A Theoretical Characterization of Semi-supervised Learning with Self-training for Gaussian Mixture Models

Samet Oymak
Electrical and Computer Engineering
University of California, Riverside
oymak@ece.ucr.edu

Talha Cihad Gulcu
tcgulcu@gmail.com

Abstract

Self-training is a classical approach in semi-supervised learning which is successfully applied to a variety of machine learning problems. Self-training algorithms generate pseudo-labels for the unlabeled examples and progressively refine these pseudo-labels which hopefully coincides with the actual labels. This work provides theoretical insights into self-training algorithms with a focus on linear classifiers. First, we provide a sample complexity analysis for Gaussian mixture models with two components. This is established by sharp non-asymptotic characterization of the self-training iterations which captures the evolution of the model accuracy in terms of a fixed-point iteration. Our analysis reveals the provable benefits of rejecting samples with low confidence and demonstrates how self-training iterations can gracefully improve the model accuracy. Secondly, we study a generalized GMM where the component means follow a distribution. We demonstrate that ridge regularization and class margin (i.e. separation between the component means) is crucial for the success and lack of regularization may prevent self-training from identifying the core features in the data.

1 Introduction

The recent widespread success of deep neural networks rely on the presence of large labeled datasets to a significant extent. Unfortunately, such good-quality datasets may not be readily available for variety of practical applications. Indeed, a grand challenge in expanding machine learning to new domains is the cost of obtaining good quality labels. This is especially true for privacy and safety sensitive tasks that are abundant in critical domains such as healthcare and defense. On the other hand, unlabeled data can be relatively cheap to obtain and may be more abundant. This necessitates semi/unsupervised learning algorithms that can go beyond supervised learning and efficiently utilize unlabeled data.

Semi-supervised learning (SSL) techniques aim to reduce the dependence on the labeled data by making use of unlabeled data. A large number of approaches for SSL involve an extra loss term accounting for unlabeled data which is expected to help the model better generalize to unseen data. Self-training, consistency training and entropy minimization are among some of the core methods (discussed in Section 1.1 in more detail) used for the purpose of SSL. Despite its popularity and practical success, we still don't have a fundamental understanding of when and why self-training algorithms work. For instance, self-training algorithms gradually utilize unlabeled data by first incorporating the most reliable pseudo-labels. Are there setups where rejecting unreliable examples provably help? Similarly, generating and overfitting to incorrect pseudo-labels is a natural concern in SSL. On the other hand, recent empirical and theory literature suggests that, for supervised learning, interpolating to training data performs surprisingly well even when the model perfectly interpolates and achieves zero training loss (Belkin et al., 2019; Hastie et al., 2019; Zhang et al., 2016). How crucial is regularization when it comes to learning with unlabeled data? Finally, for which datasets, self-training finds useful models that generalize better and what structural assumptions on the data are key to success?

Contributions. This paper takes a step towards addressing the aforementioned questions by studying algorithmic fundamentals of SSL. Specifically, we make

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

the following contributions.

• **Self-training for Gaussian Mixture Models.** One way to understand the algorithmic performance is by focusing on fundamental dataset models such as Gaussian mixtures and conducting a careful analysis capturing exact algorithmic performance. We study the problem of learning a linear classifier with self-training under a Gaussian mixture model (GMM). We precisely calculate the distributional properties of self-training iterations. Specifically we capture the evolution of the correlation between the optimal classifier and the self-training output in a non-asymptotic fashion. This reveals asymptotic and non-asymptotic formulae exactly characterizing the performance of self-training with linear models. We present associated numerical experiments demonstrating the classification performances under various scenarios which also reveals the provable benefits of rejecting weak examples.

• **Algorithmic Insights: The Role of Distribution and Regularization.** Next, we explore the importance of distributional properties by considering a more general family of mixture models where the means of mixture components are continuously distributed. This reveals that as long as there is a margin (i.e. separation) between the means, unlabeled data improves the performance, however without margin, un-regularized algorithm provably gets stuck under least-squares loss. We then show how ridge regularization and early stopping can mitigate this issue by encouraging self-training to pick up the principal eigendirections in the data in a similar fashion to power iteration. We also discuss similar benefits of regularization for logistic regression.

1.1 Prior Art

The benefits of using unlabeled data for learning models is subject of a rich literature since 70s which consider a variety of settings such as generative models (Castelli and Cover, 1995; Nigam et al., 2000), semi-supervised support vector machines (Vapnik, 1998; Joachims, 1999), graph-based models (Blum and Chawla, 2001; Belkin et al., 2006; Zhu et al., 2003), or co-training (Blum and Mitchell, 1998) and multiview models(Sindhwani et al., 2005). The relative value of labeled and unlabeled samples in a detection-estimation theoretical framework is examined in (Castelli and Cover, 1996). A line of work is related to how the presence of unlabeled data be useful to limit Radamacher complexity(Bartlett and Mendelson, 2002). For example, the compatibility of a target function with respect to a data distribution is considered by (Balcan and Blum, 2010), where the authors illustrate how enough unlabeled data can be useful to reduce the size of the search space. It is demonstrated by several papers (Oneto et al., 2011, 2015, 2016) that the additional unlabeled data can be used to improve the tightness of the Radamacher complexity (RC) based bounds. A sharper generalization error bound for multi-class learning with the help of additional unlabeled data is presented by (Li et al., 2019), along with an efficient multi-class classification algorithm using local Radamacher complexity and unlabeled samples. Apart from that, semi supervised learning (SSL) is a versatile approach for training models without using a large amount of data. SSL algorithms can achieve performance improvement with low cost, and there are a large number of SSL methods (Miyato et al., 2018; Sajjadi et al., 2016b; Laine and Aila, 2016; Tarvainen and Valpola, 2017; Berthelot et al., 2019; Xie et al., 2019; Berthelot et al., 2020; Lee, 2013; Sajjadi et al., 2016a) available in the literature.

A large portion of SSL methods relies on generating an artificial label for unlabeled data and training the model to predict those artificial labels when the unlabeled data is used as the input. Pseudo-labeling (Lee, 2013) is one of such methods where the class prediction of the model is used for training purposes. Consistency regularization is also an important component of many SSL algorithms. Consistency regularization (Tarvainen and Valpola, 2017; Sajjadi et al., 2016b; Laine and Aila, 2016) is based on the approach that the model is supposed to generate similar outputs when perturbed version of the same data is applied as the input. Adversarial transformation is used by (Miyato et al., 2018) in the loss function of consistency training, and cross-entropy loss instead of squared loss function appears in the works (Miyato et al., 2018; Xie et al., 2019). There are also hybrid algorithms combining diverse mechanisms. For example, Fix-Match (Sohn et al., 2020) combines pseudo-labeling and consistency training to generate artificial labels. Mix-Match (Berthelot et al., 2019), ReMixMatch (Berthelot et al., 2020), unsupervised data augmentation (Xie et al., 2019) are among other composite approaches. Self training in the setting of domain adaptation is covered by the papers (Long et al., 2013; Inoue et al., 2018). Class balance (Zou et al., 2018) and confidence regularization (Zou et al., 2019) for self-training are among other lines of works. Gradual domain adaptation in regularized models is analyzed by (Kumar et al., 2020). The papers (Carmon et al., 2019; Zhai et al., 2019; Najafi et al., 2019; Stanforth et al., 2019) show theoretically and empirically how semi-supervised learning procedure can achieve high robust accuracy and improve adversarial robustness.

2 Problem Setup

Let us first fix the notation. Given an event E, let $1(E)$ be the indicator function of E which is 1 if E

happens and 0 otherwise. We use $X \mid E$ to denote the conditional random variable induced by a random variable X given an event E. We will refer the vectors with unit Euclidean norm as unit norm. Given two vectors $\boldsymbol{a}, \boldsymbol{b}$, their correlation is denoted by $\rho(\boldsymbol{a}, \boldsymbol{b}) = \frac{\langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\|\boldsymbol{a}\|_{\ell_2} \|\boldsymbol{b}\|_{\ell_2}}$. Related to correlation, we define *co-tangent of the angle between two vectors* to be

$$\cot(\boldsymbol{a}, \boldsymbol{b}) = \frac{\rho(\boldsymbol{a}, \boldsymbol{b})}{\sqrt{1 - \rho(\boldsymbol{a}, \boldsymbol{b})^2}},$$

which will be useful for cleaner notation. As $\cot(\boldsymbol{a}, \boldsymbol{b}) \to \infty$, the two vectors become perfectly correlated i.e. $\rho(\boldsymbol{a}, \boldsymbol{b}) \to 1$. Let $Q(\cdot)$ be the tail of a standard normal variable and Q_X be the tail of the distribution of a random variable X. $\xrightarrow{\mathbb{P}}$ denotes convergence in probability. $a \wedge b$ and $a \vee b$ returns minimum and maximum of two scalars. Finally, $(a)_+$ returns $a \vee 0$.

Let $\mathcal{S} = (y_i, \boldsymbol{x}_i)_{i=1}^n \in \{-1, 1\} \times \mathbb{R}^p$ be independent and identically distributed (i.i.d) labeled sampled distributed as $\mathcal{D} = \mathcal{D}_{y|\boldsymbol{x}} \times \mathcal{D}_{\boldsymbol{x}}$ and let $\mathcal{U} = (\boldsymbol{x}_i)_{i=n+1}^{n+u}$ be i.i.d. unlabeled samples distributed with the marginal distribution $\mathcal{D}_{\boldsymbol{x}}$. Let $f : \mathbb{R}^p \to \mathbb{R}$ be a prediction function (e.g. a neural network) and let $\hat{y}_f(\boldsymbol{x})$ be the hard-label $(-1, 1)$ assigned to $f(\boldsymbol{x})$ defined as

$$\hat{y}_f(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad f(\boldsymbol{x}) \geq 0 \\ -1 & \text{else} \end{cases}.$$

The standard self-training approach is sufficiently general to operate on a generic algorithm. The algorithm can self-train by using its own labels $\hat{y}_f(\boldsymbol{x})$ which are also known as pseudo-labels. Self-training is often gradual, it first utilizes examples where predictions are confident and only later moves to examples which are less certain. Thus, it is a common strategy to reject weak pseudo-labels and use the more confident ones. Given a loss function ℓ, function class $\mathcal{F}$, and acceptance threshold $\Gamma \geq 0$, self-training with pseudo-labels typically solves an empirical risk minimization problem of the form

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(\boldsymbol{x}_i))}_{\mathcal{L}_{\mathcal{S}}(f)} + \lambda \underbrace{\frac{1}{u} \sum_{i=n+1}^{n+u} 1(|f(\boldsymbol{x}_i)| \geq \Gamma) \ell(\hat{y}_f(\boldsymbol{x}_i), f(\boldsymbol{x}_i))}_{\tilde{\mathcal{L}}_{\mathcal{U}}(f)}. \quad (2.1)$$

where $\mathcal{L}_{\mathcal{S}}$ and $\tilde{\mathcal{L}}_{\mathcal{U}}$ are the supervised and unsupervised empirical risks respectively. Let us also introduce our iterative learning setup. Suppose we have an algorithm $\mathcal{A}$ that takes a labeled dataset and builds a prediction model f. An obvious example for $\mathcal{A}$ is (2.1). Denote the initial model by f_0 and let $\Gamma \geq 0$ be the acceptance threshold. Given a stopping time T, the self-training algorithm we consider operates in two steps.

• **Step 1: Create Pseudo-labels:** From $\mathcal{U}$ and current iterate f_τ, determine a subset $\mathcal{U}_\tau = (\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{y}}_i)$ where $\tilde{\boldsymbol{x}}_i \in \mathcal{U}$ are the acceptable inputs that satisfy $|f_\tau(\tilde{\boldsymbol{x}}_i)| \geq \Gamma$ and $\tilde{\boldsymbol{y}}_i$ are the pseudo-labels $\tilde{\boldsymbol{y}}_i = \hat{y}_{f_\tau}(\tilde{\boldsymbol{x}}_i)$.

• **Step 2: Refine the model:** Obtain the new classifier via $f_{\tau+1} = \mathcal{A}(\mathcal{S}, \mathcal{U}_\tau)$. If $\tau < T$, go to Step 1.

We remark that $\mathcal{A}$ can treat the datasets $\mathcal{S}$ and $\mathcal{U}_\tau$ differently in a similar fashion to (2.1), e.g. by weighting labeled $\mathcal{S}$ higher than pseudo-labeled $\mathcal{U}$. In our analysis of iterative algorithms in Sections 3 and 4, we consider a slightly different version where we only use the unlabeled data for refinement in Step 2. While our approach does extend to jointly learning over $(\mathcal{S}, \mathcal{U})$, as we shall see, learning only over $\mathcal{U}$ results in cleaner and more insightful bounds.

Before moving to our main results, we remark that proof of all of the technical results are deferred to the extended manuscript (Oymak and Gulcu, 2020).

3 Understanding Self-Training for Mixtures of Two Gaussians

We start with a definition of the distribution we will study.

Definition 3.1 *(Binary Gaussian Mixture Model (GMM)) The distribution $(\boldsymbol{x}, y) \sim \mathcal{D}$ is given as follows. Fix a unit vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and scalar $\sigma \geq 0$. Let y be a Rademacher random variable ($\mathbb{P}(y = 1) = \mathbb{P}(y = -1) = 1/2$) and $\boldsymbol{x}|y \sim \mathcal{N}(y\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_p)$.*

Note that the component mean $\boldsymbol{\mu}$ is also the optimal linear classifier. If we have labeled data $\mathcal{S} = (\boldsymbol{x}_i, y_i)_{i=1}^n$, $\boldsymbol{\mu}$ can be estimated via the averaging estimator

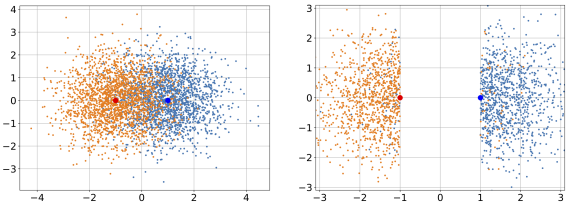
$$\boldsymbol{\beta}_{\text{init}} = \frac{1}{n} \sum_{i=1}^{n} y_i \boldsymbol{x}_i. \quad (3.1)$$

This estimator also coincide with the ridge regularized least-squares (e.g. $\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_2}^2$) when the regularization parameter $\lambda \to \infty$. Perhaps surprisingly, this estimator is known to be the Bayes optimal classifier for GMM if we have access to the labeled data alone (Mignacco et al., 2020; Lelarge and Miolane, 2019). This motivates us to investigate the analytical properties of the averaging estimator by adapting it to self-training as explained earlier. Given an initial supervised model $\boldsymbol{\beta}_{\text{init}}$ (such as (3.1)) and the unlabeled dataset $\mathcal{U} = (\boldsymbol{x}_i)_{i=n+1}^{n+u}$ sampled from GMM,

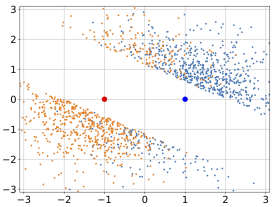
we consider the pseudo-label estimator

$$\hat{\boldsymbol{\beta}} = \text{self-train}(\boldsymbol{\beta}_{\text{init}}, \mathcal{U}) \quad \text{where} \tag{3.2}$$
$$\text{self-train}(\boldsymbol{\beta}_{\text{init}}, \mathcal{U}) = \frac{\sum_{i=1}^{u} 1(|\bar{\boldsymbol{\beta}}_{\text{init}}^T \boldsymbol{x}_i| \geq \Gamma)\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \boldsymbol{x}_i)\boldsymbol{x}_i}{\sum_{i=1}^{u} 1(|\bar{\boldsymbol{\beta}}_{\text{init}}^T \boldsymbol{x}_i| \geq \Gamma)}$$

Here $\Gamma \geq 0$ is the acceptance threshold eliminating low-confidence predictions and $\bar{\boldsymbol{\beta}}_{\text{init}} = \boldsymbol{\beta}_{\text{init}}/\|\boldsymbol{\beta}_{\text{init}}\|_{\ell_2}$ is the normalized initial model so that the choice of Γ can be invariant to the norm of $\boldsymbol{\beta}_{\text{init}}$. Acceptance threshold is commonly used in practical semi-supervised learning approaches (Xie et al., 2019; McClosky et al., 2006; Yarowsky, 1995). The impact of acceptance threshold is illustrated in Figure 1 where points are projected on two dimensions. Here the mixture center $\boldsymbol{\mu}$ is the $[1\ 0\ 0\ \ldots\ 0]$ direction. When $\Gamma = 0$, we accept all points which corresponds to a Binary GMM distribution. When Γ is non-zero, the conditional distribution of the accepted examples depend on the quality of the initial model $\boldsymbol{\beta}_{\text{init}}$. Figure 1b and 1c chooses $\Gamma = 1$ for different $\boldsymbol{\beta}_{\text{init}}$. In Figure 1b, $\boldsymbol{\beta}_{\text{init}}$ is aligned with $\boldsymbol{\mu}$ (correlation is 1) which results in a clean separation between the two classes (the red and blue dots) while rejecting 50% of the samples that lie between the mixture centers ± 1. In Figure 1c, correlation coefficient between $\boldsymbol{\beta}_{\text{init}}$ and $\boldsymbol{\mu}$ is 1/2 and $\boldsymbol{\beta}_{\text{init}}$ has a higher classification error. As a result, the two classes are not as cleanly separated despite using rejection.



(a) $\Gamma = 0$ (b) $\Gamma = 1$, $\rho(\boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu}) = 1$



(c) $\Gamma = 1$, $\rho(\boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu}) = 0.5$

Figure 1: Visualization of a Binary GMM with noise variance $\sigma^2 = 1$. Sample size is 4000. The large dots at -1 and 1 are the mixture centers $\pm\boldsymbol{\mu} = [\pm 1, 0]$. Acceptance threshold Γ removes the examples with low-correlation to the initial model $\boldsymbol{\beta}_{\text{init}}$.

Our main theorem provides a sharp non-asymptotic bound for the pseudo-label estimator (3.2). Given an acceptance threshold $\Gamma \geq 0$, we define the normalized thresholds $\bar{\Gamma}_- = \frac{\Gamma+\alpha}{\sigma}$ and $\bar{\Gamma}_+ = \frac{\Gamma-\alpha}{\sigma}$ and the quantities

$$\rho = Q(\bar{\Gamma}_+) + Q(\bar{\Gamma}_-) \quad , \quad \nu = Q(\bar{\Gamma}_-)/\rho \quad \text{and}$$
$$\Lambda = \frac{1}{\sqrt{2\pi}\rho}(e^{-\bar{\Gamma}_+^2/2} + e^{-\bar{\Gamma}_-^2/2}). \tag{3.3}$$

Interpretation: Here the ρ term captures the fraction of the examples that overcome the acceptance threshold and the ν term captures the fraction of the incorrect pseudo-labels within these examples. In terms of these, our theorem below perfectly captures the self-training evolution for GMMs with two components.

Theorem 3.2 *(Non-asymptotic Bound for GMM) Let $\boldsymbol{\mu} \in \mathbb{R}^p$ be a unit norm vector from Def. 3.1 and suppose $\boldsymbol{\beta}_{init} \in \mathbb{R}^p$ has correlation $\rho(\boldsymbol{\beta}_{init}, \boldsymbol{\mu}) = \alpha > 0$. Draw u i.i.d. unlabeled samples $(\boldsymbol{x}_i)_{i=1}^u$ from GMM. Let $\hat{\boldsymbol{\beta}}$ be defined as $\hat{\boldsymbol{\beta}} = \text{self-train}(\boldsymbol{\beta}_{init}, (\boldsymbol{x}_i)_{i=1}^u)$. Let $p \geq 3$ and fix resolution $1/2 > \varepsilon > 0$. There exists a constant $c > 0$ such that with probability at least $1 - 10e^{-c\varepsilon^2((p-3)\wedge\rho u)}$, we have*

$$\frac{1 + \sigma\alpha\Lambda - 2\nu}{\sigma\sqrt{(1-\alpha^2)\Lambda^2 + (p-2)/u\rho}}(1-\varepsilon) \leq cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu}) \leq \frac{1 + \sigma\alpha\Lambda - 2\nu}{\sigma\sqrt{(1-\alpha^2)\Lambda^2 + (p-3)/u\rho}}(1+\varepsilon). \tag{3.4}$$

Thus, fixing $\bar{u} = u/p$ and letting $p \to \infty$, we have that

$$\lim_{p\to\infty} cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu}) \xrightarrow{\mathbb{P}} \frac{1 + \sigma\alpha\Lambda - 2\nu}{\sigma\sqrt{(1-\alpha^2)\Lambda^2 + 1/\bar{u}\rho}}. \tag{3.5}$$

Theorem 3.2 shows that pseudo-label optimization as defined by (3.2) can be useful to obtain a higher correlation and thus can improve the quality of the initial direction $\boldsymbol{\beta}_{\text{init}}$. To find the optimal acceptance threshold Γ that maximizes the correlation $\cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu})$, one can differentiate (3.5) with respect to Γ and equate the result to 0. This does not give a closed form solution, but it is possible to find the optimal Γ numerically.

Let f denote the transformation that is applied to $\rho(\boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu})$ as a result of pseudo-label optimization. Theorem 3.2 provides matching upper and lower bounds for the evolution of the co-tangent. Specifically, using the relation between correlation and co-tangent, as $p \to \infty$, we have that

$$\cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu}) = F_{\bar{u}}(\cot(\boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu})) \quad \text{where} \tag{3.6}$$
$$F_{\bar{u}}(x) = \frac{1 + \sigma\frac{\Lambda x}{\sqrt{1+x^2}} - 2\nu}{\sigma\sqrt{\frac{\Lambda^2}{1+x^2} + \frac{1}{\bar{u}\rho}}}.$$

We remark that (Castelli and Cover, 1996; Lelarge and Miolane, 2019) studies mixture models and provides information theoretical bounds. Our bound complements these works by characterizing the performance

of self-training which is a widely-used practical algorithm. We also characterize the benefit of using the acceptance threshold Γ which is again a critical heuristic for the success of self-training. We suspect that one can analyze self-training performance for more general distributions and other base classifiers, going beyond the averaging estimator, by using tools from high-dimensional statistics and random matrix theory such as Gaussian min-max Theorem (Oymak et al., 2013; Thrampoulidis et al., 2015; Stojnic, 2013) and approximate message passing (Donoho et al., 2009; Bayati and Montanari, 2011).

3.1 Iterative self-training

Theorem 3.2 also allows us to analyze self-training in an iterative fashion to show further improvement with more unlabeled data. Specifically, suppose we have n labeled samples $\mathcal{S} = (\boldsymbol{x}_i)_{i=1}^n$ and $\tau \times u$ unlabeled samples $\mathcal{U} = (\boldsymbol{x}_i)_{i=n+1}^{n+\tau u}$. We first create the initial supervised model via (3.1). Then, we split $\mathcal{U}$ into τ disjoint sub-datasets $(\mathcal{U}_i)_{i=1}^{\tau}$. Starting from $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{\text{init}}$ of (3.1), we iteratively apply self-training (3.2) to obtain

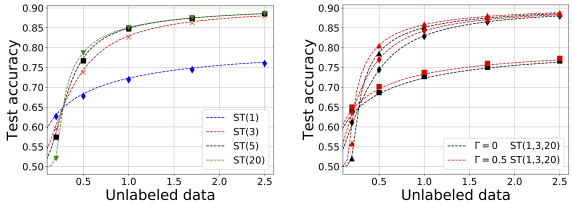
$$\boldsymbol{\beta}_i = \text{self-train}(\boldsymbol{\beta}_{i-1}, \mathcal{U}_i) \quad \text{for} \quad 1 \le i \le \tau. \tag{3.7}$$

The final model is then equal to $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_\tau$. Note that the asymptotic co-tangent of self-training with τ iterations will be given by $F^\tau(x)$ where x is the co-tangent of the initial supervised model. The following theorem establishes the asymptotic performance of this procedure.

Theorem 3.3 (Iterative self-training bound) *Set $\bar{n} = n/p$ and $\bar{u} = u/p$. Let $\mathcal{S} = (\boldsymbol{x}_i, y_i)_{i=1}^n$ and $\mathcal{U} = (\boldsymbol{x}_i)_{i=n+1}^{n+\tau u}$ be independent datasets with i.i.d. samples generated according to Binary GMM. Obtain the model $\hat{\boldsymbol{\beta}}$ via applying τ iterations of the iterative self-training* (3.7) *to the supervised model* (3.1)*. Recall the co-tangent evolution formula of* (3.6)*. We have that*

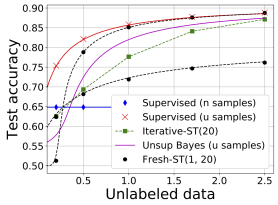
$$\lim_{p\to\infty} cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu}) \xrightarrow{\mathbb{P}} F_{\bar{u}}^\tau(\sqrt{\bar{n}}/\sigma). \tag{3.8}$$

Let us call this model Fresh-ST (ST for self-training) as each iteration requires fresh batch of unlabeled data. Figure 2a and Figure 2b illustrate the the test performance associated with this iterative approach. The parameters in these figures are as follows. We set labeled data amount to be $\bar{n} = 0.05$ and unlabeled data amount $\bar{u}$ is varied along the x axis. The noise level is $\sigma = 0.75$ and the input dimension is $p = 400$. The dashed lines are our formula (3.8). We see from Figure 2a that the test performance improves as the amount of unlabeled data increases (here $\Gamma = 0$). The self-training iterations also improve the test accuracy as long as the unlabeled



(a) The impact of self-training iterations on the model accuracy at $\Gamma = 0$.

(b) Comparing acceptance thresholds of $\Gamma = 0$ vs $\Gamma = 0.5$.



(c) Comparison of different baselines at $\Gamma = 0$.

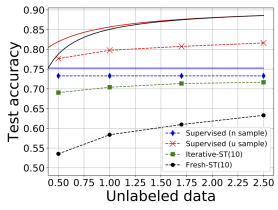
Figure 2: $p = 400$, $\bar{n} = n/p = 0.05$, $\sigma = 0.75$. x-axis is the unlabeled data amount $\bar{u} = u/p$. In Figures (a) and (b), ST(τ) refers to self-training repeated τ times with new batch of unlabeled data (same as Fresh-ST). Larger τ corresponds to the line with better accuracy. All lines are theoretical predictions except the Iterative-ST.

data amount is above the fixed point of the $F_{\bar{u}}$ function. In other words, we need $\bar{u}$ larger than a threshold u_* where u_* preserves the co-tangent of the initial supervised model i.e. $F_{u_*}(\cot(\boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu})) = \cot(\boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu})$. Clearly this threshold u_* depends on the initial supervised model (i.e. the amount of labeled training data) as well as the noise level σ. Figure 2b demonstrates that choosing a proper acceptance threshold Γ can improve the test performance over always choosing $\Gamma = 0$. We observe that benefit of optimizing Γ is more noticeable when there are fewer unlabeled data. Also optimizing Γ can shift the fixed point of the $F_{\bar{u}}$ function so that less unlabaled data is required for improvement.

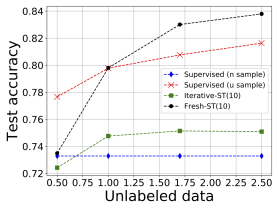
Figure 2c provides multiple baselines to compare our self-training bounds ($\Gamma = 0$) (blue, red, green curves). The blue curve is the performance of the initial model which only uses n labels. The red curve is the performance of a supervised model that uses u labeled samples. Note that, this curve is not necessarily an upper bound on the performance of the Fresh-ST however provides a natural reference. The magenta curve is the accuracy of the unsupervised Bayes optimal classifier using u input samples. Finally, the green line is the iterative self-training where we always use the same unlabeled dataset with u samples. Specifically, we apply the iterations $\boldsymbol{\beta}_{i+1} = \text{self-train}(\boldsymbol{\beta}_i, \mathcal{U})$ for $1 \le i \le \tau = 20$. Let us call this Iterative-ST. We see that, repetitively applying self-training on the same dataset improves the performance over applying it only once (i.e. green

line is above the lower dashed black line). On the other hand, we also see the positive effect of using fresh unlabeled data on the test performance from Figure 2c. Comparing the Fresh-ST with the empirical performance of Iterative-ST in Figure 2c shows that the test performance substantially benefits from resampling. For instance, only 3 iterations of resampling can be noticeably better than many iterations of Iterative-ST. Intuitively, this is due to the fact that repeated self-training on the same dataset can guide the optimization to a suboptimal fixed point of the self-training iteration. This is also known as the confirmation bias of pseudo-labeling (Arazo et al., 2019). In this example, fresh samples help get out of bad fixed points.

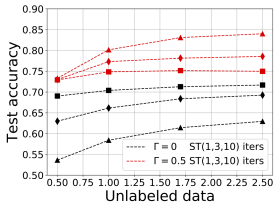
Logistic regression: We next compare our averaging-based self-training (3.2) to logistic regression. Given unlabeled data $\mathcal{U}$ and a linear classifier $\boldsymbol{\beta}_{\text{init}}$, we first obtain the dataset $\mathcal{U}'$ of acceptable inputs and associated pseudo-labels by thresholding $\boldsymbol{x}^T\boldsymbol{\beta}_{\text{init}}/\|\boldsymbol{\beta}_{\text{init}}\|_{\ell_2}$. We then solve logistic regression over $\mathcal{U}'$ to obtain a new linear classifier. The test performances of logistic-regression self-training are plotted in Figure 3. The labeled data fraction is $\bar{n} = 0.2$ and the unlabeled data amount varies along x-axis, as in the case of Figure 2. We set $\Gamma = 0$ in Figure 3a, and $\Gamma = 1/2$ in Figure 3b.



(a) Comparing logistic regression (dashed) and averaging (solid) at $\Gamma = 0$.



(b) Performance comparison of different approaches. $\Gamma = 0.5$.



(c) Logistic regression, comparison of $\Gamma = 0$ vs $\Gamma = 0.5$.

Figure 3: Experiments on logistic regression: $\bar{n} = n/p = 0.2$, $\sigma = 0.75$, $p = 400$. Fresh-ST (and ST in Fig. (c)) uses fresh batch of unlabeled data at each self-training iteration. In Fig. (c), (1,3,10) self-training iterations have markers Δ, $\diamond$, $\square$ respectively.

For both Figure 3a and Figure 3b, the black dashed line refers to Fresh-ST iterations, and green dashed line corresponds to self-training iterations with the same unlabeled data. Similarly, blue dashed line plots the test performance of supervised learning with n samples and red dashed line plots the performance of supervised learning with u samples for both figures. The dashed lines in Figure 3a are logistic regression based algorithms whereas solid lines display the performance of the corresponding averaging estimator. Observe that averaging bounds are uniformly better which is not surprising given that the averaging estimator is Bayes optimal for GMM.

We observe from Figure 3a and Figure 3b that the amount of unlabeled data has a positive effect on the test performance, and carrying out self-training iterations with fresh unlabeled data improves the performance. Comparing Figure 3a with Figure 3b, we see how the acceptance threshold Γ plays a critical role on the outcome. In fact, we find out from Figure 3b that Fresh-ST can outperform supervised learning with u samples, and regular iterative self-training can outperform regular supervised algorithm if the acceptance threshold Γ is high enough. The effect of Γ on the test performance is also demonstrated by Figure 3c, where we observe how picking an appropriate acceptance threshold boosts the test performance. We also see from Figure 3c how the test performance gets better when the number of iterations increases.

4 Importance of Regularization and Margin Between Components

We consider here a particular binary mixture model involving a scalar random variable X, and investigate the conditions and learning setups under which the use of unlabeled data improves the alignment of the classifier with the ground-truth mixture mean $\boldsymbol{\mu}$ (and hence the accuracy).

Definition 4.1 *(Generalized Mixture Model (Gen-MM)) The distribution $\mathcal{D}$ is given as follows. Fix a unit vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and scalar $\sigma \geq 0$. Let $X, y, \boldsymbol{g}$ be independent random variables where X is a scalar random variable with distribution $\mathcal{D}_X$, $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_p)$, and $\mathbb{P}(y = 1) = 1 - \mathbb{P}(y = -1) = 1/2$. The input $\boldsymbol{x}$ is generated as*

$$\boldsymbol{x} = yX\boldsymbol{\mu} + \sigma\boldsymbol{g}.$$

In this section, we provide algorithmic insights for the Gen-MM distribution which will shed light on the necessity of margin and importance of regularization. Here, our notion of margin is the gap between the class conditional distributions X and $-X$. If X is a positive random variable strictly bounded away from zero, then, we say there is a margin between the two classes since the distributions X and $-X$ are away from each other. We first focus on a simplified scenario where

we assume that we are provided an initial model $\boldsymbol{\beta}_{\text{init}}$[1] and we use $\boldsymbol{\beta}_{\text{init}}$ to label $\mathcal{U}$ and refine our estimate using pseudo-labeling. Focusing on least-squares loss and linear classifiers, in the infinite sample setup, this corresponds to the following problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}[1(|\bar{\boldsymbol{\beta}}_{\text{init}}^T \boldsymbol{x}| \geq \Gamma)(\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \boldsymbol{x}) - \boldsymbol{\beta}^T \boldsymbol{x})^2]. \tag{4.1}$$

where $\bar{\boldsymbol{\beta}}_{\text{init}} = \boldsymbol{\beta}_{\text{init}}/\|\boldsymbol{\beta}_{\text{init}}\|_{\ell_2}$. Before investigating this problem, it is worth understanding the simpler supervised loss. Denoting $\beta = \boldsymbol{\beta}^T\boldsymbol{\mu}$, the supervised quadratic loss is given by

$$\begin{aligned}\mathcal{L}_S(\boldsymbol{\beta}) &= \mathbb{E}_{\mathcal{D}}[(y - \boldsymbol{\beta}^T\boldsymbol{x})^2] = \mathbb{E}_{X,\boldsymbol{g}}[(X\boldsymbol{\beta}^T\boldsymbol{\mu} + \sigma\boldsymbol{\beta}^T\boldsymbol{g} - 1)^2] \\ &= \mathbb{E}[(X\beta - 1)^2] + \sigma^2\|\boldsymbol{\beta}^2\|_{\ell_2} \\ &= \sigma_X^2\beta^2 - 2\mu_X\beta + 1 + \sigma^2\|\boldsymbol{\beta}^2\|_{\ell_2}.\end{aligned}$$

This loss is minimized by choosing $\boldsymbol{\beta}^\star = \beta^\star\boldsymbol{\mu}$ where $\beta^\star = \mu_X/(\sigma_X^2 + \sigma^2)$. Additionally, this loss satisfies gradient dominance with respect to the global minima $\boldsymbol{\beta}^\star$ (as it will be discussed further later on), thus gradient descent on population loss will quickly find $\boldsymbol{\beta}^\star$. The question we are asking in this section is what happens when label information y is replaced by the pseudo-labels $\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T\boldsymbol{x})$. Our first theorem picks X to be the folded normal distribution (in words, X is the absolute value of a standard normal variable) and shows a negative result on pseudo-labeling.[2]

4.1 No Improvement with No Margin

Theorem 4.2 *Pick X to be the folded normal distribution (with density function $f_X(t) = \sqrt{2/\pi}\text{e}^{-t^2/2}$) and any $\Gamma \geq 0$. Let $\hat{\boldsymbol{\beta}}$ be the solution of the population self-training problem* (4.1)*. For some scalar $c > 0$ function of $(\sigma, \langle\boldsymbol{\mu}, \boldsymbol{\beta}_{init}\rangle, \Gamma)$, we have that $\hat{\boldsymbol{\beta}} = c\boldsymbol{\beta}_{init}$.*

The surprising conclusion from this theorem is that pseudo-labeling optimization (4.1) do not lead to an improved model. $\hat{\boldsymbol{\beta}}$ remains parallel to the original model $\boldsymbol{\beta}_{\text{init}}$ thus it will make the exact same label prediction as $\boldsymbol{\beta}_{\text{init}}$. Observe that folded normal distribution has no margin since the distributions of X and $-X$ both start from zero.

4.2 Guaranteed Improvement with Margin

In contrast to the result above, the following theorem shows that if there is a margin in the distribution of X, self-training does lead to an improved solution.

[1]Such an initial model can be obtained by minimizing the supervised risk $\mathcal{L}_S$ of (2.1) or via (3.1) as in Section 3.

[2]Folded normal has a nice simplifying nature during the theoretical analysis since yX becomes standard normal.

Theorem 4.3 *Fix $1 \geq \gamma \geq \sigma > 0$. Let X have unit variance and obey $M\gamma \geq X \geq \gamma$. Let $\hat{\boldsymbol{\beta}}$ be the solution of the population self-training problem* (4.1)*. For $\Gamma = 0$, setting $\rho(\boldsymbol{\beta}_{init}, \boldsymbol{\mu}) = \alpha$, we have that*

$$cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu}) \geq \frac{\sigma\text{e}^C}{4}(\gamma(1 - 6\text{e}^{-C}M)).$$

where $C = \frac{\alpha^2\gamma^2}{2\sigma^2}$. Specifically, if $\alpha\gamma > \sqrt{2\log(12M)}\sigma$, we find $cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu}) \geq 0.1\sigma\gamma\text{e}^{\frac{\alpha^2\gamma^2}{\sigma^2}}$.

Note that $\cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu})$ can be arbitrarily larger than the initial value $\cot(\boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu})$. As σ decreases, $\cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu})$ increases exponentially fast in the margin γ and the initial correlation α and $\hat{\boldsymbol{\beta}}$ becomes quickly aligned with the optimal direction $\boldsymbol{\mu}$. This should be contrasted with Theorem 4.2 where $\hat{\boldsymbol{\beta}}$ remains aligned with the initial model $\boldsymbol{\beta}_{\text{init}}$ which implies no improvement.

4.3 Provable Benefits of Regularization

In this section, we show that with proper regularization, distributional bias of the data can push the solution towards the global minima (i.e. a classifier perfectly aligned with $\boldsymbol{\mu}$). We consider two type of regularizations (recall $\bar{\boldsymbol{\beta}}_{\text{init}} = \boldsymbol{\beta}_{\text{init}}/\|\boldsymbol{\beta}_{\text{init}}\|_{\ell_2}$).

- **Ridge regression:** We consider the ridge regularized version of (4.1) given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}[1(|\bar{\boldsymbol{\beta}}_{\text{init}}^T\boldsymbol{x}| \geq \Gamma)(\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T\boldsymbol{x}) - \boldsymbol{\beta}^T\boldsymbol{x})^2] + \lambda\|\boldsymbol{\beta}\|_{\ell_2}^2. \tag{4.2}$$

- **Early-stopping:** We apply a single gradient iteration which corresponds to the averaging estimator of Section 3. This is given by the estimator

$$\hat{\boldsymbol{\beta}} = \mathbb{E}[1(|\bar{\boldsymbol{\beta}}_{\text{init}}^T\boldsymbol{x}| \geq \Gamma) \cdot \text{sgn}(\boldsymbol{\beta}_{\text{init}}^T\boldsymbol{x}) \cdot \boldsymbol{x}]. \tag{4.3}$$

In both cases, we show that regularization leads to a power iteration which emphasizes the distributional bias of the data and picks up the central direction $\boldsymbol{\mu}$. Our first result characterizes the performance of ridge regression.

Lemma 4.4 (Ridge regression) *Set $\Gamma = 0$ and let X have folded normal distribution. Define the strictly increasing function*

$$\kappa(\lambda) = \frac{1 + \sigma^2}{\sigma^2}\frac{\sigma^2 + \lambda}{1 + \sigma^2 + \lambda}.$$

Suppose $\hat{\boldsymbol{\beta}}$ is the solution of (4.2)*. We have that $cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu}) = \kappa(\lambda)cot(\boldsymbol{\beta}_{init}, \boldsymbol{\mu})$.*

Observe that $\kappa(\lambda) > 1$ and unlabeled data leads to provable improvement for any positive regularization parameter $\lambda > 0$. Our second result characterizes the performance of early-stopping (i.e. single iteration).

Lemma 4.5 (Early-stopping) *Suppose $\boldsymbol{\beta}_{init}^T \boldsymbol{\mu} = \alpha$ and let X have folded normal distribution. Suppose $\hat{\boldsymbol{\beta}}$ is the solution of (4.3). We have that*

$$cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu}) = (1 + \sigma^{-2}) cot(\boldsymbol{\beta}_{init}, \boldsymbol{\mu}). \tag{4.4}$$

Here, observe that the improvement in the co-tangent $\cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu})$ is captured by the signal-to-noise ratio. Since X is folded normal, the covariance matrix of the data obeys

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] = \sigma^2 \boldsymbol{I} + \boldsymbol{\mu}\boldsymbol{\mu}^T.$$

The eigenvalue along the signal direction $\boldsymbol{\mu}$ is $1 + \sigma^2$ whereas the orthogonal eigenvalues along the noisy directions are σ^2 and the ratio between them is $(1 + \sigma^2)/\sigma^2 = 1 + \sigma^{-2}$.

4.4 Importance of Regularization in Logistic Regression

Note that regularization is also critical for ensuring the success of self-training when it comes to classification loss functions as well. Examples include logistic loss, hinge loss and exponential loss. All of these loss functions have the common form $\ell(y, \hat{y}) = \ell(y\hat{y})$, are monotonically decreasing (Gunasekar et al., 2018), and satisfy the limit $\lim_{t\to\infty} \ell(t) = 0$. For instance hinge loss is given by $\ell(y, \hat{y}) = (1 - y\hat{y})_+$ and exponential loss is given by $\ell(y, \hat{y}) = \mathrm{e}^{-y\hat{y}}$. For logistic and exponential loss, the training loss never achieves zero and the model parameters have to indefinitely grow to minimize the loss. The pseudo-label loss function is obtained by setting $y = \mathrm{sgn}(\hat{y})$ so that the unlabeled example has loss equal to $\ell(|\hat{y}|)$.

In this section, we argue that, regularization is critical for enabling self-training/pseudo-labeling to find non-trivial models. The basic intuition is that, without regularization, the self-training loss can easily achieve zero while preserving the label decision of the original classifier. In other words, there is a trivial global minima. For instance, suppose we scale the final (i.e. logit) layer of a deep network by α. Then, this network will output the logits $\alpha\hat{y}$ rather than $\hat{y}$. For $\alpha > 0$, the class decision for $\alpha\hat{y}$ is exactly same as $\hat{y}$. However for $\alpha \geq 1$, the training loss decreases from $\ell(|\hat{y}|)$ to $\ell(\alpha|\hat{y}|)$. In general, as long as $\hat{y} \neq 0$, indefinitely enlarging α will asymptotically make the training loss zero. The following lemma formalizes this basic observation for general function classes.

Lemma 4.6 *Fix a prediction function $f : \mathbb{R}^p \to \mathbb{R}$. Consider the function class $\mathcal{F} = \{\alpha f \mid \alpha \geq 0\}$. Suppose the loss function ℓ obeys $\lim_{t\to\infty} \ell(t) = 0$ and the input distribution $\boldsymbol{x} \sim \mathcal{D}$ satisfies $\mathbb{P}_{\mathcal{D}}(f(\boldsymbol{x}) \neq 0) = 1$. Define the population self-training loss $\tilde{\mathcal{L}}(f) = \mathbb{E}_{\mathcal{D}}[\ell(|f(\boldsymbol{x})|)]$. We have that*

$$\lim_{\alpha\to\infty} \tilde{\mathcal{L}}(\alpha f) = 0.$$

Note that, the nonzero condition $\mathbb{P}_{\mathcal{D}}(f(\boldsymbol{x}) \neq 0) = 1$ helps us push the loss to zero by increasing the scale α. While this is a reasonably mild condition when the data has continuous distribution, we can also avoid this by considering an infinitesimal perturbation of f to reach a similar conclusion (e.g. using $\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}) + g$ where g is Gaussian noise with arbitrarily small variance).

Similar to least-squares, regularization techniques such as ridge-regression and early-stopping can guide self-training towards useful models by preventing degenerate solutions (which requires $\alpha \to \infty$) provided in Lemma 4.6.

5 Conclusions and Discussion

In this work, we analyzed the performance of self-training for linear classifiers and mixture distributions. We analytically showed that self-training process would converge to useful solutions for linear classifier parameters in the case of GMM. The theoretical findings demonstrate the benefits of rejecting samples with low-confidence and applying multiple self-training iterations and provides a framework for contrasting various algorithmic choices (e.g. fresh samples vs reusing samples). We also considered a variation of GMM which reveals that: (1) class margin (in terms of distance between mixture means) is critical for convergence of self-training to useful models and (2) ridge-regularization and early-stopping can enable self-training to converge to good models, in a similar fashion to power iteration converging to principal eigenvector, even without margin requirements. There are many interesting future works especially along joint statistical and algorithmic analysis of more practical self-training problems. It would be of interest to develop non-asymptotic bounds for iterative self-training schemes for more complex distributions and classifiers (e.g. logistic regression), adapt our approach to multiclass classification, and investigate the self-training behavior for nonlinear models such as deep nets.

Acknowledgments

S.O. is supported by the NSF grant CNS-1932254 and the NSF CAREER award CCF-2046816.

References

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *arXiv preprint arXiv:1908.02983*, 2019.

Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3):1–46, 2010.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(Nov):2399–2434, 2006.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116 (32):15849–15854, 2019.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060, 2019.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *Eighth International Conf. on Learning Representations*, volume 1, page 3, 2020.

Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. 2001.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh Annual Conf. on Computational Learning Theory*, pages 92–100, 1998.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11190–11201, 2019.

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

Vittorio Castelli and Thomas M Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6): 2102–2117, 1996.

David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106 (45):18914–18919, 2009.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 5001–5009, 2018.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML*, volume 99, pages 200–209, 1999.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. *arXiv preprint arXiv:2002.11361*, 2020.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013.

Marc Lelarge and Leo Miolane. Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. *arXiv preprint arXiv:1907.03792*, 2019.

Jian Li, Yong Liu, Rong Yin, and Weiping Wang. Multiclass learning using unlabeled samples: Theory and algorithm. In *Proceedings of the 28th International Joint Conf. on Artificial Intelligence (IJCAI)*, 2019.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE International Conf. on Computer Vision*, pages 2200–2207, 2013.

David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the main conference on human language technology conference of the North American Chapter of the*

Association of Computational Linguistics, pages 152–159. Association for Computational Linguistics, 2006.

Francesca Mignacco, Florent Krzakala, Yue M Lu, and Lenka Zdeborová. The role of regularization in classification of high-dimensional noisy gaussian mixture. *arXiv preprint arXiv:2002.11544*, 2020.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

Amir Najafi, Shin-ichi Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems*, pages 5542–5552, 2019.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134, 2000.

Luca Oneto, Davide Anguita, Alessandro Ghio, and Sandro Ridella. The impact of unlabeled patterns in rademacher complexity theory for kernel classifiers. In *Advances in Neural Information Processing Systems*, pages 585–593, 2011.

Luca Oneto, Alessandro Ghio, Sandro Ridella, and Davide Anguita. Local rademacher complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, 65:115–125, 2015.

Luca Oneto, Alessandro Ghio, Sandro Ridella, and Davide Anguita. Global rademacher complexity bounds: From slow to fast convergence rates. *Neural Processing Letters*, 43(2):567–602, 2016.

Samet Oymak and Talha Cihad Gulcu. Statistical and algorithmic insights for semi-supervised learning with self-training. *arXiv preprint arXiv:2006.11006*, 2020.

Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized lasso: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1002–1009. IEEE, 2013.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Mutual exclusivity loss for semi-supervised deep learning. In *2016 IEEE International Conf. on Image Processing (ICIP)*, pages 1908–1912. IEEE, 2016a.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 1163–1171, 2016b.

Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML Workshop on Learning with Multiple VIews*, volume 2005, pages 74–79. Citeseer, 2005.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Robert Stanforth, Alhussein Fawzi, Pushmeet Kohli, et al. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725*, 2019.

Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pages 1195–1204, 2017.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709, 2015.

Vladimir Vapnik. Statistical learning theory, Wiley. *New York*, 1, 1998.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. 2019.

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, 1995.

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conf. on Machine Learning (ICML-03)*, pages 912–919, 2003.

Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training.

In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 289–305, 2018.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proc. of the IEEE International Conf. on Computer Vision*, pages 5982–5991, 2019.