp} \nabla g^*(\mathbf{u}^*)\right\| = 0$ and $\max_{z\in S^{(d-1)}\cap (\mathbf{u}^*)^\perp} z^\intercal \nabla^2 g^*(\mathbf{u}^*) z = 0$. Since $\left\|\mathbf{E}_{\mathbf{x}\sim D_V} \mathbf{x}^{\otimes 3}\right\| \geq \Omega(\gamma^c)$, we know that $g^*(\mathbf{u}^*) \geq \Omega(\gamma^c)$. Since $f^* = g^* + h^*$ and h^* is a degree-3 polynomial with $o((\gamma/d)^{c_1})$-small magnitude, we know that $\mathbf{u}^*$ must be an η-approximate

solution to f^*. In particular, f^* can be completely described by its moment tensor. Thus, by estimating each entry of the moment tensor of D_X up to $\mathrm{poly}(\gamma/d)$ error, we obtain that $\hat{f}$ is a good approximation for f^*. Thus, any $(\gamma/d)^{c'}$-approximate solution for $\hat{f}$, with some c' slightly larger than c_1, must be a $(\gamma/d)^{c_1}$-approximate solution for f^*. So, any $(\gamma/d)^{c'}$-approximate solution for $\hat{f}$ must be also close to V or W.

Finally, we show that the output $\mathcal{O}$ has $\mathrm{poly}(d/\gamma)$ size and at least one of the solution in $\mathcal{O}$ is $\mathrm{poly}(\gamma/d)$ close to V. Here, we will make use of the following lemma.

Fact 7 (Theorem 9.3 in Alon (2003)) *Let $A \in \mathbb{R}^{n\times n}$ such that for all $i \in [n]$, $A_{ii} = 1$ and $|A_{ij}| \le \epsilon$ with $1/\sqrt{n} \le \epsilon < 1/2$ for all $i \neq j$. Then*

$$\textbf{\textit{rank}}(A) \ge \Omega(\frac{1}{\epsilon^2 \log(1/\epsilon)} \log(n)).$$

Denote by n the size of $\mathcal{O}$ and consider the matrix $A \in \mathbb{R}^{n\times n}$ where $A_{ij} = \mathbf{u}_i \cdot \mathbf{u}_j$. Since each $\mathbf{u}_t \in \mathbb{R}^d$ for $t \in [d]$, we know the rank of A is at most d. Furthermore, by the construction of $\mathcal{O}$, we know that for every $i, j \in [n]$, if $i = j$, then $A_{ij} = \mathbf{u}_i \cdot \mathbf{u}_j = 1$ and if $i \neq j$, then $|A_{ij}| = |\mathbf{u}_i \cdot \mathbf{u}_j| \le \epsilon := (\gamma/d)^{c_1}$. If $n \ge \epsilon^{-3} = (\gamma/d)^{-3c_1}$, then by Fact 7, we know that

$$\frac{1}{\epsilon^2 \log(1/\epsilon)} \log(n) = \Omega((d/\gamma)^{c_1}) > d,$$

which gives a contradiction. Thus, $n \le \mathrm{poly}(d/\gamma)$.

On the other hand, let $\mathbf{u}_i$ be a vector that is $\mathrm{poly}(\gamma/d)$-close to V and $\mathbf{u}_j$ be a vector that is $\mathrm{poly}(\gamma/d)$-close to W, then $|\mathbf{u}_i \cdot \mathbf{u}_j| \le \mathrm{poly}(\gamma/d)$. Thus, if $\mathcal{O}$ does not contain a point v that is $\mathrm{poly}(\gamma/d)$-close to V, there must be another point $\mathbf{u}$ that is nearly orthogonal to all points in $\mathcal{O}$ and will be added to $\mathcal{O}$ by Algorithm 2. This proves Theorem 12.

■

C.2. Proof of Theorem 13

In this section, we give the proof of Theorem 13. To begin with, we present the main algorithm and restate the Theorem 13 for convenience.

Theorem 33 (restatement of Theorem 13)

There is a learning algorithm $\mathcal{A}$ such that for every c, a suitably large constant and any instance of learning intersections of two halfspaces under factorizable distribution with γ-margin assumption if the input distribution D satisfies

1. $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}\|_F \le \gamma^c$.
2. $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}\mathbf{x}^\intercal\|_F \le \gamma^c$
3. $\left\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}^{\otimes 3}\right\|_F \le \gamma^c$

$\mathcal{A}$ runs in $\mathrm{poly}(d, 1/\gamma)$ time and outputs a list of d unit vectors $\mathcal{O}$ such that at least one direction $\mathbf{w} \in \mathcal{O}$ satisfies $\|\mathbf{w}_W\|_2 \le \mathrm{poly}(\gamma)$ with probability $\Omega(\gamma/d)$.

Algorithm 3 TENSORDIRECTIONFINDING (Efficient algorithm for finding relevant direction with matched moments)

1: **Input:** $\gamma \in (0,1)$ and i.i.d. sample access to a distribution D on $\mathbb{B}^d(1) \times \{\pm 1\}$ that is an instance of learning intersections of halfspaces under product distribution. Suppose that D satisfies the conditions in the statement of Theorem 13.
2: **Output:** $\mathcal{O}$, a list of $\mathrm{poly}(d)$, $w \in S^{d-1}$ such that at least one of $w \in \mathcal{O}$ satisfies $\|\mathrm{proj}_{V^\perp} w\| \leq \mathrm{poly}(\gamma)$.
3: Take S_1, a set of $m_1 = \mathrm{poly}(d/\gamma)$ i.i.d. samples from D_X to estimate $\mu := \mathbf{E}_{x \sim D_X} x$ with

$$\hat{\mu} := \frac{1}{m_1} \sum_{x \in S_1} x$$

up to $\mathrm{poly}(\gamma)$ error
4: Take S, a set of $N = \mathrm{poly}(d/\gamma)$ i.i.d. samples from D_X and estimate

$$\hat{T} = \frac{1}{N} \sum_{x \in S} (x - \hat{\mu})^{\otimes 3}$$

5: Let $v \sim N(0, \frac{1}{d} I)$ be a random Gaussian vector in $\mathbb{R}^d$
6: Define $\hat{M} := \hat{T} \cdot v$.
7: Compute $\mathcal{O}$, the set of d eigenvectors of $\hat{M}$ via eigen-decomposition algorithm.
8: **return** $\mathcal{O}$

The algorithm we will analyze is Algorithm 3.

Proof [Proof of Theorem 33] We consider the central third moment tensor $T^* := \mathbf{E}_{\mathbf{x} \sim D_X} (\mathbf{x} - \mu)^{\otimes 3}$. Since $\|\mu\|_2 \leq 1$, we can without loss of generality assume $\mu = 0$, because shifting D_X to μ and rescaling the distribution will only decrease the margin assumption γ by a factor of 2. Under this assumption, we have

$$T^* = \mathop{\mathbf{E}}_{\mathbf{x} \sim D_X} \mathbf{x}^{\otimes 3} = \mathop{\mathbf{E}}_{\mathbf{x} \sim D_X} (\mathbf{x}_V + \mathbf{x}_W)^{\otimes 3} = \mathop{\mathbf{E}}_{\mathbf{x} \sim D_V} \mathbf{x}_V^{\otimes 3} + \mathop{\mathbf{E}}_{\mathbf{x} \sim D_W} \mathbf{x}_W^{\otimes 3}.$$

Here the second equation follows by the fact that D_V, D_W are independent and have zero-mean. To simplify the notation, we denote by $T_V = \mathbf{E}_{\mathbf{x} \sim D_V} \mathbf{x}_V^{\otimes 3}$ and $T_W = \mathbf{E}_{\mathbf{x} \sim D_W} \mathbf{x}_W^{\otimes 3}$.

Let $\mathbf{v} \sim N(0, \frac{1}{d} I)$ be a Gaussian vector. Define random matrix $M \in \mathbb{R}^{d \times d}$ as

$$\begin{aligned} M := T^* \cdot \mathbf{v} &= \mathop{\mathbf{E}}_{\mathbf{x} \sim D_V} \mathbf{x}_V \mathbf{x}_V^\intercal (\mathbf{x}_V \cdot \mathbf{v}) + \mathop{\mathbf{E}}_{\mathbf{x} \sim D_W} \mathbf{x}_W \mathbf{x}_W^\intercal (\mathbf{x}_W \cdot \mathbf{v}) \\ &= \mathop{\mathbf{E}}_{\mathbf{x} \sim D_V} \mathbf{x}_V \mathbf{x}_V^\intercal (\mathbf{x}_V \cdot \mathbf{v}_V) + \mathop{\mathbf{E}}_{\mathbf{x} \sim D_W} \mathbf{x}_W \mathbf{x}_W^\intercal (\mathbf{x}_W \cdot \mathbf{v}_W). \end{aligned}$$

To simplify the notation, we denote by $M_V = \mathbf{E}_{\mathbf{x} \sim D_V} \mathbf{x}_V \mathbf{x}_V^\intercal (\mathbf{x}_V \cdot \mathbf{v}_V)$ and $M_W = \mathbf{E}_{\mathbf{x} \sim D_W} \mathbf{x}_W \mathbf{x}_W^\intercal (\mathbf{x}_W \cdot \mathbf{v}_W)$. We write the random vector $\mathbf{v}_V = \alpha \mathbf{v}_V^0$, where the random variable $\alpha = \|\mathbf{v}_V\|$ and the random vector $\mathbf{v}_V^0 = \mathbf{v}_V / \|\mathbf{v}_V\|$ is drawn uniformly from $S^{(d-1)} \cap V$. Denote by σ_1, σ_2 be two eigenvalues of the random matrix M_V/α such that $|\sigma_1| \geq |\sigma_2|$ and denote by $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}$ the corresponding eigenvectors to σ_1, σ_2. Notice that for $i \in [2]$,

$$M \mathbf{u}^{(i)} = M_V \mathbf{u}^{(i)} + M_W \mathbf{u}^{(i)} = M_V \mathbf{u}^{(i)} = \alpha \sigma_i \mathbf{u}^{(i)}.$$

This implies $\mathbf{u}^{(i)}$ is an eigenvector of M with eigenvalue $\alpha\sigma^{(i)}$. Furthermore, if $\alpha\sigma_i$ for some $i \in [2]$ is not close to any eigenvalue of M_W, then one of the eigenvectors of M must be in V and we can find it via PCA. Next, we show this holds for our case. By Theorem 6, we know that $\|T_V\|_F \geq \gamma^c$, for some constant $c > 0$, which implies there is some vector $\mathbf{u}^* \in S^{(d-1)} \cap V$ such that

$$T_V \cdot (\mathbf{u}^*)^{\otimes 3} = \mathop{\mathbf{E}}_{\mathbf{x}\sim D_V} (\mathbf{x}_V \cdot \mathbf{u}^*)^3 \geq \Omega(\gamma^c).$$

Since $\mathbf{v}_V^0$ is drawn uniformly from $S^{d-1} \cap V$, we know that

$$\mathbf{Pr}\left(\left\|\mathbf{u}^* - \mathbf{v}_V^0\right\| \leq \gamma^c/10\right) \geq \mathbf{Pr}\left(\sin\theta(\mathbf{u}^*, \mathbf{v}_V^0) \leq \gamma^c/10\right) \geq \Omega(\gamma^c).$$

Given this happens, we show that $\sigma_1 \geq \gamma^c$ by showing that $(\mathbf{v}_V^0)^\intercal (M_V/\alpha)\mathbf{v}_V^0 \geq \Omega(\gamma^c)$. We have

$$\begin{aligned}(\mathbf{v}_V^0)^\intercal M_V \mathbf{v}_V^0 &= T^* \cdot (\mathbf{v}_V^0)^{\otimes 3} = T^* \cdot (\mathbf{u}^*)^{\otimes 3} - T^* \cdot \left((\mathbf{v}_V^0)^{\otimes 3} - (\mathbf{u}^*)^{\otimes 3}\right) \\ &= T^* \cdot (\mathbf{u}^*)^{\otimes 3} - \mathop{\mathbf{E}}_{\mathbf{x}\sim D_V} \left((\mathbf{x}\cdot\mathbf{v}_V^0)^3 - (\mathbf{x}\cdot\mathbf{u}^*)^3\right) \\ &\geq \Omega(\gamma^c),\end{aligned}$$

where the inequality follows the fact that the polynomial function $\mathbf{E}_{\mathbf{x}\sim D_V}(\mathbf{x}\cdot\mathbf{u})^3$ is $O(1)$-Lipschitz with respect to $\mathbf{u}$. Thus, with a probability at least $\Omega(\gamma^c)$, $\sigma_1 \geq \gamma^c/2$. However, we have no structural result that can guarantee σ_2 is also large. Thus, in the rest of the proof, we consider two cases and argue that in each case, the eigenvalues of M_V are far from the eigenvalues of M_W.

In the first case, we assume that $\sigma_1 - \sigma_2 \geq \gamma^c/4$. Since M_V and M_W are independent, we consider the $d-2$ eigenvalues of M_W, $\sigma_3, \ldots, \sigma_d$. Recall that the two eigenvalues of M_V are $\alpha\sigma_1$ and $\alpha\sigma_2$. For each $i \in \{3, \ldots, d\}$, we associate each σ_i and interval $I_i = [\sigma_i - \xi, \sigma_i + \xi]$, where $\xi > 0$ will be determined later. Notice that $\alpha\sigma_1 \in I_i$ if and only if $\sqrt{d}\alpha \in [\sqrt{d}\sigma_i/\sigma_1 - \sqrt{d}\xi/\sigma_1, \sqrt{d}\sigma_i/\sigma_1 + \sqrt{d}\xi/\sigma_1]$. Since $\mathbf{v} \sim N(0, \frac{1}{d}I)$, and $\mathbf{v}_V$ is the 2-dimensional projection onto the subspace V, we know that $d\alpha^2 \sim \chi(2)$ is a χ-distribution with degree of freedom 2. Recall that $\chi(2)$ is a distribution supported over $\mathbb{R}^+$ with density function $p(x) = x\exp(-x^2/2) \leq 1, \forall x \geq 0$. This implies that

$$\mathbf{Pr}(\alpha\sigma_1 \in I_i) \leq 2\sqrt{d}\xi/\sigma_1, \forall i \in \{3, \ldots, d\} \tag{16}$$

On the other hand,

$$\mathbf{Pr}(|\alpha\sigma_1 - \alpha\sigma_2| \leq \xi) = \mathbf{Pr}(\alpha \leq \xi/|\sigma_1 - \sigma_2|) \leq O(\xi/|\sigma_1 - \sigma_2|).$$

By choosing $\xi = O(\gamma^{2c}/d)$, we know that with probability at least $1 - O(\gamma^{2c})$, $\alpha\sigma_1$, the largest eigenvalue of M_V is ξ-far from any other eigenvalues of M.

In the second case, we assume $\sigma_1 - \sigma_2 \leq \gamma^c/4$, which implies that $\sigma_2 \geq \gamma^c/4$. Recall that the two eigenvalues of M_V are $\alpha\sigma_1$ and $\alpha\sigma_2$. With the same proof as (16), we know that

$$\mathbf{Pr}(\alpha\sigma_2 \in I_i) \leq 2\sqrt{d}\xi/\sigma_2, \forall i \in \{3, \ldots, d\}. \tag{17}$$

By choosing $\xi = O(\gamma^{2c}/d)$, we know that with probability at least $1 - O(\gamma^{2c})$, $\alpha\sigma_1, \alpha\sigma_2$, the two eigenvalues of M_V are ξ-far from any other eigenvalues of M_W.

To finish the proof of Theorem 13, we make use of the well-known Davis-Kahan $\sin\theta$ theorem.

Theorem 34 (Davis-Kahan $\sin\theta$ Theorem) *Let $A = E_0 A_0 E_0^\intercal + E_1 A_1 E_1^\intercal$, $A + H = F_0 B_0 F_0^\intercal + F_1 B_1 F_1^\intercal \in \mathbb{R}^{d\times d}$ be symmetric matrices, where $(E_0, E_1), (F_0, F_1)$ are orthogonal matrices and A_0, A_1, B_0, B_1 are diagonal matrices. If every eigenvalues of A_0 are δ-far from the eigenvalues of A_1, then*

$$\|F_0^\intercal E_1\|_2 \le \frac{\|H\|_2}{\delta}.$$

Since $\mathbf{v} \sim N(0, \frac{1}{d}I)$, by Vershynin (2018), we know that with probability at least $1 - \gamma^d$, $\|\mathbf{v}\|_2 \le 1 + \log(1/\gamma)$. Since by taking $\mathrm{poly}(d/\gamma)$ samples from D_X, we are able to estimate each entry of T^* with $\hat{T}$ up to error $\mathrm{poly}(\gamma/d)$. Thus, given $\|\mathbf{v}\|_2 \le 1 + \log(1/\gamma)$, $\left\|M - \hat{M}\right\|_2 \le (\gamma/d)^{3c}$.

Recall that we have discussed two cases for the behavior of the eigenvalues of M_V. In the first case, the largest eigenvalue $\alpha\sigma_1$ is $\gamma^2 c/d$ far from any other eigenvalues of M. Using Theorem 34 by taking $A = M, A + H = \hat{M}, E_0 = \mathbf{u}^{(1)}$, we know that there is an eigenvector of $\hat{M}$, $\hat{\mathbf{u}^{(1)}}$ such that $\left\|\hat{\mathbf{u}^{(1)}}_W\right\| \le \gamma^c$. In the second case the two eigenvalues $\alpha\sigma_1, \alpha\sigma_2$ are $\gamma^2 c/d$ far from any other eigenvalues of M_W, using Theorem 34 by taking $A = M, A + H = \hat{M}, E_0 = [\mathbf{u}^{(1)}, \mathbf{u}^{(2)}]$, we know that there is an eigenvector of $\hat{M}$, $\hat{\mathbf{u}^{(1)}}$ such that $\left\|\hat{\mathbf{u}^{(1)}}_W\right\| \le \gamma^c$. Thus, with probability at least $\Omega(\gamma^c)$, Algorithm 3 outputs a list of d unit vectors such that at least one of them is γ^c close to V. ■

Remark In general, given the estimated moment tensor $\hat{M}$, there is no polynomial time algorithm that can *exactly* compute all eigenvectors of $\hat{M}$ due to the roundoff error. However, as long as the roundoff error is at most $\mathrm{poly}(\gamma/d)$, one can compute in polynomial time an eigen-decomposition that is $\mathrm{poly}(\gamma/d)$ close to $\hat{M}$. We refer the readers to Dhillon et al. (2006) for detailed guarantee of efficient algorithms for eigen-decomposition that takes the roundoff error into consideration.

Appendix D. Omitted Proofs from Section 3.2

This section is dedicated to proving Theorem 14. The main algorithm we will analyze is Algorithm 4.

D.1. Approximate Local Optimum Implies Relevant Direction

The following definition plays a core role of the proof.

Definition 35 ((α,η)-approximate solution) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a real-valued function and $\alpha, \eta > 0$. We say that $\mathbf{w} \in S^{d-1}$ is an (α,η)-approximate solution to the problem $\max_{\mathbf{w}\in S^{d-1}} f(\mathbf{w})$ if satisfies the following conditions: (1) $\|\mathrm{proj}_{\mathbf{w}^\perp} \nabla f(\mathbf{w})\|_2 \le \eta$, and (2) $|f(\mathbf{w})| \ge \alpha$.*

Suppose we have access to the exact function $f(\mathbf{u}) = \mathbf{u}^{\otimes m} \cdot T$ where $T = \mathbf{E}_{\mathbf{x}\sim D^+}[\mathbf{x}_V^{\otimes m}] - \mathbf{E}_{\mathbf{x}\sim D^-}[\mathbf{x}_V^{\otimes m}]$. If the first $m-1$ moments are highly matched, then finding an approximate local maximum $\mathbf{u}^*$ of f implies that $\mathbf{u}^*$ is close to V.

Lemma 36 *Let $\alpha, \eta, \xi > 0$, $m \in \mathbb{Z}_+$ and D be a joint distribution of $(\mathbf{x}, y)$ on $\mathbb{B}^d(1) \times \{\pm 1\}$ that is an instance of learning intersections of halfspaces under factorizable distribution with γ-margins. Suppose D does not satisfy (ξ, m)-moment matching condition. Let $f(\mathbf{u}) = \mathbf{u}^{\otimes m} \cdot T$ where $T = \mathbf{E}_{\mathbf{x}\sim D^+}[\mathbf{x}_V^{\otimes m}] - \mathbf{E}_{\mathbf{x}\sim D^-}[\mathbf{x}_V^{\otimes m}]$. Then any $\mathbf{u}^*$ that is an (α,η)-approximate solution to $\max_{\mathbf{u}\in S^{d-1}} f(\mathbf{u})$ must satisfy $\|\mathbf{u}_W^*\|_2 \le \eta/(m\alpha)$*

Algorithm 4 DIRECTIONFINDING (Relevant direction extraction with Mismatched Moments)

1: **Input:** $\alpha \in (0,1), m \in \{1,2,3\}$ and i.i.d. sample access to a distribution D on $\mathbb{B}^d(1) \times \{\pm 1\}$ that is an instance of learning intersections of halfspaces under product distribution. Suppose D does not satisfy the (α, m)-moment matching condition and D satisfies the $(\alpha^2/d^c, t)$-moment matching condition for any $t \leq m-1$ and a sufficiently large universal constant c.

2: **Output:** With $2/3$ probability, the algorithm outputs a unit vector $\mathbf{u} \in S^{d-1}$ such that $\|\mathbf{u}_W\| = O(\alpha)$.

3: Define $T = \mathbf{E}_{\mathbf{x}\sim D^+} \mathbf{x}_V^{\otimes m} - \mathbf{E}_{\mathbf{x}_V \sim D^-} \mathbf{x}_V^{\otimes m}$. We will use $\hat{T}$ as an empirical estimation of T. Set $\zeta = \alpha/d^{c-1}$. Draw $N = \mathrm{poly}(d, 1/\alpha)$ i.i.d. samples S^+ and S^- from D^+ and D^- respectively. Then estimate

$$\hat{T} = \frac{1}{N} \sum_{\mathbf{x} \in S^+} \mathbf{x}^{\otimes m} - \frac{1}{N} \sum_{\mathbf{x} \in S^-} \mathbf{x}^{\otimes m} .$$

4: Define $f : S^{d-1} \to \mathbb{R}$ as $f(\mathbf{u}) := \hat{T} \cdot \mathbf{u}^{\otimes m}$ and apply the following standard gradient descent steps. Let $T = (d/\alpha)^{c'}$ for a sufficiently large constant c' depending on c.

1. Initialize a random $\mathbf{u}_0 \sim_u S^{d-1}$. If $f(\mathbf{u}_0) < \zeta$, then reinitialize. If reinitialized T times, then output failure.
2. Repeat the following for at most T many iterations. For the t-th iteration, calculate the gradient $\mathbf{g} = \mathrm{proj}_{\mathbf{u}_t^\perp} \nabla f(\mathbf{u}_t)$, and update $\mathbf{u}_{t+1} = \frac{\mathbf{u}_t + \lambda \mathbf{g}}{\|\mathbf{u}_t + \lambda \mathbf{g}\|_2}$ where the stepsize $\lambda = c'' \min(1, 1/\|\mathrm{proj}_{\mathbf{u}} \nabla f(\mathbf{u}_t)\|_2)$ and c'' is a sufficiently small universal constant. Repeat this step until getting a unit vector $\mathbf{u}_t$ that is a (α, η)-approximate solution such that $\frac{\eta + \alpha^2/d^c}{\alpha' - \alpha^2/d^c} \geq c''\alpha$, then output such $\mathbf{u}_t$. If the condition is not satisfied in T many iterations, then output failure.

Proof [Proof of Lemma 36] Let $\mathbf{u} \in S^{d-1}$ and $W = V^{\perp}$. To simplify the notation, we write $\bar{\mathbf{u}}_V = \mathbf{u}_V / \|\mathbf{u}_V\|_2$, $\bar{\mathbf{u}}_W = \mathbf{u}_W / \|\mathbf{u}_W\|_2$ and $\mathbf{u} = s\bar{\mathbf{u}}_V + t\bar{\mathbf{u}}_W$, where $s, t \geq 0$ and $s^2 + t^2 = 1$. The gradient of f at $\mathbf{u}$ is

$$\begin{aligned}\nabla f(\mathbf{u}) &= m \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}[(\mathbf{u}_V \cdot \mathbf{x}_V)^{m-1}\mathbf{x}_V] - \mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}[(\mathbf{u}_V \cdot \mathbf{x}_V)^{m-1}\mathbf{x}_V] \right) \\ &= ms^{m-1} \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}[(\bar{\mathbf{u}}_V \cdot x_V)^{m-1}x_V] - \mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}[(\bar{\mathbf{u}}_V \cdot x_V)^{m-1}x_V] \right).\end{aligned}$$

Consider $\tilde{\mathbf{u}} = t\bar{\mathbf{u}}_V - s\bar{\mathbf{u}}_W \in S^{d-1} \cap \mathbf{u}^{\perp}$, we have

$$\begin{aligned}\nabla f(\mathbf{u}) \cdot \tilde{\mathbf{u}} &= mts^{m-1} \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}[(\bar{\mathbf{u}}_V \cdot \mathbf{x}_V)^{m-1}\mathbf{x}_V] - \mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}[(\bar{\mathbf{u}}_V \cdot \mathbf{x}_V)^{m-1}\mathbf{x}_V] \right) \cdot \bar{\mathbf{u}}_V \\ &= mts^{m-1} \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}[(\bar{\mathbf{u}}_V \cdot \mathbf{x}_V)^{m}] - \mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}[(\bar{\mathbf{u}}_V \cdot \mathbf{x}_V)^{m}] \right) \\ &= \frac{tm}{s} s^m \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}(\bar{\mathbf{u}}_V \cdot \mathbf{x}_V)^{m} - \mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}(\bar{\mathbf{u}}_V \cdot \mathbf{x}_V)^{m} \right) \\ &= \frac{tm}{s} \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}[(\mathbf{u}_V \cdot \mathbf{x}_V)^{m}] - \mathop{\mathbf{E}}_{\mathbf{x}\sim D^-}[(\mathbf{u}_V \cdot \mathbf{x}_V)^{m}] \right) = \frac{tm}{s} f(\mathbf{u})\end{aligned}$$

If $\mathbf{u}^*$ is an (α, η)-approximate solution, then

$$t \leq \left| \frac{\eta s}{m f(u)} \right| \leq \frac{\eta}{m\alpha}.$$

Thus, we have $\|\mathbf{u}^*_W\|_2 \leq \eta/(m\alpha)$. ■

However, since V is unknown to us and we only have sample access to the distribution D, we cannot hope to know the exact function f, which requires exact moment information of the distribution over V. To overcome these problems, we first show that if we replace T with its empirical estimation of the moments, the statement of Lemma 36 still holds.

Lemma 37 *Let $\alpha, \eta, \xi > 0$, $m \in \mathbb{Z}_+$ and D be a joint distribution of $(\mathbf{x}, y)$ on $\mathbb{B}^d(1) \times \{\pm 1\}$ that is consistent with an instance of learning intersections of two halfspaces under factorizable distribution with γ-margin assumption. Suppose D does not satisfy (ξ, m)-moment matching condition. Let $\hat{f}(\mathbf{u}) = \mathbf{u}^{\otimes m} \cdot \hat{T}$ where $\|\hat{T} - T\|_F \leq \epsilon$ and $T = \mathbf{E}_{\mathbf{x}\sim D^+}[\mathbf{x}_V^{\otimes m}] - \mathbf{E}_{\mathbf{x}\sim D^-}[\mathbf{x}_V^{\otimes m}]$. Then any $\mathbf{u}^*$ that is an (α, η)-approximate solution to $\max_{\mathbf{u}\in S^{d-1}} \hat{f}(\mathbf{u})$ must satisfy $\|\mathbf{u}^*_W\| \leq \frac{(\eta+\epsilon)}{m(\alpha-\epsilon)}$.*

Proof [Proof of Lemma 37] Let $\mathbf{u} \in S^{d-1}$ and $W = V^{\perp}$. To simplify the notation, we write $\bar{\mathbf{u}}_V = \mathbf{u}_V / \|\mathbf{u}_V\|$, $\bar{\mathbf{u}}_W = \mathbf{u}_W / \|\mathbf{u}_W\|$ and $\mathbf{u} = s\bar{\mathbf{u}}_V + t\bar{\mathbf{u}}_W$, where $s, t \geq 0$ and $s^2 + t^2 = 1$. We compute the gradient of $\hat{f}$ at $\mathbf{u}$. Notice that $\hat{f}(\mathbf{u}) = \hat{T} \cdot \mathbf{u}^{\otimes m}$ and

$$\nabla \hat{f}(\mathbf{u}) = \nabla f(\mathbf{x}) + \nabla(\hat{T} - T^* \cdot \mathbf{u}^{\otimes m}).$$

Therefore,

$$\begin{aligned}\nabla \hat{f}(\mathbf{u}) \cdot \tilde{\mathbf{u}} =& \nabla f(\mathbf{u}) \cdot \tilde{\mathbf{u}} + \nabla(\hat{T} - T^* \cdot \mathbf{u}^{\otimes m}) \cdot \tilde{\mathbf{u}} \\ =& \frac{tm}{s} f(\mathbf{u}) + \nabla(\hat{T} - T^* \cdot \mathbf{u}^{\otimes m}) \cdot \tilde{\mathbf{u}} \\ =& \frac{tm}{s} \hat{f}(\mathbf{u}) + \left(\frac{tm}{s} f(\mathbf{u}) - \frac{tm}{s} \hat{f}(\mathbf{u}) \right) + \nabla(\hat{T} - T^* \cdot \mathbf{u}^{\otimes m}) \cdot \tilde{\mathbf{u}} \\ =& \frac{tm}{s} \hat{f}(\mathbf{u}) + \frac{tm}{s} (\hat{T} - T) \cdot \mathbf{u}^{\otimes m} + \nabla(\hat{T} - T^* \cdot \mathbf{u}^{\otimes m}) \cdot \tilde{\mathbf{u}} \\ =& \frac{tm}{s} \hat{f}(\mathbf{u}) + \frac{tm}{s} (\hat{T} - T) \cdot \mathbf{u}^{\otimes m} + (\hat{T} - T^*) \cdot \mathbf{u}^{\otimes m-1} \otimes \tilde{\mathbf{u}} \, .\end{aligned}$$

Since $\left\| \hat{T} - T^* \right\|_F \leq \epsilon$, if $\mathbf{u}$ is an (α, η)-approximate solution, we have $\eta \geq \left| \nabla \hat{f}(\mathbf{u}) \cdot \tilde{\mathbf{u}} \right| \geq \frac{tm}{s}(\alpha - \epsilon) - \epsilon$, which implies $t \leq \left| \frac{(\eta+\epsilon)s}{m(\alpha-\epsilon)} \right| \leq \frac{(\eta+\epsilon)}{m(\alpha-\epsilon)}$. Thus, we have $\|\mathbf{u}_W^*\|_2 \leq \frac{(\eta+\epsilon)}{m(\alpha-\epsilon)}$. ■

However, since V is unknown to us, even if we have i.i.d. sample access to D, it is still unclear how to estimate T with samples. To overcome this difficulty, we show in the next lemma that if D satisfies (ϵ, t)-moment matching condition for $t \leq m - 1$, then we can efficiently estimate T up to a desired accuracy.

Lemma 38 *Let D over $\mathbb{B}^d(1) \times \{\pm 1\}$ be a distribution that is consistent with an instance of learning intersections of two halfspaces under product distribution. Consider degree m-moment tensors over $\mathbb{R}^d$, $T^* = \mathbf{E}_{\mathbf{x} \sim D_X^+} \mathbf{x}_V^{\otimes m} - \mathbf{E}_{\mathbf{x} \sim D_X^-} \mathbf{x}_V^{\otimes m}$ and $\hat{T} = \frac{1}{n} \sum_{\mathbf{x} \in S_+} \mathbf{x}^{\otimes m} - \frac{1}{n} \sum_{\mathbf{x} \in S_-} \mathbf{x}^{\otimes m}$, where $n = \mathrm{poly}(d^m, 1/\epsilon, \log(1/\delta))$. If for $i < m$, $\mathcal{D}$ satisfies the $(\epsilon/2^m, i)-$moment matching condition, then with probability at least $1 - \delta$, $\left\| \hat{T} - T^* \right\| \leq 2\epsilon$.*

Proof [Proof of Lemma 38] Denote by $T := \mathbf{E}_{\mathbf{x}\sim D_X^+} \mathbf{x}^{\otimes m} - \mathbf{E}_{\mathbf{x}\sim D_X^-} \mathbf{x}^{\otimes m} = \left(\mathbf{E}_{\mathbf{x}\sim D_X^+} - \mathbf{E}_{\mathbf{x}\sim D_X^-}\right) \mathbf{x}^{\otimes m}$.

$$
\begin{aligned}
\|T - T^*\|_F &= \left\| \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^+} - \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^-} \right) \left((\mathbf{x}_V + \mathbf{x}_W)^{\otimes m} - \mathbf{x}_V^{\otimes m} \right) \right\|_F \\
&= \left\| \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^+} - \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^-} \right) \left((\mathbf{x}_V + \mathbf{x}_W)^{\otimes m} - \mathbf{x}_V^{\otimes m} - \mathbf{x}_W^{\otimes m} \right) \right\|_F \\
&= \left\| \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^+} - \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^-} \right) \sum_{i=0}^{m} \sum_{\sigma_i} \left(\bigotimes_{j=1}^{m} \mathbf{x}_{\sigma_i(j)} \right) - \mathbf{x}_V^{\otimes m} - \mathbf{x}_W^{\otimes m} \right\|_F \\
&= \left\| \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^+} - \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^-} \right) \sum_{i=1}^{m-1} \sum_{\sigma_i} \left(\bigotimes_{j=1}^{m} \mathbf{x}_{\sigma_i(j)} \right) \right\|_F \\
&\leq \sum_{i=1}^{m-1} \sum_{\sigma_i} \left\| \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^+} - \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^-} \right) \left(\bigotimes_{j=1}^{m} \mathbf{x}_{\sigma_i(j)} \right) \right\|_F \\
&= \sum_{i=1}^{m-1} \binom{m}{i} \left\| \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^+} - \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^-} \right) \left(\mathbf{x}_V^{\otimes i} \otimes \mathbf{x}_W^{\otimes m-i} \right) \right\|_F \\
&= \sum_{i=1}^{m-1} \binom{m}{i} \left\| \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^+} \left(\mathbf{x}_V^{\otimes i} \otimes \mathbf{x}_W^{\otimes m-i} \right) - \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^-} \left(\mathbf{x}_V^{\otimes i} \otimes \mathbf{x}_W^{\otimes m-i} \right) \right) \right\|_F \\
&= \sum_{i=1}^{m-1} \binom{m}{i} \left\| \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^+} - \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^-} \right) \left(\mathbf{x}_V^{\otimes i} \right) \otimes E_{\mathbf{x}\sim D_X} \mathbf{x}_W^{\otimes m-i} \right\|_F \\
&= \sum_{i=1}^{m-1} \binom{m}{i} \left\| \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^+} - \mathop{\mathbf{E}}_{\mathbf{x}\sim D_X^-} \right) \left(\mathbf{x}_V^{\otimes i} \right) \right\|_F \left\| E_{\mathbf{x}\sim D_X} \mathbf{x}_W^{\otimes m-i} \right\|_F \leq (\epsilon/2^m) \sum_{i=1}^{m-1} \binom{m}{i} \leq \epsilon.
\end{aligned}
$$

The second equation holds by $E_{\mathbf{x}\sim D_X^+} \mathbf{x}_W^{\otimes m} = E_{\mathbf{x}\sim D_X^-} \mathbf{x}_W^{\otimes m} = E_{\mathbf{x}\sim D_X} \mathbf{x}_W^{\otimes m}$. In the third equation, we denote by $\sigma_i : [m] \to \{V, W\}$ a map of combination such that $|\{j \in [m] : \sigma_i(j) = V\}| = i$. The fifth equation holds because the tensor $\left(\mathbf{E}_{\mathbf{x}\sim D_X^+} - \mathbf{E}_{\mathbf{x}\sim D_X^-}\right) \left(\bigotimes_{j=1}^{m} \mathbf{x}_{\sigma_i(j)}\right)$ is symmetric according to the map of combination σ_i. In the seventh equation, we use the fact that $\mathbf{x}_V$ and $\mathbf{x}_W$ are independent. In the second last inequation, we use the fact that D satisfies the $(\epsilon, i)-$moment matching condition and $\|\mathbf{x}\| \leq 1$ for sure.

By Hoeffding's inequality, we know that

$$
\begin{aligned}
\mathbf{Pr}\left(\left\| \hat{T} - T \right\|_F \geq \epsilon \right) &\leq \sum_{i=1}^{d^m} \mathbf{Pr}\left(\left| T_i - \hat{T}_i \right| \geq d^{-m/2} \epsilon \right) \\
&\leq 2d^m \exp\left(-nd^{-m} \epsilon^2 \right) \leq \delta,
\end{aligned}
$$

when $n = \mathrm{poly}(d^m, 1/\epsilon, \log(1/\delta))$, where T_i is the $i-$th entry of the vectorized T. Thus, we have

$$
\left\| \hat{T} - T^* \right\|_F \leq \left\| \hat{T} - T \right\|_F + \|T - T^*\|_F \leq 2\epsilon.
$$

■

D.2. Proof of Theorem 14

Given Lemma 37, we are now ready to analyze Algorithm 4 and prove Theorem 14. For convenience, we restate Theorem 14 statement as Theorem 39.

Theorem 39 (restatement of Theorem 14) *Let $m \leq 3$, there is an algorithm $\mathcal{A}$ (Algorithm 4) such that for any instance of learning intersections of two halfspaces under factorizable distributions, if the distribution D does not satisfy the (α, m)-moment matching condition and D satisfies the $(\alpha^2 d^{-c}/2^m, t)$-moment matching condition for any $t \leq m-1$ and a sufficiently large universal constant c, then $\mathcal{A}$ draws $\mathrm{poly}(d, 1/\alpha)$ i.i.d. samples from D, runs in time $\mathrm{poly}(d, 1/\alpha)$, and outputs a unit vector $\mathbf{u} \in S^{d-1}$ such that $\|\mathbf{u}_W\| = O(\alpha)$ with probability $2/3$.*

Proof [Proof of Theorem 14] By Lemma 38, the empirical estimation $\hat{T}$ satisfies $\|\hat{T} - T\|_F \leq \alpha^2/d^c$ with high probability for a sufficiently large constant c. From Lemma 37, we have that for any (α', η)-approximate solution, $\|\mathbf{u}_W\| \leq \frac{\eta + \alpha^2/d^c}{m(\alpha' - \alpha^2/d^c)}$. Therefore, it suffices for us to show that after at most T steps of the gradient descent, we will with high probability find a (α', η)-approximate solution such that $\frac{\eta + \alpha^2/d^c}{\alpha' - \alpha^2/d^c} = O(\alpha)$.

We start by showing that the initialization will with high probability give us a $\mathbf{u}_0$ such that $f(\mathbf{u}_0) = \Omega(\alpha/\mathrm{poly}(d))$. Since the empirical estimation $\hat{T}$ satisfies $\|\hat{T} - T\|_F \leq \alpha^2/d^c$ with high probability, we must have $\|\mathrm{proj}_{V^{\otimes 3}}(\hat{T})\|_F \geq \|\mathrm{proj}_{V^{\otimes 3}}(T)\|_F - \|\mathrm{proj}_{V^{\otimes 3}}(\hat{T} - T)\|_F \geq \|T\|_F - \|\hat{T} - T\|_F = \Omega(\alpha)$. Given Fact 3 and $\hat{T}$ is inside the subspace of $V^{\otimes 3}$ (isomorphic to $(\mathbb{R}^2)^{\otimes 3}$), we have that $\max_{\mathbf{u} \in S^{d-1}} \mathbf{u}^{\otimes m} \cdot \hat{T} = \max_{\mathbf{u} \in S^{d-1} \wedge \mathbf{u} \in V} \mathbf{u}^{\otimes m} \cdot \hat{T} = \Omega(\alpha)$. We then use the following fact, which is Lemma 12 from Vempala and Xiao (2011) to show that with high probability, we will get an initialization with large $f(\mathbf{u}_0)$.

Fact 8 *Let p be a degree-m polynomial over d variables and K a convex body in $\mathbb{R}^d$. If there exists an $\mathbf{x} \in K$ such that $|p(\mathbf{x})| > \epsilon(c'd)^m$, for some suitable constant $c' > 0$, then for l random points $\mathbf{s}_i$, $\mathbf{Pr}(\forall \mathbf{s}_i : |p(\mathbf{s}_i)| \leq \epsilon) \leq 2^{-l}$.*

We apply Fact 8 with $K = \{\mathbf{u} \in \mathbb{R}^d \mid \|\mathbf{u}\|_2 \leq 1\}$. It is easy to see that if we sample the initialization $\mathbf{u}_0$ uniformly from S^{d-1} instead of K, $f(\mathbf{u}_0)$ will be larger. Therefore, we have that with probability at least $\Omega(1)$, $\mathbf{Pr}_{\mathbf{u}_0 \sim S^{d-1}} \mathbf{u}_0^{\otimes m} \cdot \hat{T} \geq \alpha/\mathrm{poly}(d)$. Therefore in Step 1, with probability $\Omega(1)$, we will have $f(\mathbf{u}_0) \geq \alpha/d^{c'}$, where c' is a universal constant.

Suppose we have such a good initialization, we start analyzing the gradient descent step in Algorithm 4. To do this, we prove the following two facts. The first fact is about the smoothness of the function f we are optimizing.

Fact 9 *The function f in Algorithm 4 satisfies $\|\nabla f(\mathbf{u}) - \nabla f(\mathbf{u}')\| \leq O(\|\mathbf{u} - \mathbf{u}'\|)$. i.e. f is a $\Omega(1)$-smooth function.*

Proof Consider $g : (\mathbb{R}^d)^m \to \mathbb{R}$ and $h : \mathbb{R}^d \to \mathbb{R}^d$ defined as $g(\mathbf{u}_1, \cdots, \mathbf{u}_m) = T \cdot (\mathbf{u}_1 \otimes \cdots, \mathbf{u}_m)$ and $h(\mathbf{u}) = \mathbf{u}$. Then we have $f(\mathbf{u}) = g(\mathbf{u}_1, \cdots, \mathbf{u}_m)$, where each $\mathbf{u}_i = g(\mathbf{u})$. Therefore,

$$\nabla f(\mathbf{u}) = \nabla g(\mathbf{u}_1, \cdots, \mathbf{u}_m) = \sum_{i=1}^{m} \frac{\partial \mathbf{u}_i}{\partial \mathbf{u}} \frac{\partial g}{\partial \mathbf{u}_i} = \sum_{i=1}^{m} I(T\mathbf{u}^{\otimes m-1}) = mT \cdot \mathbf{u}^{\otimes m-1} ,$$

where $\frac{\partial g}{\partial \mathbf{u}_i} = T\mathbf{u}^{\otimes m-1}$ follows from that T is a symmetric tensor.

Therefore, we get

$$\begin{aligned}\|\nabla f(\mathbf{u}) - \nabla f(\mathbf{u}')\|_2 =& \|mT \cdot (\mathbf{u}^{\otimes m-1} - \mathbf{u}'^{\otimes m-1})\|_2 \\ \leq& m\|T\|_F \sum_{i=1}^{m} \binom{m}{i} \|(\mathbf{u}' - \mathbf{u})^{\otimes i} \mathbf{u}^{\otimes m-i}\|_F \\ \leq& m\|T\|_F \sum_{i=1}^{m} \binom{m}{i} \|(\mathbf{u}' - \mathbf{u})\|_2^i \|\mathbf{u}\|_2^{m-i} \\ \leq& m\|T\|_F \sum_{i=1}^{m} \binom{m}{i} 2^m \|(\mathbf{u}' - \mathbf{u})\|_2 \\ \leq& m^2 m^m \|T\|_F 2^m \|(\mathbf{u}' - \mathbf{u})\|_2 = O(\|\mathbf{u}' - \mathbf{u}\|_2) \ ,\end{aligned}$$

In the first inequality, we use the symmetry of $\mathbf{u}^{\otimes i}, i \in [m]$. In the third inequality, we use the fact that $\|\mathbf{u}' - \mathbf{u}\| \leq 2$. And in the last inequality follows from $m \leq 3$ and $T = \mathbf{E}_{\mathbf{x} \sim D^+} \mathbf{x}^{\otimes m} - \mathbf{E}_{\mathbf{x} \sim D^-} \mathbf{x}^{\otimes m}$, where D^+ and D^- are supported on $\mathbb{B}^d(1)$. ■

The next fact measures the progress made in each step during the optimization step.

Fact 10 *Let $f : S^{d-1} \to R$ be an L-smooth function. For $\mathbf{u} \in S^{d-1}$, $\mathbf{g} = \mathrm{proj}_{\mathbf{u}^\perp} \nabla f(\mathbf{u})$, and $\mathbf{u}' = \frac{\mathbf{u} + \lambda \mathbf{g}}{\|\mathbf{u} + \lambda \mathbf{g}\|_2}$ where the stepsize $\lambda = c \min(1/L, 1/\left\|\mathrm{proj}_{\mathbf{u}} \nabla f(\mathbf{u})\right\|_2, 1)$ and c is a sufficiently small constant, we have $f(\mathbf{u}') - f(\mathbf{u}) = \Omega(\lambda \|\mathbf{g}\|_2^2)$.*

Proof Let the angle between $\mathbf{u}$ and $\mathbf{u}'$ be θ. Then, $\lambda \|\mathbf{g}\|_2 = \tan(\theta)$ is at most a sufficiently small constant. Notice that

$$\begin{aligned}f(\mathbf{u}') - f(\mathbf{u}) \geq& (\mathbf{u}' - \mathbf{u}) \cdot \nabla f(\mathbf{u}) - L\|\mathbf{u}' - \mathbf{u}\|^2 \\ =& \lambda \mathbf{g} \cdot \mathbf{g} + (\|\mathbf{u} + \lambda \mathbf{g}\|_2 - 1) \mathbf{u}' \cdot \nabla f(\mathbf{u}) - L(2\sin(\theta/2))^2 \\ \geq& \lambda \|\mathbf{g}\|_2^2 + (\|\mathbf{u} + \lambda \mathbf{g}\|_2 - 1)(\mathrm{proj}_{\mathbf{u}} \mathbf{u}' \cdot \mathrm{proj}_{\mathbf{u}} \nabla f(\mathbf{u}) + \mathrm{proj}_{\mathbf{u}^\perp} \mathbf{u}' \cdot \mathbf{g}) - L\lambda^2 \|\mathbf{g}\|_2^2 \\ =& \lambda \|\mathbf{g}\|_2^2 + (1/\cos(\theta) - 1)(\cos(\theta) \|\mathrm{proj}_{\mathbf{u}} \nabla f(\mathbf{u})\|_2 - \sin(\theta) \|\mathbf{g}\|_2) - L\lambda^2 \|\mathbf{g}\|_2^2 \\ \geq& \lambda \|\mathbf{g}\|_2^2 + O(\sin\theta)^2 (\cos(\theta) \|\mathrm{proj}_{\mathbf{u}} \nabla f(\mathbf{u})\|_2 - \sin(\theta) \|\mathbf{g}\|_2) - L\lambda^2 \|\mathbf{g}\|_2^2 \\ \geq& \lambda \|\mathbf{g}\|_2^2 + O(\lambda^2 \|\mathbf{g}\|_2^2) \|\mathrm{proj}_{\mathbf{u}} \nabla f(\mathbf{u})\|_2 - \sin(\theta)^2 \lambda \|\mathbf{g}\|_2^2 - L\lambda^2 \|\mathbf{g}\|_2^2 \\ \geq& \lambda \|\mathbf{g}\|_2^2 + O(\lambda \|\mathrm{proj}_{\mathbf{u}} \nabla f(\mathbf{u})\|_2) \lambda \|\mathbf{g}\|_2^2 - O(\lambda) \|\mathbf{g}\|_2^2 - (L\lambda) \lambda \|\mathbf{g}\|_2^2 \\ =& \Omega(\lambda \|\mathbf{g}\|_2^2) \ ,\end{aligned}$$

Here, in the first inequality, we use Taylor expansion for f as well as the smoothness of f. In the second inequality, we use the fact that $\lambda \left\|\mathbf{g}\right\| = \tan\theta$. And in the last equality follows from $\lambda = c \min(1/L, 1/\left\|\mathrm{proj}_{\mathbf{u}} \nabla f(\mathbf{u})\right\|_2, 1)$. ■

Now we assume for the purpose of contradiction that the algorithm did not terminate in T iterations, which implies that any $\mathbf{u}_t$ is not an (α', η)-approximate solution we desire for any $t \leq T$. From Fact 10, we have that $f(\mathbf{u}_t)$ is monotone increaseing, therefore, we must have $f(\mathbf{u}_t) \geq f(\mathbf{u}_0) \geq \alpha/d^{c-1}$ for any t. Since we know that any $\mathbf{u}_t$ is not an (α', η)-approximate solution we desire. This implies that $\frac{\eta + \alpha^2/d^c}{\alpha/d^{c-1} - \alpha^2/d^c} = \Omega(\alpha)$, and therefore we must have $\eta =$

$\|\mathrm{proj}_{\mathbf{u}_t^\perp}\nabla f(\mathbf{u}_t)\|_2 \geq \alpha^2/d^c$. This implies that we must make some progress for each step of the gradient descent. Accoriding to Fact 10, we must have that

$$\begin{aligned} f(\mathbf{u}_{t+1}) - f(\mathbf{u}_t) =& \Omega(\min(1, 1/\|\mathrm{proj}_{\mathbf{u}_t}\nabla f(\mathbf{u}_t)\|_2)\|g\|_2^2 \\ =& \Omega(\min(1, 1/\|\mathrm{proj}_{\mathbf{u}_t}\nabla f(\mathbf{u}_t)\|_2)\alpha^4/\mathrm{poly}(d) \\ =& \Omega(\min(1, 1/(mf(\mathbf{u}_t)))\alpha^4/\mathrm{poly}(d) = 1/\mathrm{poly}(d/\alpha)\ , \end{aligned}$$

where the second from the last inequality follows from we use the fact that $\mathrm{proj}_{\mathbf{u}_t}\nabla f(\mathbf{u}_t)\cdot\mathbf{u}_t = mf(\mathbf{u}_t)$. Therefore, after T iterations, we get $f(\mathbf{u}_T) \geq T\alpha^2/\mathrm{poly}(d) + f(\mathbf{u}_0) \geq T\alpha^2/\mathrm{poly}(d) = \omega(\alpha)$. While for any $\mathbf{u}\in S^{d-1}$, we should have $f(\mathbf{u}) = \mathbf{u}^{\otimes 3}\cdot\hat{T} \leq \|\hat{T}\|_F \leq \|T\|_F + \alpha^2/\mathrm{poly}(d) = O(\alpha)$. This gives a contradiction and therefore, we must find a (α', η)-approximate solution such that $\frac{\eta+\alpha^2/d^c}{\alpha'-\alpha^2/d^c} = O(\alpha)$ before T iterations. This completes the proof. ■

Appendix E. Omitted Proofs from Section 4

E.1. Proof of Lemma 15

In this section, we give the proof of Lemma 15. For convenience, we restate Lemma 15 below.

Lemma 40 (Restatement of Lemma 15) *Let D be a joint distribution of $(\mathbf{x}, y)$ on $\mathbb{B}^d(1)\times\{\pm 1\}$ that is consistent with an intersection of halfspaces with γ-margins and $\mathbf{w}\in S^{d-1}$ such that $\|\mathbf{w}_V\|_2 \leq c\gamma$, for some small constant c, where V is the relevant subspace of the intersection of halspaces. Then for any band $B_t := \{\mathbf{x}\in\mathbb{B}^d(1) \mid \mathbf{x}\cdot\mathbf{w}\in[t, t+c\gamma]\}$ where $t\in\mathbb{R}$ and c is a sufficiently small constant, the distribution of $(\mathbf{x}, y)$ conditioned on $\mathbf{x}\in B_t$ is consistent with an instance of learning a degree-2 polynomial threshold function with $\Omega(\gamma^2)$-margin.*

Proof [Proof of Lemma 15] Let $h^*(\mathbf{x}) = \mathrm{sign}(\mathbf{u}^*\cdot\mathbf{x}+t_1)\wedge\mathrm{sign}(\mathbf{v}^*\cdot\mathbf{x}+t_2)$ be the target hypothesis and V be the subspace spanned by $\mathbf{u}^*$ and $\mathbf{v}^*$ and $\mathbf{w}\in S^{d-1}$ such that $\|\mathbf{w}_V\|_2 \leq c\gamma$. Notice that if V is a one-dimensional subspace, then the statement trivially holds for every $\mathbf{w}$. Therefore, without loss of generality, we assume that V is a 2-dimensional subspace.

Without loss of generality, we assume that $|t_1|, |t_2| \leq 1$. Let $\mathbf{w}^* = \mathbf{w}_V/\|\mathbf{w}_V\|_2$. Given V is a 2-dimensional subspace, take $\mathbf{w}'\in S^{d-1}$ to be the unique direction that $\mathbf{w}'\in V$ and $\mathbf{w}\cdot\mathbf{w}' = 0$. Notice that for any $\mathbf{x}\in B_t$, we have

$$\begin{aligned} &\mathbf{u}^*\cdot\mathbf{x} + t_1 \\ =&\mathrm{proj}_{\mathbf{w}^*}\mathbf{u}^*\cdot\mathbf{x} + \mathrm{proj}_{\mathbf{w}'}\mathbf{u}^*\cdot\mathbf{x} + t_1 \\ =&\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{u}^*)\|\mathrm{proj}_{\mathbf{w}^*}\mathbf{u}^*\|_2\mathbf{w}^*\cdot\mathbf{x} + \mathrm{proj}_{\mathbf{w}'}\mathbf{u}^*\cdot\mathbf{x} + t_1 \\ =&\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{u}^*)\|\mathrm{proj}_{\mathbf{w}^*}\mathbf{u}^*\|_2\mathbf{w}\cdot\mathbf{x} + \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{u}^*)\|\mathrm{proj}_{\mathbf{w}^*}\mathbf{u}^*\|_2(\mathbf{w}^*-\mathbf{w})\cdot\mathbf{x} \\ &+ \mathrm{proj}_{\mathbf{w}'}\mathbf{u}^*\cdot\mathbf{x} + t_1\ . \end{aligned}$$

Notice that $\|\mathbf{w}_V\| \leq c\gamma$ implies that $\|\mathbf{w}^*-\mathbf{w}\|_2 \leq 2c\gamma$. Thus, for the second term, we have

$$|\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{u}^*)\|\mathrm{proj}_{\mathbf{w}^*}\mathbf{u}^*\|_2(\mathbf{w}^*-\mathbf{w})\cdot\mathbf{x}| \leq \|\mathrm{proj}_{\mathbf{w}^*}\mathbf{u}^*\|_2\|\mathbf{w}^*-\mathbf{w}\|_2\|\mathbf{x}\| \leq 2c\gamma.$$

Since the distribution satisfies γ-margin condition, we get for any $\mathbf{x} \in B_t$ and on the support of D,

$$\begin{aligned}\operatorname{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) =& \operatorname{sign}\left(\operatorname{sign}(\mathbf{w}^* \cdot \mathbf{u}^*)\|\operatorname{proj}_{\mathbf{w}^*}\mathbf{u}^*\|_2 \mathbf{w} \cdot \mathbf{x} + \operatorname{proj}_{\mathbf{w}'}\mathbf{u}^* \cdot \mathbf{x} + t_1\right)\\ =& \operatorname{sign}\left(\operatorname{proj}_{\mathbf{w}'}\mathbf{u}^* \cdot \mathbf{x} + (t_1 + \operatorname{sign}(\mathbf{w}^* \cdot \mathbf{u}^*)\|\operatorname{proj}_{\mathbf{w}^*}\mathbf{u}^*\|_2 t)\right)\\ =& \operatorname{sign}\left(\operatorname{sign}(\mathbf{w}' \cdot \mathbf{u}^*)\mathbf{w}' \cdot \mathbf{x} + (t_1 + \operatorname{sign}(\mathbf{w}^* \cdot \mathbf{u}^*)\|\operatorname{proj}_{\mathbf{w}^*}\mathbf{u}^*\|_2 t)/\|\operatorname{proj}_{\mathbf{w}'}\mathbf{u}^*\|_2\right) .\end{aligned}$$

Since $\|\operatorname{proj}_{\mathbf{w}'}\mathbf{u}^*\|_2 \leq 1$, we have for any $\mathbf{x} \in B_t$ satisfies

$$\operatorname{sign}(\mathbf{u}^* \cdot \mathbf{x} + t_1) = \operatorname{sign}(\operatorname{sign}(\mathbf{w}' \cdot \mathbf{u}^*)\mathbf{w}' \cdot \mathbf{x} + t_1')$$

with $\Omega(\gamma)$ margin, where $t_1' = (t_1 + \operatorname{sign}(\mathbf{w}^* \cdot \mathbf{u}^*)\|\operatorname{proj}_{\mathbf{w}^*}\mathbf{u}^*\|_2 t)/\|\operatorname{proj}_{\mathbf{w}'}\mathbf{u}^*\|_2$. By symmetry, we also have $\operatorname{sign}(\mathbf{v}^* \cdot \mathbf{x} + t_2) = \operatorname{sign}(\operatorname{sign}(\mathbf{w}' \cdot \mathbf{v}^*)\mathbf{w}' \cdot \mathbf{x} + t_2')$.

Therefore, it suffices for us to show that there is a degree-2 PTF function with $\Omega(\gamma^2)$ margin that is consistent with the intersection of two halfspaces $f'(\mathbf{x}) = \operatorname{sign}(\operatorname{sign}(\mathbf{w}' \cdot \mathbf{u}^*)\mathbf{w}' \cdot \mathbf{x} + t_1') \wedge \operatorname{sign}(\operatorname{sign}(\mathbf{w}' \cdot \mathbf{v}^*)\mathbf{w}' \cdot \mathbf{x} + t_2')$ with $\Omega(\gamma)$ margins. Without loss of generality, we can always assume that $|t_1'| \leq 1$, $|t_2'| \leq 1$ and $\operatorname{sign}(\mathbf{w}' \cdot \mathbf{u}^*) \neq \operatorname{sign}(\mathbf{w}' \cdot \mathbf{v}^*)$. Because, otherwise, $f'(\mathbf{x})$ is equivalent to either $\operatorname{sign}(\operatorname{sign}(\mathbf{w}' \cdot \mathbf{u}^*)\mathbf{w}' \cdot \mathbf{x} + t_1')$ or $\operatorname{sign}(\operatorname{sign}(\mathbf{w}' \cdot \mathbf{v}^*)\mathbf{w}' \cdot \mathbf{x} + t_2')$. Without loss of generality, assume that $\operatorname{sign}(\mathbf{w}' \cdot \mathbf{u}^*) = 1$ and $\operatorname{sign}(\mathbf{w}' \cdot \mathbf{v}^*) = -1$. Furthermore, we can without loss of generality assume that $-t_1' \leq t_2'$, because otherwise, f' is equivalent to the -1 constant function. Given the above assumptions, we can without loss of generality assume that the function f' is equivalent to

$$f(\mathbf{x}) = \begin{cases} -1, & \text{for } \mathbf{w}' \cdot \mathbf{x} \in (-\infty, -t_1' - c\gamma] \\ 1, & \text{for } \mathbf{w}' \cdot \mathbf{x} \in [-t_1' + c\gamma, t_2' - c\gamma] \\ -1, & \text{for } \mathbf{w}' \cdot \mathbf{x} \in [t_2' + c\gamma, \infty] \end{cases},$$

where from the margin condition, for any $\mathbf{x} \in B_t$ and from the support of D we cannot have $\mathbf{w}' \cdot \mathbf{x} \in [-t_1' - c\gamma, -t_1' + c\gamma] \cup [t_2' - c\gamma, -t_1' + c\gamma]$ and c is a sufficiently small constant. Therefore, simply take the degree-2 PTF as as $\operatorname{sign}(p(\mathbf{x}))$, where $p(\mathbf{x}) = (\operatorname{sign}(\mathbf{w}' \cdot \mathbf{u}^*)\mathbf{w}' \cdot \mathbf{x} + t_1')(\operatorname{sign}(\mathbf{w}' \cdot \mathbf{v}^*)\mathbf{w}' \cdot \mathbf{x} + t_2')$ will immediately give us a function that is consistent with f' with $c\gamma^2$ margins and c is a sufficiently small constant. This completes the proof.

■

E.2. Proof of Theorem 16

In this section, we give the main weak learning algorithm as Algorithm 5 and present the proof of Theorem 16. For convenience, we restate Theorem 16 below.

Theorem 41 *There is an algorithm $\mathcal{A}$ such that for every instance of learning intersections of two halfspaces with γ-margin assumption, given $\mathbf{w} \in S^{d-1}$ such that $\|\mathbf{w}_W\|_2 \leq c\gamma$ where c is a sufficiently small constant, $\mathcal{A}$ draws $\operatorname{poly}(d, 1/\gamma)$ examples from D, runs in $\operatorname{poly}(d, 1/\gamma)$ time and outputs a hypothesis $h : \mathbb{B}^d(1) \to \{\pm 1\}$ such that with probability at least $2/3$, $\operatorname{err}(h) \leq 1/2 - \Omega(\gamma)$.*

The algorithm in Theorem 41 is provided as Algorithm 5.

Proof [Proof of Theorem 41] From Chernoff bound, we have that the empirical estimation $\hat{P}$ in Line 4 has error at most $c/|T|$ with failure probability at most $c/|T|$, where c is a sufficiently small constant. Therefore, with at least constant probability, all estimations in Line 4 have error at most

Algorithm 5 Weak Learning Intersection of Halfspaces under Product Distribution using a Relevant Direction

Input: i.i.d. sample access to a distribution D on $\mathbb{B}^d(1) \times \{\pm 1\}$ that is an instance of learning intersections of halfspaces under product distribution with γ margins and $\mathbf{u} \in S^d(1)$ such that $\|\mathrm{proj}_{V^\perp}\mathbf{u}\|_2 \leq c\gamma$ where c is a sufficiently small constant.

Output: With at least a constant probability, the algorithm outputs a hypothesis h such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq 1/2 - \Omega(\gamma)$.

1: Let discrete set $T \subseteq [-1,1]$ such that $|T| \leq 2/(c_1\gamma)$ and for any $t^* \in [-1,1]$, there exists a $t \in T$ such that $|t - t^*| \leq c_1\gamma$, where c_1 is a sufficiently small constant depending on c.
2: **for** $t \in T$ **do**
3: Let the localization area be defined as $B_t = \{\mathbf{x} \in \mathbb{B}^d(1) \mid \mathbf{x} \cdot \mathbf{u} \in [t - c_1\gamma, t + c_1\gamma]\}$.
4: Estimate $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x} \sim B_t]$ with $\hat{P}_t := \frac{1}{n}\sum_{(\mathbf{x},y)\in S} \mathbb{1}(\mathbf{x} \in B_t)$ by drawing a set S of $\mathrm{poly}(1/\gamma)$ from D.
5: **if** $\hat{P}_t \geq 1/(2|T|)$ **then**
6: Let D_t be the distribution of $(V(\mathbf{x}), y) \sim D$ conditioned on $\mathbf{x} \in B_t$ where $V : \mathbb{R}^d \to \mathbb{R}^{(d+1)^2}$ defined as $V(\mathbf{x}) = [\mathbf{x}, 1]^{\otimes 2}$ is the degree-2 Veronese mapping.
7: Use rejection sampling to get sample access to D_t and apply the standard perceptron to learn a LTF with $c_2\gamma^2$ margins to additive error $1/4$ with success probability $1/2$, where c_2 is a sufficiently small constant depending on c_1, and let $h' : \mathbb{R}^{(d+1)^2} \to \{-1, 1\}$ be the output hypothesis.
8: Select the best $c' \in \{-1, 1\}$ and return the hypothesis h defined as

$$h(\mathbf{x}) = \begin{cases} h'(V(\mathbf{x})), & \text{if } \mathbf{x} \in B_t\ ; \\ c', & \text{otherwise,} \end{cases}$$

and terminate.

$c/|T|$. Since we only need to show that the algorithm succeeds with a constant probability, we assume that all estimations have error at most $c/|T|$ and the algorithm in Line 4 succeeds for the rest of the proof.

Notice that from the definition of T, we get $\mathbb{B}^d(1) \subseteq \bigcup_{t\in T} B_t$. Therefore, there must be a $t \in T$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x} \in B_t] \geq 1/|T|$. Thus, the "if" condition in Line 5 must be satisfied at some point. Suppose that it is satisfied for B_t, then we must have $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x} \in B_t] \geq 1/(3|T|)$. Furthermore, from Lemma 15, we must have that there is a degree-2 PTF that is consistent with the distribution D_t' with $\Omega(\gamma^2)$ margins, where D_t' is defined as the distribution of $(\mathbf{x}, y) \sim D$ conditioned on $\mathbf{x} \in B_t$. Notice that such a polynomial $p : \mathbb{R}^d \to \{\pm 1\}$ can be written in the form $p(\mathbf{x}) = \mathrm{sign}(T \cdot [\mathbf{x}, 1]^{\otimes 2})$ for some $\|T\|_f \leq 1$. Therefore, we must have that there is an LTF $h : (\mathbb{R}^{d+1})^{\otimes 2} \to \{\pm 1\}$ defined as $h(\mathbf{x}) = \mathrm{sign}(T \cdot \mathbf{x})$ that is consistent with the distribution D_t with $\Omega(\gamma^2)$ margins, where the constant where depends on c_1. Then from the correctness of the perceptron algorithm (see Cristianini (2000)), we must have that with at least constant probability (when the perceptron algorithm succeeds), $\mathbf{Pr}_{(\mathbf{x},y)\sim D_t}[h'(\mathbf{x}) \neq y] \leq 1/4$. From the definition of

D_t and D'_t, this also implies that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D'_t}[h'(V(\mathbf{x})) \neq y] \leq 1/4 \ .$$

Therefore, with at least constant probability, the error of the output hypothesis is

$$\begin{aligned}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] = &\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[h'(V(\mathbf{x})) \neq y \wedge \mathbf{x} \in B_t] + \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[c' \neq y \wedge \mathbf{x} \notin B_t] \\ \leq &\frac{1}{4}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[\mathbf{x} \in B_t] + \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[c' \neq y \wedge \mathbf{x} \notin B_t] \ .\end{aligned}$$

Since we are choosing the best constant $c' \in \{-1, 1\}$, with at least constant probability, we have $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[c' \neq y \wedge \mathbf{x} \notin B_t] \leq \frac{1}{2}\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x} \notin B_t]$. Combing with the fact that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x} \in B_t] \geq 1/(3|T|)$ and $|T| = O(1/\gamma)$, we get

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq 1/2 - \Omega(\gamma) \ .$$

This completes the proof. ■

Appendix F. Proof of Theorem 3

In this section, we present a detailed version of our main algorithm as Algorithm 6 and the full proof of Algorithm 6.

Before presenting the proof of Theorem 3, it is convenient to recall the well known Adaboost algorithm developed by Schapire and Freund (2013).

Theorem 42 (AdaBoost Schapire and Freund (2013)) *Let H be a binary hypothesis class over a space of example X. A learning algorithm $\mathcal{A}$ is said to be an α-weak learning algorithm for H if for every distribution D over H such that there exists some $h^* \in H$ such that $\mathrm{err}_D(h^*) = 0$, $\mathcal{A}$ outputs in $T(\mathcal{A})$ time a hypothesis $c : X \to \{\pm 1\}$ such that $\mathrm{err}_D(c) \leq 1/2 - \alpha$ by drawing i.i.d. examples from D. If the hypothesis c output by $\mathcal{A}$ belongs to a hypothesis class $\mathcal{C}$ with VC-dimension d, then there is an algorithm AdaBoost such that for every ϵ, δ and for every distribution D over H such that there exists some $h^* \in H$ with $\mathrm{err}_D(h^*) = 0$, it draws a set S of $\mathrm{poly}(d, 1/\epsilon, 1/\alpha, \log(1/\delta))$ examples from D, and outputs in $\mathrm{poly}(d, 1/\epsilon, 1/\alpha, \log(1/\delta), T(\mathcal{A}))$ time a hypothesis $\hat{h} : X \to \{\pm 1\}$ with $\mathrm{err}_D(\hat{h}) \leq \epsilon$ with probability at least $1 - \delta$, by running $\mathcal{A}$, $O(\log(1/\epsilon)/\alpha^2)$ times over distributions over S.*

Proof [Proof of Theorem 3] We first prove the correctness of Algorithm 6. Notice that if h^* is ϵ-close to any constant hypothesis, i.e. $\min\{\mathbf{Pr}_{\mathbf{x}\sim D}(h^*(\mathbf{x}) = +1), \min\{\mathbf{Pr}_{\mathbf{x}\sim D}(h^*(\mathbf{x}) = -1)\} \leq \epsilon/2$, then by Hoeffding's inequality Vershynin (2018), with probability at least $1-O(\delta)$, $\min\{\hat{p}, 1-\hat{p}\} < \epsilon$. In this case, Algorithm 6 outputs a constant hypothesis with error $\epsilon/2$ in $\mathrm{poly}(1/\epsilon, \log(1/\delta))$ time. In the rest of the proof, we assume $\min\{\mathbf{Pr}_{x\sim D}(h^*(x) = -1)\} > \epsilon/2$.

Since each $\mathbf{x} \sim D_X$ has $\|\mathbf{x}\| \leq 1$, we know that for $z \in \{\pm 1\}$ and $k = \{1, 2, 3\}$, the empirical distribution $\hat{D^z}$, constructed with $m_z = \mathrm{poly}(d, 1/\gamma, \log(1/\delta))$ i.i.d. examples from D^z, satisfies $\left\|\mathbf{E}_{\mathbf{x}\sim\hat{D^z}}(\mathbf{x}) - \mathbf{E}_{\mathbf{x}\sim D^z}(\mathbf{x})\right\|_F \leq \frac{1}{100}(\gamma/d)^{10c}$ Given this happens, we consider two cases for the empirical distribution $(\hat{D^+}, \hat{D^-})$. In the first case, $\forall t \in [3]$, $(\hat{D^+}, \hat{D^-})$ satisfies $(\alpha_t\gamma^{c_t}/\mathrm{poly}(d), t)$-moment matching condition, where $\alpha_1 = 1/64, \alpha_2 = 1/16, \alpha_3 = 1/2$ and $c_t \geq c$. By Hoeffding's inequality, we know that D satisfies

Algorithm 6 LEARNING INTERSECTIONS OF TWO HALFSPACES (Computationally efficient algorithm for learning intersections of two halfspaces)

Input: $\epsilon, \delta, \gamma \in (0,1)$ and i.i.d. sample access to a distribution D on $\mathbb{B}^d(1) \times \{\pm 1\}$ that is an instance of learning intersections of halfspaces under product distribution with γ-margin assumption.
Output: With probability at least $1-\delta$, the algorithm outputs a hypothesis $\hat{h} : \mathbb{R}^d \to \{\pm 1\}$, such that $\mathrm{err}(\hat{h}) \leq \epsilon$.
Draw $m_0 = O(1/\epsilon, \log(1/\delta))$ examples $S_0 := \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=0}^{m_0}$ from D and estimate $\hat{p} := \frac{1}{m_0}\mathbb{1}(y^{(i)} = 1)$. If $\hat{p} < \epsilon$ or $\hat{p} > 1-\epsilon$, return a constant hypothesis accordingly.
For $z \in \{\pm\}$, draw $m_z = \mathrm{poly}(d, \gamma, \log(1/\delta))$ i.i.d. examples $S_z := \{\mathbf{x}^{(i)}\}_{i=0}^{m_z}$ from D^z via rejection sampling.
For $z \in \{\pm\}$, let $\hat{D^z}$, the uniform distribution over S_z be the empirical distribution of D^z.
if $\forall t \in [3]$, $(\hat{D^+}, \hat{D^-})$ satisfies $(\alpha_t \gamma^{c_t}, t)$-moment matching condition, where $\alpha_1 = 1/64, \alpha_2 = 1/16, \alpha_3 = 1/2, c_1 = 4c, c_2 = 2c, c_3 = c$. **then**
    Run Algorithm 3 over D, $\mathrm{poly}(d, 1/\gamma, \log(1/\delta))$ times and denote by $\mathcal{O}$ the union of the outputs of running Algorithm 3.
else
    Find the first t such that $(\hat{D^+}, \hat{D^-})$ does not satisfy $(\alpha_t \gamma^{c_t}/\mathrm{poly}(d), t)$-moment matching condition.
    Run Algorithm 4 with parameter t over D, $O(\log(1/\delta))$ times and denote by $\mathcal{O}$ the union of the outputs of running Algorithm 4.
for $\mathbf{w} \in \mathcal{O}$ **do**
    Run Boosting algorithm to get a hypothesis $h_{\mathbf{w}}$ using the weak learning algorithm used in Section 4 using $\mathbf{w}$ as the input vector.
Draw $\mathrm{poly}(1/\epsilon, \log(d/(\gamma\delta)))$ i.i.d. examples from D and find the hypothesis $\hat{h}$ from $\hat{H} = \{h_{\mathbf{w}} \mid \mathbf{w} \in \mathcal{O}\}$ with smallest empirical error
return $\hat{h}$.

1. $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}\|_F \leq \gamma^c$.
2. $\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}\mathbf{x}^\intercal\|_F \leq \gamma^c$
3. $\left\|(\mathbf{E}_{\mathbf{x}\sim D^+} - \mathbf{E}_{\mathbf{x}\sim D^-})\mathbf{x}^{\otimes 3}\right\|_F \leq \gamma^c$

By Theorem 13, we know that with probability at least $\Omega(\gamma/d)$, Algorithm 3, outputs a list of d unit vectors $\mathbf{w}$ such that at least one of the $\mathbf{w}$ satisfies $\|\mathbf{w}_V\|_2 \leq \mathrm{poly}(\gamma)$. Thus, by running Algorithm 3 $\mathrm{poly}(d, 1/\gamma, \log(1/\delta))$ times, with probability at least $1 - O(\delta)$, one of these implementations satisfies the above guarantee, which implies that we have a list of unit vectors $\mathcal{O}$ of size $\mathrm{poly}(d, 1/\gamma, \log(1/\delta))$ such that one of the unit vectors $\mathbf{w}$ satisfies $\|\mathbf{w}_V\|_2 \leq \mathrm{poly}(\gamma)$.

In the second case, there must be some $t \in [3]$ such that $(\hat{D^+}, \hat{D^-})$ does not satisfy the $(\alpha_t \gamma^{c_t}/\mathrm{poly}(d), t)$-moment matching condition. In this case, we consider the smallest $t \in [3]$ that satisfies the above condition. By the choice of α_t, it is always holds that $\alpha_{t-1} \leq 2^{-t}\alpha_t$. This implies that D does not satisfy $(\alpha_t \gamma^{c_t}, t)$ moment matching condition but satisfies $(\alpha_t \gamma^{2c_t}/(2^t \mathrm{poly}(d)), t')$-moment matching condition, for every $t' \leq t$. Since $(\gamma^{c_t}) = \gamma^{\Omega(c)}$, by Theorem 14, with probability $\Omega(1)$, Algorithm 4 outputs a vector $\mathbf{w}$ such that $\|\mathbf{w}_V\|_2 \leq \mathrm{poly}(\gamma)$. Thus, by running Algorithm 4,

$(\log(1/\delta))$ times, we obtain a list of $O(\log(1/\delta))$ unit vectors such that at least one of the $\mathbf{w}$ satisfies $\|\mathbf{w}_V\|_2 \leq \mathrm{poly}(\gamma)$.

By Theorem 16, we know that provided a unit vector $\mathbf{w}$ such that $\|\mathbf{w}_V\|_2 \leq \mathrm{poly}(\gamma)$, Algorithm 5 is a $\Omega(\gamma)$-weak learner. And the hypothesis output by Algorithm 5 is a polynomial threshold function restricted at some band, which has a VC dimension $O(d)$. Given this happens, Theorem 42 implies Adaboost takes Algorithm 5 as a weak learner and outputs a hypothesis $h_{\mathbf{w}}$ such that $\mathrm{err}(h_{\mathbf{w}}) \leq \epsilon$ with probability at least $1 - O(\delta)$. Since one of the $\mathbf{w}$ satisfies $\|\mathbf{w}_V\|_2 \leq \mathrm{poly}(\gamma)$ and $\mathcal{O}$ has size at most $\mathrm{poly}(d, 1/\gamma, \log(1/\delta))$, a standard hypothesis testing approach outputs some $\hat{h} \in \{h_{\mathbf{w}} \mid \mathbf{w} \in \mathcal{O}\}$ with $\mathrm{err}(\hat{h}) \leq \epsilon$ with probability at least $1 - O(\delta)$. By union bound, with probability at least $1 - \delta$, Algorithm 6 outputs a hypothesis $\hat{h}$ with $\mathrm{err}(\hat{h}) \leq \epsilon$.

To conclude the proof of Theorem 3, we bound the sample complexity and the time complexity of Algorithm 6. Given $\min\{\mathbf{Pr}_{x\sim D}(h^*(x) = -1)\} > \epsilon/2$, sampling one example $\mathbf{x} \sim D^z$, for $z \in \{\pm\}$ has a sample complexity $\tilde{O}(1/\epsilon)$. Thus, constructing the empirical distribution $(\hat{D^+}, \hat{D^-})$ takes $\mathrm{poly}(d, 1/\gamma, \log(1/\delta))$ sample and time. On the other hand, by Theorem 13 and Theorem 14, every time we run Algorithm 3 and Algorithm 4, it takes us $\mathrm{poly}(d, 1/\gamma, \log(1/\delta))$ sample and time. Since we run Algorithm 3 and Algorithm 4 at most $\mathrm{poly}(d, 1/\gamma, \log(1/\delta))$ times, we know that constructing $\mathcal{O}$ takes $\mathrm{poly}(d, 1/\gamma, \log(1/\delta))$ sample and time. Finally, since $\mathcal{O}$ has size at most $\mathrm{poly}(d, 1/\gamma, \log(1/\delta))$, by Theorem 16 and Theorem 42, it takes $\mathrm{poly}(d, 1/\gamma, \log(1/\delta))$ sample and time to create $\{h_{\mathbf{w}} \mid \mathbf{w} \in \mathcal{O}\}$. Finally, as a hypothesis testing approach over $\{h_{\mathbf{w}} \mid \mathbf{w} \in \mathcal{O}\}$ can be done efficiently. We know that the sample complexity and time complexity of Algorithm 6 are both $\mathrm{poly}(d, 1/\gamma, \log(1/\delta))$. ■

Appendix G. CSQ Lower Bound on Learning Intersection of Margin Halfspaces under Factorizable Distributions

We give our main theorem for CSQ hardness as the following.

Theorem 43 *Let $\gamma > 0$, $q \in \mathbb{N}$, $\tau \in (0, 1)$ and $d' = \min(d, 1/\gamma^2)$. Any CSQ algorithm that learns intersections of two halfspaces on d-dimension with γ-margin under factorizable distributions to error $1/2 - \max(d'^{-\Omega(\log(1/\gamma))}, 2^{-d'^{\Omega(1)}})$ requires q queries of tolerance at most τ, where $q/\tau^2 \geq \min(d'^{\Omega(\log(1/\gamma))}, 2^{d'^{\Omega(1)}})$.*

The high-level proof idea here follows the framework of Non-Gaussian Component Analysis (see Diakonikolas et al. (2017) and Diakonikolas et al. (2023)). To prove the CSQ hardness, it suffices for us to prove a CSQ lower bound against an easier decision problem as defined in Definition 22. For convenience, instead of considering distributions on $\mathbb{B}^d \times \{\pm 1\}$, we will consider distributions on $\mathbb{R}^d \times \{\pm 1\}$. As we will later see, the difference here is trivial as we will be able to rescale and truncate these distributions (for our construction) inside a ball at the cost of a very small total variation distance. We show that given CSQ access to a joint distribution D of $(\mathbf{x}, y)$ supported on $\mathbb{R}^d \times \{\pm 1\}$, it is hard to solve the problem $\mathcal{B}(\mathcal{D}, D)$ with the following distributions.

(a) Null hypothesis: We have $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$ and $y = 1$ with probability $1/2$ independent of $\mathbf{x}$.

(b) Alternative hypothesis: $D \in \mathcal{D}$, where $\mathcal{D}$ is a family of distributions such that for any distribution $D \in \mathcal{D}$, D is close in total variation distance to a distribution D' that is an instance of learning intersections of two halfspaces with γ-margin under factorizable distributions.

To construct the family of distribution $\mathcal{D}$, we first construct a distribution D of $(\mathbf{x}', y)$ supported on $\mathbb{B}^2(1) \times \{\pm 1\}$ that is consistent with an intersection of halfspaces with γ margin and $\mathbf{E}_{(\mathbf{x},y)\sim D}[yp(\mathbf{x})] = 0$ for any polynomial p of degree at most $O(\log(1/\gamma))$. Being supported inside the ball here will be convenient for later truncation. We give the following lemma for distribution D where the extra third property here is needed for technical reasons.

Lemma 44 *Let $\gamma > 0$, then there exists a joint distribution D of $(\mathbf{x}, y)$ supported on $\mathbb{B}^2(1) \times \{\pm 1\}$ that satisfied the following conditions:*

1. *(Realizable by an intersection of two halfspaces with γ margins) There exists an intersection of two halfspace h^* such that D is realizable by h^* with γ margins;*

2. *(Orthogonal with low-degree polynomial) For any polynomial $p : \mathbb{R} \to \mathbb{R}$ of degree at most $c\log(1/\gamma)$ where c is a sufficiently small constant and $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_2}[p(\mathbf{x})] = 0$, we have $\mathbf{E}_{(\mathbf{x},y)\sim D}[yp(\mathbf{x})] = 0$;*

3. *(Bounded chi-squared distance with Gaussian) $\chi^2(D^+, \mathcal{N}_2),\ \chi^2(D^-, \mathcal{N}_2) = O(1/\gamma)$.*

Proof To construct such a distribution D, we first construct a distribution D' supported on $\{\pm 1\}^n \times \{\pm 1\}$. Then we obtain D by projecting D' onto a 2-dimensional subspace and add $\mathcal{N}(0, \sigma I_2)$ noise on $\mathbf{x}$ where $\sigma = \Theta(\gamma)$. The purpose of the extra noise is to make the originally discrete distribution continuous so we can have bounded χ-squared distance. We introduce the following fact about such a D' from Sherstov (2009).

Fact 11 *Let $n \in \mathbb{N}$, then there exists a joint distribution D of $(\mathbf{x}, y)$ supported on $\{\pm 1\}^n \times \{\pm 1\}$ that satisfies the following conditions.*

1. *(Realizable by an intersection of two halfspaces) There exists an intersection of two halfspace c such that D is realizable by c where the weight is the halfspaces is $2^{O(\sqrt{n})}$.*

2. *(Orthogonal with low-degree polynomial) For any polynomial $p : \mathbb{R} \to \mathbb{R}$ of degree at most $c\sqrt{n}$ where c is a sufficiently small constant and $\mathbf{E}_{\mathbf{x}\sim_u\{\pm 1\}^n}[p(\mathbf{x})] = 0$, we have $\mathbf{E}_{(\mathbf{x},y)\sim D}[yp(\mathbf{x})] = 0$.*

Suppose $\|\mathbf{w}_1\|_2, \|\mathbf{w}_2\|_2 \le 2^{c_1\sqrt{n}}$ in Item 1 of Fact 11. Then we set $n = \log(1/\gamma)^2/(100c_1)^2$ in Fact 11, and let D', $\mathbf{w}_1$ and $\mathbf{w}_2$ be the corresponding distribution and weight for the intersection of halfspaces. This implies that $\|\mathbf{w}_1\|_2, \|\mathbf{w}_2\|_2 \le (1/\gamma)^{1/100}$. We then first define the distribution D'' as the distribution of $\left(\frac{1}{2(1/\gamma)^{1/100}}[\mathbf{x}\cdot\mathbf{b}_1, \mathbf{x}\cdot\mathbf{b}_2]^\intercal, y\right)$, where we sample $(\mathbf{x}, y) \sim D'$ and $\mathbf{b}_1$ and $\mathbf{b}_2$ are orthornomal basis vectors that spans the subspace spaned by $\mathbf{w}_1$ and $\mathbf{w}_2$. Then we defined a independent noise random vector $\mathbf{z} \in \mathbb{R}^2$, sampled by having $\mathbf{z} \sim \mathcal{N}_2(0, \gamma I_2/100)$ and then conditioned on $\|\mathbf{z}\|_2 \le \gamma$. Finally, we define the desired distribution D as the distribution of $(\mathbf{x} + \mathbf{z}, y)$ where $(\mathbf{x}, y) \sim D''$ and $\mathbf{z}$ as defined above.

Notice that D'' has at least 10γ margins for an intersection of halfspaces, which follows from D' has $\Omega(1)$ margins (given γ is sufficiently small). Then the extra noise $\mathbf{z}$ has $\|\mathbf{z}\|_2 \le \gamma$, therefore, D still have γ margin for an intersection of halfspaces. Furthermore, D is supported inside $\mathbb{B}^2(1)\times\{\pm 1\}$, which follows from $\|\mathbf{w}_1\|_2, \|\mathbf{w}_2\|_2 \le (1/\gamma)^{1/100}$ and $\|\mathbf{z}\|_2 \le \gamma$. This proves Item 1 of Lemma 44.

For Item 2 of Lemma 44, notice that $\mathbf{E}_{(\mathbf{x},y)\sim D''}[yp(\mathbf{x})] = 0$ for any p of degree-$c\sqrt{n}$ (for c a sufficiently small constant) since D'' is transform from D' through a linear transformation. Then Item 2 of Lemma 44 follows from that $\mathbf{z}$ are sampled independently in D.

For Item 3 of Lemma 44, notice that due to the extra noise $\mathbf{z}$ smoothed out the discrete distribution D'',

$$\begin{aligned}\chi^2(D^+,\mathcal{N}_2) &= \mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}[P_{D^+}(\mathbf{x})/P_{\mathcal{N}_2}(\mathbf{x})] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim D^+}\left[\int_{\mathbf{x}'\in\mathbb{R}^2} P_{\mathbf{z}}(\mathbf{x}-\mathbf{x}')P_{D''^+}(\mathbf{x}')d\mathbf{x}'/P_{\mathcal{N}_2}(\mathbf{x})\right] \\ &= O(1/\gamma)\ ,\end{aligned}$$

where the last inequality follows from that $P_{\mathbf{z}}$ is bounded everywhere by $O(1/\gamma)$ and the fact that D^+ is supported inside $\mathbb{B}^2(1)$. The exact same argument holds for $\chi^2(D^-,\mathcal{N}_2)$. ■

To construct the family of distribution $\mathcal{D}$ from the 2-dimensional distribution in Lemma 44, we will pick a large set of near orthogonal 2-dimensional subspaces. For each subspace $V \in \mathbb{R}^{2\times d}$ in the set, we embed the distribution D supported on $\mathbb{B}^2(1)\times\{\pm 1\}$ along this subspace. Namely, the distribution we create is defined as the following.

Definition 45 (Hidden-Subspace Distribution) *For a distribution A supported on $\mathbb{R}^m$ and a matrix $V \in \mathbb{R}^{n\times m}$ with $V^\intercal V = I_m$, we define the distribution P_V^A supported on $\mathbb{R}^n$ such that it is distributed according to A in the subspace $\mathrm{span}(\mathbf{v}_1,\ldots,\mathbf{v}_m)$ and is an independent standard Gaussian in the orthogonal directions, where $\mathbf{v}_1,\ldots,\mathbf{v}_m$ denote the column vectors of V. In particular, if A is a continuous distribution with probability density function $A(\mathbf{y})$, then P_V^A is the distribution over $\mathbb{R}^n$ with probability density function*

$$P_V^A(\mathbf{x}) = A(\mathbf{v}_1\cdot\mathbf{x},\ldots,\mathbf{v}_m\cdot\mathbf{x})\exp(-\|\mathbf{x}-VV^\intercal\mathbf{x}\|_2^2/2)/(2\pi)^{(n-m)/2}\ .$$

Furthermore, for a distribution A of $(\mathbf{x}',y')$ supported on $\mathbb{R}^m\times\{\pm1\}$, we define the distribution P_V^A of $(\mathbf{x},y)$ as the distribution supported on $\mathbb{R}^n\times\{\pm1\}$ that satisfies the following.

1. *y and y' has the same marginal distribution; and*
2. *$\left(P_V^A\right)^-$ is $P_V^{(A^-)}$ and $\left(P_V^A\right)^+$ is $P_V^{(A^+)}$.*

That is, P_V^A over $\mathbb{R}^n\times\{\pm1\}$ is the product distribution whose orthogonal projection onto the subspace of V and the label space $\{\pm1\}$ is A, and onto the subspace in $\mathbb{R}^n$ perpendicular to V is the standard $(n-m)$-dimensional normal distribution. For our setting, we will consider the special case of $m=2$. We will use the subspaces $\mathbf{V}\in S$ where S is exponential in size. We give the following lemma for constructing S.

Fact 12 (Near-orthogonal Subspaces: Lemma 2.5 from Diakonikolas et al. (2021)) *Let $0<a,c<1/2$ and $m,n\in\mathbb{Z}_+$ such that $m\le n^a$. There exists a set S of $2^{\Omega(n^c)}$ matrices in $\mathbb{R}^{m\times n}$ such that every $U\in S$ satisfies $UU^\intercal = I_m$ and every pair $U,V\in S$ with $U\neq V$ satisfies $\|UV^\intercal\|_F \le O(n^{2c-1+2a})$.*

Let S be the set in $Fact$ 12. We let the alternative hypothesis distribution family be defined as $\mathcal{D} = \{P_V^A | V\in S\}$ where A is the distribution in $Lemma$ 44. We give the following lemma for $\mathcal{D}$.

Lemma 46 *For any sufficiently small $\gamma > 0$, there exists a distribution family $\mathcal{D} = \{D_V : V \in S\}$ supported on $\mathbb{R}^d \times \{\pm 1\}$ where $|\mathcal{D}| = 2^{d^{\Omega(1)}}$ satisfying the following:*

1. *For any $D_V \in \mathcal{D}$, both $(D_V)^+$ and $(D_V)^-$ are factorizable in V and $V^\perp$ and both $(D_V)^+_{V^\perp}$ and $(D_V)^-_{V^\perp}$ are $\mathcal{N}(0, I_{d-2})$;*
2. *For any $D_V \in \mathcal{D}$, there exists a intersection of halfspaces that is consistent with D_V with γ margins. Furthermore, both halfspaces are in the subspace V;*
3. *For any $D_V \in \mathcal{D}$, $\mathbf{Pr}_{(\mathbf{x},y)\sim D_V}\left[\mathrm{proj}_{\perp V}(\mathbf{x}) \geq 2\sqrt{d}\right] = 2^{-d^{\Omega(1)}}$; and*
4. *Let $g_V : \mathbb{R}^d \to \mathbb{R}$ denote the function $g_V(\mathbf{x}) = (P_{\mathbf{x}\sim D_V^+}(\mathbf{x}) - P_{\mathbf{x}\sim D_V^-}(\mathbf{x}))/P_{\mathcal{N}_d}(\mathbf{x})$. Then for any $D_U, D_V \in \mathcal{D}$, $(g_U \cdot g_V)_{\mathcal{N}_d} = O(1/\gamma)$ if $U = V$ and $|(g_U \cdot g_V)_{\mathcal{N}_d}| = d^{-\Omega(\log 1/\gamma)}$ if $U \neq V$.*

Proof Let S be the set in $Fact$ 12. We let the alternative hypothesis distribution family be defined as $\mathcal{D} = \{P_V^A \mid V \in S\}$, where A is the distribution in Lemma 44.

Item 1 follows immediately from the definition of $\mathcal{D}$. Item 2 follows from Item 1 of Lemma 44. Item 3 follows from the fact that A is bounded inside $\mathbb{B}^2(1) \times \{\pm 1\}$ and the concentration of l_2 norm for Gaussian. Namely,

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D_v}\left[\mathbf{x} \geq 2\sqrt{d}\right] \leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(0,I_{d-2})}\left[\mathbf{x} \geq (2\sqrt{d}-1)\right] = \mathop{\mathbf{Pr}}_{t\sim\chi^2(d-2)}[t \geq 3d] \leq 2^{-\Omega(d)} ,$$

where $\chi^2(d-2)$ the chi-squared distribution with $d-2$ degrees of freedom.

For Item 4, if $U = V$, then we have

$$\begin{aligned}
(g_U \cdot g_V)_{\mathcal{N}_d} &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_d}[g_V(\mathbf{x})^2] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_d}\left[\left(\left(P_{\mathbf{x}\sim D_V^+}(\mathbf{x}) - P_{\mathbf{x}\sim D_V^-}(\mathbf{x})\right)/\mathcal{N}_d(\mathbf{x})\right)^2\right] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_d}\left[\left(P_{\mathbf{x}\sim D_V^+}(\mathbf{x})/\mathcal{N}_d(\mathbf{x})\right)^2\right] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_d}\left[\left(P_{\mathbf{x}\sim D_V^-}(\mathbf{x})/\mathcal{N}_d(\mathbf{x})\right)^2\right] \\
&\quad + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_d}\left[\left(P_{\mathbf{x}\sim D_V^+}(\mathbf{x})/\mathcal{N}_d(\mathbf{x})\right)\left(P_{\mathbf{x}\sim D_V^-}(\mathbf{x})/\mathcal{N}_d(\mathbf{x})\right)\right] \\
&\leq \chi^2(D_V^+, \mathcal{N}_d) + \chi^2(D_V^+, \mathcal{N}_d) + 2\sqrt{\chi^2(D_V^+, \mathcal{N}_d)\chi^2(D_V^+, \mathcal{N}_d)} = O(1/\gamma) .
\end{aligned}$$

For the case $U \neq V$, we will need the following fact.

Fact 13 (Correlation Lemma: Lemma 2.3 from Diakonikolas et al. (2021)) *Let $g : \mathbb{R}^m \mapsto \mathbb{R}$ and $U, V \in \mathbb{R}^{m\times d}$ with $m \leq d$ be linear maps such that $UU^\intercal = VV^\intercal = I$ where I is the $m \times m$ identity matrix. Then, we have that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_d}[g(U\mathbf{x})g(V\mathbf{x})] \leq \sum_{t=0}^{\infty} \|UV^\intercal\|_2^t \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[(g^{[t]}(\mathbf{x}))^2] ,$$

where $g^{[t]}$ denote the degree-t Hermite part of g.

Using the above fact and Item 2, we have

$$(g_U \cdot g_V)_{\mathcal{N}_d} = d^{-\Omega(c\log(1/\gamma))} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_d}[g_V(\mathbf{x})^2] = d^{-\Omega(\log(1/\gamma))}O(1/\gamma) = d^{-\Omega(\log(1/\gamma))} .$$

This completes the proof. ■

Given Lemma 46, we are now ready to prove our main theorem Theorem 43.

Proof [Poof for Theorem 43] Without loss of generality, we will assume that $\min(d, 1/\gamma^2) = d$. Since if this is not the case, i.e., $d > 1/\gamma^2$, we can always give a lower bound for $d' = 1/\gamma^2$ and the lower bound immediately applies to $d > d'$ by simply adding dummy coordinates that is always 0 on the hard instance. Therefore, we just need to show a $d^{\Omega(\log(1/\gamma))}$ lower bound given $d \le 1/\gamma^2$.

Let $\mathcal{D}'$ be the alternative hypothesis distribution set in Lemma 46 with the margin parameter in Lemma 46 taken as $\alpha = 2\gamma\sqrt{d}$. We defined null hypothesis distribution D' as the joint distribution of $(\mathbf{x}, y)$ where $\mathbf{x} \sim \mathcal{N}_d$ and $y = 1$ independently with probability $1/2$. Now, consider the decision problem $\mathcal{B}(\mathcal{D}', D')$. From Lemma 46, we have $\mathrm{CD}(\mathcal{B}, d^{-\Omega(\log(1/\alpha))}, O(1/\alpha)) = 2^{d^{\Omega(1)}}$. Notice that from Lemma 24, by taking the parameter γ' in Lemma 24 as $\gamma' = d^{-c\log(1/\alpha)}$ for some constant c, we get that any CSQ algorithm for solving $\mathcal{B}(\mathcal{D}', D')$ either requires a query of tolerance $d^{-\Omega(\log(1/\alpha))}$ or $2^{d^{\Omega(1)}}$ many queries.

The problem here is that $\mathcal{B}(\mathcal{D}', D'_0)$ is supported on $\mathbb{R}^d \times \{\pm 1\}$ instead of $\mathbb{B}^d(1) \times \{\pm 1\}$. To fix this problem, we first truncate the distributions. We now define D'' as the distribution of $(\mathbf{x}, y)$ where $(\mathbf{x}, y) \sim D' \mid (\|\mathbf{x}\|_2 \le 2\sqrt{d})$. Similarly, we defined $\mathcal{D}''$ as the family of distribution, where for each distribution $D''_V \in \mathcal{D}''$, we take a $D'_V \in \mathcal{D}'$ and defined D''_V as the distribution of $(\mathbf{x}, y)$ where $(\mathbf{x}, y) \sim D'_V \mid (\|\mathrm{proj}_{\perp V}\mathbf{x}\|_2 \le 2\sqrt{d})$. Notice that from the definition of $\mathcal{D}'$ and Property 3 of Lemma 46, the total variation distance between the truncated distribution and the untruncated distribution is bounded by $2^{-d^{\Omega(1)}}$. Therefore, given the CSQ lower bound on the untruncated $\mathbb{B}(\mathcal{D}', D')$, we have that any CSQ algorithm for solving the truncated $\mathbb{B}(\mathcal{D}'', D'')$ will either require a query of tolerance $d^{-\Omega(\log(1/\alpha))} + 2^{-d^{\Omega(1)}} = \max(d^{-\Omega(\log(1/\alpha))}, 2^{-d^{\Omega(1)}}) = \max(d^{-\Omega(\log(1/\gamma))}, 2^{-d^{\Omega(1)}})$ or $2^{d^{\Omega(1)}}$ many queries, which follows from the definition of the CSQ oracle.

Now, since everything is bounded inside a radius $3\sqrt{d}$ ball, we just need to rescale it so everything is inside a unit ball. From the definition of the CSQ oracle, this does not change the CSQ lower bound. Defined D be the distribution of $(\mathbf{x}/(3\sqrt{d}), y)$ where $\mathbf{x}, y \sim D''$ and $\mathcal{D}$ such that any $D_V \in \mathcal{D}$ is defined as $(\mathbf{x}/(3\sqrt{d}), y)$ where $\mathbf{x}, y \sim D''_V$ for some $D''_V \in \mathcal{D}''$. It is immediate that any algorithm that solves $\mathcal{B}(\mathcal{D}, D)$ requires either a query of tolerance $\max(d^{-\Omega(\log(1/\gamma))}, 2^{-d^{\Omega(1)}})$ or $2^{d^{\Omega(1)}}$ many queries.

Furthermore, notice that any $D_V \in \mathcal{D}$ is an instance of learning intersections of two halfspaces with γ-margin under factorizable distributions. Therefore, any algorithm for learning intersections of two halfspaces with γ-margin under factorizable distributions will output a hypothesis with ϵ error if given such $D_i \in \mathcal{D}$. While given the null hypothesis distribution D, no learning algorithm can learn any hypothesis with an error nontrivially better than $1/2$. Therefore, given that such a learning algorithm can solve $\mathcal{B}(\mathcal{D}, D)$, it must require either a query of tolerance $\max(d^{-\Omega(\log(1/\gamma))}, 2^{-d^{\Omega(1)}})$ or $2^{d^{\Omega(1)}}$ many queries. This completes the proof. ■