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

Lies Hadjadj, Massih-Reza Amini, Sana Louhichi

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

Lies Hadjadj[1] , **Massih-Reza Amini**[1] , **Sana Louhichi**[2]

[1]Computer Science Laboratory (LIG), Université Grenoble Alpes, Grenoble, France
[2]Department of Statistics (LJK), Université Grenoble Alpes, Grenoble, France
{Lies.Hadjadj, Massih-Reza.Amini, Sana.Louhichi}@univ-grenoble-alpes.fr

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 [Balcan 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 regaring 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 halfspace trained 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 $\mathcal{X}$ is a subset of $\mathbb{R}^d$, and the output space is $\mathcal{Y} = \{-1, +1\}$. We study learning algorithms that operate in hypothesis space $\mathcal{H}_d = \{h_{\mathbf{w}} : \mathcal{X} \to \mathcal{Y}\}$ of centered halfspaces, where each $h_{\mathbf{w}} \in \mathcal{H}_d$ is a Boolean function of the form $h_{\mathbf{w}}(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$, with $\mathbf{w} \in \mathbb{R}^d$ such that $\|\mathbf{w}\|_2 \leq 1$.

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

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

3 Self-Training with Halfspaces

Given $\mathbf{S}_\ell$ and $\mathbf{X}_u$ drawn i.i.d. from a distribution $\mathcal{D}$ corrupted with $\mathcal{O}(f, \mathcal{D}_{\mathbf{x}}, \eta^{(0)})$. Algorithm 1 learns iteratively a list of halfspaces $L_m = [(\mathbf{w}^{(1)}, \gamma^{(1)}), ..., (\mathbf{w}^{(m)}, \gamma^{(m)})]$ 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 $\mathcal{R}_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(y, h_{\mathbf{w}}(\mathbf{x}))]$ over a set of labeled and already pseudo-labeled examples $\mathbf{S}^{(k)}$ from $\mathbf{S}_\ell$ and $\mathbf{X}_u$:

$$\min_{\mathbf{w}} \hat{\mathcal{R}}_{\mathbf{S}^{(k)}}(\mathbf{w}) = \frac{1}{|\mathbf{S}^{(k)}|} \sum_{(\mathbf{x},y)\in\mathbf{S}^{(k)}} \ell(y, h_{\mathbf{w}}(\mathbf{x})) \qquad (1)$$
$$\text{s.t. } ||\mathbf{w}||_2 \leq 1$$

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

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

Algorithm 1 Self-Training with Halfspaces

Input : $\mathbf{S}_\ell = (\mathbf{x}_i, y_i)_{1\leq i\leq l}$, $\mathbf{X}_u = (\mathbf{x}_i)_{l+1\leq i\leq n}$, p: number of threshold tests set to 5.

Set $k \leftarrow 0$, $\mathbf{S}^{(k)} = \mathbf{S}_\ell$, $\mathbf{U}^{(k)} = \mathbf{X}_u$, $w = \frac{|\mathbf{S}^{(k)}|}{p}$, $L = []$.

while $|\mathbf{S}^{(k)}| \geq \ell$ **do**

Let $\mathbf{w}$ a random vector in $\mathbb{R}^d$ such that $\|\mathbf{w}\|_2 \leq 1$, and let the cost function defined $\hat{\mathcal{R}}_{\mathbf{S}^{(k)}}(\mathbf{w}) = \frac{1}{|\mathbf{S}^{(k)}|}\sum_{(\mathbf{x},y)\in\mathbf{S}^{(k)}}[\ell_p(y, h_{\mathbf{w}}(\mathbf{x}))]$;

Run projected SGD on $\hat{\mathcal{R}}_{\mathbf{S}^{(k)}}(\mathbf{w})$ to obtain $\mathbf{w}^{(k)}$ such that $\|\mathbf{w}^{(k)}\|_2 \leq 1$;

Order $\mathbf{S}^{(k)}$ by decreasing order of margin from $\mathbf{w}^{(k)}$;
Set a window of indices $I = [w, 2w, ..., pw]$;
find $t = \underset{i\in I}{\operatorname{argmin}} \frac{1}{|\mathbf{S}^{(k)}_{\geq i}|}\sum_{(\mathbf{x},y)\in\mathbf{S}^{(k)}_{\geq i}} \mathbb{1}_{h_{\mathbf{w}^{(k)}}(\mathbf{x})\neq y}$;
Set $\gamma^{(k)}$ to the margin of the sample at position $I[t]$;
Let $\mathbf{U}^{(k)} = \{\mathbf{x}\in\mathbf{X}_u \big| |\langle\mathbf{w}^{(k)}, \mathbf{x}\rangle| \geq \gamma^{(k)}\}$;
if $|\mathbf{U}^{(k)}| > 0$ **then**
$\mathbf{S}^{(k)}_u = \{(\mathbf{x}, y)\big|\mathbf{x}\in\mathbf{U}^{(k)} \wedge y = \text{sign}(\langle\mathbf{w}^{(k)}, \mathbf{x}\rangle)\}$;
$\mathbf{S}^{(k+1)} \leftarrow \mathbf{S}^{(k)} \cup \mathbf{S}^{(k)}_u$;
$\mathbf{X}_u \leftarrow \mathbf{X}_u \setminus \mathbf{U}^{(k)}$;
else
$L = L \cup [(\mathbf{w}^{(k)}, \gamma^{(k)})]$;
$\mathbf{S}^{(k+1)} = \{(\mathbf{x}, y) \in \mathbf{S}^{(k)} \big| |\langle\mathbf{w}^{(k)}, \mathbf{x}\rangle| < \gamma^{(k)}\}$;
end if
Set $k \leftarrow k+1, w = \frac{|\mathbf{S}^{(k)}|}{p}$;
end while
Output : $L_m = [(\mathbf{w}^{(1)}, \gamma^{(1)}), ..., (\mathbf{w}^{(m)}, \gamma^{(m)})]$.

phase). Remind that $\gamma^{(k)}$ is the largest threshold above which the misclassification error over $\mathbf{S}^{(k)}$ increases.

The self-training algorithm 1, takes as input the labeled set $\mathbf{S}_\ell$, the unlabeled set $\mathbf{X}_u$ and p, which refers to the number of tests for threshold estimation, set to 5. After finding the weight vector $\mathbf{w}^{(k)}$ at round k, with projected SGD (step $4-5$), we order the labeled set $\mathbf{S}^{(k)}$ (with $\mathbf{S}^{(0)} = \mathbf{S}_\ell$) by decreasing order of unsigned-margin to $\mathbf{w}^{(k)}$. The threshold $\gamma^{(k)}$ is defined as the largest margin such that the error of examples in $\mathbf{S}^{(k)}$ with an unsigned-margin higher than $\gamma^{(k)}$ increases (step $8-9$). At this stage, observations $\mathbf{x} \in \mathbf{X}_u$ with an unsigned-margin greater than $\gamma^{(k)}$ (step $10-14$), are pseudo-labeled and added to the labeled set $\mathbf{S}^{(k)}$ and they are removed from the unlabeled set. This exploration phase of finding a halfspace with the largest threshold $\gamma^{(k)}$ is repeated until there are no more unlabeled samples with an unsigned-margin larger than this threshold. After this phase, the pruning phase begins by removing examples in $\mathbf{S}^{(k)}$ with an unsigned-margin strictly less than $\gamma^{(k)}$ (step $15-18$). The parameters of the halfspace and the corresponding threshold are added to the list of selected classifiers L_m and the procedure is repeated until that the size of the labeled set becomes less than l. To classify an unknown example $\mathbf{x}$, the prediction of the first halfspace with normal vector $\mathbf{w}^{(i)}$ in the list L_m, such that the unsigned-margin $|\langle\mathbf{w}^{(i)}, \mathbf{x}\rangle|$ of $\mathbf{x}$ is higher or equal to the corresponding threshold $\gamma^{(i)}$, is returned. By abuse of notation, we note that the prediction for $\mathbf{x}$ is $L_m(\mathbf{x}) = h_{\mathbf{w}^{(i)}}(\mathbf{x})$. If no such halfspace exists, the observation is classified using the prediction of the first classifier $h_{\mathbf{w}^{(1)}}$ that was trained over all the labeled and the pseudo-labeled samples without pruning.

4 Theoretical Analyses

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

4.1 Learning Objective

We denote by $\eta_{\mathbf{w}}(\mathbf{x}) = \mathbb{P}_{y\sim\mathcal{D}_y(\mathbf{x})}[h_{\mathbf{w}}(\mathbf{x}) \neq y]$ the conditional misclassification error of a hypothesis $h_{\mathbf{w}} \in \mathcal{H}_d$ with respect to $\mathcal{D}$, and $\mathbf{w}^*$ the normal vector of $h_{\mathbf{w}^*} \in \mathcal{H}_d$ that achieves the optimal misclassification error; $\boldsymbol{\eta}^* = \min_{\mathbf{w},\|\mathbf{w}\|_2\leq 1} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h_{\mathbf{w}}(\mathbf{x}) \neq y]$.

By considering the indicator function $\mathbb{1}_\pi$ defined as $\mathbb{1}_\pi = 1$ if the predicate π 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 $0 < \boldsymbol{\eta} < 1$ that upper-bounds the misclassification error of these examples.

Lemma 4.1. *For all $h_{\mathbf{w}} \in \mathcal{H}_d$, if there exist $\boldsymbol{\eta} \in]0,1[$ and $\gamma > 0$ such that $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|\langle\mathbf{w},\mathbf{x}\rangle| \geq \gamma] > 0$ and that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\eta_{\mathbf{w}}(\mathbf{x}) - \boldsymbol{\eta})\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq\gamma}] \leq 0$, then $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h_{\mathbf{w}}(\mathbf{x}) \neq y \big| |\langle\mathbf{w},\mathbf{x}\rangle| \geq \gamma] \leq \boldsymbol{\eta}$.*

Proof. For all hypotheses $h_{\mathbf{w}}$ in $\mathcal{H}_d$, we know that the error achieved by $h_{\mathbf{w}}$ in the region of margin γ from $\mathbf{w}$ satisfies $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\eta_{\mathbf{w}}(\mathbf{x}) - \boldsymbol{\eta})\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq\gamma}] \leq 0$; by rewriting the expectation, we obtain the following $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\eta_{\mathbf{w}}(\mathbf{x})\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq\gamma}] - \boldsymbol{\eta}\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|\langle\mathbf{w},\mathbf{x}\rangle| \geq \gamma] \leq 0$. We have then $\frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\eta_{\mathbf{w}}(\mathbf{x})\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq\gamma}]}{\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|\langle\mathbf{w},\mathbf{x}\rangle|\geq\gamma]} \leq \boldsymbol{\eta}$ and the result follows from the equality:

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h_{\mathbf{w}}(\mathbf{x}) \neq y \big| |\langle\mathbf{w},\mathbf{x}\rangle| \geq \gamma] = \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\eta_{\mathbf{w}}(\mathbf{x})\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq\gamma}]}{\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|\langle\mathbf{w},\mathbf{x}\rangle|\geq\gamma]}.$$

□

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

$$(\widetilde{\mathbf{w}}, \widetilde{\gamma}) \in \underset{\mathbf{w}\in\mathbb{R}^d,\gamma\geq 0}{\operatorname{argmin}} \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\eta_{\mathbf{w}}(\mathbf{x})\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq\gamma}]}{\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|\langle\mathbf{w},\mathbf{x}\rangle| \geq \gamma]}. \quad (2)$$

By defining $\widetilde{\eta}$ as:

$$\widetilde{\eta} = \inf_{\mathbf{w}\in\mathbb{R}^d,\gamma\geq 0} \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\eta_{\mathbf{w}}(\mathbf{x})\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq\gamma}]}{\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|\langle\mathbf{w},\mathbf{x}\rangle| \geq \gamma]}.$$

The following inequality holds:

$$\widetilde{\eta} \leq \inf_{\mathbf{w}\in\mathbb{R}^d} \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\eta_{\mathbf{w}}(\mathbf{x})\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq 0]}}{\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|\langle\mathbf{w},\mathbf{x}\rangle| \geq 0]} = \boldsymbol{\eta}^*.$$

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 $(\widetilde{\mathbf{w}}, \widetilde{\gamma})$ satisfying the minimization problem (2) with $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle\widetilde{\mathbf{w}}, \mathbf{x}\rangle| \geq \widetilde{\gamma}] > 0$, then $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h_{\widetilde{\mathbf{w}}}(\mathbf{x}) \neq y \big| |\langle\widetilde{\mathbf{w}}, \mathbf{x}\rangle| \geq \widetilde{\gamma}] \leq \boldsymbol{\eta}^*$.*

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

The claim above demonstrates that for examples generated by the probability distribution $\mathcal{D}$, there exists a region in $\mathcal{X}$ on either side of a margin $\widetilde{\gamma}$ to the decision boundary defined by $\widetilde{\mathbf{w}}$ 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 $\boldsymbol{\eta}^*$. 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 $\widetilde{\mathbf{w}}$ that minimizes Eq. (2) with a threshold $\gamma = 0$, and then by fixing $\widetilde{\mathbf{w}}$, finds the threshold $\widetilde{\gamma}$ for which Eq. (2) is minimum. We resolve this problem using the following claim, which links the misclassification error $\eta_\mathbf{w}$ and the perceptron loss $\ell_p(y, h_\mathbf{w}(\mathbf{x}))$: $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+; \ell_p(y, h_\mathbf{w}(\mathbf{x})) = -y\langle\mathbf{w}, \mathbf{x}\rangle \mathbb{1}_{y\langle\mathbf{w},\mathbf{x}\rangle \leq 0}$.

Claim 4.3. *For a given weight vector $\mathbf{w}$, we have:*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle\mathbf{w}, \mathbf{x}\rangle| \eta_\mathbf{w}(\mathbf{x})] = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell_p(y, h_\mathbf{w}(\mathbf{x}))] \quad (3)$$

Proof. For a fixed weight vector $\mathbf{w}$, we have that: $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell_p(y, h_\mathbf{w}(\mathbf{x}))] = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[-y\langle\mathbf{w}, \mathbf{x}\rangle \mathbb{1}_{y\langle\mathbf{w},\mathbf{x}\rangle\leq 0}]$. As we are considering misclassification errors, i.e., $-y\langle\mathbf{w}, \mathbf{x}\rangle \mathbb{1}_{y\langle\mathbf{w},\mathbf{x}\rangle\leq 0} = \mathbb{1}_{y\langle\mathbf{w},\mathbf{x}\rangle\leq 0}|\langle\mathbf{w}, \mathbf{x}\rangle|$, it comes that $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell_p(y, h_\mathbf{w}(\mathbf{x}))] = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[|\langle\mathbf{w}, \mathbf{x}\rangle| \mathbb{P}_{y\sim\mathcal{D}_{y(\mathbf{x})}}[-y\langle\mathbf{w}, \mathbf{x}\rangle > 0]]$. The result then follows from the definition of the misclassification error, i.e., $\eta_\mathbf{w}(\mathbf{x}) = \mathbb{P}_{y\sim\mathcal{D}_{y(\mathbf{x})}}[-y\langle\mathbf{w}, \mathbf{x}\rangle > 0]$. □

This claim shows that the minimization of the generalization error with ℓ_p is equivalent to minimizing $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle\mathbf{w}, \mathbf{x}\rangle| \eta_\mathbf{w}(\mathbf{x})]$. Hence, the minimization of $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\ell_p(y, h_\mathbf{w}(\mathbf{x}))]$ cannot result in bounded misclassification error, as the distribution of margins $|\langle\mathbf{w}, \mathbf{x}\rangle|$ 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 $\mathcal{D}$, let $R = \max_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}} \|\mathbf{x}\|_2$ and $\gamma > 0$, let $\widetilde{\mathbf{w}}$ and $\bar{\mathbf{w}}$ be defined as follows:*

$$\widetilde{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}, ||\mathbf{w}||_2 \leq 1} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle\mathbf{w}, \mathbf{x}\rangle| \eta_\mathbf{w}(\mathbf{x}) \big| |\langle\mathbf{w}, \mathbf{x}\rangle| \geq \gamma]$$

$$\overline{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}, ||\mathbf{w}||_2 \leq 1} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_\mathbf{w}(\mathbf{x}) \big| |\langle\mathbf{w}, \mathbf{x}\rangle| \geq \gamma].$$

We then have:

$$\frac{\gamma}{R}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_{\widetilde{\mathbf{w}}}(\mathbf{x}) \big| |\langle\widetilde{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma] \leq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_{\overline{\mathbf{w}}}(\mathbf{x}) \big| |\langle\overline{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma] \leq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_{\widetilde{\mathbf{w}}}(\mathbf{x}) \big| |\langle\widetilde{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma].$$

Proof. From the condition $|\langle\widetilde{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma$, we have:

$$\gamma\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_{\widetilde{\mathbf{w}}}(\mathbf{x}) \big| |\langle\widetilde{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma] \leq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle\widetilde{\mathbf{w}}, \mathbf{x}\rangle| \eta_{\widetilde{\mathbf{w}}}(\mathbf{x}) \big| |\langle\widetilde{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma]$$

Applying the definition of $\widetilde{\mathbf{w}}$ to the right-hand side of the above inequality gives:

$$\gamma\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_{\widetilde{\mathbf{w}}}(\mathbf{x}) \big| |\langle\widetilde{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma] \leq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle\overline{\mathbf{w}}, \mathbf{x}\rangle| \eta_{\overline{\mathbf{w}}}(\mathbf{x}) \big| |\langle\overline{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma]$$

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

$$\gamma\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_{\widetilde{\mathbf{w}}}(\mathbf{x}) \big| |\langle\widetilde{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma] \leq R\,\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_{\overline{\mathbf{w}}}(\mathbf{x}) \big| |\langle\overline{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma]$$

Then from the definition of $\overline{\mathbf{w}}$, we know:

$$R\,\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_{\overline{\mathbf{w}}}(\mathbf{x}) \big| |\langle\overline{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma] \leq R\,\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_{\widetilde{\mathbf{w}}}(\mathbf{x}) \big| |\langle\widetilde{\mathbf{w}}, \mathbf{x}\rangle| \geq \gamma]$$

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-learning 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 $\mathcal{R}_\mathcal{D}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\text{Relu}(-y\langle\mathbf{w}, \mathbf{x}\rangle)]$ 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 $\mathcal{O}(f, \mathcal{D}_\mathbf{x}, \eta^{(k)})$ for all pseudo-labeling iterations k in Algorithm 1, then we present a bound over the misclassification error of the classifier L_m 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 $\mathbf{S}^{(0)} = \mathbf{S}_\ell$ be a labeled set drawn i.i.d. from $\mathcal{D} = \mathcal{O}(f, \mathcal{D}_\mathbf{x}, \eta^{(0)})$ and $\mathbf{U}^{(0)} = \mathbf{X}_u$ an initial unlabeled set drawn i.i.d. from $\mathcal{D}_\mathbf{x}$. For all iterations $k \in [K]$ of Algorithm 1; the active labeled set $\mathbf{S}^{(k)}$ is drawn i.i.d. from $\mathcal{D} = \mathcal{O}(f, \mathcal{D}_\mathbf{x}, \eta^{(k)})$ where the corruption noise distribution $\eta^{(k)}$ is bounded by:*

$$\forall k \in [K], \quad \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta^{(k)}(\mathbf{x}) \big| \mathbf{x} \in \mathbf{S}^{(k)}] \leq \max_{j\in[K]} \boldsymbol{\eta}^{(j)}$$

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 η^* 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 $\mathbf{S}_\ell$ be a set of i.i.d. samples of size ℓ drawn from a distribution $\mathcal{D} = \mathcal{O}(f, \mathcal{D}_\mathbf{x}, \eta^{(0)})$ 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 $\mathbf{X}_u$ be an unlabeled set of size u drawn i.i.d. from $\mathcal{D}_\mathbf{x}$. Algorithm 1 terminates after K iterations, and outputs a non-proper classifier L_m of m halfspaces such that with high probability:*

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[L_m(\mathbf{x}) \neq y] \leq \boldsymbol{\eta}^* + \max_{k\in I} \epsilon^{(k)} + \pi_{K+1},$$

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 π_{K+1} a negligible not-accounted mass of $\mathcal{D}_\mathbf{x}$.

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

Lemma 5.3 (From [Duchi, 2016])**.** *Let $\hat{\mathcal{R}}$ be a convex function of any type. Consider the projected SGD iteration, which starts with $\mathbf{w}^{(0)}$ and computes for each step. $\mathbf{w}^{(t+\frac{1}{2})} = \mathbf{w}^{(t)} - \alpha^{(t)} g^{(t)}; \mathbf{w}^{(t+1)} = \operatorname{argmin}_{\mathbf{w}:||\mathbf{w}||_2\leq 1} ||\mathbf{w} - \mathbf{w}^{(t+\frac{1}{2})}||_2$. Where $g^{(t)}$ is a stochastic subgradient such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[g(\mathbf{w}, \mathbf{x})] \in \partial\hat{\mathcal{R}}(\mathbf{w}) = \{g : \hat{\mathcal{R}}(\mathbf{w}') \geq \hat{\mathcal{R}}(\mathbf{w}) + \langle \mathbb{E}[g], \mathbf{w}' - \mathbf{w}\rangle$ for all $\mathbf{w}'\}$ and $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[||g(\mathbf{w}, \mathbf{x})||_2^2] \leq M^2$. For any $\epsilon, \delta > 0$; if the projected SGD is executed $T = \Omega(\log(1/\delta)/\epsilon^2)$ times with a step size $\alpha^{(t)} = \frac{1}{M\sqrt{t}}$, then for $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}$, we have with probability at least $1 - \delta$ that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\hat{\mathcal{R}}(\bar{\mathbf{w}})] \leq \min_{\mathbf{w},\|\mathbf{w}\|_2\leq 1} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\hat{\mathcal{R}}(\mathbf{w})] + \epsilon$.*

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 $\mathbf{S}^{(k)}$ 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 $\mathbf{S}^{(k)}$ when we do not pseudo-label or from $\mathbf{U}^{(k)}$ when we pseudo-label, and that the empirical distributions $\mathbf{S}_\ell$ and $\mathbf{X}_u$ are finite sets. By the guarantees of Lemma 5.3, running SGD (step 4) on $\hat{\mathcal{R}}_{\mathbf{S}^{(k)}}$ for $T = \Omega(\log(1/\delta)/\epsilon^2)$ steps, we obtain a weight vector $\mathbf{w}^{(k)}$ such that with probability at least $1 - \delta$:

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

From Claim 4.3, we derive with high probability:

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle \mathbf{w}^{(k)}, \mathbf{x}\rangle| \eta_{\mathbf{w}^{(k)}}(\mathbf{x})] \leq \min_{\mathbf{w},\|\mathbf{w}\|_2\leq 1} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle \mathbf{w}, \mathbf{x}\rangle| \eta_{\mathbf{w}}(\mathbf{x})] + \epsilon^{(k)}.$$

Then the margin $\gamma^{(k)}$ is estimated minimizing Eq. (2) given $\mathbf{w}^{(k)}$, following Lemma 4.4 with $R^{(k)} = \max_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}} \|\mathbf{x}\|_2$ the radius of the truncated support of the marginal distribution $\mathcal{D}_\mathbf{x}$ at iteration k, we can assume that $\frac{\gamma^{(k)}}{R^{(k)}} \approx 1$, $\forall k \in [K]$, one may argue that the assumption is unrealistic knowing that the sequence of $(\gamma^{(k)})_{k=1}^m$ decreases overall, but as we show in the supplementary, we prove in Theorem A.1 that under some convergence guarantees of the pairs $\{(\mathbf{w}^{(k)}, \mathbf{w}^{(k+1)})\}_{k=1}^{m-1}$, one can show that the sequence $\{R^{(k)}\}_{k=1}^m$ decreases as a function of $\gamma^{(k)}$ respectively to k. As a result, we can derive with high probability:

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_{\mathbf{w}^{(k)}}(\mathbf{x}) \big| |\langle \mathbf{w}^{(k)}, \mathbf{x}\rangle| \geq \gamma^{(k)}] \leq \min_{\mathbf{w},\|\mathbf{w}\|_2\leq 1} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta_{\mathbf{w}}(\mathbf{x}) \big| |\langle \mathbf{w}, \mathbf{x}\rangle| \geq \gamma^{(k)}] + \epsilon^{(k)}.$$

From Claim 4.2 and giving the pair $(\mathbf{w}^{(k)}, \gamma^{(k)})$, we obtain with high probability that at round k:

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h_{\mathbf{w}^{(k)}}(\mathbf{x}) \neq y \big| |\langle \mathbf{w}^{(k)}, \mathbf{x}\rangle| \geq \gamma^{(k)}] \leq \boldsymbol{\eta}^* + \epsilon^{(k)}. \quad (4)$$

When the while loop terminates, we have accounted $m \leq K$ halfspaces in the list L_m satisfying Eq. (4). For all $k \in I$, every classifier $h_{\mathbf{w}^{(k)}}$ in L_m has guarantees on an empirical distribution mass of at least $\widetilde{\kappa} = \min_{k\in I} \mathbb{P}_{\mathbf{x}\sim\mathbf{S}^{(k)}}[|\langle \mathbf{w}^{(k)}, \mathbf{x}\rangle| \geq \gamma^{(k)}]$; the DKW (Dvoretzky-Kiefer-Wolfowitz) inequality [Dvoretzky *et al.*, 1956] implies that the true probability mass $\kappa = \min_{k\in I} \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle \mathbf{w}^{(k)}, \mathbf{x}\rangle| \geq \gamma^{(k)}]$ of this region is at least $\widetilde{\kappa} - \sqrt{\frac{\log\frac{2}{\delta}}{2|\mathbf{S}^{(n)}|}}$ with probability $1 - \delta$, where $n = \operatorname{argmin}_{k\in I} \mathbb{P}_{\mathbf{x}\sim\mathbf{S}^{(k)}}[|\langle \mathbf{w}^{(k)}, \mathbf{x}\rangle| \geq \gamma^{(k)}]$.
The pruning phase in the algorithm ensures that these regions are disjoint for all halfspaces in L_m, it follows that using the Boole–Fréchet inequality [Boole, 2015] on the conjunctions of Eq. (4) overall rounds $k \in [I]$, implies that L_m classifies at least a $(1 - m\kappa)$-fraction of the total probability mass of $\mathcal{D}$ with guarantees of Eq. (4) with high probability, let $\pi_{K+1} = \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[x \in \mathbf{S}^{(K+1)}]$ be the probability mass of the region not accounted by L_m. We argue that this region is negligible from the fact that $|\mathbf{S}^{(K+1)}| < \ell$ and $\ell \ll u$, such that setting $\boldsymbol{\epsilon} = \max_{k\in I} \epsilon^{(k)} + \pi_{K+1}$ provides the result. □

In the following, we show that the misclassification error of the classifier L_m output of Algorithm 1 is at most equal to the error of the supervised classifier obtained over the labeled training set $\mathbf{S}_\ell$, when using the same learning procedure. This result suggests that the use of unlabeled data in Algorithm 1 does not degrade the performance of the initial supervised classifier.

Theorem 5.4. *Let $\mathbf{S}_\ell$ be a set of i.i.d. samples of size ℓ drawn from a distribution $\mathcal{D} = \mathcal{O}(f, \mathcal{D}_\mathbf{x}, \eta^{(0)})$ 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 $\mathbf{X}_u$ be an unlabeled set of size u drawn i.i.d. from $\mathcal{D}_\mathbf{x}$. Let L_m be the output of Algorithm 1 on input $\mathbf{S}_\ell$ and $\mathbf{X}_u$, and let $h_{\mathbf{w}^{(0)}}$ be the halfspace*

of the first iteration obtained from the empirical distribution $\mathbf{S}^{(0)} = \mathbf{S}_\ell$*, there is a high probability that:*

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

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

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\text{Relu}(-y\langle\mathbf{w}^{(0)},\mathbf{x}\rangle)] - \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\text{Relu}(-y\langle\mathbf{w}^{*},\mathbf{x}\rangle)] \leq \frac{3\max_{\mathbf{x}\in\mathbf{S}_\ell}\|\mathbf{x}\|}{2\sqrt{P^{(0)}}}$$

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

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

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

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.

data set	d	-1	$+1$	$\ell+u$	test
one-two	64	177	182	251	108
banknote	4	762	610	919	453
odd-even	64	906	891	1257	540
pc-mac	3868	982	963	1361	584
baseball-hockey	5724	994	999	1395	598
religion-atheism	7829	1796	628	1696	728
spambase	57	2788	1813	3082	1519
weather	17	43993	12427	37801	18619
delicious2	500	9610	6495	12920	3185
mediamill2	120	15969	27938	30993	12914

Table 1: data set statistics, -1 and $+1$ refer to the size of negative and positive class respectively, and test is the size of test set.

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 $^\downarrow$ indicates that performance is significantly worse than the best

[1]Implementation available for research purpose at this address: https://github.com/Lies0zeta/Self-Training-of-Halfspaces.

Dataset	ℓ	SVM	LTF	LP	GM	ERLR	L_m
one-two	10	$61.38 \pm 13.71^{\downarrow}$	$70.87 \pm 13.24^{\downarrow}$	$48.61 \pm 3.98^{\downarrow}$	75.09 ± 1.30	$53.65 \pm 10.65^{\downarrow}$	$\mathbf{77.77 \pm 1.75}$
	50	$\mathbf{92.77 \pm 3.05}$	$88.00 \pm 3.24^{\downarrow}$	$49.35 \pm 4.20^{\downarrow}$	$84.67 \pm 4.98^{\downarrow}$	$75.78 \pm 8.74^{\downarrow}$	91.34 ± 3.21
	100	$\mathbf{96.15 \pm 1.38}$	$92.50 \pm 1.43^{\downarrow}$	$67.82 \pm 12.99^{\downarrow}$	$86.52 \pm 3.26^{\downarrow}$	$79.25 \pm 6.87^{\downarrow}$	94.62 ± 2.46
banknote	10	$57.50 \pm 7.21^{\downarrow}$	$69.40 \pm 5.53^{\downarrow}$	$55.98 \pm 2.00^{\downarrow}$	$69.04 \pm 4.60^{\downarrow}$	$56.71 \pm 4.53^{\downarrow}$	$\mathbf{77.24 \pm 3.81}$
	50	$61.67 \pm 4.86^{\downarrow}$	$82.31 \pm 2.13^{\downarrow}$	$56.28 \pm 1.89^{\downarrow}$	$75.48 \pm 5.30^{\downarrow}$	$65.95 \pm 2.01^{\downarrow}$	$\mathbf{85.64 \pm 5.36}$
	100	$71.65 \pm 6.24^{\downarrow}$	89.38 ± 3.24	$57.20 \pm 2.19^{\downarrow}$	$77.56 \pm 4.34^{\downarrow}$	$70.95 \pm 3.24^{\downarrow}$	$\mathbf{90.82 \pm 3.31}$
odd-even	10	$53.45 \pm 4.80^{\downarrow}$	$58.20 \pm 4.71^{\downarrow}$	$50.37 \pm 1.95^{\downarrow}$	60.69 ± 7.48	$50.40 \pm 2.21^{\downarrow}$	$\mathbf{63.21 \pm 7.51}$
	50	$64.75 \pm 5.65^{\downarrow}$	$76.84 \pm 2.99^{\downarrow}$	$50.37 \pm 1.95^{\downarrow}$	$62.67 \pm 5.82^{\downarrow}$	$53.17 \pm 4.80^{\downarrow}$	$\mathbf{80.61 \pm 3.10}$
	100	$75.89 \pm 6.25^{\downarrow}$	$77.68 \pm 4.56^{\downarrow}$	$53.37 \pm 1.95^{\downarrow}$	$64.25 \pm 8.18^{\downarrow}$	$59.23 \pm 6.28^{\downarrow}$	$\mathbf{84.58 \pm 2.12}$
pc-mac	10	$51.00 \pm 3.22^{\downarrow}$	$54.92 \pm 2.00^{\downarrow}$	$50.93 \pm 1.59^{\downarrow}$	$54.76 \pm 3.42^{\downarrow}$	$50.14 \pm 2.06^{\downarrow}$	$\mathbf{57.75 \pm 3.19}$
	50	$58.85 \pm 5.09^{\downarrow}$	$61.78 \pm 2.86^{\downarrow}$	$50.83 \pm 2.08^{\downarrow}$	$58.78 \pm 4.31^{\downarrow}$	$49.71 \pm 1.99^{\downarrow}$	$\mathbf{64.31 \pm 3.55}$
	100	$64.57 \pm 4.42^{\downarrow}$	67.98 ± 2.37	$50.76 \pm 2.26^{\downarrow}$	$62.49 \pm 1.88^{\downarrow}$	$50.36 \pm 2.19^{\downarrow}$	$\mathbf{68.15 \pm 5.66}$
baseball-hockey	10	$51.57 \pm 2.98^{\downarrow}$	$55.41 \pm 3.16^{\downarrow}$	$\mathbf{56.53 \pm 5.18}$	$49.86 \pm 1.77^{\downarrow}$	$49.88 \pm 1.89^{\downarrow}$	56.47 ± 5.50
	50	$58.66 \pm 6.90^{\downarrow}$	69.29 ± 4.32	$50.11 \pm 1.84^{\downarrow}$	$66.76 \pm 5.40^{\downarrow}$	$50.16 \pm 1.90^{\downarrow}$	$\mathbf{72.85 \pm 6.52}$
	100	$68.40 \pm 4.65^{\downarrow}$	$76.25 \pm 2.41^{\downarrow}$	$49.97 \pm 1.82^{\downarrow}$	$71.12 \pm 5.06^{\downarrow}$	$50.35 \pm 1.89^{\downarrow}$	$\mathbf{79.48 \pm 4.36}$
religion-atheism	10	67.30 ± 6.95	$57.30 \pm 4.89^{\downarrow}$	67.59 ± 6.36	$60.67 \pm 16.37^{\downarrow}$	$\mathbf{71.95 \pm 5.03}$	$64.25 \pm 7.24^{\downarrow}$
	50	$\mathbf{74.61 \pm 1.62}$	$71.79 \pm 1.98^{\downarrow}$	$67.43 \pm 6.05^{\downarrow}$	69.16 ± 7.88	74.16 ± 1.88	72.47 ± 2.00
	100	$\mathbf{74.66 \pm 1.59}$	73.67 ± 1.76	$62.84 \pm 19.33^{\downarrow}$	$70.45 \pm 4.39^{\downarrow}$	73.21 ± 1.75	73.77 ± 1.82
spambase	10	$61.20 \pm 5.15^{\downarrow}$	$57.80 \pm 5.29^{\downarrow}$	$60.82 \pm 0.84^{\downarrow}$	$\mathbf{74.41 \pm 6.64}$	$53.38 \pm 11.23^{\downarrow}$	$68.92 \pm 5.83^{\downarrow}$
	50	$62.59 \pm 9.42^{\downarrow}$	74.99 ± 6.04	$61.15 \pm 0.86^{\downarrow}$	$\mathbf{78.25 \pm 2.62}$	$53.63 \pm 9.86^{\downarrow}$	76.13 ± 3.08
	100	$69.43 \pm 10.19^{\downarrow}$	80.07 ± 4.08	$61.24 \pm 10.26^{\downarrow}$	$79.08 \pm 2.83^{\downarrow}$	$58.21 \pm 6.34^{\downarrow}$	$\mathbf{81.93 \pm 2.46}$
weather	10	74.85 ± 0.51	$68.09 \pm 1.73^{\downarrow}$	$\mathbf{75.49 \pm 0.34}$	75.02 ± 2.79	$40.35 \pm 17.29^{\downarrow}$	75.08 ± 4.18
	50	75.79 ± 0.28	75.30 ± 3.85	$\mathbf{77.99 \pm 0.31}$	75.68 ± 2.78	$41.55 \pm 27.39^{\downarrow}$	75.34 ± 3.80
	100	$\mathbf{77.99 \pm 0.25}$	76.27 ± 3.64	$\mathbf{77.99 \pm 0.25}$	74.92 ± 1.92	$46.00 \pm 24.87^{\downarrow}$	77.28 ± 2.99
delicious2	10	51.83 ± 9.88	$50.59 \pm 2.65^{\downarrow}$	$\mathbf{60.02 \pm 0.61}$	$49.41 \pm 3.83^{\downarrow}$	$51.83 \pm 10.42^{\downarrow}$	$51.08 \pm 1.80^{\downarrow}$
	50	$\mathbf{60.04 \pm 0.62}$	$54.78 \pm 2.57^{\downarrow}$	60.00 ± 0.59	$48.35 \pm 1.31^{\downarrow}$	$53.48 \pm 8.66^{\downarrow}$	$55.37 \pm 3.33^{\downarrow}$
	100	58.88 ± 3.70	$56.04 \pm 1.83^{\downarrow}$	$\mathbf{59.87 \pm 0.67}$	$48.92 \pm 0.94^{\downarrow}$	$54.43 \pm 7.27^{\downarrow}$	$56.54 \pm 1.87^{\downarrow}$
mediamill2	10	$62.54 \pm 2.62^{\downarrow}$	$60.98 \pm 6.85^{\downarrow}$	$36.35 \pm 0.15^{\downarrow}$	63.92 ± 1.71	$47.24 \pm 14.08^{\downarrow}$	$\mathbf{64.31 \pm 3.14}$
	50	$63.64 \pm 0.15^{\downarrow}$	$60.88 \pm 7.45^{\downarrow}$	$36.36 \pm 0.15^{\downarrow}$	$\mathbf{65.98 \pm 3.32}$	$58.58 \pm 11.88^{\downarrow}$	65.41 ± 4.83
	100	$63.64 \pm 0.15^{\downarrow}$	64.26 ± 4.79	$36.37 \pm 0.15^{\downarrow}$	67.34 ± 0.73	$63.64 \pm 0.16^{\downarrow}$	$\mathbf{67.80 \pm 2.21}$

Table 2: Mean and standard deviations of accuracy on test sets over the 20 trials for each data set. The best and the second-best performance are respectively in bold and underlined. $^{\downarrow}$ indicates statistically significantly worse performance than the best result, according to a Wilcoxon rank-sum test [Wolfe, 2012] with ($p < 0.01$).

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

[Amini *et al.*, 2009] Massih R. Amini, Nicolas Usunier, and François Laviolette. A transductive bound for the voted classifier with an application to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 65–72, 2009.

[Balcan and Blum, 2006] Maria-Florina Balcan and Avrim Blum. An augmented PAC model for semi-supervised learning. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, pages 396–419. The MIT Press, 2006.

[Ben-David *et al.*, 2008] Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Proceedings of the The 21st Annual Conference on Learning Theory (COLT-08)*, pages 33–44, 2008.

[Bhatia *et al.*, 2015] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. Sparse local embeddings for extreme multi-label classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[Boole, 2015] G. Boole. *An Investigation of the Laws of Thought: On Which Are Founded the Mathematical Theories of Logic and Probabilities*. Creative Media Partners, LLC, 2015.

[Castelli and Cover, 1995] Vittorio Castelli and Thomas M. Cover. On the exponential value of labeled samples. *Pattern Recognit. Lett.*, 16(1):105–111, 1995.

[Chapelle *et al.*, 2006] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. The MIT Press, 2006.

[Chawla and Karakoulas, 2011] Nitesh Chawla and Grigoris Karakoulas. Learning from labeled and unlabeled data: An empirical study across techniques and domains. *Journal of Artificial Intelligence Research - JAIR*, 23, 09 2011.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995.

[Diakonikolas and Kane, 2020] Ilias Diakonikolas and Daniel M. Kane. Hardness of learning halfspaces with massart noise. *CoRR*, abs/2012.09720, 2020.

[Diakonikolas *et al.*, 2019] Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent pac learning of halfspaces with massart noise. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 4749–4760. Curran Associates, Inc., 2019.

[Diakonikolas *et al.*, 2021] Ilias Diakonikolas, Russell Impagliazzo, Daniel M. Kane, Rex Lei, Jessica Sorrell, and Christos Tzamos. Boosting in the presence of massart noise. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1585–1644. PMLR, 15–19 Aug 2021.

[Dua and Graff, 2019] Dheeru Dua and Casey Graff. UCI machine learning repository. https://archive.ics.uci.edu/ml, 2019. Accessed: 2023-06-27.

[Duchi, 2016] John C. Duchi. Introductory lectures on stochastic convex optimization. *Park City Mathematics Series, Graduate Summer School Lectures*, 2016.

[Dvoretzky *et al.*, 1956] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics*, 27(3):642 – 669, 1956.

[Fan *et al.*, 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[Frei *et al.*, 2021] Spencer Frei, Difan Zou, Zixiang Chen, and Quanquan Gu. Self-training converts weak learners to strong learners in mixture models. *CoRR*, abs/2106.13805, 2021.

[Grandvalet and Bengio, 2005] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17, pages 529–536. MIT Press, 2005.

[Joachims, 1999] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 200–209, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[Johnson *et al.*, 2020] David S. Johnson, Lee Breslau, Ilias Diakonikolas, Nick Duffield, Yu Gu, MohammadTaghi Hajiaghayi, Howard J. Karloff, Mauricio G. C. Resende, and Subhabrata Sen. Near-optimal disjoint-path facility location through set cover by pairs. *Oper. Res.*, 68(3):896–926, 2020.

[Krijthe, 2017] Jesse H. Krijthe. Rssl: Semi-supervised learning in r. In Bertrand Kerautret, Miguel Colom, and Pascal Monasse, editors, *Reproducible Research in Pattern Recognition*, pages 104–115, Cham, 2017. Springer International Publishing.

[Li and Zhou, 2011] Yu-Feng Li and Zhi-Hua Zhou. Towards Making Unlabeled Data Never Hurt. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1081–1088, 2011.

[Massart and Nédélec, 2006] Pascal Massart and Elodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 10 2006.

[Maximov *et al.*, 2018] Yury Maximov, Massih-Reza Amini, and Zaïd Harchaoui. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm. *Journal of Artificial Intelligence Research*, 61:761–786, 2018.

[Montasser *et al.*, 2020] Omar Montasser, Surbhi Goel, Ilias Diakonikolas, and Nathan Srebro. Efficiently learning adversarially robust halfspaces with noise. In Hal Daumé III

and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7010–7021. PMLR, 13–18 Jul 2020.

[Niyogi, 2013] Partha Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research*, 14(1):1229–1250, 2013.

[Rigollet, 2007] Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8:1369–1392, 2007.

[Singh *et al.*, 2009] Aarti Singh, Robert Nowak, and Jerry Zhu. Unlabeled data: Now it helps, now it doesn't. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009.

[Wei *et al.*, 2020] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *CoRR*, abs/2010.03622, 2020.

[Wei *et al.*, 2021] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[Wolfe, 2012] Douglas A. Wolfe. *Nonparametrics: Statistical Methods Based on Ranks and Its Impact on the Field of Nonparametric Statistics*, pages 1101–1110. Springer US, Boston, MA, 2012.

[Zhu and Ghahramani, 2002] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, CMU-CALD-02-107, Carnegie Mellon University, 2002.

[Zhu, 2005] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.