Proceedings of Machine Learning Research vol 125:1–11, 2020 33rd Annual Conference on Learning Theory

A Nearly Optimal Variant of the Perceptron Algorithm for the Uniform Distribution on the Unit Sphere

Marco Schmalhofer SCHMALHOFER@EM.UNI-FRANKFURT.DE
Institute of Computer Science, Goethe University Frankfurt

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We show a simple perceptron-like algorithm to learn origin-centered halfspaces in $\mathbb{R}^n$ with accuracy $1-\epsilon$ and confidence $1-\delta$ in time

$$\mathcal{O}\left(\frac{n^2}{\epsilon}\left(\log\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right)$$

using

$$\mathcal{O}\left(\frac{n}{\epsilon}\left(\log\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right)$$

labeled examples drawn uniformly from the unit n-sphere. This improves upon algorithms given in Baum (1990), Long (1994) and Servedio (1999). The time and sample complexity of our algorithm match the lower bounds given in Long (1995) up to logarithmic factors.

Keywords: Halfspace learning, perceptron, uniform distribution, n-sphere

1. Introduction

Learning halfspaces from labeled examples is one of the central challenges in machine learning. In Blumer et al. (1989) it is shown that n-dimensional halfspaces can be learned to accuracy $1-\epsilon$ with confidence $1-\delta$ in the classical PAC model, and hence for arbitrary distributions, using $\mathcal{O}\left((n/\epsilon)\log(1/\epsilon)+(1/\epsilon)\log(1/\delta)\right)$ examples. Therefore it suffices to find a halfspace consistent with the given examples, which can be accomplished in time polynomial in n, $1/\epsilon$ and $1/\delta$ (e.g. by linear programming). In Ehrenfeucht et al. (1989) a lower bound of $\Omega\left((n/\epsilon)+(1/\epsilon)\log(1/\delta)\right)$ on the number of examples is derived, which also holds if the examples are drawn uniformly from the unit sphere, see Long (1995). In this case the bound is even tight (Long (2003)). In Balcan and Long (2013) polynomial time algorithms are constructed which achieve that bound even for any log-concave distribution.

The classical perceptron algorithm by Rosenblatt (Rosenblatt (1958)) determines a consistent halfspace given sufficiently many correctly classified examples (see e.g. Novikoff (1962)). Furthermore, in Baum (1990) a variant of the perceptron algorithm was provided, which learns halfspaces in time $\tilde{\mathcal{O}}(n^2/\epsilon^3)$ using $\tilde{\mathcal{O}}(n/\epsilon^3)$ examples. This was improved by Servedio (1999). The algorithm proposed there achieves time complexity $\tilde{\mathcal{O}}(n^2/\epsilon^2)$ with a sample size of $\tilde{\mathcal{O}}(n/\epsilon^2)$. The perceptron algorithm was also shown to be able to solve linear programs in polynomial time, see Dunagan and Vempala (2004). Table 1 summarizes related work considering halfspace learning. In the right column uniform means uniform on the unit sphere.

In this paper we show that the classical perceptron algorithm can be supplemented with an adaptive learning rate such that it (ϵ, δ)-learns halfspaces in time $\mathcal{O}\left(\frac{n^2}{\epsilon}\left(\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$ using $\mathcal{O}\left(\frac{n}{\epsilon}\left(\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$ examples drawn uniformly from the unit sphere. The extremely simple algorithm has nice properties, its error is monotonically decreasing, its hypothesis always has norm one even without rebalancing and it is conservative, i.e. it updates its hypotheses only for counterexamples.

Table 1: Related work on (ϵ, δ)-learning of halfspaces.

article	sample complexity	time complexity	distribution
Blumer et al. (1989)	$\mathcal{O}\left(\frac{1}{\epsilon}\left(n\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$	poly	arbitrary
Ehrenfeucht et al. (1989)	$\Omega\left(\frac{1}{\epsilon}\left(n + \log\frac{1}{\delta}\right)\right)$	–	arbitrary
Haussler et al. (1994)	$\mathcal{O}\left(\frac{n}{\epsilon}\log\frac{1}{\delta}\right)$	poly	arbitrary
Baum (1990)	$\tilde{\mathcal{O}}\left(n/\epsilon^3\right)$	$\tilde{\mathcal{O}}\left(n^2/\epsilon^3\right)$	uniform
Long (1994) (for $\delta = \epsilon$)	$\tilde{\mathcal{O}}(n/\epsilon)$	$\tilde{\mathcal{O}}\left(n^2/\epsilon + n^{3.38}\right)$	uniform
Long (1995)	$\Omega\left(\frac{1}{\epsilon}\left(n + \log\frac{1}{\delta}\right)\right)$	–	uniform
Servedio (1999)	$\tilde{\mathcal{O}}\left(n/\epsilon^2\right)$	$\tilde{\mathcal{O}}\left(n^2/\epsilon^2\right)$	uniform
Long (2003)	$\mathcal{O}\left(\frac{1}{\epsilon}\left(n + \log\frac{1}{\delta}\right)\right)$	–	uniform
Balcan and Long (2013)	$\mathcal{O}\left(\frac{1}{\epsilon}\left(n + \log\frac{1}{\delta}\right)\right)$	poly	log-concave
our paper	$\mathcal{O}\left(\frac{n}{\epsilon}\left(\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$	$\mathcal{O}\left(\frac{n^2}{\epsilon}\left(\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$	uniform

We present basic definitions and notations in Section 2 and motivate the algorithm in Section 3, where first properties and experimental results are also presented. Finally, in Section 4 we derive Theorem 2 as our main result. Its proof crucially depends on Lemma 6, which provides the conditional expectation $\mathbb{E}\left[dd^* \mid \beta\right]$, where d and d^* are the distances of a randomly drawn counterexample to the current hyperplane and the target hyperplane, respectively, assuming β is the angle between them. In Section 5 we summarize our results and suggest some open problems.

2. Preliminaries

We study the classical problem of learning *homogeneous halfspaces*

$$f_{\boldsymbol{w}} : \mathbb{R}^n \to \{-1, 1\}, \; f_{\boldsymbol{w}}(\boldsymbol{x}) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$$

represented by a *weight vector* $\boldsymbol{w} \in \mathbb{R}^n$. We denote the unknown target halfspace by f^* and its normalized weight vector by $\boldsymbol{w}^*$. The learner is given *labeled examples* of the form

$$\left(\boldsymbol{x}, f^*(\boldsymbol{x})\right) \in \mathbb{R}^n \times \{-1, 1\},$$

where each example $\boldsymbol{x}$ is drawn independently according to the *uniform distribution* on the *unit n-sphere*

$$S^{n-1} := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = 1\}.$$

An example $\boldsymbol{x}$ is called *positive* if $f^*(\boldsymbol{x}) = 1$ and *negative* otherwise. After receiving examples the learner outputs a *hypothesis* $\boldsymbol{w} \in \mathbb{R}^n$. The error of $\boldsymbol{w}$ is measured by the probability of misclassifying a randomly drawn example, i.e.

$$\operatorname{err}(\boldsymbol{w}) := \mathbb{P}\left[f^*(\boldsymbol{x}) \neq f_{\boldsymbol{w}}(\boldsymbol{x})\right],$$

where $\boldsymbol{x}$ is drawn uniformly from S^{n-1}. By rotational symmetry it is easy to see that the error of a hypothesis $\boldsymbol{w}$ is determined by the angle $\sphericalangle(\boldsymbol{w}^*, \boldsymbol{w})$ between $\boldsymbol{w}^*$ and $\boldsymbol{w}$, i.e.

$$\operatorname{err}(\boldsymbol{w}) = \frac{\sphericalangle(\boldsymbol{w}^*, \boldsymbol{w})}{\pi} = \frac{1}{\pi} \arccos \frac{\langle \boldsymbol{w}^*, \boldsymbol{w} \rangle}{\|\boldsymbol{w}\|}. \tag{1}$$

3. Adaptive Perceptron – The Algorithm

In this section we present our algorithm, which is in fact the classical perceptron learning rule supplemented with a variable learning rate $\eta > 0$. First, let us consider the single perceptron update

$$\boldsymbol{w}' = \boldsymbol{w} + \eta\, b \boldsymbol{x} \tag{2}$$

of a hypothesis $\boldsymbol{w} \neq \boldsymbol{0}$ through a counterexample $(\boldsymbol{x}, b) \in S^{n-1} \times \{-1, 1\}$. Then according to Equation (1) the error of the updated hypothesis $\boldsymbol{w}'$ is

$$\operatorname{err}(\boldsymbol{w}') = \frac{1}{\pi} \arccos \frac{\langle \boldsymbol{w}^*, \boldsymbol{w} + \eta b \boldsymbol{x} \rangle}{\|\boldsymbol{w} + \eta b \boldsymbol{x}\|}. \tag{3}$$

Now we minimize $\operatorname{err}(\boldsymbol{w}')$ as a function of η. Since $\arccos$ is monotonically decreasing we find a global minimum of $\operatorname{err}(\boldsymbol{w}')$ by maximizing its argument $g(\eta) := \langle \boldsymbol{w}^*, \boldsymbol{w} + \eta b \boldsymbol{x} \rangle \,/\, \|\boldsymbol{w} + \eta b \boldsymbol{x}\|$. We determine a zero of g' by applying the quotient rule and forgetting its denominator:

$$\begin{aligned}
& & 0 &\stackrel{!}{=} \frac{\mathrm{d}}{\mathrm{d}\eta} \left(\frac{\langle \boldsymbol{w}^*, \boldsymbol{w} + \eta b \boldsymbol{x} \rangle}{\|\boldsymbol{w} + \eta b \boldsymbol{x}\|} \right) \\
&\Leftrightarrow & 0 &= \frac{\mathrm{d} \langle \boldsymbol{w}^*, \boldsymbol{w} + \eta b \boldsymbol{x} \rangle}{\mathrm{d}\eta} \|\boldsymbol{w} + \eta b \boldsymbol{x}\| - \frac{\mathrm{d} \|\boldsymbol{w} + \eta b \boldsymbol{x}\|}{\mathrm{d}\eta} \langle \boldsymbol{w}^*, \boldsymbol{w} + \eta b \boldsymbol{x} \rangle \\
&\Leftrightarrow & 0 &= b \langle \boldsymbol{w}^*, \boldsymbol{x} \rangle \|\boldsymbol{w} + \eta b \boldsymbol{x}\| - \frac{b \langle \boldsymbol{w}, \boldsymbol{x} \rangle + \eta}{\|\boldsymbol{w} + \eta b \boldsymbol{x}\|} \langle \boldsymbol{w}^*, \boldsymbol{w} + \eta b \boldsymbol{x} \rangle
\end{aligned} \tag{4}$$

By setting $d^* := b \langle \boldsymbol{w}^*, \boldsymbol{x} \rangle$ and $d := -b \langle \boldsymbol{w}, \boldsymbol{x} \rangle \,/\, \|\boldsymbol{w}\|$ for the distances from $\boldsymbol{x}$ to the target hyperplane and the actual hyperplane we obtain from Equation (4)

$$\begin{aligned}
& & \left(\|\boldsymbol{w}\|^2 - \eta \|\boldsymbol{w}\| d \right) d^* &= \left(\eta - \|\boldsymbol{w}\| d \right) \langle \boldsymbol{w}^*, \boldsymbol{w} \rangle \\
&\Leftrightarrow & \eta &= \|\boldsymbol{w}\| \frac{d^* + d \left\langle \boldsymbol{w}^*, \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right\rangle}{\left\langle \boldsymbol{w}^*, \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right\rangle + d d^*}.
\end{aligned}$$

Now with $\langle \boldsymbol{w}^*, \boldsymbol{w} / \|\boldsymbol{w}\| \rangle = \cos\beta$, where β is the angle between $\boldsymbol{w}$ and $\boldsymbol{w}^*$, we get the locally optimal learning rate as a function of $\|\boldsymbol{w}\|$, d^*, d and β, namely

$$\eta_{opt} = \|\boldsymbol{w}\| \frac{d^* + d \cos\beta}{\cos\beta + d d^*}. \tag{5}$$

Of course this learning rate is useless in practice, since it depends on d^* and β, which are unknown to the algorithm. Nevertheless it motivates a useful choice of η: Assume $\boldsymbol{w}$ has a small error. Then

β is small and thus $\cos\beta$ close to one. The expected distances $\mathbb{E}\big[d^*\big]$ and $\mathbb{E}\big[d\big]$ should therefore be also small. Since by symmetry these expected distances are equal, so

$$\eta := \|\boldsymbol{w}\|\, 2d \tag{6}$$

is hopefully a good and certainly an easily computable choice. This directly provides our algorithm:

Algorithm 1: ADAPTIVEPERCEPTRON

Input: Number of examples s to be drawn uniformly from S^{n-1}
Output: Hypothesis $\boldsymbol{w}$
Get first labeled example $(\boldsymbol{x}, b)$ where $b = f^*(\boldsymbol{x})$;
$\boldsymbol{w} \leftarrow b\boldsymbol{x}$;
for $i = 2 \ldots s$ **do**
  Get labeled example $(\boldsymbol{x}, b)$;
  if $f_{\boldsymbol{w}}(\boldsymbol{x}) \neq b$ **then**
    $\boldsymbol{w}' \leftarrow \boldsymbol{w} \; - \; 2\langle \boldsymbol{w}, \boldsymbol{x}\rangle\, \boldsymbol{x}$;
    $\boldsymbol{w} \leftarrow \boldsymbol{w}'$;
  end
end
return $\boldsymbol{w}$;

Suprisingly, the error of ADAPTIVEPERCEPTRON is monotonically decreasing. Moreover, all hypotheses are unit vectors. These two properties turn out to be crucial for the analysis.

Proposition 1 (first properties) *For any hypothesis $\boldsymbol{w}$ determined by* ADAPTIVEPERCEPTRON*,*

(a) $\|\boldsymbol{w}\| = 1$*, as well as*

(b) $err(\boldsymbol{w}') \leq err(\boldsymbol{w}) \leq 1/2$.

Proof After the first example $\boldsymbol{x}$ we have $\|\boldsymbol{w}\| = \|b\boldsymbol{x}\| = 1$. Also $\mathrm{err}(\boldsymbol{w}) = \frac{1}{\pi}\arccos\langle \boldsymbol{w}^*, \boldsymbol{w}\rangle \leq 1/2$, since $\langle \boldsymbol{w}^*, b\boldsymbol{x}\rangle \geq 0$. Moreover, after updating $\boldsymbol{w}$ with a counterexample $\boldsymbol{x}$ we have

$$\left\|\boldsymbol{w}'\right\|^2 = \|\boldsymbol{w} - 2\langle \boldsymbol{w}, \boldsymbol{x}\rangle \boldsymbol{x}\|^2 = \|\boldsymbol{w}\|^2 - 4\langle \boldsymbol{w}, \boldsymbol{x}\rangle \langle \boldsymbol{w}, \boldsymbol{x}\rangle + 4\langle \boldsymbol{w}, \boldsymbol{x}\rangle^2 \|\boldsymbol{x}\|^2 = \|\boldsymbol{w}\|^2 = 1.$$

For the error of $\boldsymbol{w}'$ we have

$$\begin{aligned} \mathrm{err}(\boldsymbol{w}') &= \frac{1}{\pi}\arccos\,\langle \boldsymbol{w}^*, \boldsymbol{w}'\rangle \\ &= \frac{1}{\pi}\arccos(\langle \boldsymbol{w}^*, \boldsymbol{w}\rangle - 2\langle \boldsymbol{w}, \boldsymbol{x}\rangle \langle \boldsymbol{w}^*, \boldsymbol{x}\rangle) \\ &\leq \frac{1}{\pi}\arccos\,\langle \boldsymbol{w}^*, \boldsymbol{w}\rangle = \mathrm{err}(\boldsymbol{w}), \end{aligned}$$

since $\arccos$ is monotonically decreasing and $\langle \boldsymbol{w}, \boldsymbol{x}\rangle \langle \boldsymbol{w}^*, \boldsymbol{x}\rangle \leq 0$ for the counterexample $\boldsymbol{x}$. ■

We conducted experiments for dimension $n = 2^{10}$ and up to $s = 10^{10}$ examples, comparing the learning curves of the perceptron algorithm of Baum (1990), the average algorithm of Servedio (1999) and ADAPTIVEPERCEPTRON. It turns out that for up to $s \approx n$ examples the three algorithms do not differ significantly with ADAPTIVEPERCEPTRON even lagging behind. However after 10^5 examples asymptotics seems to take over, ADAPTIVEPERCEPTRON clearly pulls ahead and continues to stay in front from there on (see Figure 1).

We also investigated the "hypothetical" OPTADAPTIVEPERCEPTRON, which in each step uses the locally optimal learning rate η_{opt} (see Equation (5)). Observe that ADAPTIVEPERCEPTRON and OPTADAPTIVEPERCEPTRON are almost indistinguishable.

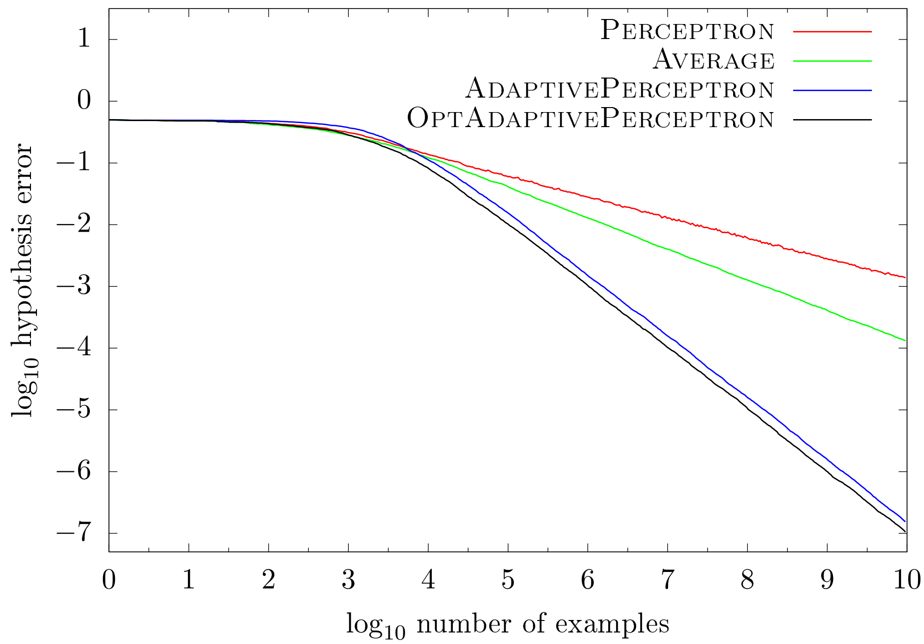

Figure 1: log-log plot of the learning curve for different algorithms at $n = 1024$ dimensions.

4. Adaptive Perceptron – Analysis

In this section we prove our main result:

Theorem 2 *After $s = \Theta\big(\frac{n}{\epsilon}(\log\frac{1}{\epsilon} + \log\frac{1}{\delta})\big)$ examples* ADAPTIVEPERCEPTRON *outputs a hypothesis which with probability at least $1-\delta$ has error at most ϵ for each $\epsilon, \delta \in (0, 1]$, $n \geq 2$.*

Note that our simulation is consistent with the results of Theorem 2, i.e. the error of ADAPTIVEPERCEPTRON behaves asymptotically as s^{-1}, whereas the error of PERCEPTRON and AVERAGE is roughly $s^{-1/3}$ and $s^{-1/2}$, respectively (see Figure 1 again).

Proof [Theorem 2] Let $\boldsymbol{w}_k$ be the k-th hypothesis determined by ADAPTIVEPERCEPTRON for $k \geq 1$. We write $\beta_k = \sphericalangle(\boldsymbol{w}^*, \boldsymbol{w}_k)$ for the angle between $\boldsymbol{w}^*$ and $\boldsymbol{w}_k$. The theorem follows if the expected cosine of β_k is "exponentially close" to one, i.e. we later show the following lemma.

Lemma 3 (expected cosine of hypothesis angle) *For each $k \geq 1$ we have*

$$\mathbb{E}\big[\cos\beta_k\big] \geq 1 - e^{-\frac{2}{3n}(k-1)},$$

where the expectation is taken over the random sequence of examples.

4.1. Proof of Theorem 2 with Lemma 3

We show the theorem with the help of two inequalities.

Lemma 4 *For each $0 < z \leq \frac{\pi}{2}$ we have*

$$\frac{\sin z}{z} - \cos z \geq \frac{1}{3}(1 - \cos z), \tag{7}$$

and for each $0 \leq y \leq 1$ we have

$$\arccos y \leq 2\sqrt{1-y}. \tag{8}$$

Proof Let $0 < z \leq \pi/2$. We show $\frac{\sin z}{z} - \frac{2}{3}\cos z - \frac{1}{3} \geq 0$ using Taylor approximations of sine and cosine. With Lagrange remainder we have

$$\sin z = z - \frac{1}{6}z^3 + \frac{\cos \xi_1}{120}z^5 \quad \text{and} \quad \cos z = 1 - \frac{1}{2}z^2 + \frac{\cos \xi_2}{24}z^4$$

with constants $0 \leq \xi_1, \xi_2 \leq z$. Since $\cos \xi_1 \geq 0$ and $\cos \xi_2 \leq 1$, we may conclude

$$\sin z \geq z - \frac{1}{6}z^3 \quad \text{and} \quad \cos z \leq 1 - \left(\frac{1}{2} - \frac{1}{24}z^2\right) z^2 \leq 1 - \frac{1}{3}z^2, \tag{$*$}$$

where the last inequality follows from the fact that $0 < z < 2$. Hence we have

$$\frac{\sin z}{z} - \frac{2}{3}\cos z - \frac{1}{3} \geq 1 - \frac{1}{6}z^2 - \frac{2}{3} + \frac{2}{9}z^2 - \frac{1}{3} = \frac{1}{18}z^2 \geq 0$$

and (7) is proven. To prove (8) one can solve $(*)$ for z. This yields

$$z \leq \sqrt{3 - 3\cos z} \leq 2\sqrt{1 - \cos z}.$$

Substituting $z := \arccos y$ shows (8). ■

Since $\arccos$ is concave in $[0, 1]$, we can apply Jensen's inequality in addition to applying Lemma 3 and thus bound the expected error of the k-th hypothesis by

$$\begin{aligned}
\mathbb{E}\big[\text{err}(\boldsymbol{w}_k)\big] &= \frac{1}{\pi}\mathbb{E}\big[\arccos(\cos \beta_k)\big] \\
&\leq \frac{1}{\pi}\arccos\Big(\mathbb{E}\big[\cos \beta_k\big]\Big) && \text{(Jensen's inequality)} \\
&\leq \frac{2}{\pi}\sqrt{1 - \mathbb{E}\big[\cos \beta_k\big]} && \text{(Inequality (8))} \\
&\leq \frac{2}{\pi}e^{-\frac{1}{3n}(k-1)} && \text{(Lemma 3)} \\
&\leq e^{-\frac{k}{3n}}. && (n \geq 2)
\end{aligned}$$

Hence after

$$k \geq k_0 := 3n \ln \frac{2}{\epsilon\delta} \tag{9}$$

counterexamples have occured, we have $\mathbb{E}\big[\mathrm{err}(\boldsymbol{w}_k)\big] \leq \frac{\epsilon\delta}{2}$. Now suppose the output hypothesis $\boldsymbol{w}_k$ has error greater than ϵ. Then due to monotonicity (Proposition 1, (b)) the error has always been greater than ϵ. So if we draw

$$s := \frac{2k_0}{\epsilon} = \Theta\left(\frac{n}{\epsilon}\left(\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right) \tag{10}$$

examples, we expect to have at least $2k_0$ counterexamples in this case. We apply the Chernoff bound in the following version to bound the probability of getting less than k_0 counterexamples in this case.

Lemma 5 (Lower tail Chernoff bound) *If $Y_1, \ldots, Y_s$ are $\{0,1\}$-valued random variables with $\mathbb{P}\big[Y_i = 1 \mid Y_1, \ldots, Y_{i-1}\big] \geq p$, then for all $0 < c < 1$,*

$$\mathbb{P}\big[\sum_i Y_i < (1-c)sp\big] \leq e^{-c^2 sp/2}.$$

Proof The lemma follows from the standard Chernoff bound. ■

To bound the probability of getting less than k_0 counterexamples in s trials let the 0-1 variable Y_i indicate whether the i-th example is a counterexample. Set $p := \epsilon$, $c := 1/2$ and we obtain

$$\mathbb{P}\big[\sum_i Y_i < k_0\big] \leq (\epsilon\delta/2)^{3n/4} \leq \delta/2.$$

If the error of the output hypothesis $\boldsymbol{w}_k$ is greater than ϵ, less than k_0 counterexamples were encountered or the random variable $\mathrm{err}(\boldsymbol{w}_k)$ exceeds its expected value by at least a factor of $2/\delta$. Hence, by the union bound and Markov's inequality, the probability that the output hypothesis $\boldsymbol{w}_k$ has error greater than ϵ is at most $\delta/2 + \delta/2 = \delta$, which proves Theorem 2. ■

4.2. Proof of Lemma 3 with Lemma 6

Proof [Lemma 3] Since $\boldsymbol{w}^*$ and $\boldsymbol{w}_k$ have norm one (Proposition 1, (a)), the cosine of β_k is given as

$$\cos\beta_k = \langle \boldsymbol{w}^*, \boldsymbol{w}_k \rangle\,. \tag{11}$$

Recalling the update rule of ADAPTIVEPERCEPTRON for $k \geq 2$, we have

$$\begin{aligned} \langle \boldsymbol{w}^*, \boldsymbol{w}_k \rangle &= \langle \boldsymbol{w}^*, \boldsymbol{w}_{k-1} - 2\,\langle \boldsymbol{w}_{k-1}, \boldsymbol{x}_k \rangle\, \boldsymbol{x}_k \rangle \\ &= \langle \boldsymbol{w}^*, \boldsymbol{w}_{k-1} \rangle - 2\,\langle \boldsymbol{w}_{k-1}, \boldsymbol{x}_k \rangle\, \langle \boldsymbol{w}^*, \boldsymbol{x}_k \rangle\,, \end{aligned} \tag{12}$$

where $\boldsymbol{x}_k$ is the k-th counterexample. We combine Equations (11) and (12) and set $d_k^* := \pm\,\langle \boldsymbol{w}^*, \boldsymbol{x}_k \rangle$, $d_k := \mp\,\langle \boldsymbol{w}_{k-1}, \boldsymbol{x}_k \rangle$ for the distances from $\boldsymbol{x}_k$ to the target hyperplane and the current hyperplane of the algorithm. This yields

$$\cos\beta_k = \cos\beta_{k-1} + 2d_k d_k^*. \tag{13}$$

By symmetry the probability distribution of $d_k d_k^*$ only depends on the hypothesis angle β_{k-1}. Thus we can form total expectation to obtain

$$\mathbb{E}\big[\cos\beta_k\big] = \mathbb{E}\big[\cos\beta_{k-1}\big] + 2\,\mathbb{E}\Big[\mathbb{E}\big[d_k d_k^* \mid \beta_{k-1}\big]\Big]. \tag{14}$$

The following key lemma provides the conditional expectation.

Lemma 6 (expected product of distances) *Let $\boldsymbol{w}$ be a hypothesis with $\sphericalangle(\boldsymbol{w}^*, \boldsymbol{w}) = \beta > 0$. Assume $\boldsymbol{x}$ is a randomly drawn counterexample. Let d^* and d be the distances from $\boldsymbol{x}$ to the target hyperplane and the hyperplane represented by $\boldsymbol{w}$. Then we have*

$$\mathbb{E}\big[dd^* \mid \beta\big] = \frac{1}{n}\left(\frac{\sin\beta}{\beta} - \cos\beta\right).$$

We show Lemma 6 later. In combination with Lemma 4 we see

$$\mathbb{E}\big[d_k d_k^* \mid \beta_{k-1}\big] \overset{\text{Lemma 6}}{=} \frac{1}{n}\left(\frac{\sin\beta_{k-1}}{\beta_{k-1}} - \cos\beta_{k-1}\right) \overset{\text{Inequality (7)}}{\geq} \frac{1}{3n}(1-\cos\beta_{k-1}).$$

Thus, we may bound $\mathbb{E}\big[\cos\beta_k\big]$ in Equation (14) as follows

$$\begin{aligned}
\mathbb{E}\big[\cos\beta_k\big] &\geq \mathbb{E}\big[\cos\beta_{k-1}\big] + 2\,\mathbb{E}\left[\frac{1}{3n}(1-\cos\beta_{k-1})\right] \\
&\geq \frac{2}{3n} + \left(1 - \frac{2}{3n}\right)\mathbb{E}\big[\cos\beta_{k-1}\big].
\end{aligned}$$

Now expand this inequality recursively and notice that $\cos\beta_1 = \langle \boldsymbol{w}^*, \boldsymbol{w}_1\rangle = \langle \boldsymbol{w}^*, b_1\boldsymbol{x}_1\rangle \geq 0$ for the first example $(\boldsymbol{x}_1, b_1)$. Hence we have for all $k \geq 1$,

$$\mathbb{E}\big[\cos\beta_k\big] \geq \frac{2}{3n}\sum_{i=0}^{k-2}\left(1-\frac{2}{3n}\right)^i = 1 - (1-2/3n)^{k-1} \geq 1 - e^{-\frac{2}{3n}(k-1)}.$$

■

4.3. Proof of Lemma 6

Proof [Lemma 6] Let $\boldsymbol{w}$, β, $\boldsymbol{x}$ and d, d^* be given as stated in the lemma. Without loss of generality (rotational symmetry) let $\boldsymbol{w}^* = (1, 0, \ldots, 0)$ and $\boldsymbol{w} = (\cos\beta, -\sin\beta, 0, \ldots, 0)$. Note that $\sphericalangle(\boldsymbol{w}^*, \boldsymbol{w}) = \beta$, $\|\boldsymbol{w}^*\| = \|\boldsymbol{w}\| = 1$ and $dd^* = -\langle \boldsymbol{w}, \boldsymbol{x}\rangle\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle = x_1(x_2\sin\beta - x_1\cos\beta)$. Also by symmetry we may assume that $\boldsymbol{x}$ is a positive counterexample. Now consider $\boldsymbol{x}$ to be the angular part of a standard normal vector $\boldsymbol{u} = r\boldsymbol{x}$, where r is its length. Note that r and $\boldsymbol{x}$ are independent and thus it holds

$$\mathbb{E}_{\boldsymbol{x}}\big[-\langle \boldsymbol{w}, \boldsymbol{x}\rangle\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle\big]\,\mathbb{E}_r\big[r^2\big] = \mathbb{E}_{\boldsymbol{u}}\big[-\langle \boldsymbol{w}, \boldsymbol{u}\rangle\langle \boldsymbol{w}^*, \boldsymbol{u}\rangle\big].$$

Since r^2 has a chi-squared distribution with n degrees of freedom, its expected value is $\mathbb{E}_r\big[r^2\big] = n$. Hence it remains to show that $\mathbb{E}_{\boldsymbol{u}}\big[-\langle \boldsymbol{w}, \boldsymbol{u}\rangle\langle \boldsymbol{w}^*, \boldsymbol{u}\rangle\big] = \sin(\beta)/\beta - \cos\beta$. This can be done by calculating a simple Gaussian integral:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{u}}\big[-\langle \boldsymbol{w}, \boldsymbol{u}\rangle \langle \boldsymbol{w}^*, \boldsymbol{u}\rangle \big] \\
&\quad = \frac{2\pi}{\beta} \int_{\substack{\boldsymbol{u}\in\mathbb{R}^n, u_1\geq 0,\\ u_1\cos\beta - u_2\sin\beta<0}} \frac{1}{\sqrt{(2\pi)^n}}\, e^{-\frac{1}{2}\left(u_1^2+\cdots+u_n^2\right)}\, u_1(u_2\sin\beta - u_1\cos\beta)\,\mathrm{d}(u_1\ldots u_n) \\
&\quad = \frac{1}{\beta} \int_{\substack{u_1\geq 0,\\ u_1\cos\beta - u_2\sin\beta<0}} e^{-\frac{1}{2}\left(u_1^2+u_2^2\right)}\, u_1(u_2\sin\beta - u_1\cos\beta)\,\mathrm{d}(u_1,u_2) \\
&\quad = \frac{1}{\beta} \int_{r=0}^{\infty}\int_{\varphi=0}^{\beta} e^{-r^2/2}\, r\sin\varphi(r\cos\varphi\sin\beta - r\sin\varphi\cos\beta)\; r\,\mathrm{d}\varphi\,\mathrm{d}r \\
&\quad = \frac{1}{\beta} \int_{r=0}^{\infty} r^3 e^{-r^2/2}\,\mathrm{d}r \int_{\varphi=0}^{\beta} \sin\varphi\cos\varphi\sin\beta - \sin^2\varphi\cos\beta\;\mathrm{d}\varphi.
\end{aligned}$$

Now substituting the two integrals

$$\int_{r=0}^{\infty} r^3 e^{-r^2/2}\,\mathrm{d}r = \left[-r^2 e^{-r^2/2}\right]_0^{\infty} + \int_0^{\infty} 2re^{-r^2/2}\,\mathrm{d}r = \left[-2e^{-r^2/2}\right]_0^{\infty} = 2$$

and

$$\begin{aligned}
&\int_{\varphi=0}^{\beta} \sin\varphi\cos\varphi\sin\beta - \sin^2\varphi\cos\beta\;\mathrm{d}\varphi \\
&\qquad\qquad = \left[-\frac{1}{2}\cos^2\varphi\right]_0^{\beta}\sin\beta - \left[\frac{1}{2}(\varphi - \sin\varphi\cos\varphi)\right]_0^{\beta}\cos\beta = \frac{1}{2}(\sin\beta - \beta\cos\beta)
\end{aligned}$$

shows the claim. ∎

5. Conclusions and Open Problems

The classical perceptron algorithm – with adaptive learning rate – turns out to be nearly optimal for learning homogeneous halfspaces against the uniform distribution on the unit sphere. The algorithm is fast, extremely simple, strictly error-decreasing and even conservative, i.e. it performs updates only on counterexamples.

Experiments suggest that OPTADAPTIVEPERCEPTRON performs only slightly better than ADAPTIVEPERCEPTRON. It would be interesting to investigate if there exist (conservative) learning algorithms which perform better than ADAPTIVEPERCEPTRON.

Moreover, it would be interesting to search for possible generalizations. For which distributions does Theorem 2 still hold? Is it possible to find a version of the algorithm which fits a given class of distributions?

Acknowledgments

I would like to thank Georg Schnitger and Martin Hoefer for a number of helpful discussions. Also thanks to all anonymous referees for their useful comments, especially for observing a simplification of the proof of Lemma 6 and for bringing Dunagan and Vempala (2004) to our attention.

References

Maria-Florina Balcan and Philip M. Long. Active and passive learning of linear separators under log-concave distributions. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 288–316. JMLR.org, 2013. URL http://proceedings.mlr.press/v30/Balcan13.html.

Eric B. Baum. The perceptron algorithm is fast for nonmalicious distributions. *Neural Computation*, 2(2):248–260, 1990. doi: 10.1162/neco.1990.2.2.248. URL https://doi.org/10.1162/neco.1990.2.2.248.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989. doi: 10.1145/76359.76371. URL https://doi.org/10.1145/76359.76371.

John Dunagan and Santosh S. Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. In László Babai, editor, *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pages 315–320. ACM, 2004. ISBN 1-58113-852-0. doi: 10.1145/1007352.1007404. URL https://doi.org/10.1145/1007352.1007404.

Andrzej Ehrenfeucht, David Haussler, Michael J. Kearns, and Leslie G. Valiant. A general lower bound on the number of examples needed for learning. *Inf. Comput.*, 82(3):247–261, 1989. doi: 10.1016/0890-5401(89)90002-3. URL https://doi.org/10.1016/0890-5401(89)90002-3.

David Haussler, Nick Littlestone, and Manfred K. Warmuth. Predicting \0,1\-functions on randomly drawn points. *Inf. Comput.*, 115(2):248–292, 1994. doi: 10.1006/inco.1994.1097. URL https://doi.org/10.1006/inco.1994.1097.

Philip M. Long. Halfspace learning, linear programming, and nonmalicious distributions. *Inf. Process. Lett.*, 51(5):245–250, 1994. doi: 10.1016/0020-0190(94)90003-5. URL https://doi.org/10.1016/0020-0190(94)90003-5.

Philip M. Long. On the sample complexity of PAC learning half-spaces against the uniform distribution. *IEEE Trans. Neural Networks*, 6(6):1556–1559, 1995. doi: 10.1109/72.471352. URL https://doi.org/10.1109/72.471352.

Philip M. Long. An upper bound on the sample complexity of pac-learning halfspaces with respect to the uniform distribution. *Inf. Process. Lett.*, 87(5):229–234, 2003. doi: 10.1016/S0020-0190(03)00311-9. URL https://doi.org/10.1016/S0020-0190(03)00311-9.

A. B. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622, New York, NY, USA, 1962. Polytechnic Institute of Brooklyn.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, pages 65–386, 1958.

Rocco A. Servedio. On PAC learning using winnow, perceptron, and a perceptron-like algorithm. In Shai Ben-David and Philip M. Long, editors, *Proceedings of the Twelfth Annual Conference on Computational Learning Theory, COLT 1999, Santa Cruz, CA, USA, July 7-9, 1999*, pages 296–307. ACM, 1999. ISBN 1-58113-167-4. doi: 10.1145/307400.307474. URL https://doi.org/10.1145/307400.307474.