AN EFFICIENT TESTER-LEARNER FOR HALFSPACES

Aravind Gollakota[*] Apple
Adam R. Klivans[†] UT Austin
Konstantinos Stavropoulos[‡] UT Austin
Arsen Vasilyan[§] MIT

ABSTRACT

We give the first efficient algorithm for learning halfspaces in the testable learning model recently defined by Rubinfeld and Vasilyan (RV23). In this model, a learner certifies that the accuracy of its output hypothesis is near optimal whenever the training set passes an associated test, and training sets drawn from some target distribution must pass the test. This model is more challenging than distribution-specific agnostic or Massart noise models where the learner is allowed to fail arbitrarily if the distributional assumption does not hold. We consider the setting where the target distribution is the standard Gaussian in d dimensions and the label noise is either Massart or adversarial (agnostic). For Massart noise, our tester-learner runs in polynomial time and outputs a hypothesis with (information-theoretically optimal) error $\mathsf{opt}+\epsilon$ (and extends to any fixed strongly log-concave target distribution). For adversarial noise, our tester-learner obtains error $O(\mathsf{opt}) + \epsilon$ in polynomial time. Prior work on testable learning ignores the labels in the training set and checks that the empirical moments of the covariates are close to the moments of the base distribution. Here we develop new tests of independent interest that make critical use of the labels and combine them with the moment-matching approach of (GKK23). This enables us to implement a testable variant of the algorithm of (DKTZ20a; DKTZ20b) for learning noisy halfspaces using nonconvex SGD.

1 INTRODUCTION

Learning halfspaces in the presence of noise is one of the most basic and well-studied problems in computational learning theory. A large body of work has obtained results for this problem under a variety of different noise models and distributional assumptions (see e.g. (BH21) for a survey). A major issue with common distributional assumptions such as Gaussianity, however, is that they can be hard or impossible to verify in the absence of any prior information.

The recently defined model of testable learning (RV23) addresses this issue by replacing such assumptions with efficiently testable ones. In this model, the learner is required to work with an arbitrary input distribution $D_{\mathcal{XY}}$ and verify any assumptions it needs to succeed. It may choose to reject a given training set, but if it accepts, it is required to output a hypothesis with error close to $\mathsf{opt}(\mathcal{C}, D_{\mathcal{XY}})$, the optimal error achievable over $D_{\mathcal{XY}}$ by any function in a concept class $\mathcal{C}$. Further, whenever the training set is drawn from a distribution $D_{\mathcal{XY}}$ whose marginal is truly a well-behaved target distribution D^* (such as the standard Gaussian), the algorithm is required to accept with high probability. Such an algorithm, or tester-learner, is then said to testably learn $\mathcal{C}$ with respect to target marginal D^*. (See Definition 2.1.) Note that unlike ordinary distribution-specific agnostic learners, a tester-learner must take some nontrivial action *regardless* of the input distribution.

The work of (RV23; GKK23) established foundational algorithmic and statistical results for this model and showed that testable learning is in general provably harder than ordinary distribution-specific agnostic learning. As one of their main algorithmic results, they showed tester-learners for the class of halfspaces over $\mathbb{R}^d$ that succeed whenever the target marginal is Gaussian (or one of a more general class of distributions), achieving error $\mathsf{opt} + \epsilon$ in time and sample complexity

[*]aravindg@cs.utexas.edu
[†]klivans@cs.utexas.edu
[‡]kstavrop@cs.utexas.edu
[§]vasilyan@mit.edu

$d^{\Theta(1/\epsilon^2)}$. This matches the running time of ordinary distribution-specific agnostic learning of halfspaces over the Gaussian using the standard approach of (KKMS08). Their testers are simple and label-oblivious, and are based on checking whether the low-degree empirical moments of the unknown marginal match those of the target D^*.

These works essentially resolve the question of designing tester-learners achieving error $\mathsf{opt} + \epsilon$ for halfspaces, matching known hardness results for (ordinary) agnostic learning (GGK20; DKZ20; DKPZ21). Their running time, however, necessarily scales exponentially in $1/\epsilon$.

A long line of research has sought to obtain more efficient algorithms at the cost of relaxing the optimality guarantee (ABL17; DKS18; DKTZ20a; DKTZ20b). These works give polynomial-time algorithms achieving bounds of the form $\mathsf{opt}+\epsilon$ and $O(\mathsf{opt})+\epsilon$ for the Massart and agnostic setting respectively under structured distributions (see Section 1.1 for more discussion). The main question we consider here is whether such guarantees can be obtained in the testable learning framework.

Our contributions. In this work we design the first tester-learners for halfspaces that run in fully polynomial time in all parameters. We match the optimality guarantees of fully polynomial-time learning algorithms under Gaussian marginals for the Massart noise model (where the labels arise from a halfspace but are flipped by an adversary with probability at most η) as well as for the agnostic model (where the labels can be completely arbitrary). In fact, for the Massart setting our guarantee holds with respect to any chosen target marginal D^* that is isotropic and strongly log-concave, and the same is true of the agnostic setting albeit with a slightly weaker guarantee.

Theorem 1.1 (Formally stated as Theorem 4.1)**.** *Let $\mathcal{C}$ be the class of origin-centered halfspaces over $\mathbb{R}^d$, and let D^* be any isotropic strongly log-concave distribution. In the setting where the labels are corrupted with Massart noise at rate at most $\eta < \frac{1}{2}$, $\mathcal{C}$ can be testably learned w.r.t. D^* up to error $\mathsf{opt} + \epsilon$ using $\mathrm{poly}(d, \frac{1}{\epsilon}, \frac{1}{1-2\eta})$ time and sample complexity.*

Theorem 1.2 (Formally stated as Theorem 5.1)**.** *Let $\mathcal{C}$ be as above. In the adversarial noise or agnostic setting where the labels are completely arbitrary, $\mathcal{C}$ can be testably learned w.r.t. $\mathcal{N}(0, I_d)$ up to error $O(\mathsf{opt}) + \epsilon$ using $\mathrm{poly}(d, \frac{1}{\epsilon})$ time and sample complexity.*

Our techniques. The tester-learners we develop are significantly more involved than prior work on testable learning. We build on the nonconvex optimization approach to learning noisy halfspaces due to (DKTZ20a; DKTZ20b) as well as the structural results on fooling functions of halfspaces using moment matching due to (GKK23). Unlike the label-oblivious, global moment tests of (RV23; GKK23), our tests make crucial use of the labels and check *local* properties of the distribution in regions described by certain candidate vectors. These candidates are approximate stationary points of a natural nonconvex surrogate of the 0-1 loss, obtained by running gradient descent. When the distribution is known to be well-behaved, (DKTZ20a; DKTZ20b) showed that any such stationary point is in fact a good solution (for technical reasons we must use a slightly different surrogate loss). Their proof relies crucially on structural geometric properties that hold for these well-behaved distributions, an important one being that the probability mass of any region close to the origin is proportional to its geometric measure.

In the testable learning setting, we must efficiently check this property for candidate solutions. Since these regions may be described as intersections of halfspaces, we may hope to apply the moment-matching framework of (GKK23). Naïvely, however, they only allow us to check in polynomial time that the probability masses of such regions are within an additive constant of what they should be under the target marginal. But we can view these regions as sub-regions of a known band described by our candidate vector. By running moment tests on the distribution *conditioned* on this band and exploiting the full strength of the moment-matching framework, we are able to effectively convert our weak additive approximations to good multiplicative ones. This allows us to argue that our stationary points are indeed good solutions.

Independent and Subsequent Works. In this paper we provide the first efficient tester-learners for halfspaces when the noise is either adversarial or Massart. In independent and concurrent work by (DKK+23), an efficient tester-learner for homogeneous halfspaces achieving error $O(\mathsf{opt}) + \epsilon$ for Gaussian target marginals is also provided, but they do not provide any results for arbitrary strongly log-concave target marginals (see Theorem 5.1) or a guarantee of $\mathsf{opt} + \epsilon$ for Massart noise. In subsequent work by (GKSV23), our techniques were used to provide tester-learners that are not tailored to a single target distribution, but are guaranteed to accept any member of a large family of distributions. Although their main results are more general, their approach crucially extends our

approach here. Moreover, on the technical side, the proof we give here shows how to make use of the moment-matching approach of (GKK23) to provide fully polynomial-time ef cient tester-learners, which might be of independent interest.

1.1 RELATED WORK

We provide a partial summary of some of the most relevant prior and related work on ef cient algorithms for learning halfspaces in the presence of adversarial label or Massart noise, and refer the reader to (BH21) for a survey.

In the distribution-speci c agnostic setting where the marginal is assumed to be isotropic and log-concave, (KLS09) showed an algorithm achieving error $O(\mathrm{opt}^{1=3})+$ for the class of origin-centered halfspaces. (ABL17) later obtained $O(\mathrm{opt})+$ using an approach that introduced the principle of *iterative localization*, where the learner focuses attention on a band around a candidate halfspace in order to produce an improved candidate. (Dan15) used this principle to obtain a PTAS for agnostically learning halfspaces under the uniform distribution on the sphere, and (BZ17) extended it to more general s-concave distributions. Further works in this line include (YZ17; Zha18; ZSA20; ZL21). (DKTZ20b) introduced the simplest approach yet, based entirely on nonconvex SGD, and showed that it achieves $O(\mathrm{opt})+$ for origin-centered halfspaces over a wide class of structured distributions. Other related works include (DKS18; DKTZ22).

In the Massart noise setting with noise rate bounded by , work of (DGT19) gave the rst ef cient distribution-free algorithm achieving error $+$; further improvements and followups include (DKT21; DTK22). However, the optimal error opt achievable by a halfspace may be much smaller than , and it has been shown that there are distributions where achieving error competitive with opt as opposed to is computationally hard (DK22; DKMR22). As a result, the distribution-speci c setting remains well-motivated for Massart noise. Early distribution-speci c algorithms were given by (ABHU15; ABHZ16), but a key breakthrough was the nonconvex SGD approach introduced by (DKTZ20a), which achieved error $\mathrm{opt}+$ for origin-centered halfspaces ef ciently over a wide range of distributions. This was later generalized by (DK22).

1.2 TECHNICAL OVERVIEW

Our starting point is the nonconvex optimization approach to learning noisy halfspaces due to (DKTZ20a; DKTZ20b). The algorithms in these works consist of running SGD on a natural nonconvex surrogate L for the 0-1 loss, namely a smooth version of the ramp loss. The key structural property shown is that if the marginal distribution is structured (e.g. log-concave) and the slope of the ramp is picked appropriately, then any w that has large angle with an optimal w cannot be an approximate stationary point of the surrogate loss, i.e. that $\|\nabla L_{}(w)\|$ must be large. This is proven by carefully analyzing the contributions to the gradient norm from certain critical regions of $\mathrm{span}(w; w)$, and crucially using the distributional assumption that the probability masses of these regions are proportional to their geometric measures. (See Fig. 3.) In the testable learning setting, the main challenge we face in adapting this approach is checking such a property for the unknown distribution we have access to.

A preliminary observation is that the critical regions of $\mathrm{span}(w; w)$ that we need to analyze are rectangles, and are hence functions of a small number of halfspaces. Encouragingly, one of the key structural results of the prior work of (GKK23) pertains to "fooling" such functions. Concretely, they show that whenever the true marginal D_x matches moments of degree at most $\Theta(1= ^2)$ with a target D that satis es suitable concentration and anticoncentration properties, then $|\mathbb{E}_{D_x}[f] \quad \mathbb{E}_{D}[f]|$ for any f that is a function of a small number of halfspaces. If we could run such a test and ensure that the probabilities of the critical regions over our empirical marginal are also related to their areas, then we would have a similar stationary point property. However, the dif culty is that since we wish to run in fully polynomial time, we can only hope to fool such functions up to that is a constant. Unfortunately, this is not suf cient to analyze the probability masses of the critical regions we care about as they may be very small.

The chief insight that lets us get around this issue is that each critical region in fact of a very speci c form, namely a rectangle that is axis-aligned with $R = \{x : hw; xi 2 [;]\}$ and $hv; xi 2 [;]\}$ for some values; ; and some v orthogonal to w. Moreover, we know w, meaning

we can ef ciently estimate the probability $P_{D_X}[\langle w, x\rangle \in [\ ;\ ]]$ up to constant multiplicative factors without needing moment tests. Denoting the band $\{x : \langle w, x\rangle \in [\ ;\ ]\}$ by T and writing $P_{D_X}[R] = P_{D_X}[\langle v, x\rangle \in [\ ;\ ] \mid x \in T]\, P_{D_X}[T]$, it turns out that we should expect $P_{D_X}[\langle v, x\rangle \in [\ ;\ ] \mid x \in T] = (1)$, as this is what would occur under the structured target distribution D . (Such a "localization" property is also at the heart of the algorithms for approximately learning halfspaces of, e.g., (ABL17; Dan15).) To check this, it suf ces to run tests that ensure that $P_{D_X}[\langle v, x\rangle \in [\ ;\ ] \mid x \in T]$ is within an additive constant of this probability under D .

We can now describe the core of our algorithm (omitting some details such as the selection of the slope of the ramp). First, we run SGD on the surrogate Lo ss to arrive at an approximate stationary point and candidate vector w (technically a list of such candidates). Then, we de ne the band T based on w, and run tests on the empirical distribution conditioned on T. Speci cally, we check that the low-degree empirical moments conditioned on T match those of D conditioned on T, and then apply the structural result of (GKK23) to ensure conditional probabilities of the form $P_{D_X}[\langle v, x\rangle \in [\ ;\ ] \mid x \in T]$ match $P_D[\langle v, x\rangle \in [\ ;\ ] \mid x \in T]$ up to a suitable additive constant. This suf ces to ensure that even over our empirical marginal, the particular stationary point we have is indeed close in angular distance to an optimal w.

A nal hurdle that remains, often taken for granted under structured distributions, is that closeness in angular distance $\theta(w, w^*)$ does not immediately translate to closeness in terms of agreement, $P[\mathrm{sign}(\langle w, x\rangle) \neq \mathrm{sign}(\langle w^*, x\rangle)]$, over our unknown marginal. Nevertheless, we show that when the target distribution is Gaussian, we can run polynomial-time tests that ensure that an angle of $= \theta(w, w^*)$ translates to disagreement of at most $O(t)$. When the target distribution is a general strongly log-concave distribution, we show a slightly weaker relationship: for any $k \geq 2$, we can run tests requiring time $d^{\Theta(k)}$ that ensure that an angle of translates to disagreement of at most $O(\sqrt[k]{}^{\,1-1/k})$. In the Massart noise setting, we can make $\theta(w, w^*)$ arbitrarily small, and so obtain our $\mathrm{opt} +$ guarantee for any target strongly log-concave distribution in polynomial time. In the adversarial noise setting, we face a more delicate tradeoff and can only make $\theta(w, w^*)$ as small as (opt). When the target distribution is Gaussian, this is enough to obtain nal error $O(\mathrm{opt}) +$ in polynomial time. When the target distribution is a general strongly log-concave distribution, we instead obtain $\Theta(\mathrm{opt}) +$ in quasipolynomial time.

2 Preliminaries

Notation and setup Throughout, the domain will be $X = \mathbb{R}^d$, and labels will lie in $Y = \{\pm 1\}$. The unknown joint distribution over $X \times Y$ that we have access to will be denoted by D_{XY}, and its marginal on X will be denoted by D_X. The target marginal on X will be denoted by D . We use the following convention for monomials: for a multi-index $= (\ _1, \ldots, \ _d) \in \mathbb{Z}^d_{\geq 0}$, x denotes $\prod_i x_i^{\ i}$, and $| \ | = \sum_i \ _i$ denotes its total degree. We use C to denote a concept class mapping $\mathbb{R}^d$ to $\{\pm 1\}$, which throughout this paper will be the class of halfspaces or functions of halfspaces over $\mathbb{R}^d$. We use $\mathrm{opt}(C, D_{XY})$ to denote the optimal error $\inf_{f \in C} P_{(x,y) \sim D_{XY}}[f(x) \neq y]$, or just opt when C and D_{XY} are clear from context. We recall the de nitions of the noise models we consider. In the Massart noise model, the labels satisfy $P_{y \sim D_{XY}|x}[y \neq \mathrm{sign}(\langle w^*, x\rangle) \mid x] = (x)$, where $(x) < \frac{1}{2}$ for all x. In the adversarial label noise or agnostic model, the labels may be completely arbitrary. In both cases, the learner's goal is to produce a hypothesis with error competitive with opt. We now formally de ne testable learning. The following de nition is an equivalent reframing of the original de nition (RV23, Def 4), folding the (label-aware) tester and learner into a single tester-learner.

De nition 2.1 (Testable learning, (RV23)). Let C be a concept class mapping $\mathbb{R}^d$ to $\{\pm 1\}$. Let D be a certain target marginal on $\mathbb{R}^d$. Let $\ ; \ > 0$ be parameters, and let $: [0, 1] \to [0, 1]$ be some function. We say C can be testably learned w.r.t. D up to error $(\mathrm{opt}) +$ with failure probability if there exists a tester-learner A meeting the following speci cation. For any distribution D_{XY} on $\mathbb{R}^d \times \{\pm 1\}$, A takes in a large sample S drawn from D_{XY}, and either rejects S or accepts and produces a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$. Further, the following conditions must be met:

(a) (Soundness.) Whenever A accepts and produces a hypothesis h, with probability at least 1 (over the randomness of S and A), h must satisfy $P_{(x,y) \sim D_{XY}}[h(x) \neq y] \leq (\mathrm{opt}(C, D_{XY})) +$.

(b) (Completeness.) Whenever D_{XY} truly has marginal D, A must accept with probability at least $1 - \delta$ (over the randomness of S and A).

3 TESTING PROPERTIES OF STRONGLY LOG-CONCAVE DISTRIBUTIONS

In this section we define the testers that we will need for our algorithm. All the proofs from this section can be found in Appendix B. We begin with a structural lemma that strengthens the key structural result of (GKK23), stated here as Proposition A.3. It states that even when we restrict an isotropic strongly log-concave D to a band around the origin, moment matching suffices to fool functions of halfspaces whose weights are orthogonal to the normal of the band.

Proposition 3.1. Let D be an isotropic strongly log-concave distribution. Let $v \in S^{d-1}$ be any fixed direction. Let p be a constant. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function of p halfspaces of the form in Eq. (A.2), with the additional restriction that its weight vectors $w^i \in S^{d-1}$ satisfy $\langle v^i, w\rangle = 0$ for all i. For some $\tau \in [0, 1]$, let T denote the band $\{x : |\langle w, x\rangle| \le \tau\}$. Let D be any distribution such that $D_{|T}$ matches moments of degree at most $\Theta(1/\tau^2)$ with $D_{|T}$ up to an additive slack of $\tau^{\Theta(k)}$. Then $|\mathbb{E}_D[f \mid T] - \mathbb{E}_D[f \mid T]| \le \varepsilon$.

We now describe some of the testers that we use. First, we need a tester that ensures that the distribution is concentrated in every single direction. More formally, the tester checks that the moments of the distribution along any direction are small.

Proposition 3.2. For any isotropic strongly log-concave D, there exists some constant C_1 and a tester T_1 that takes a set $S \subseteq \mathbb{R}^d \times \{-1, 1\}$, an even $k \in \mathbb{N}$, a parameter $\varepsilon \in (0, 1)$ and runs and in time $\mathrm{poly}(d^k, |S|, \log\frac{1}{\delta})$. Let D denote the uniform distribution over S. If T_1 accepts, then for any $v \in S^{d-1}$

$$\mathbb{E}_{(x,y)\sim D}[(\langle v, x\rangle)^k] \le (C_1 k)^{k/2}. \tag{3.1}$$

Moreover, if S is obtained by taking at least $d^k \cdot (\log\frac{1}{\delta})^{C_1 k}$ i.i.d. samples from a distribution whose $\mathbb{R}^d$-marginal is D, the test T_1 passes with probability at least $1 - \delta$.

Secondly, we will use a tester that makes sure the distribution is not concentrated too close to a specific hyperplane. This is one of the properties we will need to use in order to employ the localization technique of (ABL17).

Proposition 3.3. For any isotropic strongly log-concave D, there exist some constants C_2, C_3 and a tester T_2 that takes a set $S \subseteq \mathbb{R}^d \times \{-1, 1\}$ a vector $w \in S^{d-1}$, parameters $\tau, \varepsilon \in (0, 1)$ and runs in time $\mathrm{poly}(d, |S|, \log\frac{1}{\delta})$. Let D denote the uniform distribution over S. If T_2 accepts, then

$$\mathbb{P}_{(x,y)\sim D}[|\langle w, x\rangle| \le \tau] \in (C_2\tau, C_3\tau). \tag{3.2}$$

Moreover, if S is obtained by taking at least $\frac{100}{K_1\tau^2}\log\frac{1}{\delta}$ i.i.d. samples from a distribution whose $\mathbb{R}^d$-marginal is D, the test T_2 passes with probability at least $1 - \delta$.

Finally, in order to use the localization idea of (ABL17) in a manner similar to (DKTZ20b), we need to make sure that the distribution is well-behaved also within a band around to a certain hyperplane. The main property of the distribution that we establish is that functions of constantly many halfspaces have expectations very close to what they would be under our distributional assumption. As we show later in this work, having the aforementioned property allows us to derive many other properties that strongly log-concave distributions have, including many of the key properties that make the localization technique successful.

Proposition 3.4. For any isotropic strongly log-concave D and a constant C_4, there exists a constant C_5 and a tester T_3 that takes a set $S \subseteq \mathbb{R}^d \times \{-1, 1\}$ a vector $w \in S^{d-1}$, parameters $\tau, \varepsilon \in (0, 1)$ and runs in time $\mathrm{poly}(d^{O(\frac{1}{\varepsilon^2})}, \frac{1}{\tau}, |S|, \log\frac{1}{\delta})$. Let D denote the uniform distribution over S, let T denote the band $\{x : |\langle w, x\rangle| \le \tau\}$ and let $\mathcal{F}_w$ denote the set of $\{-1, 1\}$-valued functions of C_4 halfspaces whose weight vectors are orthogonal to w. If T_3 accepts, then

$$\max_{f\in\mathcal{F}_w} \left|\mathbb{E}_{x\sim D}[f(x) \mid x \in T] - \mathbb{E}_{(x,y)\sim D}[f(x) \mid x \in T]\right| \le \varepsilon; \tag{3.3}$$

$$\max_{v \in S^{d-1}: \langle w, v \rangle = 0} \mathbb{E}_{x \sim D}[(\langle v, x \rangle)^2 \mid x \in T] \geq \mathbb{E}_{(x,y) \sim D}[(\langle v, x \rangle)^2 \mid x \in T] \quad . \tag{3.4}$$

Moreover, if S is obtained by taking at least $\frac{1}{\epsilon}\left(\frac{1}{\alpha}\right)^{d^{\frac{1}{2}\log^{C_5}(\frac{1}{\alpha})}} \log\left(\frac{1}{\delta}\right)^{C_5}$ i.i.d. samples from a distribution whose S^{d-1}-marginal is D, the test T_3 passes w.p. at least $1 - \delta$.

4 Testably learning halfspaces with Massart noise

In this section we prove that we can testably learn halfspaces with Massart noise with respect to isotropic strongly log-concave distributions (see Definition A.1).

Theorem 4.1 (Tester-Learner for Halfspaces with Massart Noise). Let D_{XY} be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ and let D be an isotropic strongly log-concave distribution over $\mathbb{R}^d$. Let C be the class of origin centered halfspaces in $\mathbb{R}^d$. Then, for any $\eta < 1/2$, $\epsilon > 0$ and $\delta \in (0, 1)$, there exists an algorithm (Algorithm 1) that testably learns C w.r.t. D up to excess error ϵ and error probability at most δ in the Massart noise model with rate at most η, using time and a number of samples from D_{XY} that are polynomial in $d, 1/\epsilon, \frac{1}{1-2\eta}$ and $\log(1/\delta)$.

Algorithm 1: Tester-learner for halfspaces

Input: Training sets S_1, S_2, parameters ϵ, δ
Output: A near-optimal weight vector w, or rejection
Run PSGD on the empirical loss L_ϵ over S_1 to get a list L of candidate vectors.
Test whether L contains an ϵ-approximate stationary point w' of the empirical loss L_ϵ over S_2.
  Reject if no such w' exists.
for each candidate w^0 in $\{w', -w'\}$ do
  Let $B_{w^0}(\cdot)$ denote the band $\{x : |\langle w^0, x \rangle| \leq \cdot\}$. Let F_{w^0} denote the class of functions of at most two halfspaces with weights orthogonal to w^0
  Let $\gamma^0 = \Theta(\epsilon)$.
  Run $T_1(S_2, k = 2, \cdot)$ to verify that the empirical marginal is approximately isotropic.
    Reject if T_1 rejects.
  Run $T_2(S_2, w^0, \cdot, \gamma^0)$ to verify that $P_S[B_{w^0}(\cdot)] = \Theta(\epsilon)$. Reject if T_2 rejects.
  Run $T_3(S_2, w^0, \cdot = \cdot = 6, \cdot, \gamma^0)$ and $T_3(S, w^0, \cdot = \cdot = 2, \cdot, \gamma^0)$ for a suitable constant to verify that the empirical distribution conditioned on $B_{w^0}(\cdot = 6)$ and $B_{w^0}(\cdot = 2)$ fools F_{w^0} up to ϵ. Reject if T_3 rejects.
  Estimate the empirical error of w^0 on S.
If all tests have accepted, output $w^0 \in \{w', -w'\}$ with the best empirical error.

To show our result, we revisit the approach of (DKTZ20a) for learning halfspaces with Massart noise under well-behaved distributions. Their result is based on the idea of minimizing a surrogate loss that is non convex, but whose stationary points correspond to halfspaces with low error. They also require that their surrogate loss is sufficiently smooth, so that one can find a stationary point efficiently. While the distributional assumptions that are used to demonstrate that stationary points of the surrogate loss can be discovered efficiently are mild, the main technical lemma, which demostrates that any stationary point suffices, requires assumptions that are not necessarily testable. We establish a label-dependent approach for testing, making use of tests that are applied during the course of our algorithm.

We consider a slightly different surrogate loss than the one used in (DKTZ20a). In particular, for $\gamma > 0$, we let

$$L_\gamma(w) = \mathbb{E}_{(x,y) \sim D_{XY}}\left[\ell\left(-y \frac{\langle w, x \rangle}{\|w\|_2}\right)\right], \tag{4.1}$$

where $\ell : \mathbb{R} \to [0, 1]$ is a smooth approximation to the ramp function with the properties described in Proposition C.1 (see Appendix C), obtained using a piecewise polynomial of degree 3. Like the standard logistic function, our loss function has derivative exactly 0 away from the origin (for $|t| > \gamma/2$). This makes the analysis of the gradient of L_γ easier, since the contribution from points lying outside a certain band is exactly 0.

The smoothness allows us to run PSGD to obtain stationary points efficiently, and we now state the convergence lemma we need.

Proposition 4.2 (PSGD Convergence, Lemmas 4.2 and B.2 in (DKTZ20b)). Let L be as in Equation equation 4.1 with $\in (0,1]$, ℓ as described in Proposition C.1 and D_{XY} such that the marginal D_X on $\mathbb{R}^d$ satisfies Property equation 3.1 for $k = 2$. Then, for any > 0 and $\in (0,1)$, there is an algorithm whose time and sample complexity is $O(\frac{d}{4} + \frac{\log(1/)}{4\ 4})$, which, having access to samples from D_{XY}, outputs a list L of vectors $w \in S^{d-1}$ with $|L| = O(\frac{d}{4} + \frac{\log(1/)}{4\ 4})$ so that there exists $w \in L$ with

$$\|r_w L_{(w)}\|_2 \quad ; \text{ with probability at least } 1 \quad .$$

In particular, the algorithm performs Stochastic Gradient Descent on L on Projected on S^{d-1} (PSGD).

It now suffices to show that, upon performing PSGD on L, for some appropriate choice of , we acquire a list of vectors that testably contain a vector which is approximately optimal. We first prove the following lemma, whose distributional assumptions are relaxed compared to the corresponding structural Lemma 3.2 of (DKTZ20a). In particular, instead of requiring the marginal distribution to be "well-behaved", we assume that the quantities of interest (for the purposes of our proof) have expected values under the true marginal distribution that are close, up to multiplicative factors, to their expected values under some "well-behaved" (in fact, strongly log-concave) distribution. While some of the quantities of interest have values that are miniscule and estimating them up to multiplicative factors could be too costly, it turns out that the source of their vanishing scaling can be completely attributed to factors of the form $\mathbb{P}[|\langle w, x\rangle| \ ]$ (where is small), which, due to standard concentration arguments, can be approximated up to multiplicative factors, given $w \in S^{d-1}$ and > 0 (see Proposition 3.3). As a result, we may estimate the remaining factors up to sufficiently small additive constants (see Proposition 3.4) to get multiplicative overall closeness to the "well behaved" baseline. We defer the proof of the following Lemma to Appendix C.1.

Lemma 4.3. Let L be as in Equation equation 4.1 with $\in (0,1]$, ℓ as described in Proposition C.1, let $w \in S^{d-1}$ and consider D_{XY} such that the marginal D_X on $\mathbb{R}^d$ satisfies Properties equation 3.2 and equation 3.3 for $C_4 = 2$ and accuracy . Let $w \in S^{d-1}$ define an optimum halfspace and let $< 1/2$ be an upper bound on the rate of the Massart noise. Then, there are constants $c_1, c_2, c_3 > 0$ such that if $\|r_w L_{(w)}\|_2 < c_1(1 \ 2)$ and c_2, then

$$\theta(w, w^*) \quad \frac{c_3}{1 \ 2} \qquad \text{or} \qquad \theta(-w, w^*) \quad \frac{c_3}{1 \ 2}$$

Combining Proposition 4.2 and Lemma 4.3, we get that for any choice of the parameter $\in (0,1]$, by running PSGD on L, we can construct a list of vectors of polynomial size (in all relevant parameters) that testably contains a vector that is close to the optimum weight vector. In order to link the zero-one loss to the angular similarity between a weight vector and the optimum vector, we use the following Proposition (for the proof, see Appendix C.2).

Proposition 4.4. Let D_{XY} be a distribution over $\mathbb{R}^d \times \{-1,1\}$, $w \in \arg\min_{w \in S^{d-1}} \mathbb{P}_{D_{XY}}[y \neq \text{sign}(\langle w, x\rangle)]$ and $w \in S^{d-1}$. Then, for any $\theta(w, w^*)$, $\in [0, /4]$, if the marginal D_X on $\mathbb{R}^d$ satisfies Property equation 3.1 for $C_1 > 0$ and some even $k \in \mathbb{N}$ and Property equation 3.2 with set to $(C_1 k)^{\frac{k}{2(k+1)}} (\tan)^{\frac{k}{k+1}}$, then, there exists a constant $c > 0$ such that the following is true.

$$\mathbb{P}_{D_{XY}}[y \neq \text{sign}(\langle w, x\rangle)] \quad \text{opt} + c\ k^{1/2} \quad {}^{1 \frac{1}{k+1}} .$$

Proof of Theorem 4.1. Throughout the proof we consider ^{0} to be a sufficiently small polynomial in all the relevant parameters. Each of the failure events will have probability at least ^{0} and their number will be polynomial in all the relevant parameters, so by the union bound, we may pick ^{0} such that the probability of failure is at most .

The algorithm we run is Algorithm 1, with appropriate selection of parameters and given samples S_1, S_2, each of which are sufficiently large sets of independent samples from the true unknown distribution D_{XY}. For some $\in (0,1]$ to be defined later, we run PSGD on the empirical loss over S_1 as described in Proposition 4.2 with $= c_1(1 \ 2)/4$, where c_1 is given by Lemma 4.3.

Figure 1: Critical regions in the proofs of main structural lemmas (Lemmas 4.3, 5.2). We analyze the contributions of the regions labeled A_1, A_2 to the quantities A_1, A_2 in the proofs. Specifically, the regions A_1 (which have height $= 3$ so that the value of $r'(x_w)$ for any x in these regions is exactly $1/$, by Proposition C.1) form a subset of the region G, and their probability mass under D_X is (up to a multiplicative factor) a lower bound on the quantity A_1 (see Eq equation C.3). Similarly, the region A_2 is a subset of the intersection of G with the band of height , and has probability mass that is (up to a multiplicative factor) an upper bound on the quantity A_2 (see Eq equation C.4).

By Proposition 4.2, we get a list L of vectors $w \in S^{d-1}$ with $|L| = \mathrm{poly}(d, 1/)$ such that there is $w \in L$ with $\|\nabla_w L_{}(w)\|_2 < \frac{1}{2}c_1(1-2)$ under the true distribution, if the marginal is isotropic.

Having acquired the list L using samples S_1, we use the independent samples S_2 to test whether L contains an approximately stationary point of the empirical loss on S_2. If this is not the case, then we may safely reject: for large enough $|S_1|$, if the distribution is indeed isotropic strongly logconcave, there is an approximate stationary of the population loss and if $|S_2|$ is large enough, the gradient of the empirical loss on S_2 will be close to the gradient of the population loss on each of the elements of L, due to appropriate concentration bounds for log-concave distributions as well as the fact that the elements of L are independent from S_2. For the following, let w be a point such that $\|\nabla_w L_{}(w)\|_2 < c_1(1-2)$ under the empirical distribution over S_2.

In Lemma 4.3 and Proposition 4.4 we have identified certain properties of the marginal distribution that are sufficient for our purposes, given that L contains an approximately stationary point of the empirical (surrogate) loss on S_2. Our testers T_1, T_2, T_3 verify that these properties hold for the empirical marginal over our sample S_2, and it will be convenient to analyze the optimality of our algorithm purely over S_2. In particular, we will need to require that $|S_2|$ is sufficiently large, so that when the true marginal is indeed the target, our testers succeed with high probability (for the corresponding sample complexity, see Propositions 3.2, 3.3 and 3.4). Moreover, by standard generalization theory, since the VC dimension of halfspaces is $O(d)$ and for us $|S_2|$ is a large $\mathrm{poly}(d, 1/)$, both the error of our final output and the optimal error over S_2 will be close to that over D_{XY}. So in what follows, we will abuse notation and refer to the uniform distribution over S_2 as D_{XY} and the optimal error over S_2 simply as opt.

We proceed with some basic tests. Throughout the rest of the algorithm, whenever a tester fails, we reject, otherwise we proceed. First, we run testers T_2s with inputs $(w, = 2, {}^0)$ and $(w, = 6, {}^0)$ (Proposition 3.3) and T_3 with inputs $(w, = 2, c_2, {}^0)$ and with $(w, = 6, c_2, {}^0)$ (Proposition 3.4, c_2 as defined in Lemma 4.3). This ensures that for the approximate stationary point of the L, the probability within the band $B_w(= 2) = \{x : |\langle w, x\rangle| = 2\}$ is $(\ )$ (and similarly for $B_w(= 6)$) and moreover that our marginal conditioned on each of the bands fools (up to an additive constant) functions of halfspaces with weights orthogonal to w. As a result, we may apply Lemma 4.3 to w and form a list of 2 vectors $\{w, w\}$ which contains some w^0 with $\angle(w^0, w) \quad c_2 = (1-2)$ (where c_3 is as defined in Lemma 4.3).

We run T_1 (Proposition 3.2) with $k = 2$ to verify that the marginals are approximately isotropic and we use T_2 once again, with appropriate parameters for each w and its negation, to apply Proposition 4.4 and get that $\{w, -w\}$ contains a vector w' with

$$P_{D_{XY}}[y \neq \mathrm{sign}(\langle w', x\rangle)] \leq \mathrm{opt} + c\,\epsilon^{2/3}, \text{ where } \cos(w', w) := c_2 = \sqrt{1 - \epsilon^2}$$

By picking $\epsilon = (\epsilon^{3/2}(1 - \epsilon^2))$ we have $P_{D_{XY}}[y \neq \mathrm{sign}(\langle w', x\rangle)] \leq \mathrm{opt} + \epsilon$.

However, we do not know which of the weight vectors in $\{w, -w\}$ is the one guaranteed to achieve small error. In order to discover this vector, we estimate the probability of error of each of the corresponding halfspaces (which can be done efficiently, due to Hoeffding's bound) and pick the one with the smallest error. This final step does not require any distributional assumptions and we do not need to perform any further tests. □

5 Testably learning halfspaces in the agnostic setting

In this section, we provide our result on efficiently and testably learning halfspaces in the agnostic setting with respect to isotropic strongly log-concave target marginals. We defer the proofs to Appendix D. The algorithm we use is once more Algorithm 1, but we call it multiple times for different choices of the parameter, reject if any call rejects and output the vector that achieved the minimum empirical error overall, otherwise. Also, the tester T_1 is called for a general k (not necessarily $k = 2$).

Theorem 5.1 (Efficient Tester-Learner for Halfspaces in the Agnostic Setting). Let D_{XY} be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ and let D be a strongly log-concave distribution over $\mathbb{R}^d$ (Definition A.1). Let C be the class of origin centered halfspaces in $\mathbb{R}^d$. Then, for any even $k \in \mathbb{N}$, any $\epsilon > 0$ and $\delta \in (0, 1)$, there exists an algorithm that agnostically testably learns C w.r.t. D up to error $O(k^{1/2}\,\mathrm{opt}^{1 - \frac{1}{k+1}}) + \epsilon$, where $\mathrm{opt} = \min_{w \in S^{d-1}} P_{D_{XY}}[y \neq \mathrm{sign}(\langle w, x\rangle)]$, and error probability at most δ, using time and a number of samples from D_{XY} that are polynomial in $d^{O(k)}, (1/\epsilon)^{O(k)}$ and $(\log(1/\delta))^{O(k)}$. In particular, by picking some appropriate $k \approx \log^2 d$, we obtain error $O(\mathrm{opt}) + \epsilon$ in quasipolynomial time and sample complexity, i.e., $\mathrm{poly}(2^{\mathrm{polylog}\, d}, (\frac{1}{\epsilon})^{\mathrm{polylog}\, d})$.

To prove Theorem 5.1, we may follow a similar approach as the one we used for the case of Massart noise. However, in this case, the main structural lemma regarding the quality of the stationary points involves an additional requirement about the parameter ϵ. In particular, ϵ cannot be arbitrarily small with respect to the error of the optimum halfspace, because, in this case, there is no upper bound on the amount of noise that any specific point x might be associated with. As a result, picking ϵ to be arbitrarily small would imply that our algorithm only considers points that lie within a region that has arbitrarily small probability and can hence be completely corrupted with the adversarial opt budget. On the other hand, the polynomial slackness that the testability requirement introduces (through Proposition 4.4) between the error we achieve and the angular distance guarantee we can get via finding a stationary point of L (which is now coupled with opt), appears to the exponent of the guarantee we achieve in Theorem 5.1.

Lemma 5.2. Let L be as in Equation equation 4.1 with $\epsilon \in (0, 1]$, ℓ as described in Proposition C.1, let $w \in S^{d-1}$ and consider D_{XY} such that the marginal D_X on $\mathbb{R}^d$ satisfies Properties equation 3.2, equation 3.3 and equation 3.4 for with $C_4 = 2$ and accuracy parameter ϵ. Let opt be the minimum error achieved by some origin centered halfspace and let $w^* \in S^{d-1}$ be a corresponding vector. Then, there are constants $c_1, c_2, c_3, c_4 > 0$ such that if $\mathrm{opt} \leq c_1$, $\|\nabla_w L(w)\|_2 < c_2$, and $\epsilon \leq c_3$ then either $\cos(w, w^*) \geq c_4$ or $\cos(-w, w^*) \geq c_4$.

We obtain our main result for Gaussian target marginals by refining Proposition 4.4 for the specific case when the target marginal distribution is the standard multivariate Gaussian distribution. The algorithm for the Gaussian case is similar to the one of Theorem 5.1, but it runs different tests for the improved version (see Proposition D.1) of Proposition 4.4.

Theorem 5.3. In Theorem 5.1, if D is the standard Gaussian in d dimensions, we obtain error $O(\mathrm{opt}) + \epsilon$ in polynomial time and sample complexity, i.e., $\mathrm{poly}(d, 1/\epsilon, \log(1/\delta))$.

ACKNOWLEDGMENTS

We wish to thank the anonymous reviewers of ICLR 2024 for their constructive feedback. Aravind Gollakota was at UT Austin while this work was done, supported by NSF award AF-1909204 and the NSF AI Institute for Foundations of Machine Learning (IFML). Adam R. Klivans was supported by NSF award AF-1909204 and the NSF AI Institute for Foundations of Machine Learning (IFML). Konstantinos Stavropoulos was supported by NSF award AF-1909204, the NSF AI Institute for Foundations of Machine Learning (IFML), and by scholarships from Bodossaki Foundation and Leventis Foundation. Arsen Vasilyan was supported in part by NSF awards CCF-2006664, DMS-2022448, CCF-1565235, CCF-1955217, CCF-2310818, Big George Fellowship and Fintech@CSAIL. Part of this work was done while Arsen Vasilyan was visiting UT Austin.

REFERENCES

[ABHU15] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In *Conference on Learning Theory*, pages 167–190. PMLR, 2015.

[ABHZ16] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Conference on Learning Theory*, pages 152–192. PMLR, 2016.

[ABL17] Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM (JACM)*, 63(6):1–27, 2017.

[BH21] Maria-Florina Balcan and Nika Haghtalab. Noise in classification. *Beyond the Worst-Case Analysis of Algorithms*, page 361, 2021.

[BZ17] Maria-Florina F Balcan and Hongyang Zhang. Sample and computationally efficient learning algorithms under s-concave distributions. *Advances in Neural Information Processing Systems*, 30, 2017.

[Dan15] Amit Daniely. A ptas for agnostically learning halfspaces. In *Conference on Learning Theory*, pages 484–502. PMLR, 2015.

[DGT19] Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent pac learning of halfspaces with massart noise. *Advances in Neural Information Processing Systems*, 32, 2019.

[DK22] Ilias Diakonikolas and Daniel Kane. Near-optimal statistical query hardness of learning halfspaces with massart noise. In *Conference on Learning Theory*, pages 4258–4282. PMLR, 2022.

[DKK+22] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning general halfspaces with general massart noise under the gaussian distribution. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 874–885, 2022.

[DKK+23] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Sihan Liu, and Nikos Zarifis. Efficient testable learning of halfspaces with adversarial label noise. *arXiv preprint arXiv:2303.05485*, 2023.

[DKMR22] Ilias Diakonikolas, Daniel Kane, Pasin Manurangsi, and Lisheng Ren. Cryptographic hardness of learning halfspaces with massart noise. *Advances in Neural Information Processing Systems*, 2022.

[DKPZ21] Ilias Diakonikolas, Daniel M Kane, Thanasis Pittas, and Nikos Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the sq model. In *Conference on Learning Theory*, pages 1552–1584. PMLR, 2021.

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1061–1073, 2018.

[DKT21] Ilias Diakonikolas, Daniel Kane, and Christos Tzamos. Forster decomposition and learning halfspaces with noise. *Advances in Neural Information Processing Systems*, 34:7732–7744, 2021.

[DKTZ20a] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory*, pages 1486–1513. PMLR, 2020.

[DKTZ20b] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Non-convex sgd learns halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 33:18540–18549, 2020.

[DKTZ22] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning general halfspaces with adversarial label noise via online gradient descent. In *International Conference on Machine Learning*, pages 5118–5141. PMLR, 2022.

[DKZ20] Ilias Diakonikolas, Daniel Kane, and Nikos Zarifis. Near-optimal sq lower bounds for agnostically learning halfspaces and relus under gaussian marginals. *Advances in Neural Information Processing Systems*, 33:13586–13596, 2020.

[DTK22] Ilias Diakonikolas, Christos Tzamos, and Daniel M Kane. A strongly polynomial algorithm for approximate forster transforms and its application to halfspace learning. *arXiv preprint arXiv:2212.03008*, 2022.

[GGK20] Surbhi Goel, Aravind Gollakota, and Adam Klivans. Statistical-query lower bounds via functional gradients. *Advances in Neural Information Processing Systems*, 33:2147–2158, 2020.

[GKK23] Aravind Gollakota, Adam R Klivans, and Pravesh K Kothari. A moment-matching approach to testable learning and a new characterization of rademacher complexity. *Proceedings of the fifty-fifth annual ACM Symposium on Theory of Computing*, 2023. To appear.

[GKSV23] Aravind Gollakota, Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Tester-learners for halfspaces: Universal algorithms. *arXiv preprint arXiv:2305.11765*, 2023.

[KKMS08] Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[KLS09] Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(12), 2009.

[RV23] Ronitt Rubinfeld and Arsen Vasilyan. Testing distributional assumptions of learning algorithms. *Proceedings of the fifty-fifth annual ACM Symposium on Theory of Computing*, 2023. To appear.

[SW14] Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45, 2014.

[YZ17] Songbai Yan and Chicheng Zhang. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. *Advances in Neural Information Processing Systems*, 30, 2017.

[Zha18] Chicheng Zhang. Efficient active learning of sparse halfspaces. In *Conference on Learning Theory*, pages 1856–1880. PMLR, 2018.

[ZL21] Chicheng Zhang and Yinan Li. Improved algorithms for efficient active learning halfspaces with massart and tsybakov noise. In *Conference on Learning Theory*, pages 4526–4527. PMLR, 2021.

[ZSA20] Chicheng Zhang, Jie Shen, and Pranjal Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. *Advances in Neural Information Processing Systems*, 33:7184–7197, 2020.

A STRONGLY LOG-CONCAVE DISTRIBUTIONS

We also formally define the class of strongly log-concave distributions, which is the class that our target marginal D is allowed to belong to, and collect some useful properties of such distributions. We will state the definition for isotropic D (i.e. with mean 0 and covariance I) for simplicity.

Definition A.1 (Strongly log-concave distribution, see e.g. (SW14, Def 2.3)). We say an isotropic distribution D on $\mathbb{R}^d$ is strongly log-concave if the logarithm of its density q is a strongly concave function. Equivalently, q can be written as

$$q(x) = r(x) \cdot \varphi_I(x) \tag{A.1}$$

for some log-concave function r and some constant $\sigma > 0$, where φ_I denotes the density of the spherical Gaussian $\mathcal{N}(0, \sigma^2 I)$.

Proposition A.2 (see e.g. (SW14)). Let D be an isotropic strongly log-concave distribution on $\mathbb{R}^d$ with density q.

(a) Any orthogonal projection of D onto a subspace is also strongly log-concave.

(b) There exist constants $U \geq 1$ such that $q(x) \leq U$ for all x, and $q(x) \geq 1/U$ for all $\|x\| \leq R$.

(c) There exist constants $U' \geq 1$ and σ such that $q(x) \leq U' \cdot \varphi_{\sigma I}(x)$ for all x.

(d) There exist constants K_1, K_2 such that for any $\delta \in [0, 1]$ and any $v \in S^{d-1}$, $P[|\langle v, x\rangle| \geq \delta] \in (K_1 \delta, K_2 \delta)$.

(e) There exists a constant K_3 such that for any $k \in \mathbb{N}$, $\mathbb{E}[|\langle v, x\rangle|^k] \leq (K_3 k)^{k/2}$.

(f) Let $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}^d_{\geq 0}$ be a multi-index with total degree $|\alpha| = \sum_i \alpha_i = k$, and let $x^\alpha = \prod_i x_i^{\alpha_i}$. There exists a constant K_4 such that for any such α, $\mathbb{E}[|x^\alpha|] \leq (K_4 k)^{k/2}$.

For (a), see e.g. (SW14, Thm 3.7). The other properties follow readily from Eq. (A.1), which allows us to treat the density as subgaussian.

A key structural fact that we will need about strongly log-concave distributions is that approximately matching moments of degree at most $O(1/\varepsilon^2)$ with such a D is sufficient to fool any function of a constant number of halfspaces up to an additive ε.

Proposition A.3 (Variant of (GKK23, Thm 5.6)). Let p be a fixed constant, and let $\mathcal{F}$ be the class of all functions of p halfspaces mapping $\mathbb{R}^d$ to $\{\pm 1\}$ of the form

$$f(x) = g(\mathrm{sign}(\langle v^1, x\rangle + \theta_1), \ldots, \mathrm{sign}(\langle v^p, x\rangle + \theta_p)) \tag{A.2}$$

for some $g : \{\pm 1\}^p \to \{\pm 1\}$ and weights $v^i \in S^{d-1}$. Let D be any target marginal such that for every i, the projection $\langle v^i, x\rangle$ has subgaussian tails and is anticoncentrated: $P[|\langle v^i, x\rangle| > t] \leq \exp(-\Omega(t^2))$, and (b) for any interval $[a, b]$, $P[\langle v^i, x\rangle \in [a, b]] \leq O(|b - a|)$. Let D be any distribution such that for all monomials $x^\alpha = \prod_i x_i^{\alpha_i}$ of total degree $|\alpha| = \sum_i \alpha_i \leq k$,

$$|\mathbb{E}_D[x^\alpha] - \mathbb{E}_D[x^\alpha]| \leq \frac{c^{|\alpha|}\,|\alpha|^{|\alpha|}}{d^{\gamma\sqrt{k}}}$$

for some sufficiently small constant c (in particular, it suffices to have $d^{\Omega(k)}$ moment closeness for every α). Then

$$\max_{f \in \mathcal{F}} |\mathbb{E}_D[f] - \mathbb{E}_D[f]| \leq O\left(\frac{1}{d^{\gamma\sqrt{k}}}\right).$$

Note that this is a variant of the original statement of (GKK23, Thm 5.6), which requires that the 1D projection of D along any direction satisfy suitable concentration and anticoncentration. Indeed, an inspection of their proof reveals that it suffices to verify these properties for projections only along the directions $\{v^i\}_{i \in [p]}$ as opposed to all directions. This is because to fool a function of the form above, their proof only analyzes the projected distribution $(\langle v^1, x\rangle, \ldots, \langle v^p, x\rangle)$ on $\mathbb{R}^p$, and requires only concentration and anticoncentration for each individual projection $\langle v^i, x\rangle$.

B PROOFS FOR SECTION 3

B.1 PROOF OF PROPOSITION 3.1

Our plan is to apply Proposition A.3. To do so, we must verify that $D_{|T}$ satisfies the assumptions required. In particular, it suffices to verify that the 1D projection along any direction orthogonal to w has subgaussian tails and is anticoncentrated. Let $v \in S^{d-1}$ be any direction that is orthogonal to w. By Proposition A.2(d), we may assume that $P_D[T] \geq \Omega(1)$.

To verify subgaussian tails, we must show that for any t, $P_{D_{|T}}[|\langle v, x\rangle| > t] \leq \exp(-Ct^2)$ for some constant C. The main fact we use is Proposition A.2(c), i.e. that any strongly log-concave density is pointwise upper bounded by a Gaussian density times a constant. Write

$$P_{D_{|T}}[|\langle v, x\rangle| > t] = \frac{P_D[|\langle v, x\rangle| > t \text{ and } \langle w, x\rangle \in [-\gamma, \gamma]]}{P_D[\langle w, x\rangle \in [-\gamma, \gamma]]}.$$

The claim now follows from the fact that the numerator is upper bounded by a constant times the corresponding probability under a Gaussian density, which is at most $O(\exp(-C't^2))$ for some constant C', and that the denominator is $\Omega(1)$.

To check anticoncentration, for any interval $[a, b]$, write

$$P_{D_{|T}}[\langle v, x\rangle \in [a, b]] = \frac{P_D[\langle v, x\rangle \in [a, b] \text{ and } \langle w, x\rangle \in [-\gamma, \gamma]]}{P_D[\langle w, x\rangle \in [-\gamma, \gamma]]}.$$

After projecting onto $\mathrm{span}(v, w)$ (an operation that preserves logconcavity), the numerator is the probability mass under a rectangle with side lengths $|a - b|$ and 2γ, which is at most $O(|b - a|)$ as by Proposition A.2(b) the density is pointwise upper bounded by a constant. The claim follows since the denominator is $\Omega(1)$.

Now we are ready to apply Proposition A.3. We see that $D_{|T}$ if matches moments of degree at most k with $D_{|T}$ up to an additive slack of $\delta^{O(k)}$, then $|E_D[f \mid T] - E_D[f \mid T]| \leq O(1/\sqrt{k})$. Rewriting in terms of ε gives the theorem.

B.2 PROOF OF PROPOSITION 3.2

The tester T_1 does the following:

1. For all $\alpha \in \mathbb{Z}^d_{\geq 0}$ with $|\alpha| = k$:
   (a) Compute the corresponding moment $E_{(x,y)\sim D}[x^\alpha] := \frac{1}{|S|}\sum_{x \in S} x^\alpha$.
   (b) If $E_{(x,y)\sim D}[x^\alpha] - E_{x \sim D}[x^\alpha] > \frac{1}{d^k}$ then reject.
2. If all the checks above passed, accept.

First, we claim that for some absolute constant C_1, if the tester above accepts, we have $E_{(x,y)\sim D}[(\langle v, x\rangle)^k] \leq (C_1 k)^{k/2}$ for any $v \in S^{d-1}$. To show this, we first recall that by Proposition A.2(e) it is the case that $E_{(x,y)\sim D}[(\langle v, x\rangle)^k] \leq (K_3 k)^{k/2}$. But we have

$$\begin{aligned} E_{(x,y)\sim D}[(\langle v, x\rangle)^k] - E_{(x,y)\sim D}[(\langle v, x\rangle)^k] &\leq \sum_{\alpha : |\alpha| = k} E_{(x,y)\sim D}[x^\alpha] - E_{x \sim D}[x^\alpha] \\ &\leq d^k \max_{\alpha : |\alpha| = k} E_{(x,y)\sim D}[x^\alpha] - E_{x \sim D}[x^\alpha] \leq 1 \end{aligned}$$

Together with the bound $E_{(x,y)\sim D}[(\langle v, x\rangle)^k] \leq (K_3 k)^{k/2}$, the above implies that $E_{(x,y)\sim D}[(\langle v, x\rangle)^k] \leq (C_1 k)^{k/2}$ for some constant C_1.

Now, we need to show that if the elements of S are chosen i.i.d. from D, and $|S| \geq d^k \cdot \log\frac{1}{\delta}^k \cdot C_1$ then the tester above accepts with probability at least $1 - \delta$. Consider any specific

multi-index $\in \mathbb{Z}^d_{\geq 0}$ with $|\| = k$. Now, by Proposition A.2(f) we have the following:

$$\left|\mathop{\mathbb{E}}_{x\sim D}[x^\alpha]\mathop{\mathbb{E}}_{z\sim D}[z^\beta]^{2\log(1/\varepsilon)} - \sum_{\ell=0}^{2\log(1/\varepsilon)} \mathop{\mathbb{E}}_{x\sim D}[(x^\alpha)^\ell]\mathop{\mathbb{E}}_{z\sim D}[z^\beta]^{2\log(1/\varepsilon)-\ell}\right| \leq \sum_{\ell=0}^{2\log(1/\varepsilon)} (K_4\ell k)^{\ell k/2}(K_4 k)^{k(2\log(1/\varepsilon)-\ell)/2} \leq 2\log(1/\varepsilon)(2K_4\log(1/\varepsilon)k)^{k\log(1/\varepsilon)}$$

This, together with Markov's inequality implies that

$$\mathbb{P}\left[\left|\frac{1}{|S|}\sum_{x\in S} x^\alpha - \mathop{\mathbb{E}}_{x\sim D}[x^\alpha]\right| > \frac{1}{d^k}\right] \leq \frac{d^k(3K_4 k\log(1/\varepsilon))^{k/2}}{|S|}^{2\log(1/\varepsilon)}$$

Since S is obtained by taking at least $|S| \geq d^k \cdot (k\log\frac{1}{\varepsilon})^{C_1}$, for sufficiently large C_1 we see that the above is upper-bounded by $\frac{1}{d^k}$. Taking a union bound over all $\in \mathbb{Z}^d_{\geq 0}$ with $|\| = k$, we see that with probability at least $1-\varepsilon$ the tester T_1 accepts, finishing the proof.

B.3 Proof of Proposition 3.3

Let K_1 be as in part (d) of Proposition A.2. The tester T_2 computes the fraction of elements in S that are in T. If this fraction is $K_1/2$-close to $\mathbb{P}_{x\sim D}[\langle w, x\rangle \geq t]$, the algorithm accepts. The algorithm rejects otherwise.

Now, from (d) of Proposition A.2 we have that $\mathbb{P}_{x\sim D}[\langle w, x\rangle \geq t] \in [K_1, K_2]$. Therefore, if the fraction of elements in S that belong in T is $K_1/100$-close to $\mathbb{P}_{x\sim D}[\langle w, x\rangle \geq t]$, then this quantity is in $[K_1/2, (K_2 + K_1/2)]$ as required.

Finally, if $|S| \geq \frac{100}{K_1^2}\log\frac{1}{\varepsilon}$ by standard Hoeffding bound, with probability at least $1-\varepsilon$ we indeed have that the fraction of elements in S that are in T is $K_1/2$-close to $\mathbb{P}_{x\sim D}[\langle w, x\rangle \geq t]$.

B.4 Proof of Proposition 3.4

The tester T_3 does the following:

1. Runs the tester T_2 from Proposition 3.3. If T_2 rejects, T_3 rejects as well.
2. Let $S_{|T}$ be the set of elements in S for which $x \in T$.
3. Let $k = O(1/\varepsilon^2)$ be chosen as in Proposition 3.1.
4. For all $\in \mathbb{Z}^d_{\geq 0}$ with $|\| = k$:
   (a) Compute the corresponding moment $\mathbb{E}_{(x,y)\sim D}[x^\alpha \mid x \in T] := \frac{1}{|S_{|T}|}\sum_{x\in S_{|T}} x^\alpha$.
   (b) If $\mathbb{E}_{(x,y)\sim D}[x^\alpha \mid x \in T] - \mathbb{E}_{x\sim D}[x^\alpha \mid x \in T] > \frac{\sigma(k)}{d^k}$ then reject, where the polylogarithmic in d $\sigma(k)$ is chosen to satisfy the additive slack condition in Proposition 3.1.
5. If all the checks above passed, accept.

First, we argue that if the checks above pass, then Equations 3.3 and 3.4 will hold. If the tester passes, Equation 3.3 follows immediately from the guarantees in step (4b) of T_3 together with Proposition 3.1. Equation 3.4, in turn, is proven as follows:

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[(\langle v, x\rangle)^2] - \mathop{\mathbb{E}}_{(x,y)\sim D}[(\langle v, x\rangle)^2] \leq \sum_{|\alpha|=2} \left|\mathop{\mathbb{E}}_{(x,y)\sim D}[x^\alpha] - \mathop{\mathbb{E}}_{x\sim D}[x^\alpha]\right|$$

$$\leq d^2 \max_{|\alpha|=2} \left|\mathop{\mathbb{E}}_{(x,y)\sim D}[x^\alpha] - \mathop{\mathbb{E}}_{x\sim D}[x^\alpha]\right|$$

Now, we need to show that if the elements of S are chosen i.i.d. from D, and $|S| = \ldots$ then the tester above accepts with probability at least $\frac{1}{3}$. Consider any specific multi-index $\alpha \in \mathbb{Z}^d_{\geq 0}$ with $|\alpha| = k$. Now, by Proposition A.2(f) we have for any positive integer ℓ the following:

$$\mathop{\mathbb{E}}_{x \sim D}\left[(x-\mu)^\alpha\right] \leq (K_4 \ell k)^{k/2}$$

But by Proposition A.2(d) we have that $\Pr_{x \sim D}[x \in T] = \Pr_{x \sim D}[|\langle x, w_{ij}\rangle| \leq K_1]$. Therefore, the density of the distribution $D_{|T}$ (which is defined as the distribution one obtains by taking D and conditioning on $x \in T$) is upper bounded by the product of the density of the distribution D and $\frac{1}{K_1}$. This allows us to bound

$$\mathop{\mathbb{E}}_{x \sim D}\left[(x-\mu)^\alpha \mid x \in T\right] \leq \frac{1}{K_1} \mathop{\mathbb{E}}_{x \sim D}\left[(x-\mu)^\alpha\right] \leq \frac{(K_4 \ell k)^{k/2}}{K_1}$$

This implies that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x \sim D}\left[\mathop{\mathbb{E}}_{z \sim D}[\langle z-\mu, v\rangle^{2\log(1/\delta)} \mid z \in T]^{\ell}\right] \\
&\leq \sum_{\ell=0}^{2\log(1/\delta)} \mathop{\mathbb{E}}_{x \sim D}\left[(x-\mu)^\alpha \mid x \in T\right] \mathop{\mathbb{E}}_{x \sim D}\left[\langle x-\mu, v\rangle^{2\log(1/\delta)-\ell}\right] \\
&\leq \frac{1}{(K_1)^{2\log(1/\delta)}} \sum_{\ell=0}^{2\log(1/\delta)} (K_4 \ell k)^{\ell k/2} (K_4 k)^{k(2\log(1/\delta)-\ell)/2} \\
&\leq \frac{1}{(K_1)^{2\log(1/\delta)}} 2\log(1/\delta)(2K_4 \log(1/\delta) k)^{\log(1/\delta) k}
\end{aligned}$$

This, together with Markov's inequality implies that

$$\Pr\left[\frac{1}{|S|}\sum_{x \in S} \langle x-\mu, v\rangle^{2\log(1/\delta)} > \frac{1}{d^k} d^{O(k)} \left(\frac{d^{O(k)}(3K_4 k \log(1/\delta))^{k/2}}{K_1 |S_{|T}|}\right)^{2\log(1/\delta)}\right] \leq \ldots$$

Now, recall that the tester T_2 in step (1) accepted, we have $|S_{|T}| \geq \frac{1}{C_2}|S|$. Since S is obtained by taking at least $|S| \geq \ldots d^{\frac{1}{2}\log^{C_5}(\frac{1}{\delta})} \log^{\frac{1}{2}\log^{C_5}(\frac{1}{\delta})}$, for sufficiently large C_5 we see that the expression above is upper-bounded by $\frac{1}{d^k}$. Taking a union bound over all $\alpha \in \mathbb{Z}^d_{\geq 0}$ with $|\alpha| = k$, we see that with probability at least $\ldots$ the tester T_3 accepts, finishing the proof.

C Proofs from Section 4

We first present the following Proposition, which ensures that we can form a loss function with certain desired properties.

Proposition C.1. There are constants $c, c' > 0$, such that for any $\ldots > 0$, there exists a continuously differentiable function $\ell : \mathbb{R} \to [0,1]$ with the following properties.

1. For any $t \in [\ldots/6, \ldots/6]$, $\ell(t) = \frac{1}{2} + \frac{t}{\ldots}$.
2. For any $t > \ldots/2$, $\ell(t) = 1$ and for any $t < \ldots/2$, $\ell(t) = 0$.
3. For any $t \in \mathbb{R}$, $\ell'(t) \in [0, c/\ldots]$, $\ell'(t) = \ell'(\ldots - t)$ and $|\ell''(t)| \leq c'/\ldots^2$.

Proof. We define ℓ as follows.

$$\ell(t) = \begin{cases} \frac{t}{\ldots} + \frac{1}{2}, & \text{if } |t| \leq \frac{\ldots}{6} \\ 1, & \text{if } t > \frac{\ldots}{2} \\ 0, & \text{if } t < \frac{\ldots}{2} \\ \ell_+(t), & t \in (\frac{\ldots}{6}, \frac{\ldots}{2}] \\ \ell_-(t), & t \in [\frac{\ldots}{2}, \frac{\ldots}{6}) \end{cases}$$

Figure 2: The function $\grave{}$ used to smoothly approximate the ramp.

for some appropriate functions $\grave{}^+ ; \grave{}$. It is sufficient that we pick $\grave{}^+$ satisfying the following conditions (then $\grave{}$ would be defined symmetrically, i.e., $\grave{}(t) = 1 \quad \grave{}^+(\quad t)$).

- $\grave{}^+(\ = 2) = 1$ and $\grave{}^{+\,0}(\ = 2) = 0$.
- $\grave{}^+(\ = 6) = 2{=}3$ and $\grave{}^{+\,0}(\ = 6) = 1{=}$.
- $\grave{}^{+\,00}$ is defined and bounded, except, possibly at $= 6$ and/or $= 2$.

We therefore need to satisfy four equations for $\grave{}^+$. So we set $\grave{}^+$ to be a degree 3 polynomial: $\grave{}^+(t) = a_1 t^3 + a_2 t^2 + a_3 t + a_4$. Whenever > 0, the system has a unique solution that satisfies the desired inequalities. In particular, we may solve the equation to get $a_1 = \quad 9{=}\ ^3; a_2 = 15{=}(2\ ^2); a_3 = \quad 3{=}(4\ )$ and $a_4 = 5{=}8$. For the resulting function (see Figure 2 below and Figure 4 in the appendix) we have that there are constants $c; c^0 > 0$ such that $\grave{}^{+\,0}(t) \ 2\ [0; c{=}\ ]$ and $|\grave{}^{+\,00}(t)| \quad c^0{=}\ ^2$ for any $t \ 2\ [\ = 6; \ = 2]$. □

C.1 Proof of Lemma 4.3

We will prove the contrapositive of the claim, namely, that there are constants $c_1, c_2, c_3 > 0$ such that if $\rceil(w; w\ ); \rceil(\ w; w\ ) > \frac{c_3}{\sqrt{1\ \ ^2}}$, and $\quad c_2$, then $\|r_w L\ (w)\|_2 \quad c_1(1\ \ ^2)$.

Consider the case where $\rceil(w; w\ ) < \ = 2$ (otherwise, perform the same argument for w). Let v be a unit vector orthogonal to w that can be expressed as a linear combination of w and w and for which $\langle v; w\ \rangle = 0$. Then $\{v; w\}$ is an orthonormal basis for $V = \text{span}(w; w\ )$. For any vector $x \ 2\ \mathbb{R}^d$, we will use the following notation $x_w = \langle w; x\rangle$, $x_v = \langle v; x\rangle$. It follows that $\text{proj}_V(x) = x_w w + x_v v$, where proj_V is the operator that orthogonally projects vectors to V.

Using the fact that $r_w(\langle w; x\rangle{=}\|w\|_2) = x \quad \langle w; x\rangle w = x \quad x_w w$ for any $w \ 2\ S^{d\ 1}$, the interchangeability of the gradient and expectation operators and the fact that $\grave{}^0$ is an even function we get that

$$r_w L\ (w) = E\left[\quad \grave{}^0(|\langle w; x\rangle|)\ y\ (x \quad x_w w)\right]$$

Since the projection operator proj_V is a contraction, we have $\|r_w L\ (w)\|_2 \quad \|\text{proj}_V r_w L\ (w)\|_2$, and we can therefore restrict our attention to a simpler, two dimensional problem. In particular, since $\text{proj}_V(x) = x_w w + x_y v$, we get

$$\|\text{proj}_V r_w L\ (w)\|_2 = \left\|E\left[\quad \grave{}^0(|x_w|)\ y\ x_v v\right]\right\|_2 = E\left[\quad \grave{}^0(|x_w|)\ y\ x_v\right]$$
$$= E\left[\quad \grave{}^0(|x_w|)\ \text{sign}(\langle w\ ; x\rangle)\ (1 \quad 2\mathbb{1}\{y \ 6{=}\ \text{sign}(\langle w\ ; x\rangle)\})\ x_v\right]$$

Figure 3: Critical regions in the proofs of main structural lemmas (Lemmas 4.3, 5.2). We analyze the contributions of the regions labeled A_1, A_2 to the quantities A_1, A_2 in the proofs. Specifically, the regions A_1 (which have height $\sigma/3$ so that the value of $\ell^0(x_w)$ for any x in these regions is exactly $1/\sigma$, by Proposition C.1) form a subset of the region G, and their probability mass under D_X is (up to a multiplicative factor) a lower bound on the quantity A_1 (see Eq equation C.3). Similarly, the region A_2 is a subset of the intersection of G with the band of height σ, and has probability mass that is (up to a multiplicative factor) an upper bound on the quantity A_2 (see Eq equation C.4).

Let $F(y, x)$ denote $1 - 2\mathbb{1}\{y \neq \mathrm{sign}(\langle w^*, x\rangle)\}$. We may write x_v as $|x_v| \, \mathrm{sign}(x_v)$ and let $G \subseteq \mathbb{R}^2$ such that $\mathrm{sign}(x_v) \, \mathrm{sign}(\langle w^*, x\rangle) = -1$ iff $x \in G$. Then, $\mathrm{sign}(x_v)\,\mathrm{sign}(\langle w^*, x\rangle) = \mathbb{1}\{x \notin G\} - \mathbb{1}\{x \in G\}$. We get

$$
\begin{aligned}
&\|\mathrm{proj}_V \nabla_w L_\sigma(w)\|_2 = \\
&= \mathbb{E}\left[\ell^0(|x_w|)\,(\mathbb{1}\{x \in G\} - \mathbb{1}\{x \notin G\})\,F(y, x)\,|x_v|\right] \\
&\geq \mathbb{E}\left[\ell^0(|x_w|)\,\mathbb{1}\{x \in G\}\,F(y, x)\,|x_v|\right] - \mathbb{E}\left[\ell^0(|x_w|)\,\mathbb{1}\{x \notin G\}\,F(y, x)\,|x_v|\right]
\end{aligned}
$$

Let $A_1 = \mathbb{E}[\ell^0(|x_w|)\,\mathbb{1}\{x \in G\}\,F(y, x)\,|x_v|]$ and $A_2 = \mathbb{E}[\ell^0(|x_w|)\,\mathbb{1}\{x \notin G\}\,F(y, x)\,|x_v|]$. (See Figure 3.) Note that $\mathbb{E}_{y|x}[F(y, x)] = 1 - 2\eta(x) \in [1 - 2\eta, 1]$, where $1 - 2\eta > 0$. Therefore, we have that $A_1 \geq (1 - 2\eta)\,\mathbb{E}[\ell^0(|x_w|)\,\mathbb{1}\{x \in G\}\,|x_v|]$ and $A_2 \leq \mathbb{E}[\ell^0(|x_w|)\,\mathbb{1}\{x \notin G\}\,|x_v|]$.

Note that due to Proposition C.1, $\ell^0(|x_w|) \leq c/\sigma$ for some constant c and $\ell^0(|x_w|) = 0$ whenever $|x_w| > \sigma/2$. Therefore, if U_2 is the band $B_w(\sigma/2) = \{x : |x_w| \leq \sigma/2\}$ we have

$$
A_2 \leq \frac{c}{\sigma}\,\mathbb{E}[\mathbb{1}\{x \notin G\}\mathbb{1}\{x \in U_2\}\,|x_v|] \tag{C.1}
$$

Moreover, for each individual x, we have $\ell^0(|x_w|)\,\mathbb{1}\{x \in G\}\,|x_v| \geq 0$, due to the properties of ℓ^0 (Proposition C.1). Hence, for any set $U_1 \subseteq \mathbb{R}^d$ we have that

$$
A_1 \geq (1 - 2\eta)\,\mathbb{E}[\ell^0(|x_w|)\,\mathbb{1}\{x \in G\}\,\mathbb{1}\{x \in U_1\}\,|x_v|]
$$

Setting $U_1 = B_w(\sigma/6) = \{x : |x_w| \leq \sigma/6\}$, by Proposition C.1, we get $\ell^0(|x_w|)\,\mathbb{1}\{x \in U_1\} = \frac{1}{\sigma}\,\mathbb{1}\{x \in U_1\}$.

$$
A_1 \geq \frac{1 - 2\eta}{\sigma}\,\mathbb{E}[\mathbb{1}\{x \in G\}\,\mathbb{1}\{x \in U_1\}\,|x_v|] \tag{C.2}
$$

We now observe that by the definitions of G, U_1, U_2, for any constant $R > 0$, there exist some constants $c', c'' > 0$ such that if $\sigma/\tan\theta < c'R$ (the points in $\mathbb{R}^2$ where ∂G intersects either ∂U_1 or

@U_2 have projections onw that are (= tan)) we have that

$$1\{x \in G\} \; 1\{x \in U_1\} \quad 1\{|\langle x_v, x\rangle| \in [c'R, 2c'R]\} \; 1\{x \in U_1\} \quad \text{and}$$
$$1\{x \in G\} \; 1\{x \in U_2\} \quad 1\{|\langle x_v, x\rangle| \quad c''/\tan \} \; 1\{x \in U_2\}$$

By equations equation C.1 and equation C.2, we get the following bounds whose graphical representations can be found in Figure 3.

$$A_1 \quad \frac{c'R(1 \quad ^2)}{} \; \mathbb{E}[1\{|\langle x_v, x\rangle| \in [c'R, 2c'R]\} \; 1\{x \in U_1\}] \tag{C.3}$$

$$A_2 \quad \frac{c \quad c''}{\tan} \; \mathbb{E}[1\{|\langle x_v, x\rangle| \quad c''/\tan \} \; 1\{x \in U_2\}] \tag{C.4}$$

So far, we have used no distributional assumptions. Now, consider the corresponding expectations under the target marginal (which we assumed to be strongly log-concave).

$$I_1 = \mathbb{E}_{D}[1\{|\langle x_v, x\rangle| \in [c'R, 2c'R]\} \; 1\{x \in U_1\}]$$
$$I_2 = \mathbb{E}_{D}[1\{|\langle x_v, x\rangle| \quad c''/\tan \} \; 1\{x \in U_2\}]$$

Any strongly log-concave distribution enjoys the "well-behaved" properties defined by (DKTZ20a), and therefore, if is picked to be small enough, then and are of order ( ) (due to upper and lower bounds on the two dimensional marginal density over within constant radius balls – aka anti-anticoncentration and anticoncentration). Moreover, by Proposition A.2, we have and are both of order ( ). Hence we have that there exist constants such that for the conditional expectations we have

$$\mathbb{E}_{D} \; 1\{|\langle x_v, x\rangle| \in [c'R, 2c'R]\} \; 1\{x \in U_1\} \quad c_1'$$
$$\mathbb{E}_{D} \; 1\{|\langle x_v, x\rangle| \quad c''/\tan \} \; 1\{x \in U_2\} \quad c_2'$$

By assumption, Property equation 3.3 holds and, therefore, if , we get that

$$\mathbb{E}_{D_X} \; 1\{|\langle x_v, x\rangle| \in [c'R, 2c'R]\} \; 1\{x \in U_1\} \quad c_1'=2$$
$$\mathbb{E}_{D_X} \; 1\{|\langle x_v, x\rangle| \quad c''/\tan \} \; 1\{x \in U_2\} \quad c_2'=2$$

Moreover, by Property equation 3.2, we have that (under the true marginal) and are both ( ). Hence, in total, we get that for some constants , we have

$$A_1 \quad c_1 \; (1 \quad ^2) \text{ and } A_2 \quad c_2 \; \frac{}{\tan}$$

Hence, if we pick = ((1 ^{2}) tan), we get the desired result.

C.2 Proof of Proposition 4.4

For the following all the probabilities and expectations are over . First we observe that

$$\mathbb{P}[y \neq \operatorname{sign}(\langle w, x\rangle)] \quad \mathbb{P}[y \neq \operatorname{sign}(\langle w, x\rangle) \cap y = \operatorname{sign}(\langle w^*, x\rangle)] + \mathbb{P}[y \neq \operatorname{sign}(\langle w^*, x\rangle)]$$
$$\mathbb{P}[\operatorname{sign}(\langle w, x\rangle) \neq \operatorname{sign}(\langle w^*, x\rangle)] + \text{opt}.$$

Then, we observe that by assumption that satisfies Property equation 3.2, we have

$$\mathbb{P}[|\langle w, x\rangle| \quad] \quad C_3$$

and that

$$\mathbb{P}[\operatorname{sign}(\langle w, x\rangle) \neq \operatorname{sign}(\langle w^*, x\rangle) \cap |\langle w, x\rangle| >] \quad \mathbb{P}\left[|\langle v, x\rangle| \quad \frac{}{\tan}\right];$$

where is some vector perpendicular to . Using Markov's inequality, we get

$$\mathbb{P}\left[|\langle v, x\rangle| \quad \frac{}{\tan}\right] \quad \frac{(\tan)^k}{^k} \; \mathbb{E}[|\langle v, x\rangle|^k].$$

But, by assumption that $D_{\mathcal{XY}}$ satisfies Property equation 3.1, there is some constant $C_1 > 0$ such that $\mathbb{E}[|\langle \mathsf{v}, \mathsf{x}\rangle|^k] \leq (C_1 k)^{k/2}$. Thus

$$\begin{aligned}\mathbb{P}[\mathrm{sign}(\langle \mathsf{w}, \mathsf{x}\rangle) \neq \mathrm{sign}(\langle \mathsf{w}^*, \mathsf{x}\rangle)] &\leq \mathbb{P}[|\langle \mathsf{w}, \mathsf{x}\rangle| \leq \sigma] \\ &\quad + \mathbb{P}[\mathrm{sign}(\langle \mathsf{w}, \mathsf{x}\rangle) \neq \mathrm{sign}(\langle \mathsf{w}^*, \mathsf{x}\rangle) \cap |\langle \mathsf{w}, \mathsf{x}\rangle| > \sigma] \\ &\leq C_3\sigma + \frac{(C_1 k)^{k/2}(\tan\theta)^k}{\sigma^k}.\end{aligned}$$

By picking σ appropriately in order to balance the two terms (note that this is a different σ than the one in Lemma 4.3), we get the desired result.

D Proofs from Section 5

D.1 Proof of Theorem 5.1

We will follow the same steps as for proving Theorem 4.1. Once more, we draw a sufficiently large sample so that our testers are ensured to accept with high probability when the true marginal is indeed the target marginal D^* and so that we have generalization, i.e. the guarantee that any approximate minimizer of the empirical error (error on the uniform empirical distribution over the sample drawn) is also an approximate minimizer of the true error. The algorithm we use is once more Algorithm 1, but this time we make multiple calls for different parameters σ (and we run T_1 with higher k, as we will see shortly) and reject if any of these calls rejects. If we accept, we output the output of the execution of Algorithm 1 with the minimum empirical error.

The main difference between the Massart noise case and the agnostic case is that in the former we were able to pick σ arbitrarily small, while in the latter we face a more delicate tradeoff. To balance competing contributions to the gradient norm, we must ensure that σ is at least $\Theta(\mathsf{opt})$ while also ensuring that it is not too large. And since we do not know the value of opt, we will need to search over a space of possible values for σ that is only polynomially large in relevant parameters (similar to the approach of (DKTZ20b)). In our case, we may sparsify the space $(0, 1]$ of possible values for σ up to accuracy $\Theta((\frac{\epsilon}{\sqrt{k}})^{1+1/k})$ and form a list of $\mathrm{poly}(k/\epsilon)$ possible values for σ, one of which will satisfy $c_1\sigma - \Theta((\frac{\epsilon}{\sqrt{k}})^{1+1/k}) \leq \mathsf{opt} \leq c_1\sigma$. hence, we perform the same (testing-learning) process for each of the possible values of σ and get a list of candidate vectors which is still of polynomial size.

The final step is, again, to use Proposition 4.4, after running tester T_1 with parameter k (Proposition 3.2) and tester T_2 with appropriate parameters for each of the candidate weight vectors. We get that our list contains a vector w with

$$\mathbb{P}_{D_{XY}}[y \neq \mathrm{sign}(\langle \mathsf{w}, \mathsf{x}\rangle)] \leq \mathsf{opt} + c \cdot k^{1/2} \cdot \theta^{1-1/(k+1)},$$

where $\rceil(\mathsf{w}, \mathsf{w}^*) \leq \theta := c_2\sigma$ for σ such that $c_1\sigma - \Theta((\frac{\epsilon}{\sqrt{k}})^{1+1/k}) \leq \mathsf{opt} \leq c_1\sigma$.

$$\mathbb{P}_{D_{XY}}[y \neq \mathrm{sign}(\langle \mathsf{w}, \mathsf{x}\rangle)] \leq \mathsf{opt} + c\sqrt{k} \cdot \left(\frac{c_2}{c_1}\mathsf{opt} + \Theta\left(\frac{\epsilon}{\sqrt{k}}\right)^{1+\frac{1}{k}}\right)^{1-\frac{1}{k+1}} \leq O(\sqrt{k} \cdot \mathsf{opt}^{1-\frac{1}{k+1}}) + \epsilon\,.$$

However, we do not know which of the weight vectors in our list is the one guaranteed to achieve small error. In order to discover this vector, we estimate the probability of error of each of the corresponding halfspaces (which can be done efficiently, due to Hoeffding's bound) and pick the one with the smallest error. This final step does not require any distributional assumptions and we do not need to perform any further tests.

In order to obtain our $\tilde{O}(\mathsf{opt})$ quasipolynomial time guarantee, observe first that we may assume without loss of generality that $\mathsf{opt} \geq 1/d^C$ for some C; if instead $\mathsf{opt} = o(1/d^2)$, say, then a sample of $O(d)$ points will with high probability be noiseless, and so simple linear programming will recover a consistent halfspace that will generalize. Moreover, we may assume that $\mathsf{opt} \leq 1/10$, since otherwise achieving $O(\mathsf{opt})$ is trivial (we may output an arbitrary halfspace). Let us adapt our algorithm so that we run tester T_1 (see Proposition 3.2) multiple times for all $k = 1, 2, \ldots, \lceil \log^2 d \rceil$ (this only changes our time and sample complexity by a $\mathrm{polylog}(d)$ factor). Then Proposition 4.4

holds for some k^* such that $k^* \in [\log(1/\mathsf{opt}), 2\log(1/\mathsf{opt})]$, since the interval has length at least 1 (and therefore it contains some integer) and $2\log(1/\mathsf{opt}) \leq 2C\log d \leq \log^2 d$ (for large enough d). Therefore, by picking the best candidate we get a guarantee of order

$$\begin{aligned}
\sqrt{k^*}\cdot \mathsf{opt}^{1-1/k} &= \sqrt{k^*}\cdot \mathsf{opt}^{-1/k}\ \mathsf{opt} \\
&= \sqrt{k^*}\cdot 2^{\frac{1}{k}\log\frac{1}{\mathsf{opt}}}\cdot \mathsf{opt} \\
&\leq \sqrt{2\log(1/\mathsf{opt})}\cdot 2\cdot \mathsf{opt} \qquad \text{(since } \log(1/\mathsf{opt}) \leq k^* \leq 2\log(1/\mathsf{opt})\text{)} \\
&= \Theta(\mathsf{opt})\,.
\end{aligned}$$

This concludes the proof of Theorem 5.1.

D.2 Proof of Lemma 5.2

In the agnostic case, the proof is analogous to the proof of Lemma 4.3. However, in this case, the difference is that the random variable $F(y,\mathbf{x}) = 1 - 2\ \{y \neq \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)\}$ does not have conditional expectation on $\mathbf{x}$ that is lower bounded by a constant. Instead, we need to consider an additional term A_3 correcponding to the part $2\ \{y \neq \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle)\}$ and the term A_1 will not be scaled by the factor $(1-2\eta)$ as in Lemma 4.3. Hence, with similar arguments we have that

$$\|\nabla_{\mathbf{w}} \mathcal{L}_\sigma(\mathbf{w})\|_2 \geq A_1 - A_2 - A_3\,,$$

where $A_1 \geq \tilde{c}_1$, $A_2 \leq \tilde{c}_2 \cdot \frac{\sigma}{\tan\theta}$ and (using properties of ℓ'_σ as in Lemma 4.3 and the Cauchy-Schwarz inequality)

$$\begin{aligned}
A_3 &= 2\,\mathbb{E}[\ell'_\sigma(|\mathbf{x}_\mathbf{w}|)\cdot\ \{\mathbf{x}\in\mathcal{G}\}\cdot\ \{y \neq \mathrm{sign}(\langle \mathbf{w},\mathbf{x}\rangle)\}\cdot|\mathbf{x}_\mathbf{v}|] \leq \\
&\leq \frac{2c}{\sigma}\cdot \mathbb{E}[\ \{\mathbf{x}\in\mathcal{U}_2\}\cdot\ \{y \neq \mathrm{sign}(\langle \mathbf{w},\mathbf{x}\rangle)\}\cdot|\mathbf{x}_\mathbf{v}|] \leq \\
&\leq \frac{2c}{\sigma}\cdot \sqrt{\mathbb{E}[\ \{\mathbf{x}\in\mathcal{U}_2\}\cdot(\mathbf{x}_\mathbf{v})^2]}\cdot \sqrt{\mathbb{E}[\ \{y \neq \mathrm{sign}(\langle \mathbf{w},\mathbf{x}\rangle)\}]} = \\
&= \frac{2c\sqrt{\mathsf{opt}}}{\sigma}\cdot \sqrt{\mathbb{E}[\langle \mathbf{v},\mathbf{x}\rangle^2 \mid \mathbf{x}\in\mathcal{U}_2]\cdot\mathbb{P}[\mathbf{x}\in\mathcal{U}_2]}\,.
\end{aligned}$$

Similarly to our approach in the proof of Lemma 4.3, we can use the assumed properties equation 3.2 and equation 3.4 to get that

$$A_3 \leq \tilde{c}_3 \frac{\sqrt{\mathsf{opt}}}{\sqrt{\sigma}}\,,$$

which gives that in order for the gradient loss to be small, we require $\mathsf{opt} \leq \Theta(\sigma)$.

D.3 Proof of Theorem 5.3

Before presenting the proof of Theorem 5.3, we prove the following Proposition, which is, essentially, a stronger version of Proposition 4.4 for the specific case when the target marginal distribution D^* is the standard multivariate Gaussian distribution. Proposition D.1 is important to get an $O(\mathsf{opt})$ guarantee for the case where the target distribution is the standard Gaussian.

Proposition D.1. *Let $D_{\mathcal{XY}}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$, $\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\mathbb{S}^d\ 1} \mathbb{P}_{D_{\mathcal{X}\,Y}}[y \neq \mathrm{sign}(\langle \mathbf{w},\mathbf{x}\rangle)]$ and $\mathbf{w} \in \mathbb{S}^{d-1}$. Let $\theta \geq\ ](\mathbf{w},\mathbf{w}^*)$ and suppose that $\theta \in [0,\pi/4]$. Then, for a sufficiently large constant C, there is a tester that given $\delta \in (0,1)$, θ, $\mathbf{w}$ and a set S of samples from $D_{\mathcal{X}}$ with size at least $\ \frac{d}{\theta}\log\frac{1}{\delta}\ ^C$, runs in time $\mathrm{poly}\ \left(\frac{1}{\theta}, d, \log\frac{1}{\delta}\right)$ and with probability $1-\delta$ satisfies the following specifications:*

- *If the distribution $D_{\mathcal{X}}$ is $\mathcal{N}(0, I_d)$, the tester accepts.*
- *If the tester accepts, then we have:*

$$\Pr_{\mathbf{x}\sim S}[\mathrm{sign}(\langle \mathbf{w}^*,\mathbf{x}\rangle) \neq \mathrm{sign}(\langle \mathbf{w},\mathbf{x}\rangle)] \leq O(\theta)$$

Proof. The testing algorithm does the following:

1. **Given:** Integer d, set $S \subset \mathbb{R}^d$, $\mathbf{w} \in \mathbb{S}^{d-1}$, $\theta \in (0, \pi/4]$ and $\delta \in (0,1)$.
2. Let $\mathrm{proj}_{\perp \mathbf{w}} : \mathbb{R}^d \to \mathbb{R}^{d-1}$ denote the operator that projects a vector $\mathbf{x} \in \mathbb{R}^d$ to it's projection into the $(d-1)$-dimensional subspace of $\mathbb{R}^d$ that is orthogonal to $\mathbf{w}$.
3. For i in $0, \pm 1, \cdots, \pm \frac{\sqrt{2 \log \frac{1}{\theta}}}{\theta}$
   (a) $S_i \leftarrow \{\mathbf{x} \in S : \langle \mathbf{w}, \mathbf{x} \rangle \in [i\theta, (i+1)\theta]\}$
   (b) If $\frac{|S_i|}{|S|} > 2\theta$, then reject.
   (c) If $\frac{1}{|S_i|} \sum_{\mathbf{x} \in S_i} (\mathrm{proj}_{\perp \mathbf{w}}(\mathbf{x}))(\mathrm{proj}_{\perp \mathbf{w}}(\mathbf{x}))^T - I_{(d-1)} \Big\|_{\mathrm{op}} > 0.1$, reject.
4. If $\frac{1}{|S|} \sum_{\mathbf{x} \in S} \mathbb{1}_{|\langle \mathbf{w}, \mathbf{x} \rangle| > \sqrt{2 \log \frac{1}{\theta}}} > 5\theta$, then reject.
5. If reached this step, accept.

If the tester accepts, then we have the following properties for some sufficiently large constant $C' > 0$. For the following, consider the vector $\mathbf{v} \in \mathbb{R}^d$ to be the vector that is perpendicular to $\mathbf{w}$, lies within the plane defined by $\mathbf{w}$ and $\mathbf{w}^*$ and $\langle \mathbf{v}, \mathbf{w}^* \rangle \leq 0$.

1. $\mathbb{P}_{\mathbf{x} \sim S}[|\langle \mathbf{w}, \mathbf{x} \rangle| \in [\theta i, \theta(i+1)]] \leq C'\theta$, for any $i \in \left\{0, \pm 1, \ldots, \pm \frac{1}{\theta} \sqrt{2 \log \frac{1}{\theta}}\right\}$.
2. $\mathbb{P}_{\mathbf{x} \sim S_i}\left[|\langle \mathbf{v}, \mathbf{x} \rangle| > \frac{\theta}{\tan \theta} \cdot i\right] \leq C'/i^2$, for any $i \in \left\{0, \pm 1, \ldots, \pm \frac{1}{\theta} \sqrt{2 \log \frac{1}{\theta}}\right\}$.
3. $\mathbb{P}_{\mathbf{x} \sim S}\left[|\langle \mathbf{w}, \mathbf{x} \rangle| \geq \sqrt{2 \log \frac{1}{\theta}}\right] \leq C'\theta$.

Then, for $k = \frac{1}{\theta} \sqrt{2 \log \frac{1}{\theta}}$ and $\mathrm{Strip}_i = \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{w}, \mathbf{x} \rangle| \in [\theta i, \theta(i+1)]\}$, we have that

$$\begin{aligned}
&\Pr_{\mathbf{x} \sim S}[\mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \neq \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)] \leq \\
&\sum_{i=-k}^{k} \mathbb{P}_{\mathbf{x} \sim S}[\mathbf{x} \in \mathrm{Strip}_i] \cdot \mathbb{P}_{\mathbf{x} \sim S}\left[|\langle \mathbf{v}, \mathbf{x} \rangle| > \frac{\theta}{\tan \theta} \cdot i \;\middle|\; \mathbf{x} \in \mathrm{Strip}_i\right] + \mathbb{P}_{\mathbf{x} \sim S}\left[|\langle \mathbf{w}, \mathbf{x} \rangle| \geq \sqrt{2 \log \frac{1}{\theta}}\right] \leq \\
&\sum_{i=-k}^{k} \frac{|S_i|}{|S|} \cdot \mathbb{P}_{\mathbf{x} \sim S_i}\left[|\langle \mathbf{v}, \mathbf{x} \rangle| > \frac{\theta}{\tan \theta} \cdot i\right] + C'\theta \leq (C')^2 \theta \cdot \left(1 + \sum_{i \neq 0} \frac{2}{i^2}\right) + C'\theta = O(\theta)
\end{aligned}$$

Now, suppose the distribution $D_{\mathcal{X}}$ is indeed the standard Gaussian $\mathcal{N}(0, I_d)$. We would like to show that our tester accepts with probability at least $1 - \delta$. Since $D = \mathcal{N}(0, I_d)$, we see that for $\mathbf{x} \sim D$ we have that $\mathbf{x} \cdot \mathbf{w}$ is distributed as $\mathcal{N}(0,1)$. This implies that

- For all $i \in 0, \pm 1, \cdots, \pm \frac{\sqrt{2 \log \frac{1}{\theta}}}{\theta}$ we have
  - $\Pr_{\mathbf{x} \sim \mathcal{N}(0, I_d)}[\langle \mathbf{w}, \mathbf{x} \rangle \in [i\theta, (i+1)\theta]] \leq \frac{1}{\sqrt{2\pi}}\theta$
  - $\Pr_{\mathbf{x} \sim \mathcal{N}(0, I_d)}[\langle \mathbf{w}, \mathbf{x} \rangle \in [i\theta, (i+1)\theta]] \geq \theta \cdot \min_{x \in \left[-\sqrt{2 \log \frac{1}{\theta}} - \theta, \sqrt{2 \log \frac{1}{\theta}} + \theta\right]} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \geq \frac{\theta^2}{10}$
- $\Pr_{\mathbf{x} \sim \mathcal{N}(0, I_d)}[\langle \mathbf{w}, \mathbf{x} \rangle \in [i\theta, (i+1)\theta]] \leq \frac{1}{\sqrt{2\pi}}\theta$
- $\Pr_{\mathbf{x} \sim \mathcal{N}(0, I_d)}\left[\langle \mathbf{w}, \mathbf{x} \rangle > 2\sqrt{\log \frac{1}{\theta}}\right] = \int_{2\sqrt{\log \frac{1}{\theta}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \leq \theta \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = \frac{\theta}{2}$

Therefore, via the standard Hoeffding bound, we see that for sufficiently large absolute constant C we have with probability at least $1 - \frac{\delta}{4}$ over the choice of S that

- For all $i \in \{0, \pm 1, \cdots, \pm \frac{\sqrt{2 \log \frac{1}{\theta}}}{\theta}\}$ we have
  - $\Pr_{\mathbf{x} \sim S}[\langle \mathbf{w}, \mathbf{x} \rangle \in [i\theta, (i+1)\theta]] \leq \theta$
  - $\Pr_{\mathbf{x} \sim S}[\langle \mathbf{w}, \mathbf{x} \rangle \in [i\theta, (i+1)\theta]] \geq \frac{\theta^2}{20}$
- $\Pr_{\mathbf{x} \sim S}\left[\langle \mathbf{w}, \mathbf{x} \rangle > 2\sqrt{\log \frac{1}{\theta}}\right] \leq \theta$
- $\Pr_{\mathbf{x} \sim S}\left[\langle \mathbf{w}, \mathbf{x} \rangle < -2\sqrt{\log \frac{1}{\theta}}\right] \leq \theta$

Finally, we would like to show that conditioned on the above, the probability of rejection in step (3b) is small.

Fact D.2. *Given a set $S \subset \mathbb{R}^{d-1}$ of i.i.d. samples from $\mathcal{N}(0, I_d)$, with probability at least $1 - \mathrm{poly}\left(\frac{|S|}{d}\right)$ we have*

$$\frac{1}{|S|} \left\| \sum_{\mathbf{x} \in S} \mathbb{1}_{\langle \mathbf{w}, \mathbf{x} \rangle \in [i\theta, (i+1)\theta]} \mathbf{x}\mathbf{x}^T - I_{(d-1)} \right\|_{op} \leq 0.1$$

Now, since each sample $\mathbf{x}_i$ is drawn i.i.d. from $\mathcal{N}(0, I_d)$, we have that $\langle \mathbf{w}, \mathbf{x}_i \rangle$ and $\mathrm{proj}_{\perp \mathbf{w}}(\mathbf{x}_i)$ are all independent from each other for all i. Since all the events we conditioned on depend on $\{\langle \mathbf{w}, \mathbf{x}_i \rangle\}$ we see that $\{\mathrm{proj}_{\perp \mathbf{w}}(\mathbf{x}_i)\}$ are still distributed as i.i.d. samples from $\mathcal{N}(0, I_{(d-1)})$.

Recall that one of the events we have already conditioned on is that $\Pr_{\mathbf{x} \sim S}[\langle \mathbf{w}, \mathbf{x} \rangle \in [i\theta, (i+1)\theta]] \geq \frac{\theta^2}{20}$ for all $i \in \{0, \pm 1, \cdots, \pm \frac{\sqrt{2 \log \frac{1}{\theta}}}{\theta}\}$. This allows us to lower bound by $\theta^2/20$ the ratio $|S_i|/|S|$. And since, as we described, for all these elements $\mathbf{x}_i$ the vectors $\mathrm{proj}_{\perp \mathbf{w}}(\mathbf{x}_i)$ are distributed as i.i.d. samples from $\mathcal{N}(0, I_{(d-1)})$, we can use Fact D.2 to conclude that for sufficiently large absolute constant C, when $|S| = \left(\frac{d}{\theta} \log \frac{1}{\delta}\right)^C$ we have with probability $1 - \frac{\delta}{4}$ for all $i \in \{0, \pm 1, \cdots, \pm \frac{\sqrt{2 \log \frac{1}{\theta}}}{\theta}\}$ that

$$\frac{1}{|S_i|} \left\| \sum_{\mathbf{x} \in S_i} (\mathrm{proj}_{\perp \mathbf{w}}(\mathbf{x}))(\mathrm{proj}_{\perp \mathbf{w}}(\mathbf{x}))^T - I_{(d-1)} \right\|_{\mathrm{op}} \leq 0.1$$

Overall, this allows us to conclude that with probability at least $1 - \delta$ the tester accepts. □

We now present the proof of Theorem 5.3.

In the proof of Theorem 5.1, when the target distribution is the standard Gaussian in d dimensions, we may apply Proposition D.1 (and run the corresponding tester), instead of Proposition 4.4, in order to ensure that our list will contain a vector $\mathbf{w}$ with

$$\begin{aligned} \mathbb{P}_{D_{XY}}[y \neq \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)] &\leq \mathbb{P}_{D_{XY}}[y \neq \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)] + \mathbb{P}_{D_{XY}}[\mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle) \neq \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)] \\ &\leq \mathsf{opt} + O(\theta) \end{aligned}$$

where $\measuredangle(\mathbf{w}, \mathbf{w}^*) \leq \theta := c_2 \sigma$ and σ is such that $c_1 \sigma - \Theta(\epsilon) \leq \mathsf{opt} \leq c_1 \sigma$, which gives the desired $O(\mathsf{opt}) + \epsilon$ bound. To get the value of σ with the desired property, we once again sparsified the space $(0, 1]$ of possible values for σ, this time up to accuracy $\Theta(\epsilon)$.

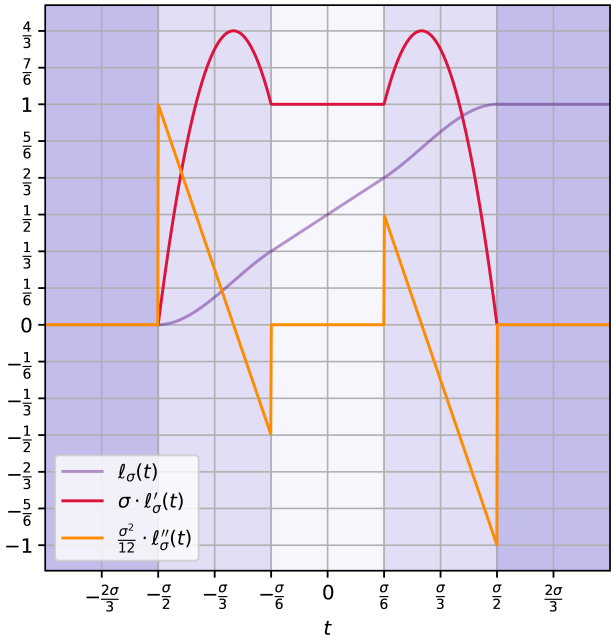

Figure 4: Figure illustrating the (normalized) first two derivatives of the function ℓ_σ used to define the non convex surrogate loss $\mathcal{L}_\sigma$. The normalization is appropriate since ℓ'_σ and ℓ''_σ are homogeneous in $1/\sigma$ and $1/\sigma^2$ respectively. In particular, we see that $\ell'_\sigma \leq \Theta(1/\sigma)$ and $|\ell''_\sigma| \leq \Theta(1/\sigma^2)$ everywhere.