Generalization Guarantees of Self-Training of Halfspaces under Label Noise Corruption

Lies Hadjadj¹, Massih-Reza Amini¹, Sana Louhichi²
¹Computer Science Laboratory (LIG), Université Grenoble Alpes, Grenoble, France
²Department of Statistics (LJK), Université Grenoble Alpes, Grenoble, France

Abstract
We investigate the generalization properties of a self-training algorithm with halfspaces. The approach learns a list of halfspaces iteratively from labeled and unlabeled training data, in which each iteration consists of two steps: exploration and pruning. In the exploration phase, the halfspace is found sequentially by maximizing the unsigned-margin among unlabeled examples and then assigning pseudo-labels to those that have a distance higher than the current threshold. These pseudo-labels are allegedly corrupted by noise. The training set is then augmented with noisy pseudo-labeled examples, and a new classifier is trained. This process is repeated until no more unlabeled examples remain for pseudo-labeling. In the pruning phase, pseudo-labeled samples that have a distance to the last halfspace greater than the associated unsigned-margin are then discarded. We prove that the misclassification error of the resulting sequence of classifiers is bounded and show that the resulting semi-supervised approach never degrades performance compared to the classifier learned using only the initial labeled training set. Experiments carried out on a variety of benchmarks demonstrate the efficiency of the proposed approach compared to state-of-the-art methods.

1 Introduction
In recent years, several attempts have been made to establish a theoretical foundation for semi-supervised learning. These studies are mainly interested in the generalization ability of semi-supervised learning techniques [Rigollet, 2007; Maximov et al., 2018] and the utility of unlabeled data in the training process [Castelli and Cover, 1995; Singh et al., 2009; Li and Zhou, 2011; Wei et al., 2021]. The majority of these works are based on the concept called compatibility in [Balcancan and Blum, 2006], and try to exploit the connection between the marginal data distribution and the target function to be learned. The common conclusion of these studies is that unlabeled data will only be useful for training if such a relationship exists.

The three key types of relations considered in the literature are cluster assumption, manifold assumption, and low-density separation [Zhu, 2005; Chapelle et al., 2006]. The cluster assumption states that data contains homogeneous labeled clusters, and unlabeled training examples allow to recognize these clusters. In this case, the marginal distribution is viewed as a mixture of class conditional distributions, and semi-supervised learning has been shown to be superior to supervised learning in terms of achieving smaller finite-sample error bounds in some general cases, and in some others, it provides a faster rate of error convergence [Castelli and Cover, 1995; Rigollet, 2007; Maximov et al., 2018; Singh et al., 2009].

In this line, [Ben-David et al., 2008] showed that the access to the marginal distribution over unlabeled training data would not provide sample size guarantees better than those obtained by supervised learning unless one assumes very strong assumptions about the conditional distribution over the class labels. Manifold assumption stipulates that the target function is in a low-dimensional manifold. [Niyogi, 2013] establishes a context through which such algorithms can be analyzed and potentially justified; the main result of this study is that unlabeled data may help the learning task in certain cases by defining the manifold. Finally, low-density separation states that the decision boundary lies in low-density regions. A principal way, in this case, is to employ a margin maximization strategy which results in pushing away the decision boundary from the unlabeled data [Chapelle et al., 2006]. Semi-supervised approaches based on this paradigm mainly assign pseudo-labels to high-confident unlabeled training examples with respect to the predictions and include these pseudo-labeled samples in the learning process. [Wei et al., 2020] demonstrated that, under the expansion assumption stipulating that a low-probability subset of data must grow to a neighborhood mostly surely regarding the subset, self-training will reach high accuracy with regard to ground-truth labels.

In this line, [Frei et al., 2021] showed that a strong classifier may be learned from a weaker one in the context of general mixture models with benign concentration and anti-concentration properties. However, [Chawla and Karakoulas, 2011] investigated empirically the problem of label noise bias introduced during the pseudo labeling process in this case and showed that the use of unlabeled examples could have
a minimal gain or even degrade performance, depending on the generalization ability of the initial classifier trained over the labeled training data.

In this paper, we study the generalization ability of a self-training algorithm with halfspaces that operates in two steps. In the first step, halfspaces are found iteratively over the set of labeled and unlabeled training data by maximizing the unsigned-margin of unlabeled examples and then assigning pseudo-labels to those with a distance greater than a found threshold. These pseudo-labels are supposed to be corrupted by label noise and pseudo-labeled examples are added to the training set, and a new classifier is then learned. This process is repeated until there are no more unlabeled examples to pseudo-label. In the second step, pseudo-labeled examples with an unsigned-margin greater than the last found threshold are removed from the training set.

Our contribution is twofold: (a) we present a first generalization bound for self-training with halfspaces in the case where class labels of examples are supposed to be corrupted by a Massart noise model; (b) we show that the use of unlabeled data in the proposed self-training algorithm does not degrade the performance of the first self-trained halfspace over the labeled training data.

In the remainder of the paper, Section 2 presents the definitions and the learning objective. In Section 3, we present in detail the adaptation of the self-training algorithm for halfspaces with a preliminary analysis in Section 4. Section 5 presents a bound over the misclassification error of the classifier outputted by the proposed algorithm and demonstrates that this misclassification error is upper-bounded by the misclassification error of the fully supervised halfspace. In Section 6, we present experimental results, and we conclude this work in Section 7.

2 Framework and Notations

We consider binary classification problems where the input space $X$ is a subset of $\mathbb{R}^d$, and the output space is $Y = \{-1, +1\}$. We study learning algorithms that operate in hypothesis space $H_d = \{h_w : X \rightarrow Y\}$ of centered halfspaces, where each $h_x \in H_d$ is a Boolean function of the form $h_x(x) = \text{sign}(\langle w, x \rangle)$, with $w \in \mathbb{R}^d$ such that $\|w\|_2 \leq 1$.

Our analysis succeeds the recent theoretical advances in robust supervised learning of polynomial algorithms for training halfspaces under large margin assumption [Diakonikolas et al., 2019; Montasser et al., 2020; Diakonikolas and Kane, 2020; Johnson et al., 2020; Diakonikolas et al., 2021], where the label distribution has been corrupted with the Massart noise model [Massart and Nédélec, 2006]. These studies derive a PAC bound for generalization error for supervised classifiers that depends on the corruption rate of the labeled training set and shed light on a new perspective for analyzing the self-training algorithm. Similarly, in our analysis, we suppose that self-training can be seen as learning with an imperfect expert. Whereat each iteration, labels of the pseudo-labeled set have been corrupted with a Massart noise [Massart and Nédélec, 2006] oracle defined as:

**Definition 2.1** ([Massart and Nédélec, 2006] noise oracle).

Let $\mathcal{C} = \{f : X \rightarrow Y\}$ be a class of Boolean functions over $X \subseteq \mathbb{R}^d$, with $f$ an unknown target function in $\mathcal{C}$, and $0 \leq \eta < 1/2$. Let $\eta(x) : X \rightarrow [0, \eta]$ be an unknown parameter function, and $D_\eta$ any marginal distribution over $X$. The corruption oracle $O(f, D_\eta, \eta)$ works as follow: each time $O(f, D_\eta, \eta)$ is invoked, it returns a pair $(x, y)$ where $x$ is generated i.i.d. from $D_\eta$; $y = -f(x)$ with probability $\eta(x)$ and $y = f(x)$ with probability $1 - \eta(x)$.

Let $D$ denote the joint distribution over $X \times Y$ generated by the above oracle with an unknown parameter function $\eta^{(0)}$ defined as $\eta^{(0)}(x) : X \rightarrow [0, \eta]$. We suppose that the training set is composed of $l$ labeled samples $S_l = \{(x_i, y_i)_{1 \leq i \leq l} \subseteq (X \times Y)^l$ and $u$ unlabeled samples $X_u = \{(x_i)_{l+1 \leq i \leq l+u} \subseteq X^u$, where $l < u$. Furthermore, we suppose that each pair $(x, y) \in X \times Y$ is i.i.d. with respect to the probability distribution $D$, we denote by $D_x$ the marginal of $D$ on $x$, and $D_y(x)$ the distribution of $y$ conditional on $x$. Finally, for any integer $d$, let $[d] = \{0, ..., d\}$.

3 Self-Training with Halfspaces

Given $S_l$ and $X_u$ drawn i.i.d. from a distribution $D$ corrupted with $O(f, D_\eta, \eta^{(0)})$. Algorithm 1 learns iteratively a list of halfspaces $L_m = [(w_1, \gamma_1), ..., (w_m, \gamma_m)]$ with each round consisting of exploration and pruning steps.

The goal of the exploration phase is to discover the halfspace with the highest margin on the set of unlabeled samples that are not still pseudo-labeled. This is done by first, learning a halfspace that minimizes the empirical surrogate loss of $R_D(w) = \mathbb{E}_{(x, y) \sim D}[\ell(y, h_w(x))]$ over a set of labeled and already pseudo-labeled examples $S^{(k)}$ from $S_l$ and $X_u$:

$$\min_w R_{S^{(k)}}(w) = \frac{1}{|S^{(k)}|} \sum_{(x, y) \in S^{(k)}} \ell(y, h_w(x))$$

s.t. $\|w\|_2 \leq 1$

At round $k = 0$, we have $S^{(0)} = S_l$. Once the halfspace with parameters $w^{(k)}$ is found, a threshold $\gamma^{(k)}$ is defined as the highest unsigned-margin in $S^{(k)}$, is set such that the empirical loss over the set of examples in $S^{(k)}$ with unsigned-margin above $\gamma^{(k)}$, is the lowest. In the pseudo-code of the algorithm, $S^{(k)}$ refers to the subset of examples in $S^{(k)}$ having an unsigned margin greater or equal to $\omega \times i$. Unlabeled examples $x \in X_u$ that are not pseudo-labeled are assigned labels, i.e., $y = \text{sign}(\langle w^{(k)}, x \rangle)$ if $\|\langle w^{(k)}, x \rangle\| \geq \gamma^{(k)}$. These pseudo-labeled examples are added to $S^{(k)}$ and removed from $X_u$, and a new halfspace minimizing Eq. (1) is found. Examples in $S^{(k)}$ are supposed to be misclassified by the oracle $O(f, D_x, \eta^{(k)})$ following Definition 2.1 with the parameter function $\eta^{(k)}$ that refers to the conditional probability of corruption in $S^{(k)}$ defined as $\eta^{(k)}(x) = \mathbb{P}_{y \sim S^{(k)}_y}[f(x) \neq y] \leq \eta^{(k)}$.

Once the halfspace with parameters $w^{(k)}$ and threshold $\gamma^{(k)}$ are found such that there are no more unlabeled samples having an unsigned-margin larger than $\gamma^{(k)}$, the pair $(w^{(k)}, \gamma^{(k)})$ is added to the list $L_m$, and samples from $S^{(k)}$ having an unsigned-margin above $\gamma^{(k)}$ are removed (pruning
Algorithm 1 Self-Training with Halvespaces

Input: \( S_t = (x_i, y_i)_{1 \leq i \leq t}, S_u = (x_i)_{t+1 \leq i \leq n}, p: \) number of threshold tests set to 5.

Set \( k \leftarrow 0, S(k) = S_t, U(k) = X_u, \quad w = \frac{|S(k)|}{p}, L = \emptyset. \)

while \( |S(k)| \geq \ell \) do

Let \( w \) a random vector in \( \mathbb{R}^d \) such that \( \|w\|_2 \leq 1 \), and let the cost function defined

\[ R_{S(k)}(w) = \frac{1}{|S(k)|} \sum_{(x,y) \in S(k)} \ell(y, h_w(x)); \]

Run projected SGD on \( R_{S(k)}(w) \) to obtain \( w(k) \) such that \( \|w(k)\|_2 \leq 1; \)

Order \( S(k) \) by decreasing order of unsigned-margin to \( (k) \); Set a window of indices \( I = \{w, 2w, \ldots, pw\} \); find \( \ell = \arg \min \|x\|_{S(k)} \sum_{(x,y) \in S(k)} \ell(y, h_w(x)); \)

Set \( \gamma \) to the margin of the sample at position \( I[\ell]; \)

Let \( U(k) = \{x \in X_u \mid \langle w(k), x \rangle \geq \gamma \}; \)

if \( |U(k)| > 0 \) then

\( S_u(k) = \{ (x, y) \mid x \in U(k) \land y = \text{sign} (\langle w(k), x \rangle) \}; \)
\( S(k+1) = S(k) \cup S_u(k); \)
\( X_u \leftarrow X_u \setminus U(k); \)
else

\( L = L \cup \{ (w(k), \gamma(k)) \}; \)
\( S(k+1) = \{ (x, y) \in S(k) \mid \langle w(k), x \rangle < \gamma(k) \}; \)
end if

Set \( k \leftarrow k + 1, w = \frac{|S(k)|}{p}; \)
end while

Output: \( L_m = \langle \langle w^{(1)}, \gamma^{(1)} \rangle, \ldots, \langle w^{(m)}, \gamma^{(m)} \rangle \rangle. \)

4 Theoretical Analyses

Our goal is to find a hypothesis \( h_w \in H_d \) such that with high probability, the misclassification error \( \mathbb{P}_{(x,y) \sim D}[h_w(x) \neq y] \) is minimized, and, that with high probability the performance of the found solution is better or equal to any hypothesis in \( H_d \) obtained only from the labeled training set, \( S_t. \)

4.1 Learning Objective

We denote by \( \eta_w(x) = \mathbb{P}_{x \sim D}[h_w(x) \neq y] \) the conditional misclassification error of a hypothesis \( h_w \in H_d \) with respect to \( D \), and \( w^* \) the normal vector of \( h_w \in H_d \) that achieves the optimal misclassification error;

\[ \eta = \min_{w} \mathbb{P}_{(x,y) \sim D}[h_w(x) \neq y]. \]

By considering the indicator function \( I_\tau \) defined as \( I_\tau = 1 \) if the predicate \( \tau \) is true and 0 otherwise; we prove in the following lemma that the probability of misclassification of halfspaces over examples with an unsigned-margin greater than a threshold \( \gamma > 0 \) is bounded by the same quantity \( 0 < \eta < 1 \) that upper-bounds the misclassification error of these examples.

Lemma 4.1. For all \( h_w \in H_d \), if there exist \( \eta \in [0,1] \) and \( \gamma > 0 \) such that \( \mathbb{P}_{x \sim D}[h_w(x) \neq y] \geq \gamma \) and that \( \mathbb{E}_{x \sim D}[\ell(h_w(x) - \eta)|h_w(x)\geq \gamma] \leq 0 \), then \( \mathbb{P}_{(x,y) \sim D}[h_w(x) \neq y] \geq \gamma \).

Proof. For all hypotheses \( h_w \in H_d \), we know that the error achieved by \( h_w \) in the region of margin \( \gamma \) from \( h_w \) satisfies \( \mathbb{E}_{x \sim D}[\ell(h_w(x) - \eta)|h_w(x)\geq \gamma] \leq 0 \); by rewriting the expectation, we obtain the following

\[ \mathbb{E}_{x \sim D}[\ell(h_w(x)|h_w(x)\geq \gamma] - \mathbb{E}_{x \sim D}[\ell(h_w(x)\geq \gamma] \leq 0. \]

We have then

\[ \mathbb{E}_{x \sim D}[\ell(h_w(x)|h_w(x)\geq \gamma] \leq \eta \]

and the result follows from the equality:

\[ \mathbb{P}_{(x,y) \sim D}[h_w(x) \neq y] \geq \gamma = \frac{\mathbb{E}_{x \sim D}[\ell(h_w(x)|h_w(x)\geq \gamma]}{\mathbb{E}_{x \sim D}[\ell(h_w(x)\geq \gamma]. \]

Suppose that there exists a pair \((\tilde{w}, \tilde{\gamma})\) minimizing:

\[ \tilde{\gamma} = \arg \min_{w \in \mathbb{R}^d} \frac{\mathbb{E}_{x \sim D}[\ell(h_w(x)|h_w(x)\geq \gamma]}{\mathbb{E}_{x \sim D}[\ell(h_w(x)\geq \gamma]. \]

By defining \( \tilde{\gamma} \) as:

\[ \tilde{\eta} = \inf_{w \in \mathbb{R}^d} \frac{\mathbb{E}_{x \sim D}[\ell(h_w(x)|h_w(x)\geq \gamma]}{\mathbb{E}_{x \sim D}[\ell(h_w(x)\geq \gamma]. \]

The following inequality holds:

\[ \tilde{\eta} \leq \inf_{w \in \mathbb{R}^d} \frac{\mathbb{E}_{x \sim D}[\ell(h_w(x)|h_w(x)\geq \gamma]}{\mathbb{E}_{x \sim D}[\ell(h_w(x)\geq \gamma] = \eta^*. \]
This inequality paves the way for the following claim, which is central to our self-training strategy.

**Claim 4.2.** Suppose that there exists a pair \((\bar{w}, \gamma)\) satisfying the minimization problem (2) with \(P_{x \sim D_u}[\langle w, x \rangle \geq \gamma] > 0\), then \(P_{(x, y) \sim D}[\langle \bar{w}, x \rangle \neq y | \langle \bar{w}, x \rangle \geq \gamma] \leq \eta^*\).

**Proof.** The requirements of Lemma 4.1 are satisfied with \((w, \gamma) = (\bar{w}, \gamma)\) and \(\eta = \eta^*\). This claim is then proved using the conclusion of Lemma 4.1 together with the fact that \(\eta \leq \eta^*\).

The claim above demonstrates that for examples generated by the probability distribution \(D\), there exists a region in \(\mathcal{X}\) on either side of a margin \(\gamma\) to the decision boundary defined by \(\bar{w}\) solution of (Eq. 2); where the probability of misclassification error of the corresponding halfspace in this region is upper-bounded by the optimal misclassification error \(\eta^*\). This result is consistent with semi-supervised learning studies that consider the margin as an indicator of confidence and search the decision boundary on low-density regions [Joachims, 1999; Amini et al., 2009].

### 4.2 Problem Resolution

We use a block coordinate minimization method for solving the optimization problem (2). This strategy consists in first finding a halfspace with parameters \(\bar{w}\) that minimizes Eq. (2) with a threshold \(\gamma = 0\), and then by fixing \(\bar{w}\), finds the threshold \(\gamma\) for which Eq. (2) is minimum. We resolve this problem using the following claim, which links the misclassification error \(\eta_w\) and the perceptron loss \(\ell_p(y, h_w(x))\):

**Claim 4.3.** For a given weight vector \(w\), we have:

\[
E_{x \sim D_u}[\langle w, x \rangle | \eta_w(x)] = E_{(x, y) \sim D}[\ell_p(y, h_w(x))] \tag{3}
\]

**Proof.** For a fixed weight vector \(w\), we have that:

\[E_{(x, y) \sim D}[\ell_p(y, h_w(x))] = E_{(x, y) \sim D}[-y \langle w, x \rangle I_{y \langle w, x \rangle \leq 0}]\]

As we are considering misclassification errors, i.e.,

\[-y \langle w, x \rangle I_{y \langle w, x \rangle \leq 0} = I_{y \langle w, x \rangle \leq 0} \langle w, x \rangle\]

It comes that

\[E_{(x, y) \sim D}[\ell_p(y, h_w(x))] = E_{(x, y) \sim D}[\langle w, x \rangle | \eta_w(x)] = E_{y \sim D_u}[\langle w, x \rangle | \eta_w(x)] = 0\]

The result then follows from the definition of the misclassification error, i.e.,

\[\eta_w(x) = P_{y \sim D_u}[\langle w, x \rangle > 0] = 0\]

This claim shows that the minimization of the generalization error with \(\ell_p\) is equivalent to minimizing \(E_{x \sim D_u}[\langle w, x \rangle | \eta_w(x)]\). Hence, the minimization of \(E_{x \sim D_u}[\gamma_p(y, h_w(x))\) cannot result in bounded misclassification error, as the distribution of margins \(\langle w, x \rangle\) might vary widely between samples in \(\mathcal{X}\). In the following lemma, we show that it is possible to achieve bounded misclassification error under margin condition and \(L_2\)-norm constraint.

**Lemma 4.4.** For a fixed distribution \(D\), let \(R = \max_{x \sim D_u} \|x\|_2^2\) and \(\gamma > 0\), let \(\bar{w}\) and \(\bar{w}\) be defined as follows:

\[\bar{w} = \arg \min_{w, \|w\|_2 \leq 1} E_{x \sim D_u}[\langle w, x \rangle | \eta_w(x)] | \langle w, x \rangle \geq \gamma]\]

\[\bar{w} = \arg \min_{w, \|w\|_2 \leq 1} E_{x \sim D_u}[\eta_w(x)] | \langle w, x \rangle \geq \gamma]\]

**We then have:**

\[\frac{\gamma}{R} E_{x \sim D_u}[\eta_w(x)] | \langle \bar{w}, x \rangle \geq \gamma] \leq E_{x \sim D_u}[\eta_w(x)] | \langle \bar{w}, x \rangle \geq \gamma] \leq E_{x \sim D_u}[\eta_w(x)] | \langle \bar{w}, x \rangle \geq \gamma].\]

**Proof.** From the condition \(\langle \bar{w}, x \rangle \geq \gamma\), we have:

\[\gamma E_{x \sim D_u}[\eta_w(x)] | \langle \bar{w}, x \rangle \geq \gamma] \leq E_{x \sim D_u}[\eta_w(x)] | \langle \bar{w}, x \rangle \geq \gamma]

Applying the definition of \(\eta\) to the right-hand side of the above inequality gives:

\[\gamma E_{x \sim D_u}[\eta_w(x)] | \langle \bar{w}, x \rangle \geq \gamma] \leq E_{x \sim D_u}[\eta_w(x)] | \langle \bar{w}, x \rangle \geq \gamma]

Using the Cauchy–Schwarz inequality and the definition of \(R\), we get:

\[\gamma E_{x \sim D_u}[\eta_w(x)] | \langle \bar{w}, x \rangle \geq \gamma] \leq R E_{x \sim D_u}[\eta_w(x)] | \langle \bar{w}, x \rangle \geq \gamma]

Then from the definition of \(\bar{w}\), we know:

\[R E_{x \sim D_u}[\eta_w(x)] | \langle \bar{w}, x \rangle \geq \gamma] \leq R E_{x \sim D_u}[\eta_w(x)] | \langle \bar{w}, x \rangle \geq \gamma]

Dividing the two inequalities above by \(R\) gives the result.

Lemma 4.4 guarantees that the approximation of the perceptron loss to the misclassification error is more accurate for examples that have a comparable distance to the halfspace. This result paves the way to our implementation of the self-training algorithm.

The proposed self-training algorithm operates iteratively, where at each round \(k\) only points with a large margin found at the previous iteration are considered for the minimization of \(R_{D}(w) = E_{(x, y) \sim D}[\text{Relu}(-y \langle w, x \rangle)]\) using the structural Risk Minimization (SRM) principle.

### 5 Corruption Noise Modeling and Generalization Guarantees

In the following, we relate the process of pseudo-labeling to the corruption noise model \(O(f, D_u, \eta^{(k)})\) for all pseudo-labeling iterations \(k\) in Algorithm 1, then we present a bound over the misclassification error of the classifier \(L_{opt}\) outputted by the algorithm and demonstrate that this misclassification error is upper-bounded by the misclassification error of the fully supervised halfspace.

**Claim 5.1.** Let \(S^{(0)} = S_t\) be a labeled set drawn i.i.d. from \(D = O(f, D_k, \eta^{(0)})\) and \(U^{(0)} = X_u\) an initial unlabeled set drawn i.i.d. from \(D_u\). For all iterations \(k \in [K]\) of Algorithm 1; the active labeled set \(S^{(k)}\) is drawn i.i.d. from \(D = O(f, D_k, \eta^{(k)})\) where the corruption noise distribution \(\eta^{(k)}\) is bounded by:

\[\forall k \in [K], \ E_{x \sim D_u}[\eta^{(k)}(x)] x \in S^{(k)} \leq \max_{j \in [K]} \eta^{(j)}\]
The proof is provided in the supplementary material. We can now bound the generalization error of the classifier $L_m$ outputted by Algorithm 1 with respect to the optimal misclassification error $\eta^*$ in the case where projected SGD is used for the minimization of Eq. (1). Note that in this case the time complexity of the algorithm is polynomial with respect to the dimension $d$, the upper bound on the bit complexity of examples, the total number of iterations, and the upper bound on SGD steps.

**Theorem 5.2.** Let $S_I$ be a set of i.i.d. samples of size $\ell$ drawn from a distribution $D = \mathcal{O}(f, D_x, \eta^{(0)})$ on $\mathbb{R}^d \times \{-1, +1\}$, where $f$ is an unknown concept function and $\eta^{(0)}$ an unknown parameter function bounded by 1/2, let $X_u$ be an unlabeled set of size $u$ drawn i.i.d. from $D_x$. Algorithm 1 terminates after $K$ iterations, and outputs a non-proper classifier $L_m$ of $m$ halfspaces such that with high probability:

$$\Pr_{(x, y) \sim D}[L_m(x) \neq y] \leq \eta^* + \max_{k \leq I} \epsilon^{(k)} + \pi_{K+1},$$

where $I$ is the set of rounds $k \in [K]$ at which the halfspaces were added to $L_m$, $\epsilon^{(k)}$ is the projected SGD convergence error rate at round $k$, and $\pi_{K+1}$ a negligible not-accounted mass of $D_x$.

The proof of Theorem 5.2 is based on the following property of projected SGD.

**Lemma 5.3 (From [Duchi, 2016]).** Let $R$ be a convex function of any type. Consider the projected SGD iteration, which starts with $\mathbf{w}^{(0)}$ and computes for each step.

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \alpha(t)g(t), \quad \mathbf{w}^{(t)} = \argmin_{\mathbf{w} \in \mathcal{D}} \|\mathbf{w} - \mathbf{w}^{(t-1)}\|_2.$$ 

Where $g(t)$ is a stochastic subgradient such that $E_{x \sim D_x}[g(x, \mathbf{w}^{(t)})] \in \partial R(\mathbf{w}) = \{g : R(\mathbf{w}') \geq R(\mathbf{w}) + \langle g, \mathbf{w}' - \mathbf{w} \rangle \ 	ext{for all} \ \mathbf{w}' \text{and} \ E_{x \sim D_x}[\|g(x, \mathbf{w})\|_2^2] \leq M^2$. For any $\epsilon, \delta > 0$: if the projected SGD is executed $T = \Omega(\log(1/\delta)/\epsilon^2)$ times with a step size $\alpha(t) = \frac{1}{M^2}t$, then for $\mathbf{w} = \frac{1}{T} \sum_{t=1}^T \mathbf{w}^{(t)}$, we have with probability at least $1 - \delta$ that $E_{x \sim D_x}[\hat{R}(\mathbf{w})] \leq \min_{\mathbf{w}} E_{x \sim D_x}[\hat{R}(\mathbf{w})] + \epsilon$.

**Proof of Theorem 5.2.** We consider the steps of Algorithm 1. At iteration $k$ of the while loop, we consider the active training set $S^{(k)}$ consisting of examples not handled in previous iterations.

We first note that the algorithm terminates after at most $K$ iterations. From the fact that at every iteration $k$, we discard a non-empty set from $S^{(k)}$ when we do not pseudo-label or from $U^{(k)}$ when we pseudo-label, and that the empirical distributions $S_I$ and $X_u$ are finite sets. By the guarantees of Lemma 5.3, running SGD (step 4) on $R_{S^{(k)}}$ for $T = \Omega(\log(1/\delta)/\epsilon^2)$ steps, we obtain a weight vector $\mathbf{w}^{(k)}$ such that with probability at least $1 - \delta$:

$$E_{x \sim D_x}[\hat{R}_{S^{(k)}}(\mathbf{w}^{(k)})] \leq \min_{\mathbf{w}, \|\mathbf{w}\|_2 \leq 1} E_{x \sim D_x}[\hat{R}_{S^{(k)}}(\mathbf{w})] + \epsilon^{(k)},$$

From Claim 4.3, we derive with high probability:

$$E_{x \sim D_x}[\langle \mathbf{w}^{(k)}, x \rangle \eta_{\mathbf{w}^{(k)}}(x)] \leq \min_{\mathbf{w}, \|\mathbf{w}\|_2 \leq 1} E_{x \sim D_x}[\langle \mathbf{w}, x \rangle \eta_{\mathbf{w}}(x)] + \epsilon^{(k)}.$$
of the first iteration obtained from the empirical distribution $S^{(0)} = S_E$, there is a high probability that:

$$\mathbb{P}_{(x,y) \sim D}[L_m(x) \neq y] \leq \mathbb{P}_{(x,y) \sim D}[h_{w^{(0)}}(x) \neq y]$$

Proof. By the guarantees of Lemma 5.3, the classifier $h_{w^{(0)}}$ obtained on running SGD on $\bar{R}_{S^{(0)}}$ with projection to the unit $l_2$-ball for $P^{(0)}$ steps satisfies:

$$\mathbb{E}_{(x,y) \sim D}[\text{Relu}(-y<w^{(0)}, x>)] - \mathbb{E}_{(x,y) \sim D}[\text{Relu}(-y<w^*, x>)] \leq \frac{3 \max_{x \in S_u} \|x\|}{2\sqrt{P^{(0)}}}$$

Let $k$ be the iteration at which the first pair $(w^{(1)}, \gamma^{(1)})$ is added to $L_m$. The first pruning phase in Algorithm 1 results in a set $S^{(k)} \subseteq S_t \cup \bigcup_{i=1}^{k-1} S_u^{(i)}$. Claim 5.1 ensures that the probability of corruption in the pseudo-labeled set $\bigcup_{i=1}^{k-1} S_u^{(i)}$ is bounded by $\max_{j \in [k]} \eta^{(j)} \leq \eta^* + \epsilon$.

In other words, the weight vector $w^{(1)}$ is obtained from an empirical distribution that includes both the initial labeled set $S_t$ and a pseudo-labeled set from $X_u$. Particularly, if this pseudo-labeled set is not empty, then its pseudo-labeling error is nearly optimal, which implies that $\mathbb{P}_{(x,y) \sim D}[h_{w^{(1)}}(x) \neq y] \leq \mathbb{P}_{(x,y) \sim D}[h_{w^{(0)}}(x) \neq y]$.

Ultimately, $L_m$ classifies a large fraction of the probability mass of $D$ with nearly optimal guarantees (i.e., Eq. (4) in proof of Theorem 5.2) and the rest using $h_{w^{(1)}}$ with an error of misclassification at most equal to $\mathbb{P}_{(x,y) \sim D}[h_{w^{(0)}}(x) \neq y]$. $\square$

6 Empirical Results

Datasets. We mainly consider data sets from [Chapelle et al., 2006]. Some of these collections such as baseball-hockey, pc-mac and religion-atheism are binary classification tasks extracted from the 20-newsgroups data set.

We used tf-idf representation for all textual data sets above. spambase is a collection of spam e-mails from the UCI repository [DuA and Graff, 2019]. one-two, odd-even are handwritten digits recognition tasks originally from optical recognition of handwritten digits database also from UCI repository, one-two is digits ’1’ versus ’2’; odd-even is the artificial task of classifying odd ’1, 3, 5, 7, 9’ versus even ’0, 2, 4, 6, 8’ digits. weather is a data set from Kaggle which contains about ten years of daily weather observations from many locations across Australia, and the objective is to classify next-day rain target variable.

We have also included data sets from extreme classification repository [Bhatia et al., 2015] mediamill2 and delicious2 by selecting the label which gives the best ratio in class distribution. The statistics of these data sets are given in Table 1.

Baseline methods. We implemented the halfspace or Linear Threshold Function (LTF) using TensorFlow 2.0 in python aside with Algorithm 1$^1$ ($L_m$), we ran a Support Vector Machine (SVM) [Cortes and Vapnik, 1995] with a linear kernel from the LIBLINEAR library [Fan et al., 2008] as another supervised classifier. We compared results with a semi-supervised Gaussian naive Bayes model (GM) [Chapelle et al., 2006] from the scikit-learn library. The working hypothesis behind (GM) is the cluster assumption stipulating that data contains homogeneous labeled clusters, which can be detected using unlabeled training samples. We also compared results with label propagation (LP) [Zhu and Ghahramani, 2002] which is a semi-supervised graph-based technique. We used the implementation of LP from the scikit-learn library.

This approach follows the manifold assumption that the decision boundary is located on a low-dimensional manifold and that unlabeled data may be utilized to identify it. We also included entropy regularized logistic regression (ERLR) proposed by [Grandvalet and Bengio, 2005] from [Krijthe, 2017]. This approach is based on low-density separation that stipulates that the decision boundary lies on low-density regions. In the implementation of [Krijthe, 2017], the initial supervised classifier is a logistic regression that has a similar performance to the SVM classifier. We evaluated these methods using relatively small labeled training sets $\ell \in \{10, 50, 100\}$, and for all methods, we used the default hyper-parameters as cross-validation methods would not be effective in this case.

Experimental Setup. In our experiments, we have randomly chosen 70% of each data collection for training and the remaining 30% for testing. We randomly selected sets of different sizes (i.e., $\ell \in \{10, 50, 100\}$) from the training set as labeled examples; the remaining was considered as unlabeled training samples. Results are evaluated over the test set using the accuracy measure. Each reported performance value is the average over the 20 random (labeled/unlabeled/test) sets of the initial collection. All experiments are carried out on a machine with an Intel Core i7 processor, 2.2Ghz quad-core, and 16Go 1600 MHz of RAM memory.

Analysis of Results. Table 2 summarizes the results. We used boldface (resp. underline) to indicate the highest (resp. the second-highest) performance rate, and the symbol $\dagger$ indicates that performance is significantly worse than the best.
result, according to a Wilcoxon rank-sum test with a p-value threshold of 0.01 [Wolfe, 2012]. From these results, it comes out that the proposed approach ($L_m$) consistently outperforms the supervised halfspace (LTF).

Furthermore, compared to other techniques, $L_m$ generally performs the best or the second-best. We also notice that in some cases, LP, GM, and ERLR outperform the supervised approaches, SVM and LTF (i.e., GM on spambase for $\ell \in \{10, 50\}$), but in other cases, they are outperformed by both SVM and LTF (i.e., GM on religion-atheism). These results suggest that unlabeled data contain useful information for classification and that existing semi-supervised techniques may use it to some extent. They also highlight that the development of semi-supervised algorithms following the given assumptions is necessary for learning with labeled and unlabeled training data but not sufficient.

These results underline the need of developing theoretically sound semi-supervised algorithms that show the method’s ability to generalize and to better understand the value of unlabeled training data in the learning process.

### 7 Conclusion

We presented a first bound over the misclassification error of a self-training algorithm that iteratively finds a list of halfspaces from partially labeled training data. Each round consists of two steps. The exploration phase’s purpose is to determine the halfspace with the largest margin and assign pseudo-labels to unlabeled observations with an unsigned-margin larger than the discovered threshold. The pseudo-labeled instances are then added to the training set, and the procedure is repeated until there are no more unlabeled instances to pseudo-label. In the pruning phase, the last halfspace with the largest threshold is preserved, ensuring that there are no more unlabeled samples with an unsigned-margin greater than this threshold and pseudo-labeled samples with an unsigned-margin greater than the specified threshold are removed. We ultimately show that the use of unlabeled data in the proposed self-training algorithm does not degrade the performance of the initially supervised classifier. An interesting future direction would be to quantify the real gain of learning with unlabeled and labeled training data compared to a fully supervised scheme.
References


