Learning Noisy Perceptrons by a Perceptron in Polynomial Time

Edith Cohen*
AT&T Labs–Research
Florham Park, NJ 07932
edith@research.att.com http://www.research.att.com/~edith

Extended Abstract

Abstract

Learning perceptrons (linear threshold functions) from labeled examples is an important problem in machine learning. We consider the problem where labels are subjected to *random classification noise*. The problem was known to be PAC-learnable via a hypothesis that consists of a polynomial number of linear thresholds (due to Blum, Frieze, Kannan, and Vempala). The question whether a hypothesis that is itself a perceptron (a single threshold function) can be found in polynomial time was open. We show that indeed, noisy perceptrons are PAC learnable with a hypothesis that is a perceptron.

1. Introduction

The *perceptron problem* (learning a linear threshold function) is one of the oldest problems studied in machine learning. Linear threshold functions (perceptrons) are well-studied representation for learning from examples whose usefulness has been demonstrated on many datasets. Perceptrons generalize many representation studied in the computational learning theory literature, for example, disjunctions of literals, boolean threshold functions [12], and decision lists [14].

The noise-free formulation of the problem is as follows. Consider two probability distributions over R^n, one of "positive" and one of "negative" examples such that there exists a hyperplane (linear threshold function) that separates the regions where the distributions have nonzero probability. Given access to labeled examples (random samples), the goal is to find a linear threshold function that classifies correctly $(1-\epsilon)$ of the examples. A straightforward transformation reduces the problem to such where the desired hyperplane goes through the origin (replace each example $\mathbf{x} = (x_1, \ldots, x_n) \in R^n$ by $(x_1, \ldots, x_n, 1) \in R^{n+1}$). By negating all negatively-labeled vectors (see [8]), we obtain the following equivalent formulation. Let χ be a distribution on R^n such that there exist a halfspace H such that χ has zero measure outside H. Suppose we are given access to random samples from χ. The objective is to identify, with very high confidence, a vector $\mathbf{g}$ such that $\mathbf{x} \cdot \mathbf{g} > 0$ for at least $(1-\epsilon)$ of the vectors $\mathbf{x} \in \chi$. This noise-free perceptron problem is known to be solvable in polynomial time using any Linear Programming algorithm. One of the oldest algorithms for this problem is the *perceptron algorithm* [15, 1], which is still commonly-used and has a good empirical behaviour [3]. The algorithm converges to a good solution, but the running time depends on a *wiggle room* parameter which could be exponentially small in the size of the binary representation of the examples. Hence, the perceptron algorithm does not have a polynomial running time.

This paper focuses on the noisy version of the perceptron problem with *random classification noise.* A noise level of $0 < \Delta < 0.5$ means that the label of each example (positive or negative) is flipped (equivalently, the sign of a vector sampled from χ is negated) with probability Δ before being presented to us. The objective here is to identify a halfspace that contains at least $(1-\epsilon)$ of the *original* distribution. Our main

*Part of this research was done while the author was visiting U.C. Berkeley and supported by NSF grant CCR-9616139

result establishes that this problem is PAC learnable. (See [10] for background on PAC learning. Knowledge of these terms in not necessary for reading this paper.)

An equivalent formulation of the noisy perceptron problem is as follows: Let χ be a distribution in R^n such that there exist a vector $\mathbf{h}$ such that for any $\mathbf{x} \in \chi$, $\mathbf{h} \cdot \mathbf{x} > 0$. The distribution χ is subject to noise, resulting in a new distribution $\tilde{\chi}$. Let $0 < \Delta < 0.5$ be the amount of random noise. $\tilde{\chi}$ is generated from χ by negating vectors with probability Δ. That is, a random sample $\tilde{\mathbf{x}} \in \tilde{\chi}$ can be seen as drawing a sample $\mathbf{x} \in \chi$ and then outputting

$$\tilde{\mathbf{x}} = \begin{cases} \mathbf{x} & \text{with probability } 1 - \Delta \\ -\mathbf{x} & \text{with probability } \Delta. \end{cases}$$

We are given access to samples from $\tilde{\chi}$. The objective is to find a vector $\mathbf{g}$ such that $\text{Prob}_{\chi}\{\mathbf{x} \cdot \mathbf{g} \leq 0\} < \epsilon$. Our main result is showing that for any χ, Δ, and ϵ, this can be done with very high probability $(1 - e^{-poly(b,n,\epsilon^{-1},(1-2\Delta)^{-1})})$ in time polynomial in b, n, ϵ^{-1}, and $(1-2\Delta)^{-1}$, where b is the maximum number of bits in a binary representation of the sampled vectors.

Some notable previous work on the problem is a result by Kearns [9] who established PAC learnability for some class of distributions (radially symmetric), by showing that the problem falls in his statistical query model. Bylander [5] provided a modification of the (non polynomial) perceptron algorithm that under some additional assumptions, converges in the presence of random classification noise. A recent major step was taken by, Blum, Frieze, Kannan, and Vempala [4]. They showed that noisy linear threshold functions (with random classification noise) can be learned (in the PAC model sense) in polynomial time via a hypothesis that consists of a *polynomial number* of linear thresholds (hyperplanes). Their hypothesis tests a sample point sequentially against a polynomial number of linear thresholds until one that agrees with the point is encountered (and the point gets a positive label) or none does (negative label). The Blum *et al.* [4] algorithm utilizes Bylander's modified perceptron algorithm repeatedly in many iterations. Each iteration runs in polynomial time and yields another condition to the classification rule that classifies well a polynomial fraction of the remaining unclassified examples. Blum *et al.* posed as an open problem the question whether, as is the case for the noise-free version, noisy perceptrons (linear threshold functions subject to random classification noise) are PAC learnable via a hypothesis that consists of *a single threshold function.* (Can a single hyperplane that separates at least $1 - \epsilon$ of the distribution be identified in polynomial time?) Note that such a hyperplane exists by the definition of the problem. Polynomial time learning of noisy perceptrons by a *single* threshold function has both theoretical and practical significance. We answer their question on the affirmative.

The noise-free perceptron problem is known to be polynomial-time solvable using any Linear Programming algorithm. The algorithm simply converts sample points $\mathbf{x} \in \chi$ to constraints $\mathbf{x} \cdot \mathbf{g} > 0$ and solves the resulting linear system. VC dimension argument establishes that a polynomial number $(O(\epsilon^{-1} \log \delta^{-1} + n\epsilon^{-1} \log \epsilon^{-1}))$ of sample points is very likely to yield a solution (probability $1 - \delta$). This LP approach does not seem to generalize to the noisy problem. In the presence of noise, the system obtained as described above is generally infeasible. The objective is to identify a vector $\mathbf{g}$ consistent with "most" of the constraints. The difficulty stems from the fact that it is NP-hard to find a solution to an infeasible linear system that minimizes the number of unsatisfied constraints. The problem remains NP-hard under some relaxations when an exact minimum is not required. Our algorithm utilizes Linear Programming via a different approach. We apply the Ellipsoid algorithm to the (non-explicit) system formed by the non-noisy constraints. The Ellipsoid algorithm solves such systems with the aid of a *separation oracle* that produces violated constraints in response to inappropriate classifiers (infeasible points). The main challenge in our work was to implement a "separation oracle" which is correct with very high probability. The main ingredients we used were leveling down the effects of noise by "averaging", restricting the distribution to "slices" of the space, and using the *outliers removal lemma* of [4]. Section 2 provides preliminaries and an outline of our algorithm. Section 3 is concerned with the "separation oracle". Section 4 discusses our use of the Ellipsoid algorithm. Section 5 mentions extensions and issues for further work.

2. Preliminaries and Outline

By *high confidence* or *high probability* we mean probability at least $1 - e^{-\text{poly}_1(n,b,\epsilon^{-1},(1-2\Delta)^{-1})}$, for any poly_1 of our choice. By *polynomial* (for the time or number of samples) we mean: no larger than some $\text{poly}_2(n, b, \epsilon^{-1}, (1 - 2\Delta)^{-1})$. Conveniently, a polynomial number of polynomial-time high-probability events is also a polynomial-time high-probability event. For notational convenience we interchange between the distribution being defined on the n dimensional sphere S^n or on the n-dimensional space R^n. When considering a vector we sometimes interchange between the vector itself and the normalized version. The correct meaning should be clear from context. For the sake of clean and clear presentation, we make several simplifying assumptions that do not affect generality. We assume that the noise level $\Delta < 0.5$ is known. We assume that χ has zero measure on any one hyperplane going through the origin, that is, for any $\mathbf{u} \neq \mathbf{0}$, $\text{Prob}_{\mathbf{x}\in\chi}\{\mathbf{x}\cdot\mathbf{u} = 0\} = 0$.

We utilize the following additive version of Chernoff bound [6] throughout our analysis.

Proposition 2.1 [Chernoff Bound] *Consider N independent and identically distributed random variables $x_1, \ldots, x_N$ defined on the interval $[0,1]$ with expectation μ. For any $\eta > 0$ we have*

$$\text{Prob}\left\{\frac{\sum_{i=1}^{N} x_i}{N} > \mu + \eta\right\} \leq e^{-2N\eta^2}$$

$$\text{Prob}\left\{\frac{\sum_{i=1}^{N} x_i}{N} < \mu - \eta\right\} \leq e^{-2N\eta^2}$$

2.1. Good and Bad Vectors

Definition 2.2 A vector $\mathbf{g}$ is *Good* if

$$\text{Prob}_{\chi}\{\mathbf{x}\cdot\mathbf{g} < 0\} < \epsilon$$

or equivalently,

$$\text{Prob}_{\tilde{\chi}}\{\mathbf{x}\cdot\mathbf{g} < 0\} < \Delta + \epsilon(1-2\Delta)\ .$$

(since $\Delta(1-\epsilon) + (1-\Delta)\epsilon = \Delta + \epsilon(1-2\Delta)$.) Figure 1 visualizes good vectors for a distribution χ on R^2. The black circles correspond to samples from χ. The good vectors are all those that exclude at most one sample. Our objective is to identify a good vector. Naturally, *bad vectors* are vectors that are not good. However, we relax the definition slightly and allow a grey area: a vector $\mathbf{v}$ is *bad* if

$$\text{Prob}_{\chi}\{\mathbf{x}\cdot\mathbf{v} < 0\} > \epsilon/2$$

or equivalently,

$$\text{Prob}_{\tilde{\chi}}\{\mathbf{x}\cdot\mathbf{v} < 0\} > \Delta + \frac{\epsilon(1-2\Delta)}{2}\ .$$

The grey area is introduced to absorb estimation inaccuracies.

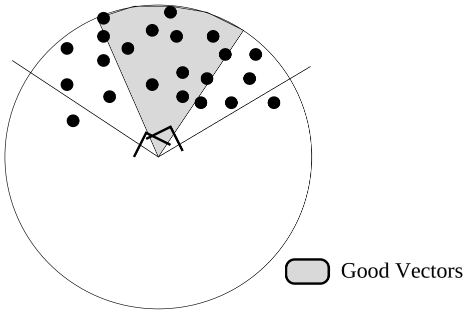

Figure 1: Visualizing good vectors ($\epsilon \leq 1/22$)

Badness or goodness of vectors can be identified with high probability using polynomially many samples. Vectors which are not good are identified as bad and vectors which are not bad are identified as good. "Grey" vectors are identified either way. The following proposition is established using Chernoff bounds.

Proposition 2.3 *For any vector $\mathbf{v} \in R^n$, badness or goodness can be identified using r samples from $\tilde{\chi}$ with confidence at least $1 - e^{-\Omega(r(\epsilon(1-2\Delta))^2)}$.*

2.2. Outliers Removal

The Outlier Removal Lemma of Blum *et al.* [4] states that for any distribution χ on R^n and $\epsilon' \geq 2^{-nb}$, we can remove ϵ' fraction of the distribution χ to obtain χ' such that for all $\mathbf{u} \in R^n$

$$\max_{\mathbf{x}\in\chi'} |\mathbf{x}\cdot\mathbf{u}| \leq \beta \mathop{\text{E}}_{\mathbf{x}\in\chi'} |\mathbf{x}\cdot\mathbf{u}|\ .$$

Where $\beta = \text{poly}(n, b, \epsilon'^{-1})$. The construction is performed in time $\text{poly}(n, b, \epsilon'^{-1})$. The removal procedure imposes a $\text{poly}(n, b, \epsilon'^{-1})$ time overhead on obtaining each sample from $\tilde{\chi}$.

Our algorithm utilizes outliers removal. The algorithm restricts the distribution to *slices* that are symmetric subsets of R^n. We need outliers removed both from the original distribution χ or from χ restricted to a slice. We always consider slices one at a time and do not select them recursively. Therefore, outliers can be removed when a slice is considered, with respect to the current slice alone. For cleanliness of presentation, we assume from here onward that outliers removal is done transparently. That is, when a slice of χ is considered, we assume that outliers are "automatically" removed. The loss of a fraction of the distribution can be absorbed by adjusting the allowed ϵ by a constant factor.

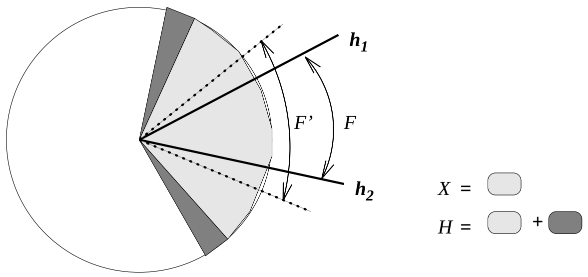

Figure 2: 2-dim example of X, H, F, F', and $\{\mathbf{h}_1, \mathbf{h}_2\}$.

2.3. The Feasible Region

The material in this section is needed for the analysis only. Standard polyhedral theory (see [7]) establishes the following: (See Figure 2 for a 2-dimensional example.)

Proposition 2.4 *There exist a set of $k = n$ vectors $\{\mathbf{h}_1, \ldots \mathbf{h}_k\} \subset S^n$ such that:*

1. *For $i = 1, \ldots, k$, $\forall \mathbf{x} \in \chi$ $\mathbf{h}_i \cdot \mathbf{x} > 0$.*
2. *Let $F = B^n \cap \text{cone}(\mathbf{h}_1, \ldots, \mathbf{h}_k)$, where B^n is the n-dimensional unit ball. We have $\text{vol}(F) \geq 2^{-poly(b,n)}$.*

In Section 3 we establish that some properties hold with high confidence for the vectors $\mathbf{h}_1, \ldots, \mathbf{h}_k$. These vectors are not provided explicitly. Note that it suffices to show that the property holds with high confidence for one generic $\mathbf{h} \in \{\mathbf{h}_1, \ldots, \mathbf{h}_k\}$. This implies that the property holds with high confidence for any fixed collection of $k = O(2^{\text{poly}})$ such vectors. Hence, we prove claims with respect to a single $\mathbf{h}$.

Denote $H = \{\mathbf{x} \in R^n | \mathbf{h}_i \cdot \mathbf{x} > 0 \text{ for } i = 1, \ldots, k\}$. The vectors $\mathbf{h}_1, \ldots, \mathbf{h}_k$, a generic $\mathbf{h}$, and the sets F and H are referred to later on in establishing correctness and convergence of our algorithm.

2.4. Sketch of our algorithm

Definition 2.5 A vector $\mathbf{w}$ is a *witness* for a bad vector $\mathbf{v}$ if

1. $\mathbf{w}$ is in H ($\mathbf{w} \cdot \mathbf{h}_i > 0$ for $i = 1, \ldots k$). (see Section 2.3).
2. $\mathbf{w}$ has a negative dot product with $\mathbf{v}$ ($\mathbf{w} \cdot \mathbf{v} < 0$).

The following theorem, which is proved in Section 3, is central to our algorithm:

Theorem 2.6 Witness Identification: *For any bad vector $\mathbf{v} \in R^n$, a witness vector $\mathbf{w}$ can be found in polynomial time with very high confidence*

In Section 4 we employ the *Ellipsoid Algorithm* [11] to find a good vector with Theorem 2.6 serving as the *separation oracle.* The Ellipsoid algorithm is applied to the following (very large and non-explicit) linear system of inequalities:

find a vector $\mathbf{g} \in R^n$ such that: for all $\mathbf{x} \in H$, $\mathbf{g} \cdot \mathbf{x} > 0$.

In each iteration the Ellipsoid algorithm considers a vector $\mathbf{v} \in R^n$ (center of current ellipsoid). The vector $\mathbf{v}$ is tested for goodness. If $\mathbf{v}$ is good, it constitutes a solution and we stop. If $\mathbf{v}$ is bad, then Theorem 2.6 finds a witness $\mathbf{w}$. The inequality $\mathbf{x} \cdot \mathbf{w} < 0$ constitutes a constraint violated by $\mathbf{v}$ and is presented to the Ellipsoid algorithm.

It is easy to see that if $\mathbf{v}$ is bad then $\epsilon(1 - 2\Delta)/2$ fraction of the samples $\mathbf{x} \in \tilde{\chi}$ qualify as witnesses. The challenge in proving Theorem 2.6 is in identifying a vector that with *very high confidence* is a witness. Property 2 of a witness can be easily verified. The difficulty is in verifying Property 1, since the vectors $\mathbf{h}_i$ (or any other feasible vectors) are not known. A confidence of at least $1 - 1/\text{poly}$ in identifying witnesses is necessary for a good success probability of the Ellipsoid algorithm. The algorithm requires correct witnesses in *all* (polynomially many) iterations.

3. Proof of the Witness Identification Theorem

We present an algorithm that for any bad vector $\mathbf{v}$ finds a witness $\mathbf{w}$ as promised in Theorem 2.6. Note that to establish Property 1 of a witness, it is sufficient to show that it holds with high confidence for a generic $\mathbf{h} \in \{\mathbf{h}_1, \ldots, \mathbf{h}_k\}$.

We sketch and motivate our approach. Consider samples $\mathbf{x} \in \tilde{\chi}$ such that $\mathbf{v} \cdot \mathbf{x} < 0$. Even though $\frac{\epsilon(1-2\Delta)}{2\Delta+\epsilon(1-2\Delta)}$ fraction of these samples are in χ and hence in H, and therefore, qualify as witnesses, we can not identify with high enough confidence even one such sample. What we can do, fortunately, is consider *centers of mass.* The centers of mass of all samples $\mathbf{x} \in \tilde{\chi}$ where $\mathbf{x} \cdot \mathbf{v} < 0$ and of all samples $\mathbf{x} \in \tilde{\chi}$ where

$\mathbf{x}\cdot\mathbf{v} > 0$ can be used to obtain an unbiased estimate of the center of mass $\mathbf{m}^-$ of all $\mathbf{x}\in\chi$ where $\mathbf{x}\cdot\mathbf{v} < 0$. The vector $\mathbf{m}^-$ is a witness: By definition it is in H and has a negative dot product with $\mathbf{v}$. The impediment is that $\mathbf{m}^-$ can not be estimated accurately enough from polynomial number of samples to guarantee the witness properties. We use two key ideas to overcome that. The first is based on the fact that after removing outliers, for $\nu \geq 1/\mathrm{poly}$, $\mathbf{w} = (1-\nu)\mathbf{m}^- + \nu\mathbf{m}^+$, ($\mathbf{m}^+$ is the center of mass of all $\mathbf{x}\in\chi$ where $\mathbf{x}\cdot\mathbf{v} < 0$) for $\nu \geq 1/\mathrm{poly}$, can we estimated sufficiently accurately with a polynomial number of samples such that with very high probability $\hat{\mathbf{w}}\cdot\mathbf{h} > 0$. Hence, we are assured that witness property 1 holds. The caveat, however, is that $\mathbf{m}^+\cdot\mathbf{v}$ may be exponentially larger than $|\mathbf{m}^-\cdot\mathbf{v}|$, and therefore, there is no good choice of ν for which $\mathbf{w}\cdot\mathbf{v} < 0$ (witness property 2 is violated). To overcome this, we use our second key idea. Observe that if all $\mathbf{x}\in\chi$ have approximately the same values of $|\mathbf{x}\cdot\mathbf{v}|$, we have $\mathbf{m}^+\cdot\mathbf{v} = \Omega(\epsilon^{-1}\mathbf{m}^-\cdot\mathbf{v})$ and the "bad case" can not happen. To achive that, we restrict χ to a sufficiently dense (relative to χ) and "narrow" (relative to $\mathbf{v}$) symmetric regions. *Symmetric regions* are such that if $\mathbf{z}$ is in the region, so is $-\mathbf{z}$. We use symmetric regions of the form $\{\mathbf{x}\in R^n | a_2 > |\mathbf{v}\cdot\mathbf{x}| > a_1\}$ and refer to them as *slices*. An alternative solution we discuss later on is to scale all vectors such that $|\mathbf{x}\cdot\mathbf{v}| = 1$.

3.1. Definitions and Properties

Let ψ be a distribution over the sphere S^n and $\mathbf{u}\in R^n$ be a vector. We define

$$\begin{aligned}
H^-(\mathbf{u}) &= \{\mathbf{x}\in R^n | \mathbf{u}\cdot\mathbf{x} < 0\}\\
H^+(\mathbf{u}) &= \{\mathbf{x}\in R^n | \mathbf{u}\cdot\mathbf{x} > 0\}\\
M^+(\psi,\mathbf{u}) &= \int_{\mathbf{x}\in H^+(\mathbf{u})} \psi(\mathbf{x})d\mathbf{x}\\
M^-(\psi,\mathbf{u}) &= \int_{\mathbf{x}\in H^-(\mathbf{u})} \psi(\mathbf{x})d\mathbf{x}\\
\mathbf{m}^+(\psi,\mathbf{u}) &= \int_{\mathbf{x}\in H^+(\mathbf{u})\cap S^n} \mathbf{x}\psi(\mathbf{x})d\mathbf{x}\\
\mathbf{m}^-(\psi,\mathbf{u}) &= \int_{\mathbf{x}\in H^-(\mathbf{u})\cap S^n} \mathbf{x}\psi(\mathbf{x})d\mathbf{x}
\end{aligned}$$

In words: $M^\pm(\psi,\mathbf{u})$ is the measure of ψ on $H^\pm(\mathbf{u})$, and the vectors $\mathbf{m}^\pm$ are the centers of mass of the part of the distribution on $H^\pm(\mathbf{u})$. Our assumptions imply

$$M^+(\chi,\mathbf{v}) + M^-(\chi,\mathbf{v}) = M^+(\tilde{\chi},\mathbf{v}) + M^-(\tilde{\chi},\mathbf{v}) = 1\ .$$

To motivate these definitions, notice that if $\mathbf{v}$ is a bad vector then $\mathbf{m}^-(\chi,\mathbf{v})\cdot\mathbf{h} > 0$ and $\mathbf{m}^-(\chi,\mathbf{v})\cdot\mathbf{v} < 0$. Hence, $\mathbf{m}^-(\chi,\mathbf{v})$ is a witness.

The quantities $M^\pm(\tilde{\chi},\mathbf{v})$ and $\mathbf{m}^\pm(\tilde{\chi},\mathbf{v})$ can be estimated by sampling from $\tilde{\chi}$. The quantities $M^\pm(\chi,\mathbf{v})$ and $\mathbf{m}^\pm(\chi,\mathbf{v})$ are related to them by the following linear equations:

$$\begin{aligned}
M^+(\tilde{\chi},\mathbf{v}) &= (1-\Delta)M^+(\chi,\mathbf{v}) + \Delta M^-(\chi,\mathbf{v}) &(1)\\
M^-(\tilde{\chi},\mathbf{v}) &= (1-\Delta)M^-(\chi,\mathbf{v}) + \Delta M^+(\chi,\mathbf{v}) &(2)\\
\mathbf{m}^+(\tilde{\chi},\mathbf{v}) &= (1-\Delta)\mathbf{m}^+(\chi,\mathbf{v}) - \boldsymbol{\Delta}\mathbf{m}^-(\chi,\mathbf{v}) &(3)\\
\mathbf{m}^-(\tilde{\chi},\mathbf{v}) &= (1-\Delta)\mathbf{m}^-(\chi,\mathbf{v}) - \boldsymbol{\Delta}\mathbf{m}^+(\chi,\mathbf{v}) &(4)
\end{aligned}$$

Solving the linear systems in Equ. 1–2 and Equ. 3–4 gives:

$$\begin{aligned}
M^+(\chi,\mathbf{v}) &= \frac{(1-\Delta)M^+(\tilde{\chi},\mathbf{v}) - \Delta M^-(\tilde{\chi},\mathbf{v})}{1-2\Delta} &(5)\\
M^-(\chi,\mathbf{v}) &= \frac{(1-\Delta)M^-(\tilde{\chi},\mathbf{v}) - \Delta M^+(\tilde{\chi},\mathbf{v})}{1-2\Delta} &(6)\\
\mathbf{m}^+(\chi,\mathbf{v}) &= \frac{(1-\Delta)\mathbf{m}^+(\tilde{\chi},\mathbf{v}) + \boldsymbol{\Delta}\mathbf{m}^-(\tilde{\chi},\mathbf{v})}{1-2\Delta} &(7)\\
\mathbf{m}^-(\chi,\mathbf{v}) &= \frac{(1-\Delta)\mathbf{m}^-(\tilde{\chi},\mathbf{v}) + \boldsymbol{\Delta}\mathbf{m}^+(\tilde{\chi},\mathbf{v})}{1-2\Delta} &(8)
\end{aligned}$$

3.2. Estimating $M^\pm(\tilde{\chi},\mathbf{v})$ and $\mathbf{m}^\pm(\tilde{\chi},\mathbf{v})$ from samples

Let $S = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\} \subset \chi$ be N (normalized) random samples from $\tilde{\chi}$. (the magnitude of N is polynomial and will be discussed later). Let $S^+ = S\cap H^+(\mathbf{v}) = \{\mathbf{x}\in S | \mathbf{x}\cdot\mathbf{v} > 0\}$ be all samples with positive dot products with $\mathbf{v}$ and similarly, let $S^- = S\cap H^-(\mathbf{v}) = \{\mathbf{x}\in S | \mathbf{x}\cdot\mathbf{v} < 0\}$.

$$\begin{aligned}
\hat{M}^+(\tilde{\chi},\mathbf{v}) &= \frac{|S^+|}{|S|} &(9)\\
\hat{M}^-(\tilde{\chi},\mathbf{v}) &= \frac{|S^-|}{|S|} &(10)\\
\hat{\mathbf{m}}^+(\tilde{\chi},\mathbf{v}) &= \frac{\sum_{\mathbf{x}\in S^+}\mathbf{x}}{|S|} &(11)\\
\hat{\mathbf{m}}^-(\tilde{\chi},\mathbf{v}) &= \frac{\sum_{\mathbf{x}\in S^-}\mathbf{x}}{|S|} &(12)
\end{aligned}$$

Estimates $\hat{\mathbf{m}}^\pm(\chi,\mathbf{v})$ and $\hat{M}^\pm(\chi,\mathbf{v})$ can be computed from $\hat{\mathbf{m}}^\pm(\tilde{\chi},\mathbf{v})$ and $\hat{M}^\pm(\tilde{\chi},\mathbf{v})$ in the obvious manner according to Equ. 5–8.

Shorthands. When χ and $\mathbf{v}$ are clear from the context, we use the notation $M^+ \equiv M^+(\chi,\mathbf{v})$, $\mathbf{m}^+ \equiv \mathbf{m}^+(\chi,\mathbf{v})$, $M^+_\sim \equiv M^+(\tilde{\chi},\mathbf{v})$, and $\mathbf{m}^+_\sim \equiv \mathbf{m}^+(\tilde{\chi},\mathbf{v})$. The quantities M^-, $\mathbf{m}^-$, $M^-_\sim$, and $\mathbf{m}^-_\sim$, are defined similarly. For the estimates we use $\hat{\mathbf{m}}^+_\sim \equiv \hat{\mathbf{m}}^+(\tilde{\chi},\mathbf{v})$, $\hat{\mathbf{m}}^-_\sim \equiv \hat{\mathbf{m}}^-(\tilde{\chi},\mathbf{v})$, and similarly define $\hat{M}^\pm_\sim$ and $\hat{m}^\pm$.

3.3. Accuracy of Estimates

Consider a distribution ψ on S^n and an arbitrary vector $\mathbf{u} \in S^n$. We apply Chernoff bounds to bound the accuracy and confidence of estimating from samples the expected magnitudes of projections on $\mathbf{u}$.

Proposition 3.1 *Let ψ be a distribution over the sphere S^n and let $\mathbf{u} \in S^n$ be a vector. Suppose that outliers were removed with respect to $\mathbf{u}$, that is,*

$$\max_{\mathbf{x}\in\psi} |\mathbf{x}\cdot\mathbf{u}| \leq \beta \mathop{\mathrm{E}}_{\mathbf{x}\in\psi}\{|\mathbf{x}\cdot\mathbf{u}|\} .$$

Let $\mathbf{a} = \mathrm{E}_{\mathbf{x}\in\psi}\{\mathbf{x}\}$ be the center of mass of ψ. Consider a set S of $r = |S|$ samples drawn from ψ. Let

$$\hat{\mathbf{a}}_r = \frac{\sum_{\mathbf{x}\in S}\mathbf{x}}{|S|}$$

be an estimate of $\mathbf{a}$. For all $\gamma > 0$,

$$|(\mathbf{a}-\hat{\mathbf{a}}_r)\cdot\mathbf{u}| \leq \gamma \mathop{\mathrm{E}}_{\mathbf{x}\in\psi}\{|\mathbf{x}\cdot\mathbf{u}|\} ,$$

with confidence at least $1 - e^{-\Omega(r\frac{\gamma^2}{\beta^2}))}$.

Proof: The random variable $\frac{\mathbf{x}\cdot\mathbf{u}}{\max_{\mathbf{x}\in\psi}|\mathbf{x}\cdot\mathbf{u}|}$ has values in the range $[-1,1]$ and we need to estimate its expectation to within $\pm\gamma/\beta$. This is equivalent to estimating the expectation of an r.v. in $[0,1]$ to within $\eta = \frac{\gamma}{2\beta}$. From Chernoff bounds (see Proposition 2.1), r samples provide confidence of

$$1 - 2e^{-2r\eta^2} = 1 - 2e^{-r\frac{\gamma^2}{2\beta^2}} .$$

∎

Consider the noisy distribution $\tilde{\chi}$ and a vector $\mathbf{v}$. Let S be a set of samples from $\tilde{\chi}$ as in Section 3.2.

Let $\mathbf{u} \in R^n$ be an arbitrary vector.

Proposition 3.2 *If $|S| = N$ samples are used then for any $\gamma > 0$, with probability at least $1-e^{-\Omega(\frac{NM^+_\sim\gamma^2}{\beta^2})}$,*

$$|(\hat{\mathbf{m}}^+_\sim - \mathbf{m}^+_\sim)\cdot\mathbf{u}| \leq \gamma \mathop{\mathrm{E}}_{\mathbf{x}\in\tilde{\chi}\cap H^+(\mathbf{v})}\{|\mathbf{u}\cdot\mathbf{x}|\}M^+_\sim .$$

The symmetric claims hold for $\hat{\mathbf{m}}^-_\sim$ estimating $\mathbf{m}^-_\sim$.

Proof: Immediate from Proposition 3.1 since with high probability $|S^+| = \Omega(|S|M^+_\sim)$. ∎

Using $\mathrm{E}_{\mathbf{x}\in\tilde{\chi}}\{|\mathbf{u}\cdot\mathbf{x}|\} =$

$$\mathop{\mathrm{E}}_{\mathbf{x}\in\tilde{\chi}\cap H^+(\mathbf{v})}\{|\mathbf{u}\cdot\mathbf{x}|\}M^+_\sim + \mathop{\mathrm{E}}_{\mathbf{x}\in\tilde{\chi}\cap H^-(\mathbf{v})}\{|\mathbf{u}\cdot\mathbf{x}|\}M^-_\sim$$

we obtain the following:

Corollary 3.3 *With probability at least*

$$1 - e^{-\Omega(\frac{N\min\{M^+_\sim, M^-_\sim\}\gamma^2}{\beta^2})} ,$$

$$\begin{aligned} |(\hat{\mathbf{m}}^+_\sim - \mathbf{m}^+_\sim)\cdot\mathbf{u}| &\leq \gamma \mathop{\mathrm{E}}_{\mathbf{x}\in\tilde{\chi}}\{|\mathbf{u}\cdot\mathbf{x}|\} \\ |(\hat{\mathbf{m}}^-_\sim - \mathbf{m}^-_\sim)\cdot\mathbf{u}| &\leq \gamma \mathop{\mathrm{E}}_{\mathbf{x}\in\tilde{\chi}}\{|\mathbf{u}\cdot\mathbf{x}|\} \end{aligned}$$

3.4. Obtaining a Witness

We motivate and sketch our approach. The vector $\mathbf{m}^-(\chi,\mathbf{v})$ has the two desired "witness" properties. It lies in the cone of χ and hence in H ($\mathbf{m}^-(\chi,\mathbf{v})\cdot\mathbf{h} > 0$) and $\mathbf{m}^-(\chi,\mathbf{v})\cdot\mathbf{v} < 0$. A desired result would be to show that if $\mathbf{v}$ is bad, these properties also hold for an estimate $\hat{\mathbf{m}}^-(\chi,\mathbf{v})$ achieved using a polynomial number of samples.

Unfortunately, this claim as is not necessarily true. The difficulty is hinted on Section 3.3. A polynomial number of samples can "guarantee" (with high confidence) accuracy of only $\pm\mathrm{E}_\chi(|\mathbf{x}\cdot\mathbf{h}|)/\mathrm{poly}$ for the estimate $|(\hat{\mathbf{m}}^- - \mathbf{m}^-)\cdot\mathbf{h}|$ and similarly, an accuracy of $\pm\mathrm{E}_\chi(|\mathbf{x}\cdot\mathbf{v}|)/\mathrm{poly}$ for $|(\hat{\mathbf{m}}^- - \mathbf{m}^-)\cdot\mathbf{v}|$. On the other hand, there are choices of χ and $\mathbf{v}$ where $\mathbf{m}^-\cdot\mathbf{h}$ (resp., $-\mathbf{m}^-\cdot\mathbf{v}$) are exponentially smaller than $\mathrm{E}\{|\mathbf{x}\cdot\mathbf{h}|\}$ (resp., $\mathrm{E}\{|\mathbf{x}\cdot\mathbf{v}|\}$), and therefore, we can not guarantee that the properties $\hat{\mathbf{m}}^-(\chi,\mathbf{v})\cdot\mathbf{h} > 0$ and $\hat{\mathbf{m}}^-(\chi,\mathbf{v})\cdot\mathbf{v} < 0$ hold.

We overcome this problem by restricting the distribution to an appropriate slice of R^n and estimating, instead of $\mathbf{m}^-(\chi,\mathbf{v})$, a linear combination of $\mathbf{m}^-(\chi,\mathbf{v})$ and $\mathbf{m}^+(\chi,\mathbf{v})$. This yields a witness such that with high confidence, its estimate obtained by a polynomial number of samples is also a witness. We elaborate on our approach:

- Observe that $\mathbf{m}^+\cdot\mathbf{h} + \mathbf{m}^-\cdot\mathbf{h} = \mathrm{E}\{|\mathbf{x}\cdot\mathbf{h}|\}$. Our estimates fail when $\mathbf{m}^-\cdot\mathbf{h}$ is exponentially smaller than $\mathrm{E}\{|\mathbf{x}\cdot\mathbf{h}|\}$. This can be corrected by replacing $\mathbf{m}^-$ by a linear combination of $\mathbf{m}^-$ and $\mathbf{m}^+$. The estimate for such a combination will guarantee membership in H with high probability. We use a vector of the form

$$\mathbf{m}^{-,\nu}(\chi,\mathbf{v}) = (1-\frac{\nu}{1-2\Delta})\mathbf{m}^- + \frac{\nu}{1-2\Delta}\mathbf{m}^+ ,$$

which can be estimated using

$$\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v}) = \frac{(1-\Delta-\nu)\hat{\mathbf{m}}^-_{\sim} + (\Delta+\nu)\hat{\mathbf{m}}^+_{\sim}}{1-2\Delta} .$$

Now we have $\mathbf{m}^{-,\nu}\cdot\mathbf{h} \geq \frac{\nu}{1-2\Delta}\,\mathrm{E}\{|\mathbf{x}\cdot\mathbf{h}|\}$. Therefore, when $\nu = \Omega(1/\mathrm{poly})$, a polynomial number of samples guarantees $\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{h} > 0$.

- However, there is a caveat. The linear combination $\mathbf{m}^{-,\nu}(\chi,\mathbf{v})$ on the original space may be such that $\mathbf{m}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{v} > 0$, and hence, $\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v})$ is unlikely to be a witness. To correct this, we restrict the distributions $\tilde{\chi}$ and χ to a *dense* (of measure at least 1/poly) *slice*,

$$\{\mathbf{x}\in S^n \,|\, a \leq |\mathbf{x}\cdot\mathbf{v}| \leq a(1+\mu)\}$$

(for a properly selected $\mu = \Theta(\epsilon(1-2\Delta)/\Delta)$). We select a slice such that $\mathbf{v}$ is bad on the slice (such a slice always exists and can be identified easily). We consider the new distribution χ' obtained by restricting χ to the slice. Note that $a \leq \mathrm{E}_{\mathbf{x}\in\chi'}\{|\mathbf{x}\cdot\mathbf{v}|\} \leq (1+\mu)a$. Therefore, the dot product $\hat{\mathbf{m}}^{-,\nu}(\chi',\mathbf{v})\cdot\mathbf{v}$ can be estimated up to additive error of a/poly. Moreover, we claim that for appropriate ν, $-\mathbf{m}^{-,\nu}(\chi',\mathbf{v})\cdot\mathbf{v} = \Omega(a/\mathrm{poly})$, and hence, with high confidence, $\hat{\mathbf{m}}^{-,\nu}(\chi',\mathbf{v})\cdot\mathbf{v} < 0$. To see the latter claim, note that $a(1+\mu) \geq -\mathbf{m}^-(\chi',\mathbf{v})\cdot\mathbf{v}/M^-(\chi',\mathbf{v}) \geq a$ and $a(1+\mu) \geq \mathbf{m}^+(\chi',\mathbf{v})\cdot\mathbf{v}/M^+(\chi',\mathbf{v}) \geq a$. Therefore, $-\mathbf{m}^{-,\nu}(\chi',\mathbf{v})\cdot\mathbf{v} > a\left((1-\frac{\nu}{1-2\Delta})M^-(\chi',\mathbf{v}) - \frac{\nu}{1-2\Delta}(1+\mu)M^+(\chi',\mathbf{v})\right) = \Omega(a\epsilon)$. The latter inequalities follow using the badness of $\mathbf{v}$ on the slice: $M^-(\chi',\mathbf{v}) = \Omega(\epsilon)$ $M^+(\chi',\mathbf{v}) < 1-\Omega(\epsilon)$ and choosing appropriate $\mu < 1$ and $\nu = \Theta(\epsilon(1-2\Delta))$.

In Section 3.5 we elaborate on choosing a slice and restricting χ to it. In Section 3.6 we consider linear combinations and conclude the proof of Theorem 2.6.

3.5. Choosing a Slice and Restricting χ

Definition 3.4 Consider the vector $\mathbf{v}$ and $0 < a_1 < a_2 < 1$. The (a_1,a_2)-*slice* of the sphere S^n is the set $\{\mathbf{x}\in S^n \,|\, a_1 \leq |\mathbf{x}\cdot\mathbf{v}| \leq a_2\}$. The *restricted distribution* χ_{a_1,a_2} is the distribution χ restricted to the (a_1,a_2)-slice. Let $0 \leq M_{a_1,a_2} \leq 1$ be the measure of χ on the (a_1,a_2)-slice. Figure 3 visualizes slices on the 2-dimensional sphere.

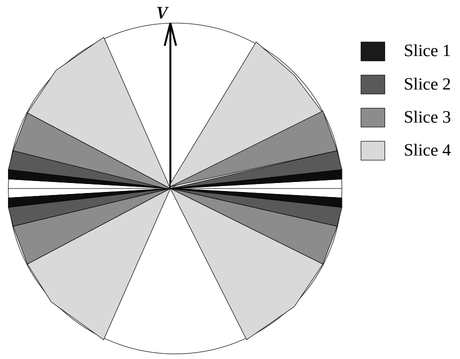

Figure 3: Visualizing slices on the 2-dimensional sphere

Since slices are symmetric and the noise is uniform, $\tilde{\chi}_{a_1,a_2}$ is the same distribution obtained by subjecting χ_{a_1,a_2} to random noise Δ.

To sample from $\tilde{\chi}_{a_1,a_2}$ we sample from $\tilde{\chi}$ repeatedly until we obtain a vector $\mathbf{x}$ that is in the (a_1,a_2)-slice. On average we need $O(M^{-1}_{a_1,a_2})$ samples from $\tilde{\chi}$ to obtain a sample from $\tilde{\chi}_{a_1,a_2}$.

Choosing an appropriate slice. A (a_1,a_2)-slice is *appropriate* when

- For $\mu = \min\{1/4, \frac{\epsilon(1-2\Delta)}{20\Delta}\}$, $a_2 = (1+\mu)a_1$.
- The slice is *fat*, That is, M_{a_1,a_2} is at least $\Omega(\frac{\epsilon(1-2\Delta)}{\mathrm{poly}_1})$, where poly_1 is the negated logarithm of the smallest possible value of $\mathbf{v}\cdot\mathbf{x}$ (for $\mathbf{x}\in\chi$) multiplied by $\log^{-1}(1+\mu)$. We will obtain the vectors $\mathbf{v}$ from the Ellipsoid algorithm. It follows that $M^{-1}_{a_1,a_2}$ is polynomial.
- The slice exhibits the badness behavior of $\mathbf{v}$. That is, if $\mathbf{v}$ is bad, then $M^-(\tilde{\chi}_{a_1,a_2},\mathbf{v}) >$

$$(\Delta+\frac{1}{4}\epsilon(1-2\Delta)(M^+(\tilde{\chi}_{a_1,a_2},\mathbf{v})+M^-(\tilde{\chi}_{a_1,a_2},\mathbf{v})) . \tag{13}$$

The proof of the following proposition is omitted:

Proposition 3.5 *If $\mathbf{v}$ is bad, then an appropriate slice always exists and can be identified with high confidence in polynomial time.*

The following proposition is immediate.

Proposition 3.6

$$M^-(\tilde{\chi}_{a_1,a_2},\mathbf{v})a_1 \le -\mathbf{m}^-(\tilde{\chi}_{a_1,a_2},\mathbf{v})\cdot\mathbf{v} \le M^-(\tilde{\chi}_{a_1,a_2},\mathbf{v})a_2$$

$$M^+(\tilde{\chi}_{a_1,a_2},\mathbf{v})a_1 \le \mathbf{m}^+(\tilde{\chi}_{a_1,a_2},\mathbf{v})\cdot\mathbf{v} \le M^+(\tilde{\chi}_{a_1,a_2},\mathbf{v})a_2$$

From here on, when we *restrict* $\tilde{\chi}$ *to a slice* (with respect to $\mathbf{v}$) an appropriate slice is selected as above and outliers are removed. We consider the new distribution χ_{a_1,a_2} of measure 1 on the slice. A sample from $\tilde{\chi}_{a_1,a_2}$ requires an expected number of $M^{-1}_{a_1,a_2}$ samples from $\tilde{\chi}$. That is, there is a polynomial overhead of $M^{-1}_{a_1,a_2}$ in the sampling from $\tilde{\chi}_{a_1,a_2}$.

3.6. Linear Combinations of Centers of Mass

In this subsection we assume that $\tilde{\chi}$ and χ are restricted distribution to an appropriate (a_1,a_2)-slice with respect to $\mathbf{v}$, as outlined in Subsection 3.5.

Consider the following linear combination for $0 \le \nu < (1-2\Delta)$: $\mathbf{m}^{-,\nu}(\chi,\mathbf{v}) =$

$$(1-\frac{\nu}{1-2\Delta})\mathbf{m}^-(\chi,\mathbf{v}) + \frac{\nu}{1-2\Delta}\mathbf{m}^+(\chi,\mathbf{v}) \qquad (14)$$

This combination can be viewed as overcompensating by ν for the noise. We choose

$$\nu = \epsilon(1-2\Delta)/20\ . \qquad (15)$$

From Equ. 7– 8 we obtain $\mathbf{m}^{-,\nu}(\chi,\mathbf{v}) =$

$$\frac{(1-\Delta-\nu)\mathbf{m}^-(\tilde{\chi},\mathbf{v}) + (\Delta+\nu)\mathbf{m}^+(\tilde{\chi},\mathbf{v})}{1-2\Delta} \qquad (16)$$

We compute the dot product values of the vector $\mathbf{m}^{-,\nu}(\chi,\mathbf{v})$ with $\mathbf{v}$ and with $\mathbf{h}$. It turns out that $\mathbf{m}^{-,\nu}(\chi,\mathbf{v})$ is a witness.

1. Using Proposition 3.6, Equ. 13, and substituting ν (Equ. 15) and $\mu = \min\{1/4, \frac{\epsilon(1-2\Delta)}{20\Delta}\}$ (Section 3.5) we obtain

$$\mathbf{m}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{v} \le -\frac{\epsilon}{8}a_1\ .$$

2. Using Equ. 14 we obtain

$$\mathbf{m}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{h} \ge \frac{\epsilon}{10}\mathop{\mathrm{E}}_{\mathbf{x}\in\chi}\{\mathbf{x}\cdot\mathbf{h}\}\ .$$

Consider the estimate: $\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v}) =$

$$\frac{(1-\Delta-\nu)\hat{\mathbf{m}}^-(\tilde{\chi},\mathbf{v}) + (\Delta+\nu)\hat{\mathbf{m}}^+(\tilde{\chi},\mathbf{v})}{1-2\Delta} \qquad (17)$$

We first bound the estimation accuracy of the dot products of $\hat{\mathbf{m}}^+_{\sim} = \hat{\mathbf{m}}^+(\tilde{\chi},\mathbf{v})$ and $\hat{\mathbf{m}}^-_{\sim} = \hat{\mathbf{m}}^+(\tilde{\chi},\mathbf{v})$ with $\mathbf{v}$ and $\mathbf{h}$ (using Section 3.3). From there we deduce the estimation accuracy for the dot products $\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{v}$ and $\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{h}$ and establish that with a polynomial amount of sampling, the estimation errors can be made small with respect to the dot product values. Hence, with high confidence, $\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v})$ is a witness.

Proposition 3.7 *If the estimates* $\hat{\mathbf{m}}^-(\tilde{\chi},\mathbf{v})$ *and* $\hat{\mathbf{m}}^+(\tilde{\chi},\mathbf{v})$ *are obtained using* N *samples then with confidence at least*

$$1-e^{-\Omega(\frac{N(1-2\Delta)^2\gamma^2}{\beta^2})}\ ,$$

$$|\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{u} - \mathbf{m}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{u}| \le \gamma \mathop{\mathrm{E}}_{\mathbf{x}\in\chi}\{|\mathbf{x}\cdot\mathbf{u}|\}\ .$$

Proof: Immediate from Corollary 3.3 and Equ. 17. We assume without loss of generality that $\min\{M^+_{\sim}, M^-_{\sim}\} \ge \Delta \ge 1/4$. (It is easy to introduce additional random noise and increase the noise level to $\max\{\Delta, 1/4\}$). ∎

To utilize Proposition 3.7, we bound $\mathrm{E}_{\tilde{\chi}}\{|\mathbf{u}\cdot\mathbf{x}|\}$ for $\mathbf{u}=\mathbf{v}$ and for $\mathbf{u}=\mathbf{h}$. Note that $\mathrm{E}_{\mathbf{x}\in\tilde{\chi}}\{|\mathbf{u}\cdot\mathbf{x}|\} = \mathrm{E}_{\mathbf{x}\in\chi}\{|\mathbf{u}\cdot\mathbf{x}|\}$.

If $\mathbf{u}=\mathbf{v}$ then, using Proposition 3.6,

$$\mathop{\mathrm{E}}_{\mathbf{x}\in\chi}\{|\mathbf{v}\cdot\mathbf{x}|\} = \mathbf{m}^+(\tilde{\chi},\mathbf{v})\cdot\mathbf{v} - \mathbf{m}^-(\tilde{\chi},\mathbf{v})\cdot\mathbf{v} \le a_2 \le 1.5a_1\ . \qquad (18)$$

If $\mathbf{u}=\mathbf{h}$ then

$$\mathop{\mathrm{E}}_{\mathbf{x}\in\chi}\{|\mathbf{h}\cdot\mathbf{x}|\} = \mathop{\mathrm{E}}_{\mathbf{x}\in\chi}\{\mathbf{h}\cdot\mathbf{x}\}\ . \qquad (19)$$

We need to choose N large enough such that with high confidence the estimation error is small with respect to the magnitudes, that is,

$$|\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{v} - \mathbf{m}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{v}| < \frac{\epsilon a_1}{8} \qquad (20)$$

$$|\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{h} - \mathbf{m}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{h}| < \frac{\epsilon}{10}\mathop{\mathrm{E}}_{\mathbf{x}\in\chi}\{\mathbf{x}\cdot\mathbf{h}\}\ . \qquad (21)$$

From Propositions 3.7 and Equ. 18 and 19 we obtain that for any $\gamma > 0$, with confidence at least

$$1-e^{-\Omega(\frac{N(1-2\Delta)^2\gamma^2}{\beta^2})}\ ,$$

we have

$$|\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{v}-\mathbf{m}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{v}|<1.5\gamma a_1 \quad (22)$$

$$|\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{h}-\mathbf{m}^{-,\nu}(\chi,\mathbf{v})\cdot\mathbf{h}|<\gamma \mathop{\mathrm{E}}_{\mathbf{x}\in\chi}\{\mathbf{x}\cdot\mathbf{h}\} \quad (23)$$

From Ineq. 20 and 22 we get $\gamma \leq \frac{\epsilon}{12}$ and from Ineq. 21 and 23 we obtain $\gamma \leq \frac{\epsilon}{10}$. Therefore, if $\gamma = \epsilon/13$ then for N samples, $\hat{\mathbf{m}}^{-,\nu}(\chi,\mathbf{v})$ is a witness with confidence

$$1-e^{-\Omega(\frac{N(1-2\Delta)^2\epsilon^2}{\beta^2})} .$$

Hence, a polynomial size N guarantees very high confidence. This concludes the proof of Theorem 2.6.

4. Employing the Ellipsoid Algorithm

Let H and F be as in Subsection 2.3. Consider the linear system of inequalities:

$$\text{Find } \mathbf{g}\in R^n, \text{ subject to: } \quad \forall\mathbf{x}\in H,\ \mathbf{g}\cdot\mathbf{x}>0 \ . \quad (24)$$

The set F is the feasible region of this system. Note that the number of inequalities is very large and none are explicitly provided, since H is not known.

The Ellipsoid algorithm [11] can solve such a system, given a *separation oracle* (see, e.g. [7, 16] for background). When the separation oracle is presented with a point $\mathbf{v}\in R^n$, it either determines that $\mathbf{v}$ is feasible and terminates, or returns a violated constraint. In our case, a violated constraint is a vector $\mathbf{w}\in H$ such that $\mathbf{w}\cdot\mathbf{v}<0$.

The Ellipsoid algorithm is defined for non-strict inequalities, but can be easily modified to handle strict inequalities as is the case here (strict inequalities mean that we need to find an interior point in the feasible region of the equivalent system with non-strict inequalities). In our problem, we need to avoid the trivial vector $\mathbf{v}=\mathbf{0}$. (It is the only vector that does not yield a violated constraint and is not a good vector.) The trivial vector can be excluded by perturbing slightly when encountered or by artificially adding the constraint $w_i > 1/n$ or $w_i < -1/n$ for some $1\leq i\leq n$. (if $F\neq\{\mathbf{0}\}$, then there must be a solution for at least one choice of i and one of the two constraints).

The feasible set F of the system (24) contains vectors that separate χ with *no error* (F contains good vectors with $\epsilon = 0$). Our objective is the more moderate goal of finding a good vector for inverse-polynomially large ϵ (see Definition 2.2). The feasible set F is contained in the set of good vectors. We use the following separation oracle.

Modified Separation Oracle. Let $\mathbf{v}$ be the center of the current Ellipsoid.

- Test if $\mathbf{v}$ is a good vector. Proposition 2.3 establishes that this can be performed with high confidence in polynomial time. If $\mathbf{v}$ is a good vector, stop the Ellipsoid algorithm and return $\mathbf{v}$. Note that $\mathbf{v}$ is not necessarily in F.
- Otherwise, $\mathbf{v}$ is bad. Apply Theorem 2.6 to find, with high confidence and in polynomial time, a witness $\mathbf{w}$ for the badness of $\mathbf{v}$. By definition, the witness is $\mathbf{w}\in H$ such that $\mathbf{v}\cdot\mathbf{w}<0$. This is a violated inequality of the system (24) and it allows to continue executing the Ellipsoid Algorithm.

The algorithm is applied in conjunction with the modified separation oracle.

The Ellipsoid algorithm initializes by bounding a portion or all of feasibility region by an Ellipsoid E_0. In our problem, the feasibility region is initially bounded by the unit ball B^n. Therefore, an initial Ellipsoid of volume $O((2e\pi/n)^{n/2}/\sqrt{\pi n})$ is used. (All we need is for the volume to be $\leq 2^{\text{poly}}$)

Correctness:. Either a good vector is found during the execution of the Ellipsoid algorithm, or else, it is carried to termination and finds a vector in F, which is a good vector for any ϵ. Each application of the separation oracle is correct with high confidence, therefore, if there are polynomially many iterations, all of them are correct with high confidence. (For correctness we can tolerate even e^{poly} many iterations).

Polynomial time:. To establish a polynomial running time it suffices to show that a good vector is found within polynomially many iterations.

In each iteration, the Ellipsoid algorithm applies the separation oracle to the center $\mathbf{v}$ of the current Ellipsoid E_i. The violated constraint returned by the oracle is used to obtain a new Ellipsoid E_{i+1} that also encloses the feasibility region, but its volume $\mathrm{vol}(E_{i+1})$ is smaller by at least $\Omega(1/n)$ fraction than the volume $\mathrm{vol}(E_i)$, that is, $\frac{\mathrm{vol}(E_{i+1})}{\mathrm{vol}(E_i)}\leq 1-\Omega(1/n)$.

The Ellipsoid algorithm terminates in polynomially many iteration when the volume of the feasibility polyhedron is at least exponent of a negated polynomial in the volume of E_0. This is established in Proposition 2.4.

5. Discussion

We discuss some extensions of our results and issues for further work.

When the noise level Δ is not known. Standard techniques can be used [10] to adapt our algorithm. Alternatively, an unknown Δ can be incorporated directly in our algorithm. The basic idea is to select the "worst" slice and to choose as a witness a linear combination of the form $\mathbf{w} = \hat{\mathbf{m}}^{-}{}_{\sim} + \alpha\hat{\mathbf{m}}^{+}{}_{\sim}$, with α such that $\mathbf{w} \cdot \mathbf{v} = 0$. Intuitively, this pushes the second witness property as far as we can, and can only "help" the first property.

Scaling instead of slicing. An alternative algorithm to find a witness for a bad vector $\mathbf{v}$, is to scale the distribution with the transformation $T(\mathbf{x}) = \frac{\mathbf{x}}{||\mathbf{x}|||\mathbf{x}\cdot\mathbf{v}|}$. We apply outliers removal to $T(\tilde{\chi})$. Then we treat the whole distribution as we treated a slice before.

Practical polynomial-time algorithm?. An interesting issue is finding an algorithm that has both polynomial-time guarantee and is practical. The Ellipsoid algorithm in itself is not a practical algorithm for LP problems. One approach might be to try to adapt our ideas to work with an interior-point algorithm. Another heuristic approach is to combine the perceptron algorithm, which seems to work well on "real data" with some ideas from our "witness identification" subroutines.

Non-uniform noise function $\Delta(\mathbf{x})$. Questions with regard to the tractability of the problem for various *non-uniform* noise models remain unanswered. Versions of the problem where the noise is adversarial (for example, $\Delta = 1$ on some regions or points and $\Delta = 0$ or arbitrary elsewhere) are known to be hard. This is equivalent to the problem of finding a solution to a linear system of inequalities that minimizes the number of unsatisfied constraints. The latter is known to be NP-hard (and so is the respective version for the perceptron problem [2]). Our algorithm extend to noise models that are *variable but known*. If the function $\Delta(\mathbf{x})$ is provided, it can be incorporated into the algorithm. Variants of our algorithm where we handle unknown noise by taking a linear combination of $\hat{\mathbf{m}}^{+}_{\sim}$ and $\hat{\mathbf{m}}^{-}_{\sim}$ that nullifies the dot product with $\mathbf{v}$, still results in a witness for many noise distributions. One intriguing question that remains open is an algorithm that handles arbitrary unknown noise, but under the constraint that $\Delta(\mathbf{x}) \leq \Delta_{\max} < 1/2$.

References.

[1] S. Agmon. The relaxation method for linear inequalities. *Canadian J. of Mathematics*, 6:382–392, 1954.

[2] E. Amaldi. *From finding maximum feasible subsystems of linear systems to feedforward neural network design.* PhD thesis, Swiss Federal Institute of Technology at Lausanne (EPFL), 1994. (Ph.D. dissertation No. 1282. Department of Mathematics).

[3] J. A. Anderson and E. Rosenfeld, editors. *Neurocomputing: Foundations of research.* MIT Press, 1988.

[4] A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. In *Proc. 37th IEEE Annual Symposium on Foundations of Computer Science*, pages 330–338. IEEE, 1996.

[5] T. Bylander. Learning linear threshold functions in the presence of classification noise. In *Proc. 7th Annual ACM conference on Computational Learning Theory*, pages 340–347. ACM, 1994.

[6] H. Chernoff. A measure of the asymptotic efficiency for test of a hypothesis based on the sum of observations. *Annals of Math. Statistics*, 23:493–509, 1952.

[7] B. Grötschel, L. Lovasz, and Schrijver A. *Geometric algorithms and combinatorial optimization.* Springer-Verlag, New York, 1988.

[8] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation.* Addison-Wesley Publishing Co., 1991.

[9] M. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proc. 25th Annual ACM Symposium on Theory of Computing*, pages 392–401. ACM, 1993.

[10] M. Kearns and U. Vazirani. *An introduction to computational learning theory.* MIT Press, Cambridge, MA, 1994.

[11] L. G. Khachiyan. A polynomial algorithm in linear programming. *Doklady Akad. Nauk. USSR*, 244:1093–1096, 1979. translated as *Soviet Mathematics Doklady*, **20**, 191–194.

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

[13] M. L. Minsky and S. A. Papert. *Perceptrons.* MIT Press, 1969.

[14] R. L. Rivest. Learning decision lists. *Machine Learning*, 2:229–246, 1987.

[15] F. Rosenblatt. *Priniciples of Neurodynamics.* Spartan Books, 1962.

[16] A. Schrijver. *Theory of Linear and Integer Programming.* John Wiley & Sons, New York, 1987.