Self-Directed Linear Classification

Ilias Diakonikolas ILIAS@CS.WISC.EDU
UW Madison

Vasilis Kontonis KONTONIS@WISC.EDU
UW Madison

Christos Tzamos TZAMOS@WISC.EDU
UW Madison, National and Kapodistrian University of Athens (NKUA)

Nikos Zarifis ZARIFIS@WISC.EDU
UW Madison

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

In online classification, a learner is presented with a sequence of examples and aims to predict their labels in an online fashion so as to minimize the total number of mistakes. In the self-directed variant, the learner knows in advance the pool of examples and can adaptively choose the order in which predictions are made. Here we study the power of choosing the prediction order and establish the first strong separation between worst-order and random-order learning for the fundamental task of linear classification. Prior to our work, such a separation was known only for very restricted concept classes, e.g., one-dimensional thresholds or axis-aligned rectangles.

We present two main results. If X is a dataset of n points drawn uniformly at random from the d-dimensional unit sphere, we design an efficient self-directed learner that makes $O(d \log \log(n))$ mistakes and classifies the entire dataset. If X is an arbitrary d-dimensional dataset of size n, we design an efficient self-directed learner that predicts the labels of 99% of the points in X with mistake bound independent of n. In contrast, under a worst- or random-ordering, the number of mistakes must be at least $\Omega(d \log n)$, even when the points are drawn uniformly from the unit sphere and the learner only needs to predict the labels for 1% of them.

Keywords: Online Learning, Self-Directed Learning, Halfspaces, Mistake Bound

1. Introduction

Online prediction has a rich history going back to the pioneering works of Robbins (1951); Hannan (1957); Blackwell et al. (1954). In the online setting, the learner aims to resolve a prediction task by learning a hypothesis from a sequence of examples one at a time. The goal is to minimize the total number of incorrect predictions (aka mistake bound) given the knowledge of the correct answers to previously queried examples (Littlestone, 1988, 1989; Blum, 1990; Littlestone and Warmuth, 1994; Maass and Turan, 1994). A standard, worst-case assumption is that an adversary controls the sequence of examples and/or labels. In this worst-case setting, a wide range of algorithms based on exponential reweighting (Vovk, 1990; Littlestone and Warmuth, 1994; Freund and Schapire, 1997; Vovk, 1995; Cesa-Bianchi and Lugosi, 2006) and online convex optimization (Hazan, 2016; Orabona, 2019) have been developed. Motivated by the fact that in many applications the sequence of examples is not adversarial, the problem of online prediction has also been studied in more benign settings, such as assuming that the examples are given to the learner by a teacher (who knows the ground-truth hypothesis) (Goldman and Mathias, 1993; Mathias, 1997; Doliwa et al., 2010; Mansouri et al., 2022)

or making regularity assumptions about the sequence of examples (Rakhlin and Sridharan, 2013; Jadbabaie et al., 2015). In this work, we study the model of Self-Directed learning (Goldman et al., 1993; Goldman and Sloan, 1994), where the learning algorithm can choose which example to label next. Self-directed learning has many applications. For example, in direct marketing (Ni and Ling, 2011), the learner must study customers' characteristics and needs and adaptively select examples (customers) to market their products. Moreover, it is related to curriculum learning — proposed in the influential work of Bengio et al. (2009) — where the training examples are sorted from "easier" to "harder" in order to help the model learn faster (see Section 1.2 for more details).

Here we consider the mistake-bound model (Littlestone, 1988, 1989) in the realizable setting, where the labels revealed to the learner in each round are consistent with a ground-truth classifier f that belongs to a known concept class $\mathcal{C}$ fixed before the online learning phase starts. We now formally define the self-directed online prediction model (Goldman and Sloan, 1994) and its random- and worst-order variants that we consider in this work.

Definition 1 (Self-Directed Online Learning (Goldman and Sloan, 1994)) *Let $f \in \mathcal{C}$ be an unknown target concept from some concept class $\mathcal{C}$ of boolean functions from $\mathbb{R}^d$ to $\{\pm 1\}$ and let $X = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ be a subset of $n \in \mathbb{N}$ points in $\mathbb{R}^d$. The learner has access to the full set of (unlabeled) points X.*
Until the labels of all examples of X have been predicted:

- *The learning algorithm picks a point $\mathbf{x} \in X$ making a prediction $z \in \{\pm 1\}$ about its label.*
- *The true label $f(\mathbf{x})$ of $\mathbf{x}$ is revealed to the learning algorithm.*

We say that the learner makes M mistakes to label X if, with probability at least 99%, it holds that the number of incorrect predictions of the learner is at most M.

In this work, we investigate the power of the above self-directed online learning setting compared to the worst- and random-order settings.

Remark 2 (Random-order and Worst-order Online Learning) *We shall refer to the setting where the point $\mathbf{x}$ during the training phase is picked uniformly at random (without replacement) from the unlabeled data X as **random-order** learning. Moreover, we shall refer to the setting where the next example $\mathbf{x}$ is chosen by an adversary as **worst-order** learning.*

Observe that in Definition 1 the learner must predict labels for *all examples* of the dataset X. We will also consider weaker learning notions allowing the learner to avoid labeling a fraction of examples in X. As we will see, these weaker learning notions are especially useful when dealing with an unstructured dataset X, containing potentially ambiguous or adversarial examples that may be inherently hard to predict.

Remark 3 (Perfect, Strong, and Weak Self-Directed Learners) We will refer to the setting of Definition 1, where the learner predicts the labels of all examples in X, as **perfect** self-directed learning. Moreover, we refer to the setting where the learning algorithm provides labels for at least $(1-\epsilon)$-fraction of X for every $\epsilon \in (0, 1]$ as **strong** self-directed learning. Finally, in **weak** self-directed learning, the learner predicts the labels of some fixed fraction of X, say 1%.

Online Linear Classification We focus on the fundamental setting of online linear classification that dates back to Rosenblatt's perceptron (Rosenblatt, 1958). A (homogeneous) *halfspace* or Linear Threshold Function (LTF) is a Boolean-valued function $f : \mathbb{R}^d \mapsto \{\pm 1\}$ of the form $f(\mathbf{x}) = \text{sign}(\mathbf{w}^* \cdot \mathbf{x})$, for a vector $\mathbf{w}^* \in \mathbb{R}^d$ (known as the weight vector). Halfspaces are a central class of Boolean functions in several areas of computer science, including complexity theory, learning theory, and optimization Novikoff (1962); Yao (1990); Goldmann et al. (1992); Freund and Schapire (1997); Vapnik (1998); Shawe-Taylor and Cristianini (2000); Rubinfeld (2006). In (realizable) online linear classification, the halving algorithm — first appeared in Barzdiņš and Freivald (1972) and further analyzed and refined in Mitchell (1982); Angluin (1987); Littlestone (1988) — makes $O(d \log n)$ mistakes for perfect classification of n points in d dimensions. This mistake bound and in particular its dependence on the dataset size n is known to be information-theoretically optimal for linear classification (Littlestone, 1988) in the worst-order online learning setting. We show (see Proposition 17) that, even when the order of examples is random, $\Omega(d \log n)$ mistakes are required. In fact, this is true even when the n points of X are drawn uniformly from the d-dimensional unit-sphere $\mathbb{S}_d$ and the learner only needs to predict the labels of 1% of them (weak-learning).

Improved Mistake Bounds via Self-Directed Learning In the self-directed learning setting, it is known (Goldman and Sloan, 1994) that for one-dimensional threshold functions one mistake suffices — significantly improving over the $\Theta(\log n)$ mistake-bound for the same class in the worst-order setting (Littlestone, 1988, 1989). Similar improvements in self-directed learning have also been shown for other simple concept classes such as monotone-monomials and axis-aligned rectangles (see Ben-David et al. (1997) for a discussion on the gaps between self-directed learning and worst-order learning). In this work, we study whether self-directed learning can improve the number of mistakes to learn more complicated, high-dimensional concept classes focusing on the fundamental class of d-dimensional LTFs. We remark that, beyond the $O(d \log n)$ mistake-bound given by the halving algorithm, no other mistake bounds are known for self-directed learning of halfspaces, even when the dataset X is assumed to be structured, e.g., the n examples of X are drawn uniformly at random from the d-dimensional unit sphere $\mathbb{S}_d$. We ask the following natural question.

Can self-directed learners bypass the $\Omega(d \log n)$ mistake barrier of worst- and random-order learning for d-dimensional LTFs?

We give a positive answer to the above question showing that allowing the learner to choose the order of examples significantly improves the mistake bound under both structured and arbitrary datasets.

1.1. Our Results and Techniques

Our first result assumes that the dataset X is structured: its n examples are drawn uniformly at random from the unit sphere $\mathbb{S}_d$. In this case, we give a self-directed learning algorithm that only makes $O(d \log \log n)$ mistakes to predict the labels of all examples in X (**perfect learning**), establishing an exponential improvement over the $\Omega(d \log n)$ mistake bound of worst- and random-order learners. Since it is known (Ben-David et al., 1997) that even self-directed learners must make $\Omega(d)$-mistakes, our $O(d \log \log n)$ is close to best-possible.

Informal Theorem 4 (Perfect Self-Directed Learner on $\mathbb{S}_d$) *There exists a perfect self-directed learner for halfspaces that makes $O(d \log \log n)$ mistakes to classify a set of n points drawn uniformly from $\mathbb{S}_d$.*

Our learner maintains a halfspace hypothesis and at each round predicts the label of the example with the largest margin from the current halfspace. At a high level, this corresponds to choosing the example for which the current hypothesis is most-confident in an "easy examples first" manner. When the prediction on such an example is incorrect, we use the so-called margin-perceptron update rule used in Dunagan and Vempala (2004) in the context of linear programming. We show that the margin-perceptron update used on examples for which the classifier is very confident (large-margin) – but made a mistake – converges super-linearly to the ground-truth halfspace $\mathbf{w}^*$. For a more detailed overview, we refer to Section 2.1.

Our second result is a self-directed learner for arbitrary datasets. In this case, we give a **strong** self-directed learner that makes $\mathrm{poly}(d)$ mistakes and labels 99% of X. Recall that, in contrast, any worst- or random-order learner makes $\Omega(d \log n)$ mistakes to label 1% of the dataset (even when the data are drawn uniformly from the unit sphere).

Informal Theorem 5 (Strong Self-Directed Learner for Arbitrary Data) *There exists a strong, self-directed learner for halfspaces that, given an arbitrary set X of n points in d dimensions, makes $\mathrm{poly}(d)$ mistakes to classify 99% of X.*

At a high level, similarly to our algorithm for uniformly spherical data, our self-directed learner for arbitrary data again picks the examples with the largest margin from the current hypothesis and uses the margin-perceptron update rule of Dunagan and Vempala (2004). The second ingredient of our learner is the Forster transform (Forster, 2002) that puts the examples in (approximately) Radially Isotropic Position, i.e., make X isotropic and also normalize all points so that they lie on the unit sphere; see Definition 13. Recent works (Hardt and Moitra (2013); Artstein-Avidan et al. (2020); Diakonikolas et al. (2022)) have provided efficient algorithms to compute the Forster transform. When the dataset is in Radially Isotropic Position, one can show that it satisfies a "soft-margin" condition (i.e., that a non-trivial fraction of X has non-trivial margin with every halfspace). We use this property to first obtain a *weak learner* that does $O(d \log d)$ mistakes to label an $\Omega(1/d)$-fraction of X. We then use a generic boosting approach to transform this weak learner into a *strong learner* that does $\mathrm{poly}(d)$ mistakes to label 99% of X, see Section 3.1 and Appendix D.2.

1.2. Related Work

Related to the setting of self-directed learning is active learning (Cohn et al., 1994), where the learner has access to a large pool of unlabeled examples and chooses the "most informative" to ask for their labels. The goal is to find a classifier with good generalization while minimizing the number of label queries. There is a long line of research on active linear classification in the distribution-specific setting (e.g., under the uniform distribution on the unit sphere) (Dasgupta et al., 2005; Hanneke, 2011; Balcan and Urner, 2016). We remark that our goal of minimizing the number of mistakes is orthogonal to that of active learning: at a high-level, our algorithms pick the examples for which the current hypothesis is most confident ("easiest examples") while in active learning one typically asks for the labels of the "hardest examples", e.g., those with the smallest margin with respect to the current guess (see, e.g., (Awasthi et al., 2015, 2016; Zhang et al., 2020)).

In deep learning, stochastic gradient descent typically trains models by considering the examples in a random order. In the influential work of Bengio et al. (2009) the authors proposed curriculum learning: training machine learning models in a "meaningful order" – from easy examples to harder ones. There is a long line of research (see the surveys (Hacohen and Weinshall, 2019; Wang et al.,

2021; Soviany et al., 2022) and references therein) giving empirical evidence that curriculum learning provides significant benefits in convergence speed and generalization over training with random order. Our results provide theoretical evidence that ordering the examples from easier to harder significantly reduces the mistakes made by the learner.

2. Self-Directed Learning on $\mathbb{S}_d$

In this section, we present our self-directed learning algorithm for datasets uniformly distributed on the unit sphere. We first state the formal version of Theorem 4.

Theorem 6 (Perfect Self-Directed Learner on $\mathbb{S}_d$) *Let $\delta \in (0, 1/2]$ and let n be larger than some sufficiently large universal constant. Let X be a set of n i.i.d. samples from $\mathbb{S}_d$ with true labels given by a homogeneous halfspace, $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. There exists a self-directed classifier that makes $O(d \log\log n \ \log(1/\delta))$ mistakes, runs in time $\mathrm{poly}(d, n)$ and classifies all points of X with probability at least $1 - \delta$.*

2.1. Roadmap of the Proof of Theorem 6

The first ingredient of our algorithm is an adaptive way to pick examples: at every step the learner predicts the labels of examples for which the current hypothesis is most confident. The intuition behind this choice is that for those examples the hypothesis is more often correct and, when it predicts incorrectly, they can be used to improve it significantly. The second ingredient is the "margin-perceptron" algorithm of Dunagan and Vempala (2004). The margin-perceptron iteration is a variant of the standard perceptron update rule that scales the update with the signed margin of the example, i.e., given an example $\mathbf{x}$ that the current hypothesis $\mathbf{w}^{(t)}$ predicts incorrectly (i.e., $\mathrm{sign}(\mathbf{w}^{(t)} \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$), we update $\mathbf{w}^{(t)}$ as follows:

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - (\mathbf{w}^{(t)} \cdot \mathbf{x})\mathbf{x} \tag{1}$$

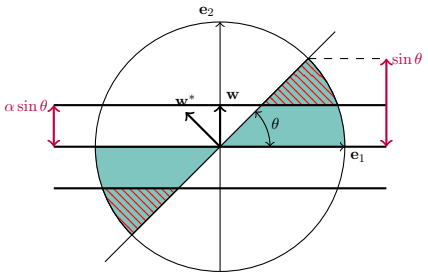

Figure 1: The probability that the maximum-margin mistake is has margin larger than $\alpha \sin\theta$ corresponds to the probability of the shaded subset of the disagreement region (shown in cyan).

Contrary to the standard perceptron update (i.e., $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \mathbf{x}$), the perceptron update rule of Equation (1) does not rely on improving the correlation with the target vector $\mathbf{w}^*$ but on *decreasing the norm* $\|\mathbf{w}^{(t)}\|_2$ and *not decreasing the correlation* with $\mathbf{w}^*$. We show that as long as the margin $|\mathbf{w}^{(t)} \cdot \mathbf{x}|$ is large the margin-perceptron update of Equation (1) will significantly reduce the angle $\theta(\mathbf{w}^{(t)}, \mathbf{w}^*)$ between $\mathbf{w}^{(t)}$ and $\mathbf{w}^*$.

Reducing $\tan(\theta(\mathbf{w}^{(t)}, \mathbf{w}^*))$ via Margin-Perceptron We first observe that given an example that $\mathbf{w}^{(t)}$ mispredicts, its maximum possible margin with the current guess $\mathbf{w}^{(t)}$ is equal to $\sin(\theta(\mathbf{w}^{(t)}, \mathbf{w}^*))$, see Figure 1. We show, see Lemma 8, that when $|\mathbf{w}^{(t)} \cdot \mathbf{x}| \geq r \sin\theta(\mathbf{w}^{(t)}, \mathbf{w}^*))$, the margin-perceptron update reduces $\tan(\theta(\mathbf{w}^{(t)}, \mathbf{w}^*))$ multiplicatively:

$$\tan^2(\theta(\mathbf{w}^{(t+1)}, \mathbf{w}^*)) \leq (1 - r^2) \tan^2(\theta(\mathbf{w}^{(t)}, \mathbf{w}^*)) \,. \tag{2}$$

Observe that, the closer r is to 1, (i.e., the larger the margin) the faster $\theta^{(t)} = \theta(\mathbf{w}^{(t)}, \mathbf{w}^*)$ will converge to 0. Using the fact that the distribution is uniform on the sphere, the probability that some halfspace with normal vector $\mathbf{u}$ disagrees with the ground-truth halfspace $\mathbf{w}^*$ is equal to $\theta(\mathbf{u}, \mathbf{w}^*)/\pi$. Thus, achieving $\theta(\mathbf{w}^{(t)}, \mathbf{w}^*) \leq O(1/n)$, implies, that on expectation, the number of mistakes on a sequence of n i.i.d. examples from $\mathbb{S}_d$ is going to $O(1)$, see Lemma 27. Therefore, in what follows, our goal will be to make $\theta^{(t)}$ smaller than $O(1/n)$. [1]

Achieving Large Margin on $\mathbb{S}_d$ Given the current hypothesis with normal vector $\mathbf{w}^{(t)}$, a uniformly random point on the d-dimensional sphere has margin roughly $\Omega(1/\sqrt{d})$. Given that the point falls in the disagreement region of $\mathbf{w}^{(t)}$ and $\mathbf{w}^*$ the margin can be shown to be roughly $\Omega(\sin\theta^{(t)}/\sqrt{d})$. Therefore, in each round, Equation (2) implies the angle θ (recall that for small θ it holds that $\theta \approx \tan\theta$) is going to decrease roughly by a factor of $\sqrt{1-\Omega(1/d)}$:

$$\theta^{(t+1)} \leq \sqrt{1-\Omega(1/d)}\,\theta^{(t)}\,.$$

In order to make $\theta^{(t)} \leq O(1/n)$ the above iteration requires roughly $d\log n$ updates which only implies a mistake bound of $O(d\log n)$. This is where the fact that at every iteration t we choose the example with maximum margin with respect to $\mathbf{w}^{(t)}$ comes into play. We show that for n examples distributed uniformly on the sphere the maximum-margin mistake with respect to $\mathbf{w}^{(t)}$ has margin roughly $\Omega(\sqrt{\log(n\theta^{(t)})/d})\ \sin(\theta^{(t)})$ when $n\theta^{(t)} \leq e^{O(d)}$ and $(1-(1/(n\theta^{(t)}))^{2/d})\ \sin(\theta^{(t)})$ when $n\theta^{(t)}$ is larger than $e^{O(d)}$. We show the following proposition; for the formal statements, see Proposition 20 and Lemma 21.

Proposition 7 (Informal: Max-Margin in the Disagreement Region) *Let $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d$ be unit vectors with angle $\theta(\mathbf{v}, \mathbf{u}) = \theta$. Let C be the indicator of the disagreement region of the two homogeneous halfspaces defined by $\mathbf{v}, \mathbf{u}$, i.e., $C = \mathbb{1}\{(\mathbf{v}\cdot\mathbf{x})(\mathbf{u}\cdot\mathbf{x}) \leq 0\}$. Let X be a dataset with n i.i.d. samples from $\mathbb{S}_d$. Then with probability at least $2/3$:*

1. *If $n\theta \leq e^d$, it holds that $\max_{\mathbf{x}\in X\cap C} |\mathbf{u}\cdot\mathbf{x}| \geq \Omega(\sqrt{\log(n\theta)/d})\sin\theta$.*
2. *Otherwise, $\max_{\mathbf{x}\in X\cap C} |\mathbf{u}\cdot\mathbf{x}| \geq (1-(1/(n\theta))^{2/d})\sin\theta$.*

We observe that when $n\theta \to \infty$ the maximum-margin over the n samples converges to its maximum value of $\sin\theta$. Since the analysis of the general case turns out to be similar to the case of $n\theta \leq e^d$, for simplicity, in this overview we will focus on this case.

(Stochastic) Super-Linear Convergence We observe that by taking the maximum-margin sample and using Proposition 7 we improved the decay of the angle to roughly:

$$\theta^{(t+1)} \leq \sqrt{1-\frac{\log(n\theta^{(t)})}{d}}\,\theta^{(t)} \leq e^{-\log(n\theta^{(t)})/(2d)}\,\theta^{(t)} = (\theta^{(t)})^{1-1/(2d)}(1/n)^{1/(2d)}\,. \tag{3}$$

Therefore, the angle after an update on the maximum-margin mistake is *the weighted geometric mean between $\theta^{(t)}$ and $1/n$* with weights $1-1/(2d)$ and $1/(2d)$. It is not hard to show that after $m = O(d\log\log n)$ such updates we have $\theta^{(m)} \leq O(1/n)$, achieving our goal. One issue is that Proposition 7 only gives "good" probability, i.e., 2/3, that such a large-margin update will happen. In Lemma 9, we show that increasing the iterations by a constant factor is enough to show that $\theta^{(m)}$ will be $O(1/n)$ with good probability.

1. More, precisely it suffices to have $\theta^{(t)} \leq O(d\log\log n/n)$ as this would imply an $O(d\log\log n)$ mistake bound on a sequence of n data.

Dealing with the Dependencies In our discussion so far we have ignored the fact that if at some step t, the algorithm searches over the whole dataset X in order to find the example $\mathbf{x}$ with the largest margin with respect to the guess $\mathbf{w}^{(t)}$, in the next step the remaining points of X are no-longer i.i.d. samples from $\mathbb{S}_d$ and therefore, many of our claims using the independence of the samples (e.g., Proposition 7) no longer work. We handle this issue by only considering a large-enough subset of examples in each step of the margin-perceptron update, i.e., instead of selecting the point of maximum-margin over the whole dataset, we select the maximum-margin example of a random subset, see Steps 3,4 in Algorithm 1, inside which we only perform a single margin-perceptron update. Since the total number of margin-perceptron updates required is only $O(d \log \log n)$ we split the dataset into $k = O(d \log \log n)$ random subsets of equal size; therefore assuming that n is larger than roughly $\Omega(d \log \log d)$, each bucket will have enough samples to guarantee that a margin-perceptron update with large margin will happen with good probability.

Finally, as we use examples of X to update the guess $\mathbf{w}^{(t)}$, when we reach the target angle, say $\theta^{(t)} \leq O(1/n)$, we cannot guarantee that $\mathbf{w}^{(t)}$ will do few mistakes on the samples that we used to train it (as $\mathbf{w}^{(t)}$ depends on those samples). To avoid this issue we split the initial dataset into two random subsets of equal size A and B and then train a linear classifier for each part. In the final step, we use the linear classifier trained on A to label the dataset B (that was not used during its training) and the linear classifier trained on B to label the examples of A, see Step 5 in Algorithm 1.

Input: An initialization $\mathbf{w}$. **Output:** A sequence of labeled data $(\mathbf{x}^{(t)}, z^{(t)})$.

1. Initialize guesses $\mathbf{w}^{(0)} \leftarrow \mathbf{w}, \quad \mathbf{v}^{(0)} \leftarrow \mathbf{w}$.
2. Initialize the set of unlabeled data $U \leftarrow X$.
3. Split U in $2k$ sets $U_1, \ldots, U_{2k}$.
4. For $t = 1, \ldots, k$:
   $\mathbf{w}^{(t)} \leftarrow$ MARGIN-PERCEPTRON(U_t, $\mathbf{w}^{(t-1)}$), $\mathbf{v}^{(t)} \leftarrow$ MARGIN-PERCEPTRON(U_{k+t}, $\mathbf{v}^{(t-1)}$).
5. For $t = 1, \ldots, k$:
   Label points of U_{k+t} with $\mathbf{w}^{(k)}$ and label points of U_t with $\mathbf{v}^{(k)}$.

MARGIN-PERCEPTRON($U, \mathbf{w}$)
Input: An initialization $\mathbf{w}$ and a set of points U. **Output:** A vector $\mathbf{w}'$.

1. Obtain U' by sorting the points of U in decreasing order of margin from $\mathbf{w}$, i.e., $|\mathbf{x}^{(i+1)} \cdot \mathbf{w}| \leq |\mathbf{x}^{(i)} \cdot \mathbf{w}|$.
2. For $\mathbf{x} \in U'$:
   (a) Predict the label of $\mathbf{x}$ with $\mathbf{w}$.
   (b) If the prediction is incorrect, exit the loop and return $\mathbf{w}' \leftarrow \mathbf{w} - (\mathbf{w} \cdot \mathbf{x})\mathbf{x}$.

Algorithm 1: *Self-Directed Learning on* $\mathbb{S}_d$

2.2. Proof of Theorem 6

We first give a proof sketch of Proposition 7 showing that maximum-margin mistakes will have margin significantly larger than the margin of an "average" mistake. In the following sketch we only show the first case of Proposition 7; for the full proof we refer to Appendix C.

Proof sketch of Proposition 7 We observe that by the rotational symmetry of the uniform distribution on the sphere, the probability that a sample $\mathbf{x}$ falls in the disagreement region C is exactly θ/π, see Figure 1. Therefore, on expectation, out of the n samples that we draw from $\mathbb{S}_d$, $n\theta/\pi$ fall in C. Since it is not hard to show that with high probability $\Omega(n\theta)$ samples fall in C (see Appendix C); for this sketch we assume that this is case with probability 1.

We now show that conditionally on observing m samples in the disagreement region C, the maximum-margin has strong *anti-concentration*. Denote by $\mathbb{S}_d(C)$ the conditional distribution on the disagreement region C. For any $\alpha \in [0,1]$, it holds:

$$\mathop{\mathbf{Pr}}_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}\sim\mathbb{S}_d(C)}\left[\max_{i=1,\ldots,m}|\mathbf{u}\cdot\mathbf{x}^{(i)}| \leq \alpha\sin(\theta/2)\right] \leq \exp\left(-m\,(1-\alpha^2)^{d/2-1}/2\right).$$

To simplify notation, set $\Delta = \alpha\sin(\theta/2)$. We first compute the probability that a single sample in C has $|\mathbf{u}\cdot\mathbf{x}| \geq \Delta$. By the symmetry of the set C and the uniform distribution on the sphere it holds $\mathbf{Pr}_{\mathbf{x}\sim\mathbb{S}_d}[\mathbf{x}\in C, |\mathbf{u}\cdot\mathbf{x}|\geq\Delta] = 2\,\mathbf{Pr}_{\mathbf{x}\sim\mathbb{S}_d}[\mathbf{x}\in C, \mathbf{u}\cdot\mathbf{x}\geq\Delta] = 2\,\mathbf{Pr}_{\mathbf{x}\sim\mathbb{S}_d}[E_1]$, where $E_1 = \{\mathbf{x} : \mathbf{x}\in C, \mathbf{u}\cdot\mathbf{x}\geq\Delta\}$ (E_1 corresponds to the upper shaded cell in Figure 1). Assume, without loss of generality that $\mathbf{u} = \mathbf{e}_2$ and $\mathbf{v} = -\sin\theta\mathbf{e}_1 + \cos\theta\mathbf{e}_2$. Using polar coordinates $\mathbf{x}_1 = r\cos\phi$, $\mathbf{x}_2 = r\sin\phi$ we have that (see Figure 1 and Appendix C), $E_1 = \{(r,\phi) : \alpha\leq r\leq 1, r\sin\phi\geq\alpha\sin\theta, \phi\leq\theta\}$. The set E_1 has coupled constraints (i.e., constraints that depend on both r,ϕ). To avoid this we show that the set $E_2 = \{(r,\phi) : \alpha\leq r\leq 1, \theta/2\leq\phi\leq\theta\}$ that has decoupled constraints is a subset of E_1. The 2-dimensional marginal of the uniform distribution on $\mathbb{S}_d$ has density $\frac{d-2}{2\pi}(1-r^2)^{d/2-2}r$ (in polar coordinates).

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathbb{S}_d}[E_2] = \frac{d-2}{2\pi}\int_\alpha^1\int_{\theta/2}^{\theta}(1-r^2)^{d/2-2}\,r\,d\phi dr = \frac{\theta}{4\pi}(1-\alpha^2)^{d/2-1}.$$

Recall that by, the symmetry of $\mathbb{S}_d$ we directly obtain that $\mathbf{Pr}_{\mathbf{x}\sim\mathbb{S}_d}[C] = \theta/\pi$. We conclude that the conditional probability $\mathbf{Pr}_{\mathbf{x}\sim\mathbb{S}_d(C)}[|\mathbf{u}\cdot\mathbf{x}|\geq\Delta] \geq (1/2)(1-\alpha^2/d)^{d/2-1}$. We can now bound by above the probability that the maximum-margin of m independent samples from $\mathbb{S}_d(C)$ is small.

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}_1,\ldots,\mathbf{x}_m\sim\mathbb{S}_d(C)}\left[\max_{i=1,\ldots,m}|\mathbf{u}\cdot\mathbf{x}_i|\leq\Delta\right] &= (1-\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathbb{S}_d(C)}[|\mathbf{u}\cdot\mathbf{x}|\geq\Delta])^m \\ &\leq \exp\left(-m\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathbb{S}_d(C)}[|\mathbf{u}\cdot\mathbf{x}|\geq\Delta]\right) \leq \exp\left(-m\,(1-\alpha^2)^{d/2-1}/2\right),\end{aligned}$$

where, for the first inequality, we used the fact $e^x \geq 1+x$. Using $m = \Omega(n\theta)$ (since we know that roughly $n\theta$ examples land in the disagreement region C) and $\alpha = \Omega(\sqrt{\log m}/\sqrt{d})$ we obtain the result (for the first case of Proposition 7). ■

We prove the following lemma, showing on each mistake the margin-perceptron has good probability of significantly (super-linearly) decreasing $\tan\theta$. We require that the current guess $\mathbf{w}$ is not exactly orthogonal with the target $\mathbf{w}^*$ (notice the assumption $1/\cos\theta \leq O(\zeta)$). We show that it is not hard to obtain such an initialization.

Lemma 8 (Stochastic Multiplicative Decay of $\tan\theta$) *Let $\mathbf{w} \in \mathbb{R}^d$ and let $\theta(\mathbf{w}^*, \mathbf{w}) = \theta \in [0, \pi/2)$ and assume that for some $\zeta > 0$, $1/\cos\theta \leq \zeta/(12\pi)$. Let C be the indicator of the disagreement region of $\mathbf{w}^*, \mathbf{w}$, i.e., $C = \mathbb{1}\{(\mathbf{w}^* \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x}) \leq 0\}$. Let $X = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ be a sample set drawn from $\mathbb{S}_d$ with n larger than a sufficiently large constant and let $\widehat{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x} \in X} |\mathbf{w} \cdot \mathbf{x}| \mathbb{1}\{\mathbf{x} \in C\}$. Let $\mathbf{w}' = \mathbf{w} - (\mathbf{w} \cdot \widehat{\mathbf{x}})\widehat{\mathbf{x}}$ and $\theta' = \theta(\mathbf{w}', \mathbf{w}^*)$.*

1. *We have that $\tan^2(\theta') \leq \tan^2(\theta)$. (Monotonicity)*
2. *With probability at least $2/3$, it holds $\tan(\theta') \leq \tan^{1-1/(8d)}(\theta)(\zeta/n)^{1/(8d)}$.*

Proof sketch First, we assume that we perform an update on an example where $(\mathbf{x} \cdot \mathbf{w})(\mathbf{x} \cdot \mathbf{w}^*) < 0$ (i.e., $\mathbf{w}$ makes a mistake on $\mathbf{x}$) and the margin of $\mathbf{x}$ is large: $|\mathbf{w} \cdot \mathbf{x}| \geq r \sin\theta$. Assuming that $\theta \in [0, \pi/2)$ we can show that $\tan\theta$ decreases multiplicatively. We first observe that the correlation with $\mathbf{w}^*$ does not decrease, $\mathbf{w}' \cdot \mathbf{w}^* = (\mathbf{w} - (\mathbf{w} \cdot \mathbf{x})\mathbf{x}) \cdot \mathbf{w}^* = \mathbf{w} \cdot \mathbf{w}^* - (\mathbf{w} \cdot \mathbf{x})(\mathbf{x} \cdot \mathbf{w}^*) \geq \mathbf{w} \cdot \mathbf{w}^*$, where we used that the hypothesis $\mathbf{w}$ disagrees with the ground-truth $\mathbf{w}^*$ on $\mathbf{x}$, i.e., $(\mathbf{w} \cdot \mathbf{x})(\mathbf{x} \cdot \mathbf{w}^*) \leq 0$. Furthermore, since $\mathbf{w} \cdot \mathbf{w}^* \geq 0$ (by the assumption that $\theta \in [0, \pi/2)$) we also have that $(\mathbf{w}' \cdot \mathbf{w}^*)^2 \geq (\mathbf{w} \cdot \mathbf{w}^*)^2$. We next show that the norm of $\mathbf{w}'$ decreases multiplicatively. We have that $\|\mathbf{w}'\|_2^2 = \|(\mathbf{w} - (\mathbf{w} \cdot \mathbf{x})\mathbf{x})\|_2^2 = \|\mathbf{w}\|_2^2 - (\mathbf{w} \cdot \mathbf{x})^2 \leq \|\mathbf{w}\|_2^2(1 - r^2 \sin^2\theta)$. Using (twice) the trigonometric identity $\tan^2\phi = (1/\cos^2\phi) - 1$, we show that $\tan^2\theta$ decreases by a factor of $(1 - r^2)$:

$$\tan^2\theta' = \frac{\|\mathbf{w}'\|_2^2}{(\mathbf{w}' \cdot \mathbf{w}^*)^2} - 1 \leq \frac{\|\mathbf{w}\|_2^2(1 - r^2\sin^2\theta)}{(\mathbf{w} \cdot \mathbf{w}^*)^2} - 1 = \frac{1 - r^2\sin^2\theta}{\cos^2\theta} - 1 = (1 - r^2)\tan^2\theta \,.$$

We note that from the above derivation, we have that whenever we use the update rule on mistakes, it holds that $\tan(\theta') \leq \tan(\theta)$. We show that with constant probability, the decrease is significantly larger. For this sketch, we assume that $n\theta$ is not exponentially large and refer to Appendix C for details. From Proposition 7, we have that with probability at least $2/3$, it holds $|\mathbf{w} \cdot \widehat{\mathbf{x}}| \geq \Omega(\sqrt{\log(n\theta)/d}) \sin\theta$. Simplifying the expression for $\tan\theta'$ similarly to Equation (3), we obtain the result. ■

We now show that given a non-increasing stochastic process ξ_t that has good probability to decrease at a super-linear rate, then after $T = O(\log\log(1/\alpha))$ iterations it holds that $\xi_T \leq \alpha$.

Lemma 9 (Super-Linear Convergence) *Fix $\kappa, \rho \in (0, 1)$. Consider a stochastic process ξ_t adapted to a filtration $\mathcal{F}_t$ that satisfies: (i) $0 \leq \xi_0 \leq M$ (Bounded Initialization); (ii) for all t: $0 \leq \xi_{t+1} \leq \xi_t$ (Monotonicity); (iii) for all t: $\mathbf{Pr}[\xi_{t+1} \leq \xi_t^{(1-\rho)}\kappa^\rho \mid \mathcal{F}_t] \geq 2/3$ (Super-Linear Decay). Then, for any T larger than $(3/2)\left((1/\rho)\max(\log\log(1/\kappa), \log\log(M+1)) + \log(e/\delta)\right)$, with probability at least $1 - \delta$, it holds that $\xi_T \leq e^2\kappa$.*

2.2.1. PUTTING EVERYTHING TOGETHER: THE PROOF OF THEOREM 6

For this sketch we shall assume that we have an initialization $\mathbf{w}$ such that $\theta(\mathbf{w}, \mathbf{w}^*)$ is sufficiently small. Let $T = cd\log\log n\log(1/\delta)$ for some sufficiently large absolute constant $c > 0$. We split X into $2k$ subsets $U_1, \ldots, U_{2k}$, with $k = n/(2T)$ and let $N = n/(2T)$ be the number of samples in each bucket. We assume that N is greater than a sufficiently large constant; otherwise, $N \leq O(T)$ and the mistake bound would be at most $O(T)$. Note that each set U_i is independent of all others.

Let $\mathbf{w}^{(0)} = \mathbf{w}$ and $\mathbf{u}^{(0)} = \mathbf{w}$. We analyze algorithm Algorithm 1 for $\mathbf{w}^{(0)}$ (as the analysis of $\mathbf{v}^{(0}$ is similar). Step 5 of Algorithm 1 runs Margin-Perceptron in each set and goes to the next set when a mistake occurs. Let $\mathbf{w}^{(t)}$ be the current hypothesis and $\theta^{(t)} = \theta(\mathbf{w}^{(t)}, \mathbf{w}^*)$. From Lemma 8, conditioned on $\mathbf{w}^{(t)}$, we have that if a mistake occurred, then we construct a new vector $\mathbf{w}^{(t+1)}$ with $\theta^{(t+1)} = \theta(\mathbf{w}^{(t+1)}, \mathbf{w}^*)$ so that $\tan\theta^{(t+1)} \leq \tan\theta^{(t)}$ and furthermore with probability at least $2/3$ we have that $\tan(\theta^{(t+1)}) \leq \tan^{1-1/(8d)}\theta^{(t)}(C''/N)^{1/(8d)}$, where $C'' > 0$ is an absolute constant. Let $\xi_t = \tan\theta^{(t)}$. We have that $0 \leq \xi_0 = \tan\theta^{(0)} \leq 1$ and that $\xi_{t+1} \leq \xi_t$. Hence, we have that $\mathbf{Pr}[\xi_{t+1} \leq \xi_t^{1-(1/8d)}(C''/N)^{1/(8d)}|\mathbf{w}^{(t)}] \geq 2/3$. Therefore, using Lemma 9, we get that $\theta_T \leq O(1/N)$, with probability at least $1 - \delta/4$. Therefore, in Step 5a of Algorithm 1, the algorithm made at most $M_1 = 2T$ mistakes. Next, we bound the number of mistakes in Step 6a. Note that $A = \cup_{i=k}^{2k} U_i$, contains $\Omega(n)$ samples. From Lemma 27, we have that with probability at least $1 - \delta/4$ conditioned on the event that $\theta^{(T)} \leq O(T/n)$, Algorithm 1, labels the points in A, with at most $O(T)$ mistakes. The same arguments show the same for the hypothesis $\mathbf{u}^{(T)}$. Therefore, the number of mistakes is at most $O(d\log\log(n))\log(1/\delta)$, with probability at least $1 - \delta$.

3. Self-Directed Learning on Arbitrary Datasets

In this section we prove our result for self-directed classification for arbitrary datasets. We first state the formal version of Theorem 5.

Theorem 10 (Strong, Self-Directed Learner for Arbitrary Data) *Let $\mathcal{C}$ be the class of LTFs on $\mathbb{R}^d$ and let X be a set of n unlabeled points in $\mathbb{R}^d$. There exists a algorithm that runs in $\mathrm{poly}(d, n)$ time, makes $\widetilde{O}(d^2\log(d/(\epsilon\delta)))$ mistakes, and, with probability at least $1 - \delta$, correctly classifies a $(1 - \epsilon)$-fraction of the points of X.*

3.1. Roadmap of the Proof Theorem 10

Boosting a Weak Self-Directed Learner The main ingredient in the proof of Theorem 10 is a *weak-learner* that does $O(d\log d)$ mistakes and correctly labels roughly $\Omega(1/d)$-fraction of the dataset X with non-trivial (say above 1%) probability of success. We show the following proposition.

Proposition 11 (A Weak, Self-Directed Learner for Arbitrary Data) *Let $\mathcal{C}$ be the class of LTFs on $\mathbb{R}^d$ and let X be a set of n unlabeled points in $\mathbb{R}^d$. There exists a universal constant c and an algorithm that runs in $\mathrm{poly}(d, n)$ time, makes $O(d\log d)$ mistakes, and, with probability at least c, correctly classifies an $\Omega(1/d)$-fraction of the points of X.*

We give a generic boosting algorithm that allows one to obtain a strong learner and prove Theorem 10. At a high-level one can iteratively use the weak-learner to label a fraction of points, remove it from the dataset, and reuse the weak-learner on the remaining data.

Lemma 12 (Boosting) *Let $\mathcal{A}$ be a distribution-free self-directed learner that makes M mistakes and correctly labels a $(1 - \alpha)$-fraction of X for some fixed $\alpha \in (0, 1)$, with probability at least $c \in (0, 1)$. Then, there exists a strong self-directed learner that makes $\widetilde{O}((M/c)\ \log(1/(\delta\epsilon))/\log(1/\alpha))$ mistakes and labels $(1 - \epsilon)$-fraction of X with probability at least $1 - \delta$.*

We remark, that this "label-then-remove" approach crucially relies on the weak-learner being able to handle arbitrary datasets (as the distribution of the remaining data is no-longer the same as the one that generated the data initially). We present the details of our boosting approach in Appendix D.2.

Weak Learning via Forster Transform and Margin Perceptron Similarly to our algorithm for spherical data, at a high-level, our algorithm relies on picking the "easiest" examples first, i.e., picking the samples with the maximum possible margin from the current hypothesis. We then use the margin-perceptron update as we did in the distribution specific setting, see Equation (1). However, as we observed in Section 2.1, picking examples that have good margin with the current hypothesis is crucial and since an arbitrary dataset X is not guaranteed to have margin, the margin-perceptron update *may make small or even zero progress*. To overcome this issue we perform a pre-processing step to ensure that the resulting dataset has *soft-margin* with respect to every halfspace while at the same time remaining linearly separable.

We observe that given any dataset X one can perform an (invertible) linear transformation $\mathbf{A}$ on the points of X and obtain a dataset that is still linearly separable: assuming that the initial dataset is separable by $\mathbf{w}^*$ then for every $\mathbf{x} \in X$ we have $\mathbf{w}^* \cdot \mathbf{x} = (\mathbf{A}^{-1}\mathbf{w}^*) \cdot (\mathbf{A}\mathbf{x})$ and therefore the vector $\mathbf{A}^{-1}\mathbf{w}^*$ corresponds to the normal vector of a linear separator of the transformed dataset. Moreover, we can preserve linear separability by rescaling each $\mathbf{x}$ to lie on the unit-sphere $\mathbf{x} \mapsto \mathbf{x}/\|\mathbf{x}\|_2$. Forster transform combines the two transformations for some invertible matrix $\mathbf{A}$, i.e., $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}/\|\mathbf{A}\mathbf{x}\|_2$ and transforms the dataset so that it is in (approximate) Radially Isotropic Position. There are several efficient algorithms (see e.g., Artstein-Avidan et al. (2020); Diakonikolas et al. (2021)) to compute such an invertible matrix $\mathbf{A}$ and more recently in Diakonikolas et al. (2022) a strongly polynomial-time algorithm for computing Forster transforms was given, see Proposition 14.

Definition 13 (Radially Isotropic Position) *Let X be a multiset of n non-zero points of $\mathbb{R}^d$. We say that X is in δ-approximate Radially Isotropic Position if:*

1. *For every $\mathbf{x} \in X$, it holds $\|\mathbf{x}\|_2 = 1$. (Unit Norm)*
2. *For any unit vector $\mathbf{u} \in \mathbb{R}^d$, it holds $(1/|X|)\sum_{\mathbf{x}\in X}(\mathbf{u}\cdot\mathbf{x})^2 \geq 1/d - \delta$. (Isotropic Position)*

Assuming that the dataset X is in Radially Isotropic Position, one can show that X has "soft-margin" with respect to every halfspace, in the sense that for every unit vector $\mathbf{w}$ it holds that at least $\Omega(1/d)$-fraction of X has margin $|\mathbf{w}\cdot\mathbf{x}| \geq \Omega(1/\sqrt{d})$, see Lemma 15. Now that we have this "soft-margin" we are able to show that the margin-perceptron will correctly label a non-trivial ($\Omega(1/d)$-fraction) part of the dataset. We refer to Appendix D.1 and Algorithm 2 for more details.

3.2. Proof of Proposition 11

We shall use the strongly polynomial time algorithmic result to compute a Forster transform (or show that one does not exist) given in the recent work of Diakonikolas et al. (2022).

Proposition 14 (Algorithmic Forster Transform, Diakonikolas et al. (2022)) *Given a set of non-zero points X, and an invertible matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, we denote by $S_{\mathbf{A}}(X) = \{\mathbf{A}\mathbf{x}/\|\mathbf{A}\mathbf{x}\|_2 : \mathbf{x} \in X\}$. There exists an algorithm, that given a set of points X in $\mathbb{Z}^d \setminus \{\mathbf{0}\}$ and some $\delta > 0$, runs in time $\mathrm{poly}(n, d, \log(1/\delta))$ and returns a subspace V of $\mathbb{R}^d$ containing at least a $\dim(V)/d$-fraction of the points X and an invertible matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ such that $S_{\mathbf{A}}(X \cap V)$ is in δ-approximate radially isotropic position.*

In the next lemma we show that a dataset in (approximate) Radially Isotropic Position, satisfies a notion of "soft-margin" in the sense that non-trivial part of the dataset has non-trivial margin with respect to every halfspace. Its proof can be found on Appendix D.

Input: An unlabeled dataset $X \subseteq \mathbb{R}^d$.
Output: A sequence of labeled data $(\mathbf{x}^{(t)}, z^{(t)})$.

1. Find subspace V of dimension k so that $|X \cap V| \geq (k/d)\, n$ and $X \cap V$ is in $1/(2d)$-approximate Radially Isotropic Position using Proposition 32. Set $U = X \cap V$.
2. Randomly initialize guess $\mathbf{w}^{(0)} \sim \mathbb{S}_k$.
3. For $t = 0, \ldots, 5k \log k$:
   (a) Obtain $U_{\mathbf{w}^{(t)}}$ by sorting the points of U in decreasing order of margin from $\mathbf{w}^{(t)}$, i.e., $|\mathbf{x}^{(i+1)} \cdot \mathbf{w}^{(t)}| \leq |\mathbf{x}^{(i)} \cdot \mathbf{w}^{(t)}|$.
   (b) Initialize the set of correctly predicted points $C \leftarrow \emptyset$.
   (c) For $\mathbf{x} \in U_{\mathbf{w}^{(t)}}$:
      i. Predict the label of $\mathbf{x}$ with $\mathbf{w}^{(t)}$.
      ii. If the prediction is incorrect, update $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - (\mathbf{w}^{(t)} \cdot \mathbf{x})\, \mathbf{x}$, add $(\mathbf{x}, -\mathrm{sign}(\mathbf{w}^{(t)} \cdot \mathbf{x}))$ to C, and exit the inner loop.
      iii. If the prediction is correct, add $(\mathbf{x}, \mathrm{sign}(\mathbf{w}^{(t)} \cdot \mathbf{x}))$ to C.
   (d) If $|C| \geq |U|/(4k)$ then return C and exit the loop.

Algorithm 2: *A Weak Self-Directed Learner for an Arbitrary Dataset X.*

Lemma 15 (Soft-Margin via Radially Isotropic Position) *Let X be a multi-set of non-zero points in $1/(2d)$-approximate Radially Isotropic Position. For every unit vector $\mathbf{u} \in \mathbb{R}^d$, we have $\mathbf{Pr}_{\mathbf{x} \sim X}[|\mathbf{u} \cdot \mathbf{x}| \geq 1/(2\sqrt{d})] \geq 1/(4d)$.*

Denote by $N = |U|$ the number of points that are returned in Step 1 of Algorithm 2, and note that $N \geq nk/d$. From Lemma 3.2.4 Vershynin (2018), we get that with probability larger than an absolute constant, the random initialitation gives a point $\mathbf{w}^{(0)}$, so that $\mathbf{w}^{(0)} \cdot \mathbf{v} \geq 1/(2\sqrt{k})$. In what follows, we condition on the initialization satisfying this correlation bound. We show that if Algorithm 2 terminates, then $1/(4d)$-fraction of points is correctly classified. Note that Algorithm 2 terminates if the algorithm makes $5d \log d$ mistakes or when $|C| \geq |U|/(4k) \geq (k/d)n/(4k) \geq n/(4d)$ (and therefore, the algorithm classifies at least $1/(4d)$-fraction of X correctly. Thus the bad event is that algorithm does $5d \log d$ mistakes and $|C| < |U|/(4k)$. We argue that this cannot happen. Let n_i be the remaining points in the i-th iteration. Note that $N = n_i + |C|$. We make use of the following lemma (a variant of which was shown in Dunagan and Vempala (2004)); its proof can be found on Appendix D. It shows that when we are using the margin-perceptron update, not many mistakes with large margin can occur.

Lemma 16 (Margin Perceptron (Dunagan and Vempala, 2004)) *Let $\mathbf{w}^*, \mathbf{w}^{(0)} \in \mathbb{R}^d$ be unit vectors such that $\mathbf{w}^* \cdot \mathbf{w}^{(0)} \geq \alpha$, for some $\alpha > 0$. Assume the following: $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \mathbf{x}^{(t)}(\mathbf{x}^{(t)} \cdot \mathbf{w}^{(t)})$ and let $t_0 \in \mathbb{Z}_+$, so that for all $t \in \mathbb{Z}_+$ with $t \leq t_0$, $|\mathbf{x}^{(t)} \cdot \mathbf{w}^{(t)}| \geq \beta \|\mathbf{w}^{(t)}\|_2$ and $(\mathbf{x}^{(t)} \cdot \mathbf{w}^{(t)})(\mathbf{x}^{(t)} \cdot \mathbf{v}) < 0$. Then, $t_0 \leq (2/\beta^2) \log(1/\alpha)$.*

Assume that after $t_1 = (5d \log d - 1)$ mistakes, $|C| < |U|/(4k)$. That means for all $t \leq t_1$ it holds $n_t = |U| - |C| \geq n(k/d - 1/(4d)) \geq |U|/2$, as $d \geq 1$. Let $\mathcal{S}_t = \{\mathbf{x}^{(i)} : |\mathbf{w}^{(t)} \cdot \mathbf{x}^{(i)}| \geq 1/(2\sqrt{k})\}$. From Lemma 15, it holds that for each t, $|\mathcal{S}_t| \geq |U|/(4k)$ and combining with the fact that

$n_t \geq |U|/2$, that means that either in each iteration, the algorithm makes no mistakes in the set $\mathcal{S}_t$, which means that $|C| \geq |U|/(4k)$ and the algorithm terminates, or that it makes one mistake in the set $\mathcal{S}_t$, which means that if $\mathbf{x}^{(t)}$ is the vector that $\mathbf{w}^{(t)}$ made a mistake then $|\mathbf{w}^{(t)} \cdot \mathbf{x}^{(t)}| \geq 1/(2\sqrt{k})$. Hence, conditional on the event that the algorithm did not terminate before the iteration t_0, by Lemma 16 if $t_0 \geq 5d \log d$, then $\mathbf{w}^{(t_0)}$ makes no mistakes in the set $\mathcal{S}_{t_0}$, so it classifies correctly $|U|/(4k)$ points, and the algorithm terminates.

Acknowledgments

Ilias Diakonikolas was supported by NSF Medium Award CCF-2107079, NSF Award CCF-1652862 (CAREER), and a DARPA Learning with Less Labels (LwLL) grant. Christos Tzamos was Supported by NSF Award CCF-2144298 (CAREER). Vasilis Kontonis was supported in part by NSF NSF Award CCF-2144298 (CAREER). Supported in part by NSF Award CCF-2144298 (CAREER). Nikos Zarifis was supported in part by NSF award 2023239, NSF Medium Award CCF-2107079, and a DARPA Learning with Less Labels (LwLL) grant.

References

D. Angluin. Learning Regular Sets from Queries and Counterexamples. *Information and Computation*, 75(2):87–106, 1987.

S. Artstein-Avidan, H. Kaplan, and M. Sharir. On radial isotropic position: Theory and algorithms. *arXiv preprint arXiv:2005.04918*, 2020.

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

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

M. Balcan and R. Urner. Active learning-modern learning theory., 2016.

J. Barzdiņš and R. Freivald. On the prediction of general recursive functions. In *Doklady Akademii Nauk*, volume 206, pages 521–524. Russian Academy of Sciences, 1972.

S. Ben-David, E. Kushilevitz, and Y. Mansour. Online learning versus offline learning. *Machine Learning*, 29:45–63, 1997.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

D. Blackwell et al. Controlled random walks. In *Proceedings of the international congress of mathematicians*, volume 3, pages 336–338, 1954.

A. Blum. Learning Boolean functions in an infinite attribute space. In *Proceedings of the Twenty-Second Annual Symposium on Theory of Computing*, pages 64–72, 1990.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine learning*, 15:201–221, 1994.

S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *International conference on computational learning theory*, pages 249–263. Springer, 2005.

I. Diakonikolas, D. Kane, and C. Tzamos. Forster decomposition and learning halfspaces with noise. *Advances in Neural Information Processing Systems*, 34:7732–7744, 2021.

I. Diakonikolas, C. Tzamos, and D. Kane. A strongly polynomial algorithm for approximate forster transforms and its application to halfspace learning. *arXiv preprint arXiv:2212.03008*, 2022.

T. Doliwa, H. Simon, and S. Zilles. Recursive teaching dimension, learning complexity, and maximum classes. In *Algorithmic Learning Theory: 21st International Conference, ALT 2010, Canberra, Australia, October 6-8, 2010. Proceedings 21*, pages 209–223. Springer, 2010.

J. Dunagan and S. Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 315–320, 2004.

J. Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

S. Goldman and D. Mathias. Teaching a smart learner. In *Proceedings of the sixth annual conference on computational learning theory*, pages 67–76, 1993.

S. Goldman, R. Rivest, and R. Schapire. Learning binary relations and total orders. *SIAM Journal on Computing*, 22(5):1006–1034, 1993.

S. A Goldman and R. H Sloan. The power of self-directed learning. *Machine Learning*, 14:271–294, 1994.

M. Goldmann, J. Håstad, and A. Razborov. Majority gates vs. general weighted threshold gates. *Computational Complexity*, 2:277–300, 1992.

G. Hacohen and D. Weinshall. On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning*, pages 2535–2544. PMLR, 2019.

J. Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3: 97–139, 1957.

S. Hanneke. Rates of convergence in active learning. *Ann. Statist.*, 39(1):333–361, February 2011.

M. Hardt and A. Moitra. Algorithms and hardness for robust subspace recovery. In *COLT 2013*, pages 354–375, 2013.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2 (3-4):157–325, 2016.

A. Jadbabaie, A. Rakhlin, S. Shahrampour, and K. Sridharan. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pages 398–406. PMLR, 2015.

N. Littlestone. Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

N. Littlestone. *Mistake bounds and logarithmic linear-threshold learning algorithms*. PhD thesis, University of California at Santa Cruz, 1989.

N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, February 1994.

P. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

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

F. Mansouri, H. Simon, A. Singla, and S. Zilles. On batch teaching with sample complexity bounded by vcd. In *Advances in Neural Information Processing Systems*, 2022.

D. Mathias. A model of interactive teaching. *journal of computer and system sciences*, 54(3): 487–501, 1997.

T. Mitchell. Generalization as search. *Artificial Intelligence*, 18:203–226, 1982.

E. Ni and C. Ling. Direct marketing with fewer mistakes. In *Advanced Data Mining and Applications: 7th International Conference, ADMA 2011, Beijing, China, December 17-19, 2011, Proceedings, Part I 7*, pages 256–269. Springer, 2011.

A. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on Mathematical Theory of Automata*, volume XII, pages 615–622, 1962.

F. Orabona. A modern introduction to online learning, 2019.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

H. Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 2, pages 131–149. University of California Press, 1951.

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

R. Rubinfeld. Sublinear time algorithms. In *Proceedings of the international congress of mathematicians (ICM), Madrid, Spain, August 22–30, 2006. Volume III: Invited lectures*. 2006.

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

J. Shawe-Taylor and N. Cristianini. *An introduction to support vector machines*. Cambridge University Press, 2000.

P. Soviany, R. Ionescu, P. Rota, and N. Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

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

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.

V. Vovk. Aggregating strategies. In *Annual Workshop on Computational Learning Theory: Proceedings of the third annual workshop on Computational learning theory, 1990*. Association for Computing Machinery, Inc, 1990.

V. Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 51–60, 1995.

X. Wang, Y. Chen, and W. Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2021.

A. Yao. On ACC and threshold circuits. In *Proceedings of the Thirty-First Annual Symposium on Foundations of Computer Science*, pages 619–627, 1990.

C. Zhang, J. Shen, and P. Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

Appendix A. Preliminaries and Notation

For $n \in \mathbb{Z}_+$, let $[n] \coloneqq \{1, \ldots, n\}$. We use small boldface characters for vectors and capital bold characters for matrices. For $\mathbf{x} \in \mathbb{R}^d$ and $i \in [d]$, $\mathbf{x}_i$ denotes the i-th coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2 \coloneqq (\sum_{i=1}^d \mathbf{x}_i^2)^{1/2}$ denotes the ℓ_2-norm of $\mathbf{x}$. We will use $\mathbf{x} \cdot \mathbf{y}$ for the inner product of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\theta(\mathbf{x}, \mathbf{y})$ for the angle between $\mathbf{x}, \mathbf{y}$. We slightly abuse notation and denote $\mathbf{e}_i$ the i-th standard basis vector in $\mathbb{R}^d$. We will use $\mathbb{1}_A$ to denote the characteristic function of the set A, i.e., $\mathbb{1}_A(\mathbf{x}) = 1$ if $\mathbf{x} \in A$ and $\mathbb{1}_A(\mathbf{x}) = 0$ if $\mathbf{x} \notin A$. We use the standard $O(\cdot), \Theta(\cdot), \Omega(\cdot)$ asymptotic notation. We also use $\widetilde{O}(\cdot)$ to omit poly-logarithmic factors. We use $\mathbf{E}_{x \sim D}[x]$ for the expectation of the random variable x according to the distribution D and $\mathbf{Pr}[\mathcal{E}]$ for the probability of event $\mathcal{E}$. To simplicity notation, we may omit the distribution when it is clear from the context. For a set X we use the $\mathbf{x} \sim X$ to denote sampling $\mathbf{x}$ uniformly at random from X. For example, $\mathbf{x} \sim \mathbb{S}_d$ means that we sample $\mathbf{x}$ uniformly at random from the d-dimensional unit sphere.

Appendix B. Random-Order Learners Make $\Omega(d \log n)$ Mistakes

In this section we show that random- and worst-order learners make at least $\Omega(d \log n)$ mistakes. This is true even for weak learning (i.e., labeling only 1% of the dataset) and even when the dataset X is drawn i.i.d. from the unit sphere $\mathbb{S}_d$. The proof relies on a distribution specific (for $\mathbb{S}_d$) PAC learning lower-bound given in Long (1995).

Proposition 17 (Mistake Lower Bound for Random-Order) *Let $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ be a set of n i.i.d. samples from $\mathbb{S}_d$ with ground-truth labels given by some halfspace with normal vector $\mathbf{w}^*$, i.e., the label of $\mathbf{x}^{(i)}$ is $\operatorname{sign}(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$. Then any algorithm that predicts the labels of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ in random order makes at least $\Omega(d \log n)$ mistakes in expectation. Moreover, this is true even if the labeling algorithm predicts labels for only 1% of the samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$.*

Proof We consider the time t in the labeling algorithm, i.e., the algorithm has predicted (and therefore also observed the correct labels) of a random subset of t examples. Since all points $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are drawn i.i.d. from the uniform distribution on the unit sphere we have that any random subset of t points is also an i.i.d. sample of uniformly random points on the sphere. We are going to show that any algorithm that has observed the labels of the random subset of size t, makes a mistake on the next example (that is also a uniformly random sample on the unit sphere) with probability at least $\Omega(d/t)$. Although this is generally given by standard VC bounds since our distribution is uniform on the sphere, we require the following result from Long (1995). In what follows we shall denote by $\mathcal{A}(\mathbf{x}; S)$ the prediction of some generic learning algorithm $\mathcal{A}$ on an example $\mathbf{x}$ given a labeled dataset S. When the training dataset is clear from the context we may also simply write $\mathcal{A}(\mathbf{x})$.

Lemma 18 (PAC Learning Halfspaces on the Unit Sphere (Long, 1995)) *Fix a ground-truth halfspace $f(\mathbf{x}) = \operatorname{sign}(\mathbf{w}^* \cdot \mathbf{x})$ for some weight vector $\mathbf{w}^* \in \mathbb{R}^d$. Let $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}$ be a set of t i.i.d. samples drawn uniformly at random on the unit sphere. The expected error of any learning algorithm $\mathcal{A}$ that has observed $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}$ (and their ground-truth labels) is at least $\Omega(d/t)$*

$$\mathop{\mathbf{E}}_{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)} \sim \mathbb{S}_d} \left[\mathop{\mathbf{Pr}}_{\mathbf{x} \sim \mathbb{S}_d} [\mathcal{A}\Big(\mathbf{x}; (\mathbf{x}^{(1)}, f(\mathbf{x}^{(1)})), \ldots, (\mathbf{x}^{(t)}, f(\mathbf{x}^{(t)}))\Big) \neq f(\mathbf{x})] \right] \geq c \frac{d}{t},$$

where c is some universal constant.

Using Lemma 18, we obtain that after predicting the labels on t examples, the expected probability that any algorithm makes an incorrect prediction on a fresh example is at least $2/3$. Given any prediction algorithm $\mathcal{A}$, we define the error of $\mathcal{A}$ to be the probability that $\mathcal{A}$ makes an incorrect prediction on a fresh sample from $\mathbb{S}_d$, i.e., $\mathrm{err}(\mathcal{A}, \mathbf{x}) = \mathbb{1}\{\mathcal{A}(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})\}$. To simplify notation, we shall denote by $S_t = \{(\mathbf{x}^{(1)}, f(\mathbf{x}^{(1)})), \ldots, (\mathbf{x}^{(t)}, f(\mathbf{x}^{(t)}))\}$ a set of t labeled examples and by $\mathcal{F}_t$ the corresponding filtration (so that S_t is adapted to $\mathcal{F}_t$). We have that

$$\mathbf{E}\left[\sum_{t=1}^{n} \mathrm{err}(\mathcal{A}(\cdot; S_t), \mathbf{x}^{(t+1)})\right] = \sum_{t=1}^{n} \mathbf{E}[\mathrm{err}(\mathcal{A}(\cdot; S_t), \mathbf{x}^{(t+1)})\Big] \geq \sum_{t=1}^{n} c\frac{d}{t} \geq cd \log n\,,$$

where for the last inequality, we used the fact that the harmonic number $\sum_{t=1}^{n} 1/t = \Omega(\log n)$. Finally, we see that the same is true if we only label only $1\%n$ points since $\sum_{t=1}^{0.01n} 1/t = \Omega(\log n)$. ■

Appendix C. Self-Directed Learning on $\mathbb{S}_d$: Proof Details

Input: An initialization $\mathbf{w}$. **Output:** A sequence of labeled data $(\mathbf{x}^{(t)}, z^{(t)})$.

1. Initialize guesses $\mathbf{w}^{(0)} \leftarrow \mathbf{w}$, $\quad \mathbf{v}^{(0)} \leftarrow \mathbf{w}$.
2. Initialize the set of unlabeled data $U \leftarrow X$, $t \leftarrow 0$.
3. Split U in $2k$ sets $U_1, \ldots, U_{2k}$.
4. For $t = 1, \ldots, k$:
   $\mathbf{w}^{(t)} \leftarrow$ MARGIN-PERCEPTRON(U_t, $\mathbf{w}^{(t-1)}$), $\mathbf{v}^{(t)} \leftarrow$ MARGIN-PERCEPTRON(U_{k+t}, $\mathbf{v}^{(t-1)}$).
5. For $t = 1, \ldots, k$:
   Label points of U_{k+t} with $\mathbf{w}^{(k)}$ and label points of U_t with $\mathbf{v}^{(k)}$.

MARGIN-PERCEPTRON($U, \mathbf{w}$)
Input: An initialization $\mathbf{w}$ and a set of points U. **Output:** A vector $\mathbf{w}'$.

1. Obtain U' by sorting the points of U in decreasing order of margin from $\mathbf{w}$, i.e., $|\mathbf{x}^{(i+1)} \cdot \mathbf{w}| \leq |\mathbf{x}^{(i)} \cdot \mathbf{w}|$.
2. For $\mathbf{x} \in U'$:
   (a) Predict the label of $\mathbf{x}$ with $\mathbf{w}$.
   (b) If the prediction is incorrect, exit the loop and return $\mathbf{w}' \leftarrow \mathbf{w} - (\mathbf{w} \cdot \mathbf{x})\mathbf{x}$.

Algorithm 3: *Self-Directed Learning on* $\mathbb{S}_d$

C.1. The Proof of Theorem 6

We restate and prove Theorem 6 in this section.

Theorem 19 *Let $\delta \in (0, 1/2]$ and let n be larger than some sufficiently large universal constant. Let X be a set of n i.i.d. samples from $\mathbb{S}_d$ with true labels given by a homogeneous halfspace, $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. There exists a self-directed classifier that makes $O(d \log\log n \ \log(1/\delta))$ mistakes, runs in time $\mathrm{poly}(d, n)$ and classifies all points of X with probability at least $1 - \delta$.*

We first show our anti-concentration result for the maximum-margin of the conditional distribution on the disagreement region C. We believe that our tight anti-concentration bound is of independent interest and may find other applications in convex geometry and learning linear classifiers.

Proposition 20 *Let C be the indicator of the disagreement region of two homogeneous halfspaces, i.e., $C = \mathbb{1}\{(\mathbf{v} \cdot \mathbf{x})(\mathbf{u} \cdot \mathbf{x}) \leq 0\}$ for some unit vectors $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d$ with angle $\theta(\mathbf{v}, \mathbf{u}) = \theta$. Let $\mathbb{S}_d$ denote the uniform distribution on the unit sphere and by $\mathbb{S}_d(C)$ the conditional distribution on the disagreement region C.*

1. *For any $\alpha \in [0, 1]$, it holds:*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)} \sim \mathbb{S}_d(C)} \left[\max_{i=1,\ldots,m} |\mathbf{u} \cdot \mathbf{x}^{(i)}| \leq \alpha \sin(\theta/2) \right] \leq \exp\left(-m \, (1 - \alpha^2)^{d/2-1}/2 \right) .$$

2. *For any $\beta \in [0, 1]$, it holds:*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)} \sim \mathbb{S}_d(C)} \left[\max_{i=1,\ldots,m} |\mathbf{u} \cdot \mathbf{x}^{(i)}| \leq (1 - \beta) \sin(\theta) \right] \leq \exp\left(- m \, (\beta/2)^{d/2}/2 \right) .$$

Proof To simplify notation, set $\Delta = \alpha \sin(\theta)$. We first compute the probability that a single sample in C has $|\mathbf{u} \cdot \mathbf{x}| \geq \Delta$. By the symmetry of the set C and the uniform distribution on the sphere $\mathbb{S}_d$

$$\mathop{\mathbf{Pr}}_{\mathbf{x} \sim \mathbb{S}_d} [\mathbf{x} \in C, |\mathbf{u} \cdot \mathbf{x}| \geq \Delta] = 2 \mathop{\mathbf{Pr}}_{\mathbf{x} \sim \mathbb{S}_d} [\mathbf{x} \in C, \mathbf{u} \cdot \mathbf{x} \geq \Delta] = 2 \mathop{\mathbf{Pr}}_{\mathbf{x} \sim \mathbb{S}_d} [E_1] .$$

where $E_1 = \{\mathbf{x} : \mathbf{x} \in C, \mathbf{u} \cdot \mathbf{x} \geq \Delta\}$. Assume, without loss of generality that $\mathbf{u} = \mathbf{e}_2$ and $\mathbf{v} = -\sin\theta \mathbf{e}_1 + \cos\theta \mathbf{e}_2$. Observe that the set E_1 can now be written as $E_1 = \{(\mathbf{x}_1, \mathbf{x}_2) : \mathbf{x}_2 \geq \Delta, \cos\theta \mathbf{x}_2 \leq \sin\theta \mathbf{x}_1\}$. Using polar coordinates $\mathbf{x}_1 = r\cos\phi$, $\mathbf{x}_2 = r\sin\phi$ we have that $E_1 = \{(r, \phi) : 0 \leq r \leq 1, r\sin\phi \geq \Delta, r\cos\theta\sin\phi \leq r\sin\theta\cos\phi\} = \{(r, \phi) : 0 \leq r \leq 1, r\sin\phi \geq \Delta, \sin(\theta - \phi) \geq 0\}$, where we used the trigonometric identity $\sin(\phi - \theta) = \sin\phi\cos\theta - \cos\phi\sin\theta$ and the fact that $\sin(-z) = -\sin(z)$. Moreover, by $r\sin\phi \geq \Delta$ we obtain that $\sin\phi \geq 0$ and therefore $\phi \in [0, \pi]$. Combining this with the fact that the angle between two halfspaces can be at most $\theta \leq \pi$, we obtain that $\sin(\theta - \phi) \geq 0$ implies that $\phi \leq \theta$. Finally, observe that the constraint $r\sin\phi \geq \Delta = \alpha\sin\theta$ implies that the radius $r \geq \alpha$. Therefore the set E_1 can be equivalently written as

$$E_1 = \{(r, \phi) : \alpha \leq r \leq 1, r\sin\phi \geq \alpha\sin\theta, \phi \leq \theta\}$$

The set E_1 has coupled constraints (i.e., constraints that depend on both r, ϕ). The set $E_2 = \{(r, \phi) : \alpha \leq r \leq 1, \theta/2 \leq \phi \leq \theta\}$ has decoupled constraints and is a subset of E_1. To see that $E_2 \subseteq E_1$, notice that for $(r, \phi) \in E_2$ it holds $r\sin\phi \geq \alpha\sin(\theta/2) = \Delta$ and $\theta - \phi \in [0, \pi/2]$ which implies that $\sin(\theta - \phi) \geq 0$ and therefore, $(r, \phi) \in E_1$. We can now directly estimate the probability of the

set E_1. The 2-dimensional projection of the uniform on the sphere has density $\frac{d-2}{2\pi}(1-r^2)^{d/2-2}r$ (in polar coordinates).

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathbb{S}_d}[E_2] &= \frac{d-2}{2\pi}\int_{\alpha}^{1}\int_{\theta/2}^{\theta}(1-r^2)^{d/2-2}\, r\, d\phi dr = \frac{\theta(d-2)}{4\pi}\int_{\alpha}^{1}(1-r^2)^{d/2-2}\, r\, dr \\ &= \frac{\theta}{4\pi}(1-\alpha^2)^{d/2-1}\,.\end{aligned}$$

By the symmetry of $\mathbb{S}_d$ we directly obtain that $\mathbf{Pr}_{\mathbf{x}\sim\mathbb{S}_d}[C] = \theta/\pi$. We conclude that the conditional probability $\mathbf{Pr}_{\mathbf{x}\sim\mathbb{S}_d(C)}[|\mathbf{u}\cdot\mathbf{x}|\geq\Delta]\geq(1/2)(1-\alpha^2/d)^{d/2-1}$. We can now bound above the probability that the maximum of m independent samples from $\mathbb{S}_d(C)$ is small.

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}_1,\ldots,\mathbf{x}_m\sim\mathbb{S}_d(C)}\left[\max_{i=1,\ldots,m}|\mathbf{u}\cdot\mathbf{x}_i|\leq\Delta\right] &= (1-\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathbb{S}_d(C)}[|\mathbf{u}\cdot\mathbf{x}|\geq\Delta])^m \\ &\leq \exp\left(-m\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathbb{S}_d(C)}[|\mathbf{u}\cdot\mathbf{x}|\geq\Delta]\right)\leq\exp\left(-m\,(1-\alpha^2)^{d/2-1}/2\right)\,,\end{aligned}$$

where, for the first inequality, we used the fact $e^x\geq 1+x$.

We now prove the second inequality that allows us to achieve correlation arbitrarily close to $\sin\theta$ albeit with worse success probability. To keep the proof similar to the previous one, we shall use continue using the parameter $\alpha = 1-\beta$ and replace it with β in the final expression for the probability. Recall that the expression of the set E_1 in polar coordinates is

$$E_1 = \{(r,\phi) : \alpha\leq r\leq 1, r\sin\phi\geq\alpha\sin\theta, \phi\leq\theta\}\,.$$

This time, we estimate directly the probability of E_1. To simplify notation, set $q=\alpha\sin\theta$. We have:

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathbb{S}_d}[E_1] = \frac{d-2}{2\pi}\int_{\sin^{-1}(q)}^{\theta}\int_{\frac{q}{\sin\phi}}^{1}(1-r^2)^{d/2-2}\, r\, dr d\phi = \frac{1}{2\pi}\int_{\sin^{-1}(q)}^{\theta}\left(1-\left(\frac{q}{\sin\phi}\right)^2\right)^{d/2-1}d\phi\,.$$

Since the quantity inside the integral is positive, we can bound its value from below by slightly increasing the lower threshold to $\sin^{-1}(s\sin\theta)$ for $s=(1+\alpha)/2$ (where we used that $\sin^{-1}(\cdot)$ is increasing. We have

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathbb{S}_d}[E_1]\geq\frac{1}{2\pi}\int_{\sin^{-1}(s\sin\theta)}^{\theta}\left(1-\left(\frac{q}{\sin\phi}\right)^2\right)^{d/2-1}d\phi\geq\frac{1}{2\pi}\int_{\sin^{-1}(s\sin\theta)}^{\theta}\left(1-\left(\frac{\alpha}{s}\right)^2\right)^{d/2-1}d\phi\,.$$

Finally, observe that since $s=(1+\alpha)/2$, it holds that $a/s\leq s$ and therefore:

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathbb{S}_d}[E_1]\geq\frac{1}{2\pi}(1-s^2)^{d/2-1}(\theta-\sin^{-1}(s\sin\theta))\geq\frac{\theta}{2\pi}\frac{1-\alpha}{2}(1-s^2)^{d/2-1}(\theta-\sin^{-1}(s\sin\theta))\,.$$

where, for the last inequality, we used the inequality $\sin^{-1}(\alpha x)\leq\alpha\sin^{-1}(x)$. Therefore, we have proved the bound

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathbb{S}_d}[E_1]\geq\frac{1}{2}s(1-s^2)^{d/2-1}=\frac{1-\alpha}{4}\left(1-\left(\frac{1+\alpha}{2}\right)^2\right)^{d/2-1}\,.$$

We can now switch back to using the parameter $\beta = 1 - \alpha$ to obtain the bound

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathbb{S}_d}[E_1] \geq \frac{\beta}{4}(1-(1-\beta/2)^2)^{d/2-1} \geq (\beta/2)^{d/2}/2\,,$$

where we used the inequality $1-(1-x)^2 \geq x$ for all $x \in [0,1]$. The final steps to obtain the upper bound for the probability that the maximum is small are the same as those of the previous case. ■

Using Proposition 20 we now show considering the original n i.i.d. samples from $\mathbb{S}_d$ the maximum-margin of those that fall in the disagreemeent region is going to be significantly larger than that of a random sample of $\mathbb{S}_d$. This is the formal version of Proposition 7.

Lemma 21 *Let $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d$ be unit vectors with angle $\theta(\mathbf{v},\mathbf{u}) = \theta$. Let C be the indicator of the disagreement region of the two homogeneous halfspaces defined by $\mathbf{v}, \mathbf{u}$, i.e., $C = \mathbb{1}\{(\mathbf{v}\cdot\mathbf{x})(\mathbf{u}\cdot\mathbf{x}) \leq 0\}$. Furthermore, let $\mathbb{S}_d$ denote the uniform distribution on the unit sphere.*

1. *For all $n, s \geq 1$ and $c \geq 2$ such that $e^{-dc/4} \leq 4\pi s/(n\theta) \leq 1$, it holds*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}\sim\mathbb{S}_d}\left[\max_{i=1,\ldots,n}|\mathbf{u}\cdot\mathbf{x}^{(i)}|\mathbb{1}\{\mathbf{x}^{(i)}\in C\} \leq \sqrt{\frac{\log(n\theta/(4\pi s))}{2c\,d}}\sin(\theta)\right] \leq 2e^{-s/2}\,.$$

2. *For all $n, s \geq 1$ such that $4\pi s/(n\theta) \leq 1$ it holds: it holds*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}\sim\mathbb{S}_d}\left[\max_{i=1,\ldots,n}|\mathbf{u}\cdot\mathbf{x}_i|\mathbb{1}\{\mathbf{x}^{(i)}\in C\} \leq \left(1-\left(\frac{4\pi s}{n\theta}\right)^{2/d}\right)\sin(\theta)\right] \leq 2e^{-s/2}\,.$$

Proof Denote by S the set of samples that fall in the disagreement region C and denote by $m = |S|$ the number of samples that fall in the disagreement region. We observe that by the rotational symmetry of the uniform distribution on the sphere, the probability of the set C is exactly θ/π. Therefore, on expectation, out of the n samples that we draw from $\mathbb{S}_d$, the number of samples that fall in C is $\mathbf{E}[|S|] = \mu = n\theta/\pi$. We first show that with high-probability we are going to observe at least $\mu/2$ samples in C. Denote by m the number of samples that fall in C. Using Chernoff's bound, we obtain that $\mathbf{Pr}[m \leq \mu/2] \leq e^{-\mu/8} \leq e^{-n\theta/8} \leq e^{-4\pi s/8} \leq e^{-s/2}$. Therefore, from now on, we condition on the event that at least $m \geq n\theta/(2\pi)$ samples fall in C. In other words, out of the n original samples from $\mathbb{S}_d$, with probability at least $1-e^{-s/2}$, we have drawn at least $n\theta/(2\pi)$ samples from the conditional distribution $\mathbb{S}_d(C)$.

We first prove the second case of Lemma 21. Using the second case of Proposition 20 we have that for $\beta = (4\pi s/(n\theta))^{2/d}$, it holds that $\mathbf{Pr}[\max_{i=1,\ldots,m}|\mathbf{u}\cdot\mathbf{x}^{(i)}| \leq (1-\beta)\sin\theta] \leq \exp(-m\,2\pi s/(n\theta))$. Since we have conditioned on the event that $m \geq n\theta/(2\pi)$, we obtain that this probability is at most e^{-s}. Combining this probability with the rejection sampling failure probability (that the number of conditional samples, i.e., those that fell in S, is smaller than $n\theta/(2\pi)$), we obtain that the total probability of failure is at most $2e^{-s/2}$.

We now prove the first case of Lemma 21. Observe first that for n, using the fact that $\sin(\theta)/2 \leq \sin(\theta/2)$ it holds that

$$\sqrt{\frac{\log(n\theta/(4\pi s))}{2c\,d}}\sin\theta \leq \sqrt{\frac{2\log(n\theta/(4\pi s))}{c\,d}}\sin(\theta/2)\,.$$

At this point, notice that by the assumption of the first case of Lemma 21 that $e^{-cd/4} \leq n\theta/(4\pi s)$, we have that $2\log(n\theta/(4\pi s))/(cd) \leq 1/2$. In particular, we never ask for margin larger than $\sin(\theta/2)/2$ (notice that the maximum possible margin is always $\sin(\theta)$). Using the first case of Proposition 20, we have that

$$\mathbf{Pr}\left[\max_{i=1,\ldots,m} |\mathbf{u}\cdot\mathbf{x}^{(i)}| \leq \sqrt{\frac{\log(n\theta/(4\pi s))}{4d}}\sin\theta\right] \leq \exp\left(-(m/2)\left(1-2\frac{\log(n\theta/(4\pi s))}{cd}\right)^{d/2-1}\right). \tag{4}$$

Next, we will use the inequality $1-x \geq e^{-2x}$ that holds for all $x \in [0,1/2]$, to obtain that

$$\left(1-\frac{2\log(n\theta/(4\pi s))}{cd}\right)^{d/2-1} \geq \exp\left(-\frac{2d-4}{cd}\log(n\theta/(4\pi s))\right) \geq \frac{4\pi s}{n\theta}$$

where for the first inequality we used the fact that $2\log(n\theta/(4\pi s))/(cd) \leq 1/2$ (so that we are able to use the inequality $1-x \geq e^{-2x}$), and for the second inequality the fact that $c \geq 1$. Using the fact that $m \geq n\theta/(4\pi)$ we have that this probability of Equation (4) is at most $e^{-s/2}$. Combining this failure probability with the probability that we do not observe at least $n\theta/(2\pi)$ samples in C we obtain that the total failure probability is at most $2e^{-s/2}$. ■

We prove the following lemma, showing on each mistake the margin-perceptron has good probability of significantly (super-linearly) decreasing $\tan\theta$. We require that the current guess $\mathbf{w}$ is not exactly orthogonal with the target $\mathbf{w}^*$ (notice the assumption $1/\cos\theta \leq O(\zeta)$). We show that it is not hard to obtain such an initialization.

Lemma 22 *Let $\mathbf{w} \in \mathbb{R}^d$ and let $\theta(\mathbf{w}^*,\mathbf{w}) = \theta \in [0,\pi/2)$ and assume that for some parameter $\zeta > 0$, $1/\cos\theta \leq \zeta/(12\pi)$. Let C be the indicator of the disagreement region of $\mathbf{w}^*,\mathbf{w}$, i.e., $C = \mathbb{1}\{(\mathbf{w}^*\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{x}) \leq 0\}$. Let $X = \{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}\}$ be a sample set drawn from $\mathbb{S}_d$ with n larger than a sufficiently large constant and let $\widehat{\mathbf{x}} = \mathrm{argmax}_{\mathbf{x}\in\mathcal{X}}|\mathbf{w}\cdot\mathbf{x}|\mathbb{1}\{\mathbf{x}\in C\}$. Denote $\mathbf{w}' = \mathbf{w}' - (\mathbf{w}\cdot\widehat{\mathbf{x}})\widehat{\mathbf{x}}$ and let $\theta' = \theta(\mathbf{w}',\mathbf{w}^*)$.*

1. *We have that $\tan^2\theta' \leq \tan^2\theta$. (Monotonicity)*
2. *With probability at least $2/3$, it holds $\tan\theta' \leq \tan^{1-1/(8d)}\theta(\zeta/n)^{1/(8d)}$.*

Proof First, we claim that this update rule decreases $\tan\theta$.

Claim 23 *Let $\mathbf{w} \in \mathbb{R}^d$ and let $\mathbf{x} \in \{\mathbf{x} : (\mathbf{w}^*\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{x}) \leq 0\}$. Furthermore, assume that $|\mathbf{w}\cdot\mathbf{x}| \geq r\sin\theta$, where $\theta(\mathbf{w}^*,\mathbf{w}) = \theta \in [0,\pi/2)$ and $r > 0$. Denote $\mathbf{w}' = \mathbf{w} - (\mathbf{w}\cdot\mathbf{x})\mathbf{x}$ and let $\theta' = \theta(\mathbf{w}',\mathbf{w}^*)$. Then $\tan^2\theta' \leq \tan^2\theta(1-r^2)$.*

Proof We first observe that the correlation with $\mathbf{w}^*$ does not decrease, $\mathbf{w}'\cdot\mathbf{w}^* = (\mathbf{w}-(\mathbf{w}\cdot\mathbf{x})\mathbf{x})\cdot\mathbf{w}^* = \mathbf{w}\cdot\mathbf{w}^* - (\mathbf{w}\cdot\mathbf{x})(\mathbf{x}\cdot\mathbf{w}^*) \geq \mathbf{w}\cdot\mathbf{w}^*$, where we used that the hypothesis $\mathbf{w}$ disagrees with the ground-truth $\mathbf{w}^*$ on $\mathbf{x}$, i.e., $(\mathbf{w}\cdot\mathbf{x})(\mathbf{x}\cdot\mathbf{w}^*) \leq 0$. Furthermore, since $\mathbf{w}\cdot\mathbf{w}^* \geq 0$ (by the assumption that $\theta \in [0,\pi/2]$) we also have that $(\mathbf{w}'\cdot\mathbf{w}^*)^2 \geq (\mathbf{w}\cdot\mathbf{w}^*)^2$. We next show that the norm of $\mathbf{w}'$ decreases multiplicatively. We have that $\|\mathbf{w}'\|_2^2 = \|(\mathbf{w}-(\mathbf{w}\cdot\mathbf{x})\mathbf{x})\|_2^2 = \|\mathbf{w}\|_2^2 - (\mathbf{w}\cdot\mathbf{x})^2 \leq \|\mathbf{w}\|_2^2(1-r^2\sin^2\theta)$. Using (twice) the trigonometric identity $\tan^2\phi = (1/\cos^2\phi) - 1$, we show that $\tan^2\theta$ decreases by a factor of $(1-r^2)$:

$$\tan^2\theta' = \frac{\|\mathbf{w}'\|_2^2}{(\mathbf{w}'\cdot\mathbf{w}^*)^2} - 1 \leq \frac{\|\mathbf{w}\|_2^2(1-r^2\sin^2\theta)}{(\mathbf{w}\cdot\mathbf{w}^*)^2} - 1 = \frac{1-r^2\sin^2\theta}{\cos^2\theta} - 1 = (1-r^2)\tan^2\theta\,.$$

■

We note that from Claim 23 whenever we use the update rule on mistakes, it gives that $\tan(\theta') \leq \tan(\theta)$. We show that with constant probability, we can argue that the decrease is significantly larger.

We split our analysis into two cases, one that $12\pi/(n\theta) \leq \exp(-d/2)$ and the later case is when $1 \geq 12\pi/(n\theta) \geq \exp(-d/2)$. We first consider the case where $12\pi/(n\theta) \leq \exp(-d/2)$. From Lemma 21, we have that with probability at least $2/3$, it holds $|\mathbf{w} \cdot \widehat{\mathbf{x}}| \geq (1 - (12\pi/(n\theta))^{2/d}) \sin\theta$. Therefore, from Claim 23, we have that

$$\tan\theta' \leq \tan\theta \left(1 - (1 - (12\pi/(n\theta))^{2/d})^2\right)^{1/2} \leq \tan\theta \left(2(12\pi/(n\theta))^{2/d}\right)^{1/2} ,$$

where we used that $1 - (1-x)^2 \leq 2x$ for $x > 0$. Note that by our assumption $12\pi/(n\theta) \leq \exp(-d/2)$, therefore $\left(2(12\pi/(n\theta))^{2/d}\right)^{1/2} \leq (12\pi/(n\theta))^{1/(2d)} \leq (12\pi/(n\theta))^{1/(8d)}$.

Next, we consider the case where $1 \geq 12\pi/(n\theta) \geq \exp(-d/2)$. In this case, from Lemma 21, we have that with probability at least $2/3$, it holds $|\mathbf{w} \cdot \widehat{\mathbf{x}}| \geq \sqrt{\log(n\theta/(12\pi))/(4d)} \sin\theta$. Therefore, from Claim 23, we have that

$$\begin{aligned}\tan\theta' \leq \tan\theta \left(1 - \frac{\log(\theta n/(12\pi))}{4d}\right)^{1/2} &\leq \tan\theta \exp\left(-\frac{\log(\theta n/(12\pi))}{8d}\right) \\ &= \tan\theta \left(12\pi/(n\theta)\right)^{1/(8d)} .\end{aligned}$$

Therefore, in both cases, with probability at least $2/3$ that $\tan\theta' \leq \tan\theta \left(12\pi/(n\theta)\right)^{1/(8d)}$. Let $\rho = 1/(8d)$. We have that

$$\tan\theta' \leq \tan^{1-\rho}\theta \left(\frac{12\pi \sin\theta}{\theta\cos\theta}\frac{1}{n}\right)^{\rho} \leq \tan^{1-\rho}(\theta) \left(\frac{\zeta}{n}\right)^{\rho} ,$$

where we used our assumption that $\cos\theta \geq (12\pi/\zeta)$. This completes the proof of Lemma 22. ■

In the following lemma we show that given a decreasing stochastic process ξ_t that has good probability to decrease in a superlinear-rate then after $T = O(\log\log(1/\alpha))$ iterations it holds that $\xi_T \leq \alpha$.

Lemma 24 (Super-Linear Convergence) *Fix $\kappa, \rho \in (0,1)$. Consider a stochastic process ξ_t adapted to a filtration $\mathcal{F}_t$ that satisfies:*

1. *$0 \leq \xi_0 \leq M$ (Bounded Initialization),*
2. *For all t: $0 \leq \xi_{t+1} \leq \xi_t$ (Monotonicity),*
3. *For all t: $\mathbf{Pr}[\xi_{t+1} \leq \xi_t^{(1-\rho)}\kappa^{\rho} \mid \mathcal{F}_t] \geq 2/3$ (Super-Linear Decay).*

Then, for any T larger than $(3/2)\left((1/\rho)\max(\log\log(1/\kappa), \log\log(M+1)) + \log(e/\delta)\right)$, with probability at least $1-\delta$, it holds that $\xi_T \leq e^2\kappa$.

Proof Define the random variable I_t to be the indicator of the event that the super-linear decay step happens at step t, i.e., that $\xi_{t+1} \leq (\xi_t)^{(1-\rho)}\kappa^\rho$. We first observe that by the fact that the stochastic process is monotone in the sense that $\xi_t \leq \xi_{t+1}$ for all t, it does not matter at which steps the super-linear decay happens (but only how many times it does so). Assume that $\sum_{k=1}^{T} I_k = m$, and denote by t_j be the subsequence of $\{1, \ldots, T\}$ of length m where the super-linear decay steps happen. Using the monotonicity of ξ_t, we have that

$$\begin{aligned}\xi_T \leq \xi_{t_m} &\leq (\xi_{t_m-1})^{1-\rho}\kappa^\rho \leq (\xi_{t_{m-1}})^{1-\rho}\kappa^\rho \leq (\xi_{t_{m-1}-1})^{(1-\rho)^2}\kappa^{\rho(1-\rho)+\rho} \\ &\leq (\xi_{t_{m-2}})^{(1-\rho)^2}\kappa^{\rho(1-\rho)+\rho}\,.\end{aligned}$$

By continuing to unroll the recurrence, we obtain $\xi_T \leq \xi_0^{(1-\rho)^m}\kappa^{1-(1-\rho)^m}$, where we used the fact that $\rho + \rho(1-\rho) + \ldots + \rho(1-\rho)^{k-1} = 1 - (1-\rho)^k$. Using the fact that $\xi_0 \leq M$, we have that $\xi_T \leq M\kappa^{1-(1-\rho)^m}$. We show the following in Appendix C.

Fact 25 *If $m \geq (1/\rho)\max(\log\log(1/\kappa), \log\log(M+1))$, then $M^{(1-\rho)^m}\kappa^{1-(1-\rho)^m} \leq e^2\kappa$.*

Proof We first show that with $m \geq \log\log(M+1)/\rho$ it holds that $M^{(1-\rho)^m} \leq e$. We first observe that this is trivially true when $M \leq 1$. For $M > 1$ we can take logarithms in both sides of $M^{(1-\rho)^m} \leq e$ and obtain $m\log(1-\rho) + \log\log M \leq 0$ or equivalently $m \geq \log\log M/\log(1/(1-\rho))$. Since $\log(1/(1-\rho) \geq 1/\rho$ for all $\rho \in (0,1)$ we obtain that for the chosen m the inequality is true. Next we show that for $m \geq \log\log(1/\kappa)/\rho$ it holds that $\kappa^{1-(1-\rho)^m} \leq e\kappa$. We first observe that we can rewrite this inequality as $(1/\kappa)^{(1-\rho)^m} \leq e$. Using the same argument as in the previous case (by replacing M with $1/\kappa > 1$), we obtain the result. ■

To complete the proof, it remains to show that many "fast-decay" updates will happen with good probability, or, in other words, that the number m defined above is at least

$$m^* \coloneqq (1/\rho)\max(\log\log(1/\kappa), \log\log(M+1)),$$

with good probability. We show that if the total number of updates $T \geq 8\log(e/\delta)m^*$, then, with probability at least $1-\delta$, $m \geq m^*$. To do this, we shall Azuma's inequality for martingales.

Lemma 26 (Azuma-Hoeffding) *Let (D_t) be a martingale with bounded increments, i.e., $D_t - D_{t-1} \leq L$. It holds that $\mathbf{Pr}[D_T \leq D_0 - \lambda] \leq e^{-\lambda^2/(2L^2T)}$.*

We define the martingale $D_T = \sum_{t=1}^{T}(I_t - \mathbf{E}[I_t \mid \mathcal{F}_{t-1}])$, with $D_0 = 0$. Using the fact that the super-linear decay step happens with probability at least $2/3$ (see Item 3 of Lemma 24) we have that with probability at least $2/3$ we have $I_t = 1$ and therefore $\sum_{t=1}^{T}\mathbf{E}[I_t \mid \mathcal{F}_{t-1}] \geq (2/3)\,T$. Moreover, we observe that the increments of D_T are bounded by 1 and therefore Azuma's inequality Lemma 26 implies that $\mathbf{Pr}[D_T \leq -\sqrt{2T\log(1/\delta)}] \leq \delta$. Equivalently, we obtain that with probability at least $1-\delta$, it holds that the number of super-linear decay steps is bounded below by $m \geq \sum_{t=1}^{T}\mathbf{E}[I_t \mid \mathcal{F}_{t-1}] \geq (2/3)\,T - \sqrt{2T\log(1/\delta)}$. For $T = (3/2)m^* + (3/2)\log(e/\delta)$ we obtain that $m \geq m^*$ with probability at least $1-\delta$.

■

The following lemma shows that a halfspace that has angle θ with the ground-truth $\mathbf{w}^*$ makes roughly $n\theta$ mistakes on a sequence of n i.i.d. examples from the uniform distribution on the sphere.

Lemma 27 *Fix $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ and assume that $\theta(\mathbf{v}, \mathbf{w}) = \theta$. Let C be the indicator of the disagreement region of $\mathbf{v}, \mathbf{w}$, i.e., $C = \mathbb{1}\{(\mathbf{v}\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{x}) \leq 0\}$. Let $X = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ be a sample set drawn i.i.d. from $\mathbb{S}_d$. Then, with probability at least $1-\delta$, the set $X \cap C$ has size at most $O(n\theta + \sqrt{n\theta \log(1/\delta)})$.*

Proof Let $Z_i = \mathbb{1}\{\mathbf{x}^{(i)} \in C\}$. From the fact that $\mathbf{Pr}[(\mathbf{v}\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{x}) \leq 0] = \theta/\pi$, we have that $\mathbf{E}[Z_i] = \theta/\pi$. We use the following version of the standard Hoeffding bound.

Fact 28 *Let $z_1, \ldots, a_n$ be i.i.d. random variables on $\{0,1\}$ with $\mathbf{E}[z_1] = p$. Then, it holds that*

$$\mathbf{Pr}\left[\sum_{i=1}^{n} z_i \geq np + \epsilon n\right] \leq \exp\left(-\frac{\epsilon^2 n}{2p(1-p)}\right) .$$

An application of Fact 28, gives that

$$\mathbf{Pr}\left[\sum_{i=1}^{N} Z_i \geq n\theta/\pi + nt\right] \leq \exp(-\pi t^2 n/\theta) .$$

Choose $t \geq \Omega(\sqrt{\theta/n \log(1/\delta)})$, and the result follows. ■

C.1.1. PUTTING EVERYTHING TOGETHER: THE PROOF OF THEOREM 6

First, assume that $n \leq Cd \log\log(d) \log(1/\delta)$, for some large enough absolute constant $C > 0$. In this case, even if the algorithm makes a wrong prediction in all the points, the mistake bound will be $n = O(d \log\log(d) \log(1/\delta))$. For the rest of the proof, we assume that $n \geq Cd \log\log(d) \log(1/\delta)$. We use the following algorithm for the initialization process.

Lemma 29 (Theorem 2 of Dasgupta et al. (2005)) *Let $\epsilon, \delta \in (0,1]$. Consider a stream of data points $\mathbf{x}^{(t)}$ drawn uniformly at random from the surface of the unit sphere in $\mathbb{R}^d$, and the corresponding labels $y^{(t)}$ are consistent with an LTF $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. There is an algorithm that, if it is applied to this stream of data, then with probability at least $1-\delta$, after $O(d(\log(1/\epsilon) + \log(1/\delta)))$ mistakes, we get a halfspace $\mathbf{w}$ with generalization error at most ϵ.*

From Lemma 29, we have that with $O(d \log(1/\delta))$ mistakes, we get with probability at least $1 - \delta/2$, a halfspace $\mathbf{w} \in \mathbb{R}^d$ with generalization error $1/10$, therefore $\theta(\mathbf{w}, \mathbf{w}^*) = \pi/10 \leq 1/2$, hence $\cos(\theta(\mathbf{w}, \mathbf{w}^*)) \geq 1/2$. Let $T = c'd \log\log(n) \log(1/\delta)$ for some sufficiently large absolute constant $c > 0$. We split X into $2k$ subsets $U_1, \ldots, U_{2k}$, with $k = n/(2T)$, so that all the subsets contain at least $N = n/T \geq C/c'$, which we can make $C > 0$ to be large enough so that C/c' is greater than a sufficiently large absolute constant. Note that each set is independent of the other.

Let $\mathbf{w}^{(0)} = \mathbf{w}$ and $\mathbf{u}^{(0)} = \mathbf{w}$. We analyze first the algorithm Algorithm 3 for $\mathbf{w}^{(0)}$. Step 5 of Algorithm 3 runs Margin-Perceptron in each set and goes to the next set when a mistake occurs. Let $\mathbf{w}^{(t)}$ be the current hypothesis and $\theta^{(t)} = \theta(\mathbf{w}^{(t)}, \mathbf{w}^*)$. From Lemma 22, conditioned on $\mathbf{w}^{(t)}$, we have that if a mistake occurred, then we construct a new vector $\mathbf{w}^{(t+1)}$ with $\theta^{(t+1)} = \theta(\mathbf{w}^{(t+1)}, \mathbf{w}^*)$ so that $\tan\theta^{(t+1)} \leq \tan\theta^{(t)}$ and furthermore with probability at least $2/3$ we have that

$$\tan(\theta^{(t+1)}) \leq \tan^{1-1/(8d)}\theta^{(t)} (C''/N)^{1/(8d)} ,$$

where $C'' > 0$ is an absolute constant. Denote $\xi_t = \tan\theta^{(t)}$, we have that $0 \leq \xi_0 = \tan\theta^{(0)} \leq 1$ and that $\xi_{t+1} \leq \xi_t$. Hence, we have that

$$\mathbf{Pr}[\xi_{t+1} \leq \xi_t^{1-(1/8d)}(C''/N)^{1/(8d)} | \mathbf{w}^{(t)}] \geq 2/3 \, .$$

Therefore, using Lemma 24, we get that $\xi_T \leq e^2 C''/N$, with probability at least $1 - \delta/4$. Therefore $\theta^T \leq e^2 C''/n = e^2 C'' T/n$. Therefore, in Step 5a of Algorithm 3, the algorithm made at most $M_1 = 2T$ mistakes. Next, we bound the number of mistakes in Step 6a. Note that $U = \cup_{i=k}^{2k} U_i$, contains $\Omega(n)$ samples. From Lemma 27, we have that with probability at least $1 - \delta/4$ conditioned on the event that $\theta^{(T)} \leq e^2 C'' T/n$, Algorithm 3, labels the points in U, with at most $O(T)$ mistakes. The same arguments show the same for the hypothesis $\mathbf{u}^{(T)}$. Therefore, the number of mistakes is at most $O(d \log\log(n)) \log(1/\delta)$, with probability at least $1 - \delta$. Combining the two cases above, we obtain that the number of mistakes is

$$M = O(d \log(1/\delta)) * \begin{cases} O(\log\log n) \text{ if } n \geq C d \log\log d \log(1/\delta) \\ O(\log\log d) \text{ otherwise} \end{cases}$$

Finally, we may simplify further the above mistake bound by noticing that if the number of unlabeled points $n \leq d$ the number of mistakes is always at most d. Therefore, the total number of mistakes is $M = O(d \max(\log\log n, 1)) \log(1/\delta)$.

Appendix D. Self Directed Learning on Arbitrary Datasets

Input: An unlabeled dataset $X \subseteq \mathbb{R}^d$.
Output: A sequence of labeled data $(\mathbf{x}^{(t)}, z^{(t)})$.

1. Find subspace V of dimension k so that $|X \cap V| \geq (k/d)\ n$ and $X \cap V$ is in $1/(2d)$-approximate Radially Isotropic Position using Proposition 32. Set $U = X \cap V$.
2. Randomly initialize guess $\mathbf{w}^{(0)} \sim \mathbb{S}_k$.
3. For $t = 0, \ldots, 5k \log k$:
   (a) Obtain $U_{\mathbf{w}^{(t)}}$ by sorting the points of U in decreasing order of margin from $\mathbf{w}^{(t)}$, i.e., $|\mathbf{x}^{(i+1)} \cdot \mathbf{w}^{(t)}| \leq |\mathbf{x}^{(i)} \cdot \mathbf{w}^{(t)}|$.
   (b) Initialize the set of correctly predicted points $C \leftarrow \emptyset$.
   (c) For $\mathbf{x} \in U_{\mathbf{w}^{(t)}}$:
      i. Predict the label of $\mathbf{x}$ with $\mathbf{w}^{(t)}$.
      ii. If the prediction is incorrect, update $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - (\mathbf{w}^{(t)} \cdot \mathbf{x})\ \mathbf{x}$, add $(\mathbf{x}, -\text{sign}(\mathbf{w}^{(t)} \cdot \mathbf{x}))$ to C, and exit the inner loop.
      iii. If the prediction is correct, add $(\mathbf{x}, \text{sign}(\mathbf{w}^{(t)} \cdot \mathbf{x}))$ to C.
   (d) If $|C| \geq |U|/(4k)$ then return C and exit the loop.

Algorithm 4: *A Weak Self-Directed Learner for any set X.*

D.1. The proof of Proposition 11

We restate and prove the following proposition giving a weak, self-directed learner for arbitrary datasets that does $O(d \log d)$ mistakes and with non-trivial success probability (say above 1%) labels roughly $\Omega(1/d)$-fraction of X.

Proposition 30 (Weak, Self-Directed Learner for Arbitrary Datasets) *Let $\mathcal{C}$ be the class of LTFs on $\mathbb{R}^d$ and let X be a set of n unlabeled points in $\mathbb{R}^d$. There exists a universal constant c and an algorithm that runs in $\mathrm{poly}(d, n)$ time, makes $O(d \log d)$ mistakes, and, with probability at least c, correctly classifies an $\Omega(1/d)$-fraction of the points of X.*

We start by defining the "linear-map plus rescaling" transformation that puts the dataset in Radially Isotropic Position.

Definition 31 (Normalized Linear Transformation) *Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be an invertible matrix. Given a non-zero vector $\mathbf{x} \in \mathbb{R}^d$, we denote by $S_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}\mathbf{x}/\|\mathbf{A}\mathbf{x}\|_2$. We will also overload notation and, given a set of points X, we denote by $S_{\mathbf{A}}(X) = \{S_{\mathbf{A}}(\mathbf{x}) : \mathbf{x} \in X\}$.*

We shall use the strongly polynomial time algorithmic result to compute a Forster transform (or show that one does not exist) given in the recent work of Diakonikolas et al. (2022).

Proposition 32 (Algorithmic Forster Transform, (Diakonikolas et al., 2022)) *There exists an algorithm, that given a set of points X in $\mathbb{Z}^d \setminus \{\mathbf{0}\}$ and some $\delta > 0$, runs in time $\mathrm{poly}(n, d, \log(1/\delta))$ and returns a subspace V of $\mathbb{R}^d$ containing at least a $\dim(V)/d$-fraction of the points X and an invertible matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ such that $S_{\mathbf{A}}(X \cap V)$ is in δ-approximate radially isotropic position.*

In the next lemma we show that a dataset in (approximate) Radially Isotropic Position, satisfies a notion of "soft-margin" in the sense that non-trivial part of the dataset has non-trivial margin with respect to every halfspace.

Lemma 33 (Soft-Margin via Radially Isotropic Position) *Let X be a multi-set of non-zero points in $1/(2d)$-approximate Radially Isotropic Position. For every unit vector $\mathbf{u} \in \mathbb{R}^d$, we have $\mathbf{Pr}_{\mathbf{x} \sim X}[|\mathbf{u} \cdot \mathbf{x}| \geq 1/(2\sqrt{d})] \geq 1/(4d))$.*

Proof Since the set X is in Radially Isotropic Position, we have that $\|\mathbf{x}\|_2 \leq 1$ for every $\mathbf{x} \in X$, and therefore, by Cauchy-Schwarz, $|\mathbf{u} \cdot \mathbf{x}| \leq 1$. For a random variable z taking values in $[0, 1]$, the following reverse Markov inequality holds, $\mathbf{Pr}[z \geq a] \geq \mathbf{E}[z] - a$ (see, e.g., Appendix B1 in Shalev-Shwartz and Ben-David (2014)). We obtain that

$$\mathop{\mathbf{Pr}}_{\mathbf{x} \in X}\left[|\mathbf{u} \cdot \mathbf{x}| \geq \frac{1}{2\sqrt{d}}\right] = \mathop{\mathbf{Pr}}_{\mathbf{x} \in X}\left[(\mathbf{u} \cdot \mathbf{x})^2 \geq \frac{1}{4d}\right] \geq \mathop{\mathbf{E}}_{\mathbf{x} \sim X}[(\mathbf{u} \cdot \mathbf{x})^2] - \frac{1}{4d} \geq \frac{1}{2d} - \frac{1}{4d} = \frac{1}{4d}\,,$$

where we used the fact that the set X is in $1/(2d)$-approximate Radially Isotropic Position to replace $\mathbf{E}_{\mathbf{x} \sim X}[(\mathbf{u} \cdot \mathbf{x})^2]$ by its lower bound $1/d - 1/(2d) = 1/(2d)$. ■

Denote by N the number of points that are returned in Step 1 of Algorithm 4, and note that $N \geq nk/d$. From Lemma 3.2.4 in Vershynin (2018), we get that with probability larger than an absolute constant, the random initialitation gives a point $\mathbf{w}^{(0)}$, so that $\mathbf{w}^{(0)} \cdot \mathbf{v} \geq 1/(2\sqrt{k})$. We start

the analysis of our algorithm. We show that if Algorithm 4 terminates, then $1/(4d)$-fraction of points is correctly classified. Note that Algorithm 4 terminates if the algorithm makes $5d \log d$ mistakes or when $|C| \geq n/(4d)$, i.e., the algorithm classified $1/d$-fraction of points correctly which is the goal of the algorithm. The only bad event is if the algorithm terminated after $5d \log d$ mistakes and $|C| < n/(4d)$. We argue that this cannot happen. Let n_i be the remaining points in i iteration. Note that $N = n_i + |C|$. We make use of the following lemma (a variant of which was shown in Dunagan and Vempala (2004)).

Lemma 34 (Margin Perceptron (Dunagan and Vempala, 2004)) *Let $\mathbf{v}, \mathbf{w}^{(0)} \in \mathbb{R}^d$ be unit vectors such that $\mathbf{v} \cdot \mathbf{w}^{(0)} \geq \alpha$, for some $\alpha > 0$. Assume the following: $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \mathbf{x}^{(t)}(\mathbf{x}^{(t)} \cdot \mathbf{w}^{(t)})$ and let $t_0 \in Z_+$, so that for all $t \in \mathbb{Z}_+$ with $t \leq t_0$, $|\mathbf{x}^{(t)} \cdot \mathbf{w}^{(t)}| \geq \beta \|\mathbf{w}^{(t)}\|_2$ and $(\mathbf{x}^{(t)} \cdot \mathbf{w}^{(t)})(\mathbf{x}^{(t)} \cdot \mathbf{v}) < 0$. Then, $t_0 \leq (2/\beta^2) \log(1/\alpha)$.*

Proof From our assumption, we have that $\mathbf{w}^{(0)} \cdot \mathbf{v} \geq \alpha$. We have that

$$\mathbf{w}^{(t+1)} \cdot \mathbf{v} = (\mathbf{w}^{(t)} - (\mathbf{w}^{(t)} \cdot \mathbf{x}^{(t)})\mathbf{x}^{(t)}) \cdot \mathbf{v} = \mathbf{w}^{(t)} \cdot \mathbf{v} - (\mathbf{w}^{(t)} \cdot \mathbf{x}^{(t)})(\mathbf{x}^{(t)} \cdot \mathbf{v}) \geq \mathbf{w}^{(t)} \cdot \mathbf{v} \,,$$

where we used that $(\mathbf{w}^{(t)} \cdot \mathbf{x}^{(t)})(\mathbf{x}^{(t)} \cdot \mathbf{v}) \leq 0$. Therefore, for all $\mathbf{w}^{(t)}$, we have inner product with the target vector at least as large as the initialization, i.e., $\mathbf{w}^{(t)} \cdot \mathbf{v} \geq \alpha$. We show that the norm of $\mathbf{w}$ decreases multiplicatively. We have that

$$\begin{aligned} \|\mathbf{w}^{(t)} - (\mathbf{w}^{(t)} \cdot \mathbf{x}^{(t)})\mathbf{x}^{(t)}\|_2^2 &= \|\mathbf{w}^{(t)}\|_2^2 - (\mathbf{w}^{(t)} \cdot \mathbf{x}^{(t)})^2 \\ &\leq \|\mathbf{w}^{(t)}\|_2^2 (1 - \beta^2) \,. \end{aligned}$$

Hence, after t iterations, we have that $\|\mathbf{w}^{(t)}\|_2 \leq (1-\beta^2)^{t/2} \leq \exp(-t\beta^2/2)$. If $t \geq (2/\beta^2)\log(1/\alpha)$, we would have $\mathbf{w}^{(t)} \cdot \mathbf{v}/\|\mathbf{w}^{(t)}\|_2 > 1$, which is a contradiction. Hence, after $t = (2/\beta^2)\log(1/\alpha)$ updates, we have that either $|\mathbf{x}^{(t)} \cdot \mathbf{w}^{(t)}| \leq \beta \|\mathbf{w}^{(t)}\|_2$ or $(\mathbf{x}^{(t)} \cdot \mathbf{w}^{(t)})(\mathbf{x}^{(t)} \cdot \mathbf{v}) \geq 0$. ■

Assume that after the $t_1 = (5d \log d - 1)$ mistake, $|C| < n/(4d)$. That means $n_t = N - |C| \geq n(k/d - 1/(4d)) \geq N/2$, as $d \geq 1$. Let $\mathcal{S}_t = \{\mathbf{x}^{(i)} : |\mathbf{w}^{(t)} \cdot \mathbf{x}^{(i)}| \geq 1/(2\sqrt{k})\}$. From Lemma 33, we have that for each t, we have $|\mathcal{S}_t| \geq N/(4k)$ and combining with the fact that $n_t \geq N/2$, that means that either in each iteration, the algorithm makes no mistakes in the set $\mathcal{S}_t$, which means that $|C| \geq N/(4k)$ and the algorithm terminates, or that it makes one mistake in the set $\mathcal{S}_t$, which means that if $\mathbf{x}^{(t)}$ is the vector that $\mathbf{w}^{(t)}$ made a mistake then $|\mathbf{w}^{(t)} \cdot \mathbf{x}^{(t)}| \geq 1/(2\sqrt{k})$. Hence, condition to the event that the algorithm did not terminate before the iteration t_0, then by Lemma 34 if $t_0 \geq 5d \log d$, then $\mathbf{w}^{(t_0)}$ makes no mistakes in the set $\mathcal{S}_{t_0}$, so it classifies correctly $N/(4k)$ points, and the algorithm terminates. To derive the result, note that $N/(4k) \geq n/(4d)$ by definition.

D.2. Boosting: Obtaining Strong Self-Directed Learners from Weak Learners

In this section we present our boosting result showing that given a weak self-directed learner that labels some non-trivial part of the dataset one can obtain a strong self-directed that labels arbitrarily large fractions of the dataset.

Lemma 35 (Boosting) *Let $\mathcal{A}$ be a distribution-free self-directed learner that makes M mistakes and correctly labels a $(1-\alpha)$-fraction of X for some fixed $\alpha \in (0,1)$, with probability at least $c \in (0,1)$. Then, there exists a strong self-directed learner that makes $\widetilde{O}((M/c)\ \log(1/(\delta\epsilon))/\log(1/\alpha))$ mistakes and labels $(1-\epsilon)$-fraction of X with probability at least $1-\delta$.*

Proof We first boost the success probability of the self-directed learner to $1-\delta$ by repeating the algorithm $\log(1/\delta)$ times. When we perform independent runs of the algorithm, we stop a run if the algorithms make more than M mistakes. By the assumption that the algorithm succeeds in labeling at least $(1-\alpha)$-fraction with probability at least c, in each run, we obtain that after $O((1/c)\log(1/\delta'))$ runs one of them will succeed with probability at least $1-\delta'$. Therefore, we can boost the success probability of the algorithm to $1-\delta'$ by doing $O((M/c)\log(1/\delta'))$ mistakes. After performing a single successful run of the algorithm we have that the number of remaining unlabeled data is αn. Similarly, after k runs the number of unlabeled data is going to be at most $\alpha^k n$. In order for the fraction of unlabeled data to become smaller than ϵn we have to pick $k = \log(1/\epsilon)/\log(1/\alpha)$. Therefore, in order to have probability of success above $1-\delta$ overall, we can do a union bound over the k repetitions of the algorithm in order to cover $1-\epsilon$-fraction of the data. Therefore, we have to pick the success probability δ' of each run of the algorithm to be $\delta' = 1/(k\delta)$. We conclude that the overhead in the number of mistakes is a factor of $O(\log(1/\delta)(\log\log(1/\epsilon) - \log\log(1/\alpha)))$. ■

D.2.1. THE PROOF OF THEOREM 10

We restate and prove Theorem 10

Theorem 36 *Let $\mathcal{C}$ be the class of LTFs on $\mathbb{R}^d$ and let X be a set of n unlabeled points in $\mathbb{R}^d$. There exists a algorithm that runs in $\mathrm{poly}(d,n)$ time, makes $\widetilde{O}(d^2\log(d/(\epsilon\delta)))$ mistakes, and, with probability at least $1-\delta$, correctly classifies a $(1-\epsilon)$-fraction of the points of X.*

Proof From Proposition 11, we get that there is a weak learner so that with probability at least c, for some absolute constant $c>0$, classifies C/d-fraction of the points of X, where $C>0$ is an absolute constant and makes $O(d\log d)$ mistakes. Applying Lemma 35 on this algorithm, we get that the total number of mistakes is $\widetilde{O}(d\log(d/(\delta\epsilon)))/\log(1/(1-1/d))$. Using the inequality $\log(1+x)\le x$ for $x>-1$, we get that the total number of mistakes is $\widetilde{O}(d^2\log(1/(\delta\epsilon)))$. ■