

Statistical and Algorithmic Insights for Semi-supervised Learning with Self-training

Samet Oymak*

Talha Cihad Gulcu†

June 19, 2020

Abstract

Self-training is a classical approach in semi-supervised learning which is successfully applied to a variety of machine learning problems. Self-training algorithm generates pseudo-labels for the unlabeled examples and progressively refines these pseudo-labels which hopefully coincides with the actual labels. This work provides theoretical insights into self-training algorithm with a focus on linear classifiers. We first investigate Gaussian mixture models and provide a sharp non-asymptotic finite-sample characterization of the self-training iterations. Our analysis reveals the provable benefits of rejecting samples with low confidence and demonstrates that self-training iterations gracefully improve the model accuracy even if they do get stuck in sub-optimal fixed points. We then demonstrate that regularization and class margin (i.e. separation) is provably important for the success and lack of regularization may prevent self-training from identifying the core features in the data. Finally, we discuss statistical aspects of empirical risk minimization with self-training for general distributions. We show how a fully unsupervised notion of generalization for self-training based clustering can be formalized via cluster margin. We then establish a connection between self-training based semi-supervision and the more general problem of learning with heterogenous data and weak supervision.

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 utilizes 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

*Email: oymak@ece.ucr.edu. Department of Electrical and Computer Engineering, University of California, Riverside.

†Email: tcgulcu@gmail.com.

and achieves zero training loss [5, 17, 47]. 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 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 (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.

- **Statistical Insights: Empirical Risk Minimization with Self-Training:** Focusing on general data distributions, we consider ERM with self-training. When the problem is purely unsupervised, we discuss how an unsupervised notion of generalization can be formalized based on the margin induced by the clusters found by self-training. Secondly, we discuss the loss landscapes of the supervised and unsupervised components of self-training. Inspired from the seminal results of [2], we connect self-training based semi-supervised learning to the more general problem of learning with heterogenous datasets and formalize how unlabeled and labeled data can be viewed as weak-supervision and strong supervision respectively.

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 [12, 30], semi-supervised support vector machines [19, 43], graph-based models [6, 9, 49], or co-training [10] and multiview models [37]. The relative value of labeled and unlabeled samples in a detection-estimation theoretical framework is examined in [13]. A line of work is related to how the presence of unlabeled data be useful to limit Radamacher complexity [3]. For example, the compatibility of a target function with respect to a data distribution is considered by [2], 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 [31–33] 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 [24], 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 [7, 8, 21, 22, 28, 35, 36, 41, 44] 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 [22] 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 [21, 36, 41] 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 [28] in the loss function of consistency training, and cross-entropy loss instead of squared loss function appears in the works [28, 44]. There are also hybrid algorithms combining diverse mechanisms. For example, Fix-Match [38] combines pseudo-labeling and consistency training to generate artificial labels. Mix-Match [8], ReMixMatch [7], unsupervised data augmentation [44] are among other composite approaches. Self training in the setting of domain adaptation is covered by the papers [18, 25]. Class balance [51] and confidence regularization [50] for self-training are among other lines of works. Gradual domain adaptation in regularized models is analyzed by [20]. The papers [11, 29, 39, 46] 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 \mathbf{a}, \mathbf{b} , their correlation is denoted by $\rho(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\|_{\ell_2} \|\mathbf{b}\|_{\ell_2}}$. Related to correlation, we define *co-tangent of the angle between two vectors* to be

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

which will be useful for cleaner notation. As $\cot(\mathbf{a}, \mathbf{b}) \rightarrow \infty$, the two vectors become perfectly correlated i.e. $\rho(\mathbf{a}, \mathbf{b}) \rightarrow 1$. 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 $\mathcal{S} = (y_i, \mathbf{x}_i)_{i=1}^n \in \{-1, 1\} \times \mathbb{R}^p$ be independent and identically distributed (i.i.d) labeled sampled distributed as $\mathcal{D} = \mathcal{D}_{y|\mathbf{x}} \times \mathcal{D}_{\mathbf{x}}$ and let $\mathcal{U} = (\mathbf{x}_i)_{i=n+1}^{n+u}$ be i.i.d. unlabeled samples distributed with the marginal distribution $\mathcal{D}_{\mathbf{x}}$. Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ be a prediction function (e.g. a neural network) and let $\hat{y}_f(\mathbf{x})$ be the hard-label $(-1, 1)$ assigned to $f(\mathbf{x})$ defined as

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

The standard self-training approach is sufficiently general to operate on a generic algorithm. The algorithm can self-train by using its own labels $\hat{y}_f(\mathbf{x})$ 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

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

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 $\mathcal{U}_\tau = (\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)$ where $\tilde{\mathbf{x}}_i \in \mathcal{U}$ are the acceptable inputs that satisfy $|f_\tau(\tilde{\mathbf{x}}_i)| \geq \Gamma$ and $\tilde{\mathbf{y}}_i$ are the pseudo-labels $\tilde{\mathbf{y}}_i = \hat{y}_{f_\tau}(\tilde{\mathbf{x}}_i)$.

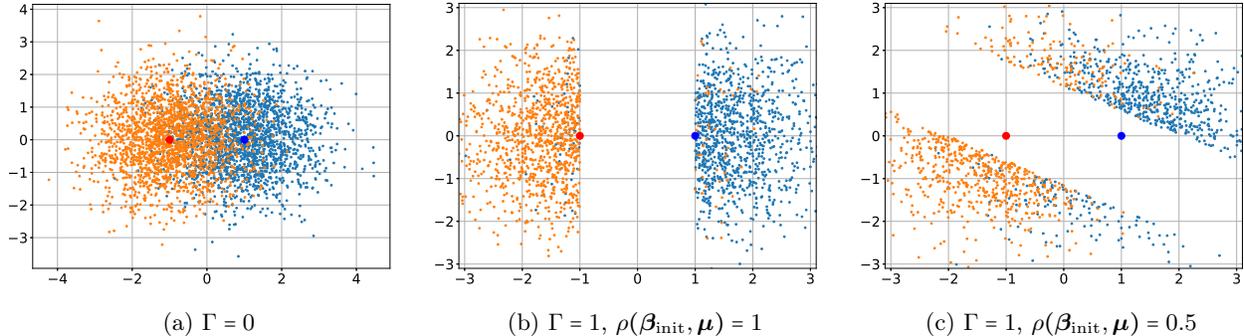


Figure 1: Visualization of a Binary GMM with noise variance $\sigma^2 = 1$. Sample size is 4000. The large dots at -1 and 1 are the mixture centers $\pm\boldsymbol{\mu} = [\pm 1, 0]$. Acceptance threshold Γ removes the examples with low-correlation to the initial model $\boldsymbol{\beta}_{\text{init}}$.

- **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}_τ 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.

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 $(\mathbf{x}, y) \sim \mathcal{D}$ 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 ($\mathbb{P}(y = 1) = 1 - \mathbb{P}(y = -1) = 1/2$) and $\mathbf{x} \sim \mathcal{N}(y\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.*

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

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

This estimator also coincide with the ridge regularized least-squares (e.g. $\arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_2}^2$) when the regularization parameter $\lambda \rightarrow \infty$. Perhaps surprisingly, this estimator is known to be the Bayes optimal classifier for GMM if we have access to the labeled data alone [23, 27]. This motivates us to investigate the analytical properties of the averaging estimator by adapting it to self-training as explained earlier. Given this initial supervised model $\boldsymbol{\beta}_{\text{init}}$ and the unlabeled dataset $\mathcal{U} = (\mathbf{x}_i)_{i=1}^u$ sampled from GMM, we consider the pseudo-label estimator

$$\hat{\boldsymbol{\beta}} = \text{self-train}(\boldsymbol{\beta}_{\text{init}}, \mathcal{U}) \quad \text{where} \quad \text{self-train}(\boldsymbol{\beta}_{\text{init}}, \mathcal{U}) = \frac{\sum_{i=1}^u \mathbf{1}(|\frac{\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}_i}{\|\boldsymbol{\beta}_{\text{init}}\|_{\ell_2}}| \geq \Gamma) \text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}_i) \mathbf{x}_i}{\sum_{i=1}^u \mathbf{1}(|\frac{\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}_i}{\|\boldsymbol{\beta}_{\text{init}}\|_{\ell_2}}| \geq \Gamma)}. \quad (3.2)$$

where $\Gamma \geq 0$ is the acceptance threshold eliminating low-confidence predictions. Acceptance threshold is commonly used in practical semi-supervised learning approaches [26, 44, 45]. 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 $[1 \ 0 \ 0 \ \dots \ 0]$ 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 β_{init} . Figure 1b and 1c chooses $\Gamma = 1$ for different β_{init} . In Figure 1b, β_{init} is aligned with μ (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 β_{init} and μ is 1/2 and β_{init} has a higher classification error. As a result, the two classes are not as cleanly separated despite using rejection.

The following theorem provides a sharp non-asymptotic bound for the pseudo-label estimator (3.2). Below, we set $\gamma_p = \mathbb{E}_{\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_p)}[\|\mathbf{g}\|_{\ell_2}]^2$. It is well-known that γ_p satisfies $p \geq \gamma_p \geq p - 1$.

Theorem 3.2 (Non-asymptotic Bound for GMM) *Let $\mu \in \mathbb{R}^p$ be a unit norm vector from Def. 3.1 and suppose $\beta_{\text{init}} \in \mathbb{R}^p$ has correlation $\rho(\beta_{\text{init}}, \mu) = \alpha > 0$. Set $\beta = \sqrt{1 - \alpha^2}$. Draw u i.i.d. unlabeled samples $(\mathbf{x}_i)_{i=1}^u$ from GMM. Let $\hat{\beta}$ be defined as $\hat{\beta} = \text{self-train}(\beta_{\text{init}}, (\mathbf{x}_i)_{i=1}^u)$. Fix resolution $1/2 > \varepsilon > 0$ and absolute constant $c > 0$. Define the normalized thresholds $\bar{\Gamma}_- = \frac{\alpha + \Gamma}{\sigma}$ and $\bar{\Gamma}_+ = \frac{\Gamma - \alpha}{\sigma}$ and the quantities*

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

With probability $1 - 10e^{-c\varepsilon^2((p-3) \wedge \rho u)}$, we have that

$$\frac{1 + \sigma\alpha\Lambda - 2\nu + (1 + \sigma)\varepsilon}{\sigma\sqrt{(\beta\Lambda - \varepsilon)_+^2 + (1 - \varepsilon)_+^2\gamma_{p-2}/u\rho}} \geq \cot(\hat{\beta}, \mu) \geq \frac{1 + \sigma\alpha\Lambda - 2\nu - (1 + \sigma)\varepsilon}{\sigma\sqrt{(\beta\Lambda + \varepsilon)^2 + (1 + \varepsilon)^2\gamma_{p-2}/u\rho}}.$$

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

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

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 β_{init} . Let f denote the transformation that is applied to $\rho(\beta_{\text{init}}, \mu)$ 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 $p \rightarrow \infty$, we have that

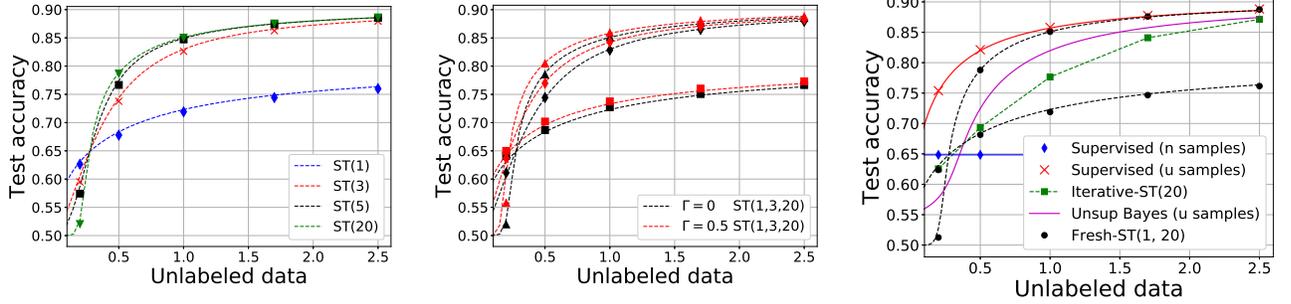
$$\cot(\hat{\beta}, \mu) = F_{\bar{u}}(\cot(\beta_{\text{init}}, \mu)) \quad \text{where} \quad F_{\bar{u}}(x) = \frac{1 + \sigma \frac{\Lambda x}{\sqrt{1+x^2}} - 2\nu}{\sigma \sqrt{\frac{\Lambda^2}{1+x^2} + \frac{1}{\bar{u}\rho}}}. \quad (3.4)$$

We remark that [13, 23] 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, instead of averaging estimator, by using tools from high-dimensional statistics and random matrix theory such as Gaussian min-max Theorem [34, 40, 42] and approximate message passing [4, 15].

3.1 Iterative self-training

Theorem 3.2 also allows us to analyze pseudo-labeling in an iterative fashion to show further improvement with more unlabeled data. Specifically, suppose we have n labeled samples $\mathcal{S} = (\mathbf{x}_i)_{i=1}^n$ and $\tau \times u$ unlabeled samples $\mathcal{U} = (\mathbf{x}_i)_{i=n+1}^{n+\tau u}$. We first create the supervised model via (3.1). Then, we split \mathcal{U} into τ disjoint sub-datasets $(\mathcal{U}_i)_{i=1}^\tau$. Starting from $\beta_0 = \beta_{\text{init}}$ of (3.1), we iteratively apply self-training via pseudo-labeling (3.2) to obtain

$$\beta_i = \text{self-train}(\beta_{i-1}, \mathcal{U}_i) \quad \text{for} \quad 1 \leq i \leq \tau. \quad (3.5)$$



(a) The impact of more self-training iterations on the model accuracy. (b) Comparing acceptance thresholds of $\Gamma = 0$ vs $\Gamma = 0.5$. (c) Comparison of different baselines at $\Gamma = 0$.

Figure 2: $p = 400$, $\bar{n} = n/p = 0.05$, $\Gamma = 0.5$, $\sigma = 0.75$. x -axis is the unlabeled data amount $\bar{u} = u/p$. In Figures (a) and (b), $ST(\tau)$ 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.

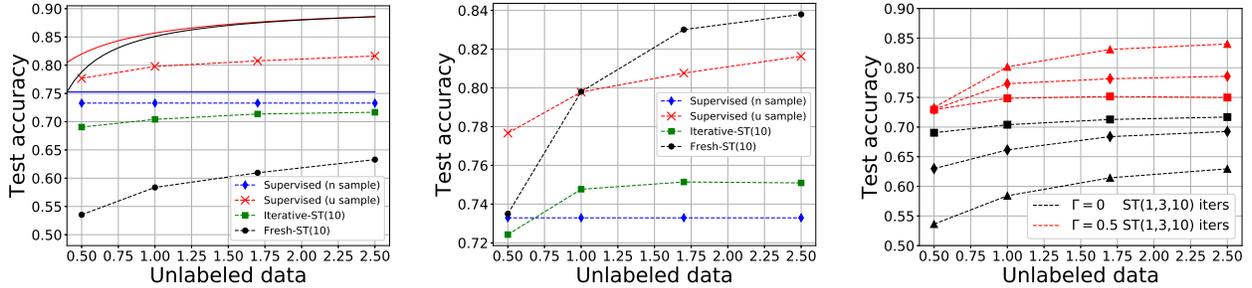
The final model is then equal to $\hat{\beta} = \beta_\tau$. Note that the asymptotic co-tangent of self-training with τ iterations will be given by $F^\tau(x)$ where x is the co-tangent of the initial supervised model. The following theorem establishes the asymptotic performance of this procedure.

Theorem 3.3 (Iterative self-training bound) *Set $\bar{n} = n/p$ and $\bar{u} = u/p$. Let $\mathcal{S} = (\mathbf{x}_i, y_i)_{i=1}^n$ and $\mathcal{U} = (\mathbf{x}_i)_{i=n+1}^{n+u}$ be independent datasets with i.i.d. samples generated according to Binary GMM. Obtain the model $\hat{\beta}$ via applying T iterations of the iterative self-training (3.5) to the supervised model (3.1). Recall the co-tangent evolution formula of (3.4). We have that*

$$\lim_{p \rightarrow \infty} \text{cot}(\hat{\beta}, \boldsymbol{\mu}) \xrightarrow{\mathbb{P}} F_{\bar{u}}^\tau(\sqrt{\bar{n}}/\sigma). \quad (3.6)$$

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 (3.6). 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 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. $F_{u_*}(\text{cot}(\beta_{\text{init}}, \boldsymbol{\mu})) = \text{cot}(\beta_{\text{init}}, \boldsymbol{\mu})$. 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 unlabeled 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 $\beta_{i+1} = \text{self-train}(\beta_i, \mathcal{U})$ for $1 \leq i \leq \tau = 20$. 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



(a) Comparing the performance of logis- (b) Performance of different approaches (c) Logistic regression, comparison of $\Gamma =$ tic regression and averaging estimator. for logistic regression. 0 vs $\Gamma = 0.5$.

Figure 3: Experiments on logistic regression: $\bar{n} = n/p = 0.2$, $\sigma = 0.75$, $p = 400$. In (a), for the same color, solid lines are the performance of the averaging estimator and the dashed lines with markers correspond to the logistic regression. 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.

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 [1]. 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 β_{init} , we first obtain the dataset \mathcal{U}' of acceptable inputs and associated pseudo-labels by thresholding $\mathbf{x}^T \beta_{\text{init}} / \|\beta_{\text{init}}\|_{\ell_2}$. 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. 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 Algorithmic Insights: Importance of Regularization and Margin

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 μ (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, \mathbf{g} be independent random variables where X is a scalar random variable with distribution \mathcal{D}_X , $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_p)$, and $\mathbb{P}(y = 1) = 1 - \mathbb{P}(y = -1) = 1/2$. The input \mathbf{x} is generated as

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

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

$$\hat{\boldsymbol{\beta}} = \frac{1}{2} \arg \min_{\boldsymbol{\beta}} \mathbb{E}[1(|\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}| \geq \Gamma \|\boldsymbol{\beta}_{\text{init}}\|_{\ell_2}) (\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}) - \boldsymbol{\beta}^T \mathbf{x})^2]. \quad (4.1)$$

Before investigating this problem, it is worth understanding the supervised loss. Setting $\boldsymbol{\beta} = \boldsymbol{\beta}^T \boldsymbol{\mu}$, the supervised quadratic loss is given by

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

The loss is minimized by choosing $\boldsymbol{\beta}^* = \beta^* \boldsymbol{\mu}$ where $\beta^* = \mu_X / (\sigma_X^2 + \sigma^2)$. Additionally, this loss satisfies gradient dominance with respect to the global minima $\boldsymbol{\beta}^*$ (as it will be discussed further later on), thus gradient descent on population loss will quickly find $\boldsymbol{\beta}^*$. The question we are asking in this section is what happens when label information y is replaced by the pseudo-labels $\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{x})$. 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 $f_X(t) = \sqrt{2/\pi} e^{-t^2/2}$) and any $\Gamma \geq 0$. Let $\hat{\boldsymbol{\beta}}$ be the solution of the population pseudo-labeling problem (4.1). For some scalar $c > 0$ depending on $\sigma, \langle \boldsymbol{\mu}, \boldsymbol{\beta}_{\text{init}} \rangle, \Gamma$, we have that $\hat{\boldsymbol{\beta}} = c\boldsymbol{\beta}_{\text{init}}$.*

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 Improvement with Margin

In contrast to the result above, the following theorem shows that if there is a margin in the distribution of X (i.e. X is strictly bounded away from zero), self-training does lead to an improved solution.

Theorem 4.3 *Fix $1 \geq \gamma \geq \sigma > 0$. Let X satisfy the second moment condition $\mathbb{E}[X^2] = 1$ and the margin condition $M\gamma \geq X \geq \gamma$. Let $\hat{\boldsymbol{\beta}}$ be the solution of the population self-training problem (4.1). For $\Gamma = 0$, setting $\rho(\boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu}) = \alpha$, we have that*

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

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

¹Such 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.

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

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

4.3 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 μ). We consider two type of regularizations.

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

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

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

$$\hat{\beta} = \mathbb{E}[1(|\beta_{\text{init}}^T \mathbf{x}| \geq \Gamma \|\beta_{\text{init}}\|_{\ell_2}) \cdot \text{sgn}(\beta_{\text{init}}^T \mathbf{x}) \cdot \mathbf{x}]. \quad (4.3)$$

In both cases, we show that regularization has a power-iteration-like affect which emphasizes the distributional bias of the data and picks up the central direction μ . Our first result characterizes the performance of the ridge regularization.

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

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 $\beta_{\text{init}}^T \mu = \alpha$ and let X have folded normal distribution. Suppose $\hat{\beta}$ is the solution of (4.3). We have that*

$$\cot(\hat{\beta}, \mu) = (1 + \sigma^{-2}) \cot(\beta_{\text{init}}, \mu). \quad (4.4)$$

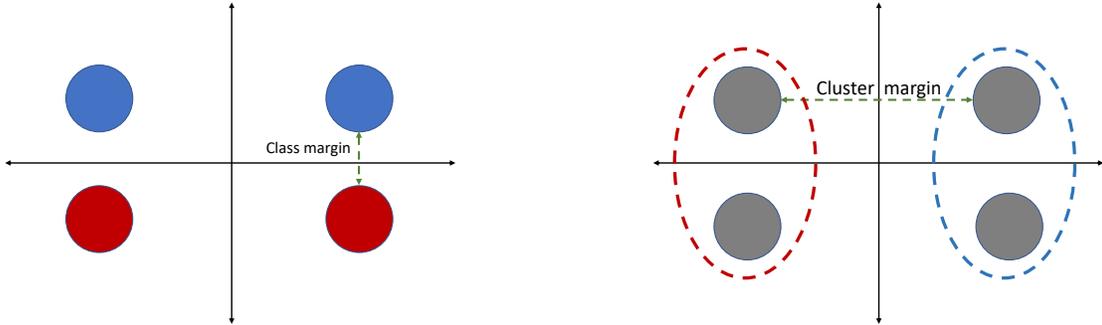
Here, observe that the improvement in the co-tangent $\cot(\hat{\beta}, \mu)$ is captured by the signal-to-noise ratio. Since X is folded normal, the covariance matrix of the data obeys

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

The eigenvalue along the signal direction μ 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 [16], and satisfy the limit $\lim_{t \rightarrow \infty} \ell(t) = 0$. For instance hinge loss is given by $\ell(y, \hat{y}) = (1 - y\hat{y})_+$ and exponential loss is given by $\ell(y, \hat{y}) = e^{-y\hat{y}}$. For logistic and exponential loss, the training loss can never achieve zero and the model



(a) Example distribution for the labeled data. x-axis separates the classes. (b) Unsupervised clustering induced by pseudo-labels. While x-axis separates the classes, y-axis maximizes the margin.

Figure 4: Supervised training would find a model maximizing the class margin. Unsupervised training would find a model maximizing the cluster margin.

parameters have to indefinitely grow to minimize the training loss. The pseudo-label loss function is obtained by setting $y = \text{sgn}(\hat{y})$ so that the unlabeled example has loss equal to $\ell(|\hat{y}|)$.

In this section, we will briefly 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 optimum. For instance, suppose we scale the final (i.e. logit) layer of a neural 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 provides a rigorous statement of this basic observation for a general function classes.

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

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

Note that, the nonzero condition $\mathbb{P}_{\mathcal{D}}(f(\mathbf{x}) \neq 0) = 1$ 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 $\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + g$ 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 $\alpha \rightarrow \infty$) provided in Lemma 4.6.

5 Statistical Insights Beyond Mixture Models

In this section, we focus on statistical aspects of empirical risk minimization with self-training for general datasets. First, we show how a fully unsupervised notion of generalization for self-training based clustering can be formalized via cluster margin. Then, we connect self-training based semi-supervised learning to the more general problem of learning with heterogenous datasets.

5.1 Unsupervised Learning with Self-Training

Classical statistical learning bounds such as Rademacher complexity arguments focus on labeled datasets and aims to show that the minimizer of empirical risk can accurately predict test labels. A natural question

is how to assess the success of pseudo-label optimization without any labels. While it is not possible for self-training optimization to find the true class distributions [2] without supervision, it is possible to argue that self-training loss can induce good clusters. This is illustrated via an example distribution in Figure 4. Figure 4a shows the distribution of the labeled data where classes are separated along x-axis. While there is a margin between the classes, clustering along y-axis leads to a larger margin. Thus, the class distributions are not the ideal clusters and self-training will not be able to find the correct class assignments without supervision. However, minimizing a proper self-training loss should be able to find the margin-maximizing clustering (i.e. Figure 4b). Following this example, let us assess the clustering quality via margin, i.e. ensuring that the input samples are away from the decision boundary. For instance, we can declare error if an input sample has margin less than γ . Such a clustering error can be defined as

$$\mathcal{E}_\gamma(f) = \mathbb{P}_{\mathbf{x} \sim \mathcal{D}}(|f(\mathbf{x})| \leq \gamma). \quad (5.1)$$

where $|f(\mathbf{x})|$ is the margin with respect to the model's own decision. Here, recall that the absolute value $|f(\mathbf{x})|$ naturally arises from the use of pseudo-labels. As discussed in Section 4.4, common loss functions such as quadratic, logistic and hinge loss has the simplifying multiplicative form $\ell(y, \hat{y}) = \ell(\hat{y}y)$ where $\hat{y} = f(\mathbf{x})$ is the prediction and y is the label. Plugging pseudo-label $\text{sgn}(f(\mathbf{x}))$ instead of the true label y leads to $\ell(|f(\mathbf{x})|)$. To encourage γ margin smoothly, let us define the margin loss $\ell_\gamma(\cdot) : [0, \infty) \rightarrow [0, 1]$

$$\ell_\gamma(x) = \begin{cases} 0 & \text{if } x \geq 2\gamma \\ -\frac{x}{\gamma} + 2 & \text{if } \gamma \leq x \leq 2\gamma \\ 1 & \text{if } 0 \leq x \leq \gamma \end{cases}.$$

Observe that $\ell_\gamma(x)$ is upper and lower bounded by the indicator functions as follows

$$1(x \leq \gamma) \leq \ell_\gamma(x) \leq 1(x \leq 2\gamma).$$

To proceed, given unlabeled samples $\mathcal{U} = (\mathbf{x}_i)_{i=1}^u$ and a function class \mathcal{F} , we show that solving the unsupervised empirical risk minimization problem

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{u} \sum_{i=1}^u \ell_\gamma(|f(\mathbf{x}_i)|). \quad (5.2)$$

can return a solution with good generalizability. Let $(\varepsilon_i)_{i=1}^n$ be i.i.d. Rademacher variables. Define the Rademacher complexity of \mathcal{F} with respect to \mathcal{U} to be

$$\mathcal{R}_u(\mathcal{F}) = \frac{1}{n} \mathbb{E}[\sup_{f \in \mathcal{F}} \sum_{i=1}^u \varepsilon_i f(\mathbf{x}_i)].$$

To make the dependence on the distribution \mathcal{D} explicit, we will also use the notation $\mathcal{R}_u^{\mathcal{D}}(\mathcal{F})$ later on. The following lemma follows from standard Rademacher complexity arguments³ to show that \hat{f} can induce a good clustering over the distribution \mathcal{D} in terms of prediction margin.

Lemma 5.1 (Self-Training Based Clustering) *Sample unlabeled data $\mathcal{U} = (\mathbf{x}_i)_{i=1}^u \stackrel{i.i.d.}{\sim} \mathcal{D}$. With probability at least $1 - \delta$ over \mathcal{U} , the clustering risk (5.1) of the solution \hat{f} of the ERM (5.2) obeys*

$$\mathcal{E}_\gamma(\hat{f}) \leq \min_{f \in \mathcal{F}} \mathcal{E}_{2\gamma}(f) + \frac{2}{\gamma} \mathcal{R}_u(\mathcal{F}) + 2\sqrt{\frac{\log(2/\delta)}{u}}.$$

In words, this bound states that the γ -clustering error induced by the empirical minimizer \hat{f} is upper bounded by the optimal 2γ -clustering error plus the Rademacher complexity term. It is also important to note that

³We are not aware of prior literature explicitly stating such a result however the proof does not require significant novelty over the standard Rademacher complexity arguments.

this bound is scale invariant. If the functions in the hypothesis set \mathcal{F} are scaled by a constant, the margin γ can be scaled by the same constant and the bound would remain perfectly intact. Thus, the bound is essentially in terms of normalized margin i.e. the margin normalized by the norm/magnitude of the functions. Recall that, if we fix γ and scale up the functions f , it is trivial to obtain 0 clustering error as discussed in Section 4.4. However, this would not learn a meaningful clustering of the data.

5.2 Learning with Weak Supervision with Relation to Self-Training

We discussed how pseudo-labels can help finding generalizable clusterings of the inputs however it is not clear how they can help towards identifying correct classes. To this aim, in this section, we discuss the fundamental principles of learning with heterogeneous distributions where the primary motivation is jointly learning from labeled and unlabeled datasets. Let $\mathcal{S} = (\mathbf{z}_i)_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}$ and $\mathcal{U} = (\tilde{\mathbf{z}}_i)_{i=1}^u \stackrel{\text{i.i.d.}}{\sim} \tilde{\mathcal{D}}$ be i.i.d. datasets with possibly different distributions. In this section, we consider the setup where \mathcal{D} is the primary distribution of interest and $\tilde{\mathcal{D}}$ provides side information about \mathcal{D} . Specifically, our goal is finding a model achieving small population risk over \mathcal{D} . Given a loss function ℓ and function class \mathcal{F} , we wish to find $f \in \mathcal{F}$ achieving small population risk

$$\mathcal{L}(f) = \mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[\ell(f(\mathbf{z}))]. \quad (5.3)$$

With this point of view, the dataset \mathcal{S} provides strong supervision and \mathcal{U} provides weak-supervision as it has a different distribution. In case of semisupervised learning, \mathcal{S} contains labeled data $(\mathbf{z}_i)_{i=1}^n = (\mathbf{x}_i, y_i)_{i=1}^n$ and \mathcal{U} contains unlabeled data $(\tilde{\mathbf{z}}_i)_{i=1}^u = (\mathbf{x}_i)_{i=n+1}^{n+u}$. Of particular interest, we focus on the scenario where weak-supervision dataset is larger than strong supervision i.e. $u \gg n$.

Numerical intuitions on heterogeneous losses: To proceed, we would like to formulate a problem which jointly uses \mathcal{S} and \mathcal{U} . We first start with some numerical intuition towards this goal with a focus on GMM distribution of Def. 3.1 with variance $\sigma^2 = 1$. Let us use linear classifier and quadratic loss. Then, for (\mathbf{x}, y) distributed as GMM, define the supervised and unsupervised population losses, which corresponds to strong and weak supervision respectively, as follows

$$\text{Supervised: } \mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2} \mathbb{E}[(y - \mathbf{x}^T \boldsymbol{\beta})^2] \quad (5.4)$$

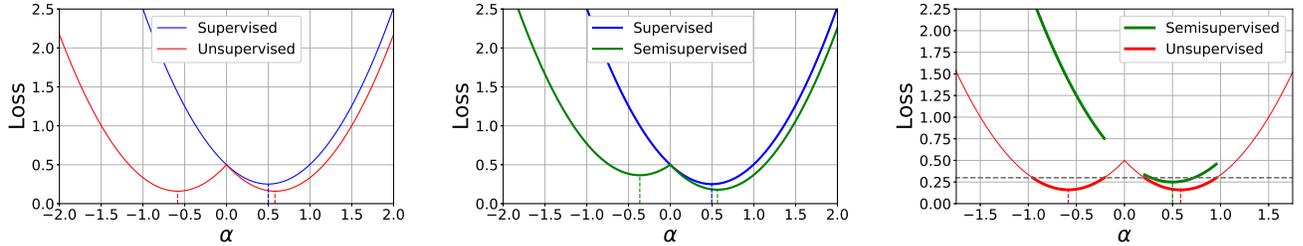
$$\text{Unsupervised: } \tilde{\mathcal{L}}(\boldsymbol{\beta}) = \frac{1}{2} \frac{\mathbb{E}[1(|\mathbf{x}^T \boldsymbol{\beta}| \geq \Gamma \|\boldsymbol{\beta}\|_{\ell_2}) (\text{sgn}(\mathbf{x}^T \boldsymbol{\beta}) - \mathbf{x}^T \boldsymbol{\beta})^2]}{\mathbb{P}(|\mathbf{x}^T \boldsymbol{\beta}| \geq \Gamma \|\boldsymbol{\beta}\|_{\ell_2})}, \quad (5.5)$$

where Γ is the acceptance threshold for self-training. In Figure 5, we plot these supervised and unsupervised population losses for parameters $\boldsymbol{\beta}$ along the $\boldsymbol{\mu}$ direction where $\pm \boldsymbol{\mu}$ are the mixture centers and $\Gamma = 0$. We choose $\boldsymbol{\beta} = \alpha \boldsymbol{\mu}$ where α is the scaling parameter (x-axis) and y -axis shows the loss associated with $\boldsymbol{\beta}$. In Figure 5a, the supervised loss curve is shown in blue which is convex and have a unique global minimum around $\alpha = 0.5$. The red curve shows the unsupervised loss $\tilde{\mathcal{L}}(\boldsymbol{\beta})$ (purely self-training/pseudo-labels). The unsupervised loss has two global minima (symmetrically located) and one of these minima are closely located to the global minimum of the supervised loss. Also observe from Figure 5a that the unsupervised loss is always less than the supervised loss over the entire α range as the pseudo-label $\text{sgn}(f(\tilde{\mathbf{z}}))$ induced by the data is guaranteed to result in a smaller or equal loss compared to that of the actual label. For semisupervised learning, we consider two types of regularization

$$\text{Regularized: } \mathcal{L}_{\text{semi}}(\boldsymbol{\beta}) = (1 - \rho)\mathcal{L}(\boldsymbol{\beta}) + \rho\tilde{\mathcal{L}}(\boldsymbol{\beta}) \quad (5.6)$$

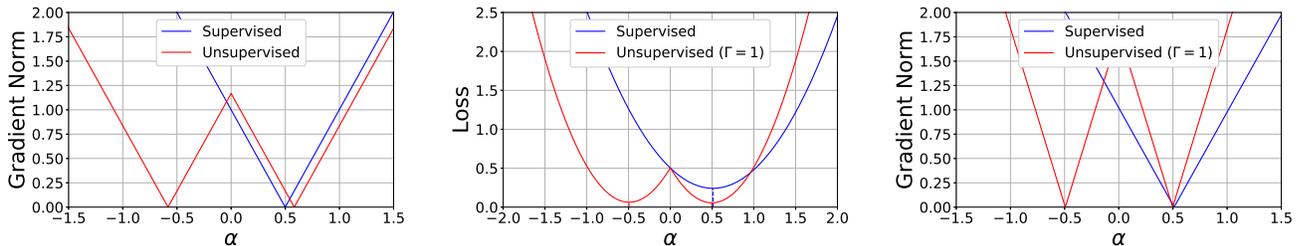
$$\text{Constrained: } \mathcal{L}_{\text{semi}}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) \quad \text{subject to } \tilde{\mathcal{L}}(\boldsymbol{\beta}) \leq \Xi. \quad (5.7)$$

Figure 5b plots the supervised loss curve along with the semisupervised loss function of 20% labeled data which can be expressed as $0.2\mathcal{L}(f) + 0.8\tilde{\mathcal{L}}_{\mathcal{U}}(f)$. Comparing Figure 5b with Figure 5a, we see that the semisupervised loss still has two local minima but only have a unique global minima. This global minima coincide with the global minima of the supervised loss i.e. both are around $\alpha \approx 0.5$. In summary, as we introduce 20% labeled data into the loss expression of the unsupervised case, the landscape difference between



(a) Landscapes of unsupervised (self-training) and supervised loss. $\Gamma = 0$. (b) Semisupervised learning with $\rho = 80\%$ self-training regularization (c) Semisupervised learning with self-training constraint $\Xi = 0.3$.

Figure 5: Loss landscape of self-training based semi-supervision for different regularization.



(a) The gradient length (in ℓ_2 norm) associated to the losses in Figure 5a. (b) Same as Figure 5a however self-training acceptance threshold is $\Gamma = 1$ (c) The gradient lengths in Figure 6b where $\Gamma = 1$.

Figure 6: Visualizing gradient sizes and the effect of acceptance threshold.

the semisupervised loss and the supervised loss becomes smaller. In general, semisupervised loss is not symmetrical around $\alpha = 0$ as soon as we have some fraction of labeled data in the loss and it has a unique global minima obeying $\alpha > 0$. Figure 5c shows the landscape of the constrained formulation (5.7). Here, the critical takeaway is that unsupervised loss has two global minima however constrained Ξ greatly narrows down the search space. Specifically, for semisupervised loss, the regions $\alpha < 0$ and $\alpha > 0$ are easily distinguishable. We use this intuition to formalize the benefit of weak supervision in Theorem 5.3. Figure 6a shows the gradient norms associated with the supervised and unsupervised losses in Figure 5a which shows the multiple local minima behavior for unsupervised loss (5.5). Figure 6b is same as 5a however we choose $\Gamma = 1$. Interestingly, the global minima of the unsupervised loss over $\alpha > 0$ axis coincides much better with the global minima of supervised loss. This shows how acceptance threshold can improve the loss landscape and the compatibility between actual labels and pseudo-labels. Figure 6c is the gradient norms corresponding to Figure 6b which shows that gradients of the unsupervised loss behave sharper compared to Fig 6a⁴ and the global minima over $\alpha > 0$ has a better match to supervised gradients.

Theoretical analysis: Following this intuition, we consider a constrained empirical risk minimization which first constrains the solution space to a smaller set of functions that achieve small loss on \mathcal{U} and then searches over this smaller set using \mathcal{S} . Let $\tilde{\ell}$ be the loss function to be used on the \mathcal{U} dataset. For instance, if \mathcal{U} is unlabeled, $\tilde{\ell}$ can be the loss function with respect to pseudo-labels. Define the empirical loss functions

$$\mathcal{L}_{\mathcal{S}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(z_i)) \quad \text{and} \quad \tilde{\mathcal{L}}_{\mathcal{U}}(f) = \frac{1}{u} \sum_{i=1}^u \tilde{\ell}(f(\tilde{z}_i)).$$

⁴This is likely due to the change in the distribution of the data after rejecting weak samples (which have small norms along μ direction).

We then solve the constrained problem

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \mathcal{L}_{\mathcal{S}}(f) \quad \text{subject to} \quad \tilde{\mathcal{L}}_{\mathcal{U}}(f) \leq \Xi, \quad (5.8)$$

where $\Xi > 0$ is the hyperparameter governing the strength of the constraint.

Landscape compatibility: To formalize our analysis we need to characterize how weak-supervision \mathcal{U} can help towards finding a solution for the population risk \mathcal{L} . Following our earlier discussion on semisupervised loss landscape, intuitively, this could be achieved by relating the loss landscapes associated with \mathcal{U} and \mathcal{S} .

Let $\tilde{\mathcal{L}}$ be the population risk of the $\tilde{\mathcal{D}}$ distribution i.e. $\tilde{\mathcal{L}}(f) = \mathbb{E}_{\tilde{\mathbf{z}} \sim \tilde{\mathcal{D}}}[\tilde{\ell}(f(\tilde{\mathbf{z}}))]$. We would like to ensure that the loss landscapes of \mathcal{L} and $\tilde{\mathcal{L}}$ have commonality to a certain extent. The basic idea is that there should be $f \in \mathcal{F}$ which achieves small population loss in both objectives. The following definition connects the sublevel sets of both loss functions and will be helpful for formalizing this commonality.

Definition 5.2 (Sublevel set and loss commonality) *Given $\varepsilon > 0$, function class \mathcal{F} and loss function \mathcal{L} , the ε -sublevel set of \mathcal{L} is defined as*

$$\mathcal{F}_{\mathcal{L}, \varepsilon} = \{f \mid f \in \mathcal{F} \quad \text{and} \quad \mathcal{L}(f) \leq \min_{f \in \mathcal{F}} \mathcal{L}(f) + \varepsilon\}.$$

Given another loss function $\tilde{\mathcal{L}}$, let $\tilde{\varepsilon} = \tilde{\varepsilon}(\mathcal{L}, \tilde{\mathcal{L}}, \varepsilon)$ be the smallest number such that

$$\mathcal{F}_{\mathcal{L}, \varepsilon} \cap \mathcal{F}_{\tilde{\mathcal{L}}, \tilde{\varepsilon}} \neq \emptyset. \quad (5.9)$$

In light of this definition, weak-supervision $\tilde{\mathcal{D}}$ would help \mathcal{D} when the sublevel sets of the loss functions \mathcal{L} and $\tilde{\mathcal{L}}$ intersects. This intuition is visualized in Figure 5c. Related notions of compatibility are used in earlier works [2, 14, 48] for semi-supervised learning. The following theorem establishes a statistical learning bound based on Rademacher complexity analysis by building on this intuition.

Theorem 5.3 (Learning with Weak-Supervision) *Fix $\varepsilon > 0$ and let $\tilde{\varepsilon}$ be as in Definition 5.2. Choose the constraint hyperparameter in (5.8) to be $\Xi = \bar{\Xi} + \min_{f \in \mathcal{F}} \tilde{\mathcal{L}}(f)$ with $\bar{\Xi} \geq 2\tilde{\varepsilon}$. Draw datasets $\mathcal{U} = (\tilde{\mathbf{z}}_i)_{i=1}^u$ i.i.d. $\tilde{\mathcal{D}}$ and $\mathcal{S} = (\mathbf{z}_i)_{i=1}^n$ i.i.d. \mathcal{D} . Assume $\ell, \tilde{\ell} : \mathbb{R} \rightarrow [0, 1]$ are L -Lipschitz loss functions. Suppose sample sizes n and u (are sufficiently large to) satisfy the following Rademacher complexity bounds*

$$\underbrace{2LR_u^{\tilde{\mathcal{D}}}(\mathcal{F}) + \frac{t}{\sqrt{u}}}_{\text{weakly supervised}} \leq \tilde{\varepsilon} \quad \text{and} \quad \underbrace{2LR_n^{\mathcal{D}}(\mathcal{F}_{\tilde{\mathcal{L}}, 2\bar{\Xi}}) + \frac{t}{\sqrt{n}}}_{\text{strongly supervised}} \leq \varepsilon. \quad (5.10)$$

Then, with probability $1 - 4e^{-t^2}$, the solution \hat{f} of the constrained problem (5.8) satisfies

$$\mathcal{L}(\hat{f}) \leq \min_{f \in \mathcal{F}} \mathcal{L}(f) + 3\varepsilon.$$

This theorem shows that as long as weak supervision has enough samples to narrow down the initial large search space to a small sublevel set, strong supervision can provably find a generalizing solution with very few samples where the sample complexity is only dictated by the Rademacher complexity of the sublevel set $\mathcal{F}_{\tilde{\mathcal{L}}, 2\bar{\Xi}}$. Recall that if the strong supervision loss \mathcal{L} and the weak supervision loss (e.g. unsupervised self-training) $\tilde{\mathcal{L}}$ have similar sub-level sets, then, $\tilde{\varepsilon}$ can be chosen to be very small which leads to a small search space for supervised loss in (5.8) and only few labels are sufficient for generalization. In essence, the technical idea (loss commonality) of this theorem is inspired from [2] which focuses on semi-supervised learning, however we show that the landscape compatibility can shed light on the statistical analysis of the more general problem class of weakly-supervised learning involving heterogeneous datasets.

6 Conclusions

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. Finally, we discussed the connections between semisupervised learning and learning with weak-supervision and heterogeneous data from a statistical learning perspective. 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), adapting our approach to multiclass classification, and investigating the self-training behavior for nonlinear models such as deep nets.

References

- [1] ARAZO, E., ORTEGO, D., ALBERT, P., O’CONNOR, N. E., AND MCGUINNESS, K. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *arXiv preprint arXiv:1908.02983* (2019).
- [2] BALCAN, M.-F., AND BLUM, A. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)* 57, 3 (2010), 1–46.
- [3] BARTLETT, P. L., AND MENDELSON, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3, Nov (2002), 463–482.
- [4] BAYATI, M., AND MONTANARI, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory* 57, 2 (2011), 764–785.
- [5] BELKIN, M., HSU, D., MA, S., AND MANDAL, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* 116, 32 (2019), 15849–15854.
- [6] BELKIN, M., NIYOGI, P., AND SINDHWANI, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research* 7, Nov (2006), 2399–2434.
- [7] BERTHELOT, D., CARLINI, N., CUBUK, E. D., KURAKIN, A., SOHN, K., ZHANG, H., AND RAFFEL, C. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *Eighth International Conference on Learning Representations* (2020), vol. 1, p. 3.
- [8] BERTHELOT, D., CARLINI, N., GOODFELLOW, I., PAPERNOT, N., OLIVER, A., AND RAFFEL, C. A. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems* (2019), pp. 5050–5060.
- [9] BLUM, A., AND CHAWLA, S. Learning from labeled and unlabeled data using graph mincuts.
- [10] BLUM, A., AND MITCHELL, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory* (1998), pp. 92–100.
- [11] CARMON, Y., RAGHUNATHAN, A., SCHMIDT, L., DUCHI, J. C., AND LIANG, P. S. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems* (2019), pp. 11190–11201.
- [12] CASTELLI, V., AND COVER, T. M. On the exponential value of labeled samples. *Pattern Recognition Letters* 16, 1 (1995), 105–111.
- [13] CASTELLI, V., AND COVER, T. M. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on information theory* 42, 6 (1996), 2102–2117.
- [14] DARNSTÄDT, M., SIMON, H. U., AND SZÖRÉNYI, B. Unlabeled data does provably help.
- [15] DONOHO, D. L., MALEKI, A., AND MONTANARI, A. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences* 106, 45 (2009), 18914–18919.

- [16] GUNASEKAR, S., LEE, J., SOUDRY, D., AND SREBRO, N. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246* (2018).
- [17] HASTIE, T., MONTANARI, A., ROSSET, S., AND TIBSHIRANI, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560* (2019).
- [18] INOUE, N., FURUTA, R., YAMASAKI, T., AND AIZAWA, K. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 5001–5009.
- [19] JOACHIMS, T. Transductive inference for text classification using support vector machines. In *Icml* (1999), vol. 99, pp. 200–209.
- [20] KUMAR, A., MA, T., AND LIANG, P. Understanding self-training for gradual domain adaptation. *arXiv preprint arXiv:2002.11361* (2020).
- [21] LAINE, S., AND AILA, T. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242* (2016).
- [22] LEE, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML* (2013), vol. 3, p. 2.
- [23] LELARGE, M., AND MIOLANE, L. Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. *arXiv preprint arXiv:1907.03792* (2019).
- [24] LI, J., LIU, Y., YIN, R., AND WANG, W. Multi-class learning using unlabeled samples: Theory and algorithm. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)* (2019).
- [25] LONG, M., WANG, J., DING, G., SUN, J., AND YU, P. S. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision* (2013), pp. 2200–2207.
- [26] MCCLOSKEY, D., CHARNIAK, E., AND JOHNSON, M. 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* (2006), Association for Computational Linguistics, pp. 152–159.
- [27] MIGNACCO, F., KRZAKALA, F., LU, Y. M., AND ZDEBOROVÁ, L. The role of regularization in classification of high-dimensional noisy gaussian mixture. *arXiv preprint arXiv:2002.11544* (2020).
- [28] MIYATO, T., MAEDA, S.-I., KOYAMA, M., AND ISHII, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2018), 1979–1993.
- [29] NAJAFI, A., MAEDA, S.-I., KOYAMA, M., AND MIYATO, T. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems* (2019), pp. 5542–5552.
- [30] NIGAM, K., MCCALLUM, A. K., THRUN, S., AND MITCHELL, T. Text classification from labeled and unlabeled documents using em. *Machine learning* 39, 2-3 (2000), 103–134.
- [31] ONETO, L., ANGUIA, D., GHIO, A., AND RIDELLA, S. The impact of unlabeled patterns in rademacher complexity theory for kernel classifiers. In *Advances in neural information processing systems* (2011), pp. 585–593.
- [32] ONETO, L., GHIO, A., RIDELLA, S., AND ANGUIA, D. Local rademacher complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks* 65 (2015), 115–125.
- [33] ONETO, L., GHIO, A., RIDELLA, S., AND ANGUIA, D. Global rademacher complexity bounds: From slow to fast convergence rates. *Neural Processing Letters* 43, 2 (2016), 567–602.
- [34] OYMAK, S., THRAMOULIDIS, C., AND HASSIBI, B. The squared-error of generalized lasso: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (2013), IEEE, pp. 1002–1009.
- [35] SAJJADI, M., JAVANMARDI, M., AND TASDIZEN, T. Mutual exclusivity loss for semi-supervised deep learning. In *2016 IEEE International Conference on Image Processing (ICIP)* (2016), IEEE, pp. 1908–1912.
- [36] SAJJADI, M., JAVANMARDI, M., AND TASDIZEN, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in neural information processing systems* (2016), pp. 1163–1171.
- [37] SINDHWANI, V., NIYOGI, P., AND BELKIN, M. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML workshop on learning with multiple views* (2005), vol. 2005, Citeseer, pp. 74–79.

- [38] SOHN, K., BERTHELOT, D., LI, C.-L., ZHANG, Z., CARLINI, N., CUBUK, E. D., KURAKIN, A., ZHANG, H., AND RAFFEL, C. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685* (2020).
- [39] STANFORTH, R., FAWZI, A., KOHLI, P., ET AL. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725* (2019).
- [40] STOJNIC, M. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291* (2013).
- [41] TARVAINEN, A., AND VALPOLA, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems* (2017), pp. 1195–1204.
- [42] THRAMOULIDIS, C., OYMAK, S., AND HASSIBI, B. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory* (2015), pp. 1683–1709.
- [43] VAPNIK, V. Statistical learning theory wiley. *New York 1* (1998).
- [44] XIE, Q., DAI, Z., HOVY, E., LUONG, M.-T., AND LE, Q. V. Unsupervised data augmentation for consistency training.
- [45] YAROWSKY, D. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics* (1995), pp. 189–196.
- [46] ZHAI, R., CAI, T., HE, D., DAN, C., HE, K., HOPCROFT, J., AND WANG, L. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555* (2019).
- [47] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B., AND VINYALS, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530* (2016).
- [48] ZHOU, X., AND BELKIN, M. Semi-supervised learning. In *Academic Press Library in Signal Processing*, vol. 1. Elsevier, 2014, pp. 1239–1269.
- [49] ZHU, X., GHAHRAMANI, Z., AND LAFFERTY, J. D. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)* (2003), pp. 912–919.
- [50] ZOU, Y., YU, Z., LIU, X., KUMAR, B., AND WANG, J. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 5982–5991.
- [51] ZOU, Y., YU, Z., VIJAYA KUMAR, B., AND WANG, J. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 289–305.

A Proofs for Section 3

The following lemma provides a straightforward guarantee on the estimation of $\boldsymbol{\mu}$.

Lemma A.1 (Simple supervised estimator) *Suppose we have n i.i.d. labeled examples $(\mathbf{x}_i, y_i)_{i=1}^n$ from the GMM model. Consider the supervised estimator*

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

With probability $1 - 2e^{-\varepsilon^2 p/2} - 2e^{-\varepsilon^2 n/2}$, we have that

$$\frac{1 + \sigma\varepsilon}{(1 - \varepsilon)\sigma\sqrt{p/n}} \geq \text{cot}(\boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu}) \geq \frac{1 - \sigma\varepsilon}{(1 + \varepsilon)\sigma\sqrt{p/n}}.$$

Proof Observe that, $\boldsymbol{\beta}_{\text{init}}$ is distributed as

$$\boldsymbol{\beta}_{\text{init}} = \boldsymbol{\mu} + \mathbf{h} \quad \text{where} \quad \mathbf{h} \sim \mathcal{N}\left(0, \frac{\sigma^2 \mathbf{I}_p}{n}\right).$$

Next, writing $\mathbf{h} = h\boldsymbol{\mu} + \mathbf{h}^\perp$ where $h \sim \mathcal{N}(0, 1)$, we have that $|h| \leq \sigma\varepsilon$ with probability at least $1 - 2e^{-\varepsilon^2 n/2}$ and $(1 - \varepsilon)\sigma\sqrt{p/n} \leq \|\mathbf{h}^\perp\|_{\ell_2} \leq (1 + \varepsilon)\sigma\sqrt{p/n}$ with probability at least $1 - 2e^{-\varepsilon^2 p/2}$. Combining, we find

$$\frac{1 + \sigma\varepsilon}{(1 - \varepsilon)\sigma\sqrt{p/n}} \geq \frac{1 + |h|}{\|\mathbf{h}^\perp\|_{\ell_2}} \geq \text{cot}(\boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu}) \geq \frac{1 - |h|}{\|\mathbf{h}^\perp\|_{\ell_2}} \geq \frac{1 - \sigma\varepsilon}{(1 + \varepsilon)\sigma\sqrt{p/n}}. \quad \blacksquare$$

Lemma A.2 *Let $g, h \sim \mathcal{N}(0, 1)$ and $f : \mathbb{R} \rightarrow \mathbb{R}$ be a bounded function. Then*

$$\mathbb{E}[(h + \sigma g)g] = \sigma \mathbb{E}[(h + \sigma g)h].$$

Proof Let $z = \sigma h - g$ and observe that z is independent of $h + \sigma g$. Thus, we note that

$$\mathbb{E}[f(h + \sigma g)g] = \mathbb{E}[f(h + \sigma g)(\sigma h - z)] = \mathbb{E}[f(h + \sigma g)\sigma h] = \sigma \mathbb{E}[f(h + \sigma g)h],$$

which is the desired statement. \blacksquare

A.1 Proof of Theorem 3.2

Proof Due to symmetry of the input clusters around 0, without losing generality, we can assume samples belong to the + cluster i.e. $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p)$. Set $\beta = \sqrt{1 - \alpha^2}$. Let us assume $\boldsymbol{\beta}_{\text{init}} = \alpha\boldsymbol{\mu} + \beta\mathbf{v}$ for some unit norm \mathbf{v} orthogonal to $\boldsymbol{\mu}$ and analyze $\hat{\boldsymbol{\beta}}$. Decompose the Gaussian noise vector \mathbf{g} as follows

$$\mathbf{g} = g_0\boldsymbol{\mu} + g\mathbf{v} + \mathbf{g}^\perp.$$

Here $g_0, g \sim \mathcal{N}(0, 1)$ and $\mathbf{g}^\perp \sim \mathcal{N}(0, \mathbf{I}_p - \boldsymbol{\mu}\boldsymbol{\mu}^T - \mathbf{v}\mathbf{v}^T)$. Additionally set $h = \boldsymbol{\beta}_{\text{init}}^T \mathbf{g} = \alpha g_0 + \beta g$. Let $\mathbf{g}_i, g_i, h_i, g_{0,i}, \mathbf{g}_i^\perp$ denote the variables associated with the i th sample. Proceeding, note that

$$\begin{aligned} \mathbb{P}(|\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}| \geq \Gamma) &= \mathbb{P}(|\alpha + \sigma h| \geq \Gamma) = \mathbb{P}(\{h \geq \bar{\Gamma}_+\} \cup \{h \leq -\bar{\Gamma}_-\}) \\ &= Q(\bar{\Gamma}_+) + Q(\bar{\Gamma}_-) \\ &:= \rho \end{aligned}$$

Set $s = \sum_{i=1}^u \mathbf{1}(|\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}_i| \geq \Gamma)$. Additionally define E_i to be the event that the pseudo-label prediction is wrong on the i th sample i.e. $E_i = \{|\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}_i| \geq \Gamma\} \cap \{\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}_i) \neq y_i\}$. Similar to above

$$\mathbb{P}(E_i) = \mathbb{P}(\alpha + \sigma h \leq -\Gamma) = Q(\bar{\Gamma}_-).$$

Define $f = \sum_{i=1}^u 1(E_i)$ (the total number of accepted examples with wrong pseudo-label predictions). Chernoff bound yields that with probability $1 - 4e^{-\frac{\varepsilon^2 \rho u}{3}}$, s and f obeys

$$|f - uQ(\bar{\Gamma}_-)| \leq \varepsilon u \rho \quad \text{and} \quad |s - u\rho| \leq \varepsilon u \rho. \quad (\text{A.1})$$

Define the conditional distribution $\mathbf{x}' \sim \mathbf{x} \mid |\beta_{\text{init}}^T \mathbf{x}| \geq \Gamma$. Let $\{\mathbf{x}'_i\}_{i=1}^s$ be the s accepted instances out of u (i.e. $|\beta_{\text{init}}^T \mathbf{x}'_i| \geq \Gamma$) with this distribution and write $\mathbf{x}'_i = \boldsymbol{\mu} + \sigma \mathbf{g}'_i$. Then, we can decompose $\mathbf{g}'_i = \mathbf{g}_i^\perp + g_i^\perp \mathbf{v} + g'_{0,i} \boldsymbol{\mu}$ where

$$\mathbf{g}_i^\perp \sim \mathcal{N}(0, \mathbf{I} - \boldsymbol{\mu} \boldsymbol{\mu}^T - \mathbf{v} \mathbf{v}^T) \quad \text{and} \quad |\alpha + \sigma h'_i| \geq \Gamma.$$

and $h'_i := \beta g'_i + \alpha g'_{0,i}$. This implies h'_i is distributed as $h' \sim h \mid \{h \geq \bar{\Gamma}_+\} \cup \{h \leq -\bar{\Gamma}_-\}$ where $h \sim \mathcal{N}(0, 1)$. Without losing generality, suppose $(\mathbf{x}'_i)_{i=1}^f$ are instances with wrong pseudo-label prediction and the rest are instances with correct pseudo-label prediction. To proceed, we estimate $\hat{\boldsymbol{\beta}}$ as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \frac{1}{s} \sum_{i=1}^s \text{sgn}(\beta_{\text{init}}^T \mathbf{x}'_i) \mathbf{x}'_i \\ &= \frac{1}{s} \sum_{i=1}^s \text{sgn}(\alpha + \sigma h'_i) (\boldsymbol{\mu} + \sigma \mathbf{g}_i^\perp + \sigma g_i^\perp \mathbf{v} + \sigma g'_{0,i} \boldsymbol{\mu}) \\ &= \underbrace{\left(1 - 2\frac{f}{s}\right) \boldsymbol{\mu}}_{a_1 \boldsymbol{\mu}} + \underbrace{\frac{\sigma}{s} \sum_{i=1}^s \text{sgn}(\alpha + \sigma h'_i) \mathbf{g}_i^\perp}_{\boldsymbol{\beta}^\perp} + \underbrace{\frac{\sigma}{s} \sum_{i=1}^s \text{sgn}(\alpha + \sigma \alpha g'_{0,i} + \sigma \beta g'_i) (g_i^\perp \mathbf{v} + g'_{0,i} \boldsymbol{\mu})}_{a_2 \boldsymbol{\mu} + a_3 \mathbf{v}}. \end{aligned} \quad (\text{A.2})$$

where $\boldsymbol{\beta}^\perp$ is orthogonal to $\mathbf{v}, \boldsymbol{\mu}$. Using $0 \leq \varepsilon \leq 1/2$ and recalling $Q(\bar{\Gamma}_-)/\rho = \nu$, the scalar a_1 can be bounded as

$$1 - 2\nu + 8\varepsilon \geq 1 - \frac{2Q(\bar{\Gamma}_-) - 2\rho\varepsilon}{\rho(1 + \varepsilon)} \geq a_1 = 1 - \frac{2f}{s} \geq 1 - \frac{2Q(\bar{\Gamma}_-) + 2\rho\varepsilon}{\rho(1 - \varepsilon)} \geq 1 - 2\nu - 8\varepsilon. \quad (\text{A.3})$$

The $\boldsymbol{\beta}^\perp$ term can be bounded by noting that \mathbf{g}^\perp is independent of h'_i which implies

$$\boldsymbol{\beta}^\perp = \frac{\sigma}{s} \sum_{i=1}^s \text{sgn}(1 + \sigma h'_i) \mathbf{g}_i^\perp \sim \mathcal{N}\left(0, \sigma^2 \frac{\mathbf{I} - \boldsymbol{\mu} \boldsymbol{\mu}^T - \mathbf{v} \mathbf{v}^T}{s}\right).$$

Consequently, using Gaussianity, $\boldsymbol{\beta}^\perp$ obeys $\sigma \sqrt{\gamma_{p-2}/s} = \mathbb{E}[\|\boldsymbol{\beta}^\perp\|_{\ell_2}]$. Via Lipschitz concentration and (A.1), this implies with probability at least $1 - 2e^{-\varepsilon^2 \gamma_{p-2}^2/2}$,

$$\begin{aligned} (1 + \varepsilon) \sigma \gamma_{p-2} / \sqrt{s} &\geq \|\boldsymbol{\beta}^\perp\|_{\ell_2} \geq (1 - \varepsilon) \sigma \gamma_{p-2} / \sqrt{s} \implies \\ (1 + 3\varepsilon) \sigma \sqrt{\gamma_{p-2}/u\rho} &\geq \|\boldsymbol{\beta}^\perp\|_{\ell_2} \geq (1 - 3\varepsilon) \sigma \sqrt{\gamma_{p-2}/u\rho}. \end{aligned} \quad (\text{A.4})$$

Finally, what remains is bounding the scalars a_2 and a_3 in (A.2). We accomplish this by going back to the original problem rather than the conditional sum which allows us to use Gaussianity. Specifically, we consider the summation

$$\frac{1}{u} \sum_{i=1}^u 1(|\alpha + \sigma h_i| \geq \Gamma) \text{sgn}(\alpha + \sigma h_i) (g_i \mathbf{v} + g_{0,i} \boldsymbol{\mu}) := a_2 \boldsymbol{\mu} + a_3 \mathbf{v}. \quad (\text{A.5})$$

First, note that we have the following expectation over h

$$\begin{aligned} \mathbb{E}[1(|\alpha + \sigma h| \geq \Gamma) \text{sgn}(\alpha + \sigma h) h] &= \rho^{-1} \int_{\bar{\Gamma}_+}^{\infty} \frac{1}{\sqrt{2\pi}} x e^{-x^2/2} dx + \rho^{-1} \int_{-\infty}^{-\bar{\Gamma}_-} \frac{1}{\sqrt{2\pi}} x e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}\rho} (e^{-\bar{\Gamma}_+^2/2} + e^{-\bar{\Gamma}_-^2/2}) := \Lambda. \end{aligned}$$

Now, using Lemma A.2 and $h = \alpha g_0 + \beta g$, this expectation will be proportionally split between the $\boldsymbol{\mu}$ associated variable g_0 and \mathbf{v} associated variable g . Specifically, we have

$$\begin{aligned} \mathbb{E}[1(|\alpha + \sigma h| \geq \Gamma) \text{sgn}(\alpha + \sigma h) g_0] &= \frac{\alpha}{\beta} \mathbb{E}[1(|\alpha + \sigma h| \geq \Gamma) \text{sgn}(\alpha + \sigma h) g] \\ \alpha \mathbb{E}[1(|\alpha + \sigma h| \geq \Gamma) \text{sgn}(\alpha + \sigma h) g_0] + \beta \mathbb{E}[1(|\alpha + \sigma h| \geq \Gamma) \text{sgn}(\alpha + \sigma h) g] &= \Lambda. \end{aligned}$$

This implies

$$\begin{aligned}\mathbb{E}[a_2] &= \mathbb{E}[1(|\alpha + \sigma h| \geq \Gamma) \text{sgn}(\alpha + \sigma h) g_0] = \alpha \sigma \Lambda \\ \mathbb{E}[a_3] &= \mathbb{E}[1(|\alpha + \sigma h| \geq \Gamma) \text{sgn}(\alpha + \sigma h) g] = \beta \sigma \Lambda.\end{aligned}\tag{A.6}$$

To proceed, we need to show concentration of the average in (A.5) which can be accomplished by noticing the subgaussianities

$$\|1(|\alpha + \sigma h| \geq \Gamma) \text{sgn}(\alpha + \sigma h) g_0\|_{\psi_2}, \|1(|\alpha + \sigma h| \geq \Gamma) \text{sgn}(\alpha + \sigma h) g\|_{\psi_2} \lesssim 1.$$

These immediately follow from the bounded moments. The subgaussian concentration implies that, with probability at least $1 - 4e^{-cu\varepsilon^2}$, we have that

$$\max(|a_2 - \mathbb{E}[a_2]|, |a_3 - \mathbb{E}[a_3]|) \leq \sigma\varepsilon.\tag{A.7}$$

Combining all of the estimates (A.3), (A.4), (A.6), with the advertised probability, we can write

$$\hat{\beta} = a_0 \mu + a_3 \mathbf{v} + \beta^\perp,$$

where $a_0 = a_1 + a_2$ and the components satisfy the following two sided bounds

$$\begin{aligned}|a_0 - (1 + \sigma\alpha\Lambda - 2\nu)| &\leq 8\varepsilon + \sigma\varepsilon \\ |a_3 - \sigma\beta\Lambda| &\leq \sigma\varepsilon \\ \|\beta^\perp\|_{\ell_2} - \sigma\sqrt{\gamma_{p-2}/u\rho} &\leq 3\varepsilon\sigma\sqrt{\gamma_{p-2}/u\rho}.\end{aligned}$$

This implies that, with probability at least $1 - 4e^{-cu\varepsilon^2} - 4e^{-\varepsilon^2\rho u/3} - 2e^{\varepsilon^2(p-3)/2}$, we obtain the advertised bound of

$$\frac{1 + \sigma\alpha\Lambda - 2\nu + (8 + \sigma)\varepsilon}{\sigma\sqrt{(\beta\Lambda - \varepsilon)_+^2 + (1 - 3\varepsilon)_+^2\gamma_{p-2}/u\rho}} \geq \cot(\hat{\beta}, \mu) \geq \frac{1 + \sigma\alpha\Lambda - 2\nu - (8 + \sigma)\varepsilon}{\sigma\sqrt{(\beta\Lambda + \varepsilon)^2 + (1 + 3\varepsilon)^2\gamma_{p-2}/u\rho}}$$

After mapping $\varepsilon \leftrightarrow \varepsilon/8$, we find that with the advertised probability, we have that

$$\frac{1 + \sigma\alpha\Lambda - 2\nu + (1 + \sigma)\varepsilon}{\sigma\sqrt{(\beta\Lambda - \varepsilon)_+^2 + (1 - \varepsilon)_+^2\gamma_{p-2}/u\rho}} \geq \cot(\hat{\beta}, \mu) \geq \frac{1 + \sigma\alpha\Lambda - 2\nu - (1 + \sigma)\varepsilon}{\sigma\sqrt{(\beta\Lambda + \varepsilon)^2 + (1 + \varepsilon)^2\gamma_{p-2}/u\rho}}.$$

Convergence in probability immediately follows from this non-asymptotic bound. ■

B Proofs for Section 4

Throughout, we assume β_{init} is unit Euclidian norm without losing generality. This is to simplify the subsequent notation.

B.1 Proof of Theorem 4.2

Proof Let us recall the distribution of the data. Given label y , we have that $\mathbf{x} = g\mu + \sigma g$ where $g := yX$. Noticing $g \sim \mathcal{N}(0, 1)$, this means that the marginal distribution of \mathbf{x} is $\mathcal{N}(0, \Sigma)$ where covariance matrix is $\Sigma = \sigma^2 \mathbf{I} + \mu\mu^T$. Thus, our interest is understanding the minimizer $\hat{\beta}$ of (4.1). For this, we have the following lemma that applies to arbitrary covariance matrices.

Lemma B.1 *Let $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ where Σ is a full-rank positive-semidefinite matrix and set $\Gamma \geq 0$. Then, the minimizer of (4.1) obeys*

$$\hat{\beta} = \frac{c_\Gamma \beta_{\text{init}}}{\|\sqrt{\Sigma} \beta_{\text{init}}\|_{\ell_2}}.$$

For the special case of $\Gamma = 0$, $c_\Gamma = \sqrt{2/\pi}$.

Proof Let $\mathbf{x}' \sim \mathbf{x} \mid |\mathbf{x}^T \boldsymbol{\beta}_{\text{init}}| \geq \Gamma$. Let $\boldsymbol{\Sigma}'$ be the covariance of \mathbf{x}' . Let us differentiate the loss with respect to $\boldsymbol{\beta}$. This yields

$$\mathbb{E}[\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}') \mathbf{x}'] = \mathbb{E}[\mathbf{x}' \mathbf{x}'^T] \hat{\boldsymbol{\beta}} \implies \hat{\boldsymbol{\beta}} = \boldsymbol{\Sigma}'^{-1} \mathbb{E}[\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}') \mathbf{x}'].$$

Let us write $\mathbf{x} = \sqrt{\boldsymbol{\Sigma}} \bar{\mathbf{x}}$ so that $\bar{\mathbf{x}} \sim \mathcal{N}(0, \mathbf{I})$. Set $\mathbf{a} = \sqrt{\boldsymbol{\Sigma}} \boldsymbol{\beta}_{\text{init}}$ and $\bar{\mathbf{a}} = \mathbf{a} / \|\mathbf{a}\|_{\ell_2}$. Decompose $\bar{\mathbf{x}} = \bar{\mathbf{a}} h + \bar{\mathbf{x}}^\perp$ where $\bar{\mathbf{x}}^\perp$ is independent of $h \sim \mathcal{N}(0, 1)$. Also let $h' \sim h \mid |h| \geq \Gamma / \|\mathbf{a}\|_{\ell_2}$. With this, we note that $\mathbf{x}' \sim \sqrt{\boldsymbol{\Sigma}} \bar{\mathbf{x}}^\perp + \sqrt{\boldsymbol{\Sigma}} \bar{\mathbf{a}} h'$. Consequently, using independence of h' and $\bar{\mathbf{x}}^\perp$, we obtain

$$\mathbf{s} = \mathbb{E}[\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}') \mathbf{x}'] = \mathbb{E}[\text{sgn}(h') \sqrt{\boldsymbol{\Sigma}} \bar{\mathbf{a}} h'] = \sqrt{\boldsymbol{\Sigma}} \bar{\mathbf{a}} \mathbb{E}[|h'|] = \frac{\boldsymbol{\Sigma} \boldsymbol{\beta}_{\text{init}}}{\|\mathbf{a}\|_{\ell_2}} \mathbb{E}[|h'|]. \quad (\text{B.1})$$

Secondly, observe that $\boldsymbol{\Sigma}' = \mathbb{E}[\mathbf{x}' \mathbf{x}'^T] = \sqrt{\boldsymbol{\Sigma}} (\mathbf{I} + (\mathbb{E}[h'^2] - 1) \bar{\mathbf{a}} \bar{\mathbf{a}}^T) \sqrt{\boldsymbol{\Sigma}}$. This yields

$$\boldsymbol{\Sigma}'^{-1} \mathbf{s} = \boldsymbol{\Sigma}^{-1/2} (\mathbf{I} + (\mathbb{E}[h'^2] - 1) \bar{\mathbf{a}} \bar{\mathbf{a}}^T)^{-1} \boldsymbol{\Sigma}^{-1/2} \frac{\boldsymbol{\Sigma} \boldsymbol{\beta}_{\text{init}}}{\|\mathbf{a}\|_{\ell_2}} \mathbb{E}[|h'|] \quad (\text{B.2})$$

$$= \boldsymbol{\Sigma}^{-1/2} (\mathbf{I} + (\mathbb{E}[h'^2] - 1) \bar{\mathbf{a}} \bar{\mathbf{a}}^T)^{-1} \bar{\mathbf{a}} \mathbb{E}[|h'|] \quad (\text{B.3})$$

$$= \boldsymbol{\Sigma}^{-1/2} \bar{\mathbf{a}} \mathbb{E}[h'^2]^{-1} \mathbb{E}[|h'|] \quad (\text{B.4})$$

$$= \frac{\mathbb{E}[|h'|] \boldsymbol{\beta}_{\text{init}}}{\mathbb{E}[h'^2] \|\sqrt{\boldsymbol{\Sigma}} \boldsymbol{\beta}_{\text{init}}\|_{\ell_2}} \quad (\text{B.5})$$

For the special case of $\Gamma = 0$, $h = h' \sim \mathcal{N}(0, 1)$ which implies $\mathbb{E}[|h'|] / \mathbb{E}[h'^2] = \sqrt{2/\pi}$. ■

This result also implies that for the original covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I} + \boldsymbol{\mu} \boldsymbol{\mu}^T$, $\hat{\boldsymbol{\beta}}$ will have the same direction as $\boldsymbol{\beta}_{\text{init}}$ with an additional scaling that depends on problem parameters $\sigma, \boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu}$. ■

B.2 Proof of Theorem 4.3

Proof We will argue that $\boldsymbol{\beta}_{\text{init}}$ successfully labels a large fraction of the unlabeled data under the margin condition $\gamma > 0$. The covariance of the input is again given by

$$\boldsymbol{\Sigma} = \sigma^2 \mathbf{I} + \boldsymbol{\mu} \boldsymbol{\mu}^T.$$

The population model is given by

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1} \mathbb{E}[\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}) \mathbf{x}] = \boldsymbol{\Sigma}^{-1} \mathbb{E}[\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}') \mathbf{x}'].$$

where $\mathbf{x}' = \mathbf{y} \mathbf{x} = X + \sigma \mathbf{g}$ ⁵. To proceed, we will analyze $\hat{\boldsymbol{\beta}}$ along the $\boldsymbol{\mu}$ direction and its orthogonal subspace. Without losing generality let us assume $\boldsymbol{\beta}_{\text{init}}$ is unit length and apply orthogonal decomposition $\boldsymbol{\beta}_{\text{init}} = \alpha \boldsymbol{\mu} + \sqrt{1 - \alpha^2} \boldsymbol{\mu}^\perp$ where $\|\boldsymbol{\mu}^\perp\|_{\ell_2}^2 = 1$. Also decompose $\mathbf{g} = g \boldsymbol{\mu} + g^\perp \boldsymbol{\mu}^\perp + \mathbf{g}_r$. We can write

$$\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}' = \alpha(X + \sigma g) + \sigma \sqrt{1 - \alpha^2} g^\perp \quad \text{where } g^\perp \sim \mathcal{N}(0, 1) \quad (\text{B.6})$$

To proceed, we decompose

$$\hat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{\Sigma}^{-1} \mathbb{E}[\text{sgn}(\boldsymbol{\mu}^T \mathbf{x}') \mathbf{x}']}_{\hat{\boldsymbol{\beta}}_m} + \underbrace{\boldsymbol{\Sigma}^{-1} \mathbb{E}[r(\mathbf{x}') \mathbf{x}']}_{\hat{\boldsymbol{\beta}}_p}$$

where $r(\mathbf{x}') = \text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{x}') - \text{sgn}(\boldsymbol{\mu}^T \mathbf{x}')$. The first component precisely returns the supervised model i.e.

$$\hat{\boldsymbol{\beta}}_m = \boldsymbol{\Sigma}^{-1} \mathbb{E}[\text{sgn}(X + \sigma g) ((X + \sigma g) \boldsymbol{\mu} + \mathbf{g}_r)] \quad (\text{B.7})$$

$$= \boldsymbol{\Sigma}^{-1} \mathbb{E}[|X + \sigma g|] \boldsymbol{\mu} \quad (\text{B.8})$$

$$= \frac{\mathbb{E}[|X + \sigma g|] \boldsymbol{\mu}}{1 + \sigma^2} \quad (\text{B.9})$$

Next, we focus on the perturbation term. Observe that \mathbf{g}_r is independent of $r(\mathbf{x}')$ thus, we have that

$$\hat{\boldsymbol{\beta}}_p = \boldsymbol{\Sigma}^{-1} \mathbb{E}[r(\mathbf{x}') \mathbf{x}'] = \boldsymbol{\Sigma}^{-1} \mathbb{E}[r(\mathbf{x}') ((X + \sigma g) \boldsymbol{\mu} + \sigma g^\perp \boldsymbol{\mu}^\perp)] \quad (\text{B.10})$$

$$= \frac{\boldsymbol{\mu}}{1 + \sigma^2} \mathbb{E}[r(\mathbf{x}') (X + \sigma g)] + \frac{\boldsymbol{\mu}^\perp}{\sigma} \mathbb{E}[r(\mathbf{x}') g^\perp]. \quad (\text{B.11})$$

⁵Here we slightly abuse the notation by using $\mathbf{g} \leftrightarrow \mathbf{y} \mathbf{g}$ via the rotational invariance of the standard normal \mathbf{g} .

Setting $\hat{\alpha} = \rho(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu})$, this shows that

$$\frac{\hat{\alpha}}{\sqrt{1 - \hat{\alpha}^2}} \geq \frac{\sigma}{1 + \sigma^2} \frac{\mathbb{E}[|X + \sigma g|] - \mathbb{E}[|r(\mathbf{x}')|(X + \sigma g)]}{\mathbb{E}[|r(\mathbf{x}')g^\perp|]}$$

Lemma B.2 states that $\mathbb{P}(r(\mathbf{x}') \neq 0) \leq 2Q(C) \leq e^{-C^2/2}$ where $C = \frac{\alpha\gamma}{\sigma}$. Using this, Lemma B.3 and $|r(\mathbf{x}')| \leq 2$, we have the followings.

- $\mathbb{E}[|X + \sigma g|] \geq \mathbb{E}[X] \geq \gamma$.
- $\mathbb{E}[|r(\mathbf{x}')X|] \leq 4e^{-C^2/2}M\gamma$.
- $\mathbb{E}[|r(\mathbf{x}')\sigma g|] \leq 2\sigma e^{-C^2/2}$.
- $\mathbb{E}[|r(\mathbf{x}')g^\perp|] \leq 2e^{-C^2/2}$.

Plugging these, we find

$$\cot(\hat{\boldsymbol{\beta}}, \boldsymbol{\mu}) = \frac{\hat{\alpha}}{\sqrt{1 - \hat{\alpha}^2}} \geq \frac{\sigma e^{C^2/2}}{2(1 + \sigma^2)} (\gamma(1 - 4e^{-C^2/2}M) - 2\sigma e^{-C^2/2}).$$

The advertised results follows from this bound by specializing to $\sigma \leq \gamma$ and $M \geq 1$ and then applying the change of variable $C \leftrightarrow C^2/2$. ■

B.3 Proof of Lemma 4.4

Proof The proof is similar to that of Theorem 4.2. Following same notation as the proof of Theorem 4.2, setting $\Gamma = 0$ and plugging covariance $\boldsymbol{\Sigma}$, the solution is given by

$$\hat{\boldsymbol{\beta}} = ((\lambda + \sigma^2)\mathbf{I} + \boldsymbol{\mu}\boldsymbol{\mu}^T)^{-1} \mathbf{s}.$$

$\mathbf{s} = c(\sigma^2\mathbf{I} + \boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{\beta}_{\text{init}}$ following from (B.1). Writing $\boldsymbol{\beta}_{\text{init}} = \alpha\boldsymbol{\mu} + \sqrt{1 - \alpha^2}\boldsymbol{\mu}^\perp$ and noticing $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^\perp$ are eigenvectors of $\boldsymbol{\Sigma}$, we obtain

$$c^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{s} = \frac{1 + \sigma^2}{1 + \sigma^2 + \lambda}\alpha\boldsymbol{\mu} + \frac{\sigma^2}{\sigma^2 + \lambda}\sqrt{1 - \alpha^2}\boldsymbol{\mu}^\perp. \quad (\text{B.12})$$

This implies the correlation guarantee given by Lemma 4.4 noticing the ratio of the $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^\perp$ terms above. ■

B.4 Proof of Lemma 4.5

Proof The proof is similar to that of Theorem 4.2. Following same notation as the proof of Theorem 4.2 and recalling (B.1), we have

$$\hat{\boldsymbol{\beta}} = \mathbb{P}(|\mathbf{x}^T \boldsymbol{\beta}_{\text{init}}| \geq \Gamma) \frac{\boldsymbol{\Sigma} \boldsymbol{\beta}_{\text{init}}}{\|\boldsymbol{\alpha}\|_{\ell_2}} \mathbb{E}[|h'|] = c\boldsymbol{\Sigma}\boldsymbol{\beta}_{\text{init}}.$$

Writing $\boldsymbol{\beta}_{\text{init}} = \alpha\boldsymbol{\mu} + \sqrt{1 - \alpha^2}\boldsymbol{\mu}^\perp$ and noticing $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^\perp$ are eigenvectors of $\boldsymbol{\Sigma}$, we obtain

$$c^{-1}\hat{\boldsymbol{\beta}} = (1 + \sigma^2)\alpha\boldsymbol{\mu} + \sigma^2\sqrt{1 - \alpha^2}\boldsymbol{\mu}^\perp. \quad (\text{B.13})$$

This implies the correlation guarantee given by (4.4) noticing the ratio of the $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^\perp$ terms above. ■

B.5 Proof of Lemma B.2

Lemma B.2 (Properties of rejection) Fix unit norm vectors $\boldsymbol{\beta}_{\text{init}}, \boldsymbol{\mu} \in \mathbb{R}^p$ with $\rho(\boldsymbol{\mu}, \boldsymbol{\beta}_{\text{init}}) = \alpha$. Let $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I})$ and X be a strictly positive random variable obeying $X \geq \gamma = \sigma\tilde{\gamma} > 0$. Set $\mathbf{x} = X\boldsymbol{\mu} + \sigma\mathbf{g}$ and let \mathbf{z} be the random vector with conditional distribution $\mathbf{x} \mid |\mathbf{x}^T \boldsymbol{\beta}_{\text{init}}| \geq \Gamma$. We have that

- When $\Gamma = 0$: $\mathbb{P}(\{\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{z}) \neq \text{sgn}(\boldsymbol{\mu}^T \mathbf{z})\}) \leq 2Q(\alpha\tilde{\gamma})$.

- General $\Gamma > 0$: Using the change of variable $\Gamma = \alpha\sigma\bar{\Gamma}$, we have

$$\mathbb{P}(\{\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{z}) \neq \text{sgn}(\boldsymbol{\mu}^T \mathbf{z})\}) \leq 2 \frac{Q(\bar{\gamma})Q(\frac{\alpha\bar{\Gamma}}{\sqrt{1-\alpha^2}}) + Q(\alpha(\bar{\gamma} + \bar{\Gamma}))}{Q_X(\bar{\Gamma})}.$$

Proof Represent $\mathbf{g} = \sigma g \boldsymbol{\mu} + \mathbf{g}^\perp$ and set $g' = \langle \boldsymbol{\mu}^\perp, \mathbf{g}^\perp \rangle$ where $\boldsymbol{\beta}_{\text{init}} = \alpha \boldsymbol{\mu} + \sqrt{1-\alpha^2} \boldsymbol{\mu}^\perp$. We analyze the event $E = \{\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{z}) \neq \text{sgn}(\boldsymbol{\mu}^T \mathbf{z})\}$. Clearly

$$\mathbb{P}(E) \leq \underbrace{\mathbb{P}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{z} < 0)}_{P(\boldsymbol{\beta}_{\text{init}})} + \underbrace{\mathbb{P}(\boldsymbol{\mu}^T \mathbf{z} < 0)}_{P(\boldsymbol{\mu})}. \quad (\text{B.14})$$

After bounding $\mathbb{P}(E)$, we also have that $|\mathbb{E}[\text{sgn}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{z}) - \text{sgn}(\boldsymbol{\mu}^T \mathbf{z})]| \leq 2\mathbb{P}(E)$.

When $\Gamma = 0$: First, using $\boldsymbol{\mu}^T \mathbf{x} = X + \sigma g$, we bound

$$P(\boldsymbol{\mu}) = P(\sigma g < -X)P(g < -\gamma/\sigma) = Q(\gamma/\sigma).$$

Secondly, recalling (B.6), we bound

$$P(\boldsymbol{\beta}_{\text{init}}) = P(\sqrt{1-\alpha^2}g' + \alpha g > \alpha X) = Q(\alpha\gamma/\sigma)$$

When $\Gamma > 0$: For $\Gamma > 0$, we condition on the event $A = \{|\mathbf{x}^T \boldsymbol{\beta}_{\text{init}}| \geq \Gamma\}$ which is equivalent to

$$|\alpha X + \sigma h| \geq \Gamma \quad \text{where} \quad h = \alpha g + \sqrt{1-\alpha^2}g'.$$

Following (B.14), we are interested in

$$\mathbb{P}(E) \leq \frac{\mathbb{P}(\boldsymbol{\beta}_{\text{init}}^T \mathbf{x} < 0 \cap A)}{P(A)} + \frac{\mathbb{P}(\boldsymbol{\mu}^T \mathbf{x} < 0 \cap A)}{P(A)}.$$

First, note that

$$P(A) \geq \mathbb{P}(\alpha X + \sigma h \geq \Gamma) \quad (\text{B.15})$$

$$\geq \mathbb{P}(\alpha X + \sigma h \geq \Gamma \mid h > 0)\mathbb{P}(h > 0) \geq \frac{1}{2}\mathbb{P}(\alpha X \geq \Gamma) \quad (\text{B.16})$$

$$= \frac{Q_X(\Gamma/\alpha)}{2} \quad (\text{B.17})$$

Secondly, we have

$$P(A)\mathbb{P}(\boldsymbol{\beta}_{\text{init}}) \leq \mathbb{P}(\alpha X + \sigma h \leq -\Gamma) = \mathbb{P}(h \leq \frac{-\alpha X - \Gamma}{\sigma}) \leq Q(\frac{\alpha\gamma + \Gamma}{\sigma})$$

Finally, we are interested in the probability $\mathbb{P}(X + \sigma g < 0 \cap A)$. Intersection event implies two things

- $E_1 = \{X + \sigma g < 0\}$.
- $E_2 = \{g' \geq \frac{\Gamma}{\sigma\sqrt{1-\alpha^2}}\}$. This follows from the fact that $X + \sigma g < 0$ and event A as follows

$$\Gamma \leq \alpha X + \alpha\sigma g + \sqrt{1-\alpha^2}\sigma g' \leq \sqrt{1-\alpha^2}\sigma g' \implies g' \geq \frac{\Gamma}{\sigma\sqrt{1-\alpha^2}}$$

Using independence of g, g' , we obtain

$$\mathbb{P}(X + \sigma g < 0 \cap A) \leq \mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2) \leq Q(\frac{\gamma}{\sigma})Q(\frac{\Gamma}{\sigma\sqrt{1-\alpha^2}})$$

which upper bounds $\mathbb{P}(A)\mathbb{P}(\boldsymbol{\mu})$. Combining these, we obtain the desired conclusion. \blacksquare

B.6 Proof of Lemma B.3

Lemma B.3 Let $g \sim \mathcal{N}(0, 1)$ and E be an event with probability $\mathbb{P}(E) = Q'(\alpha) = 2Q(\alpha)$ where Q' is the tail of folded normal distribution. We have that

$$\mathbb{E}[1(E)|g] \leq \sqrt{2/\pi}e^{-\alpha^2/2}.$$

Proof Let f be the density function of folded normal. Observe that

$$\begin{aligned} \mathbb{E}[1(E)|g] &= \int_0^\infty \mathbb{P}(\{|g| > x\} \cap E) dx \\ &= \int_0^\alpha \mathbb{P}(E) dx + \int_\alpha^\infty Q'(x) dx \\ &= Q'(\alpha)\alpha + \int_\alpha^\infty Q'(x) dx \\ &= \int_\alpha^\infty x f(x) dx \\ &= \int_\alpha^\infty \sqrt{2/\pi} x e^{-x^2/2} dx \\ &= \sqrt{2/\pi} e^{-\alpha^2/2}. \end{aligned}$$

■

B.7 Proof of Theorem 3.3

Proof Lemma A.1 shows that asymptotically $\cot(\beta_{\text{init}}, \boldsymbol{\mu}) \xrightarrow{\mathbb{P}} \sqrt{\bar{n}}/\sigma$. Since $(\mathcal{U}_i)_{i=1}^T$ are disjoint subsets, β_i is independent of \mathcal{U}_{i+1} and each iteration of self-training will apply $F_{\bar{u}}$ function on the co-tangent of the current iterate as a consequence of Theorem 3.2. This leads to the advertised bound. ■

B.8 Proof of Lemma 4.6

Proof Using right-continuity of cumulative distribution function, for any $\varepsilon > 0$, there exists $\delta > 0$ such that $\mathbb{P}_{\mathcal{D}}(f(\mathbf{x}) \leq \delta) \geq 1 - \varepsilon$. Thus, writing the expected loss as an integral over inputs $f(\mathbf{x}) < \delta$ and $f(\mathbf{x}) \geq \delta$,

$$\tilde{\mathcal{L}}(\alpha f) \leq \varepsilon \ell(0) + (1 - \varepsilon) \ell(\alpha \delta),$$

which implies $\lim_{\alpha \rightarrow \infty} \tilde{\mathcal{L}}(\alpha f) \leq \varepsilon \ell(0)$. Since this is true for any $\varepsilon \geq 0$, the limit is zero. ■

C Proofs for Section 5

C.1 Proof of Lemma 5.1

Proof Define the Rademacher complexity of the composition

$$\mathcal{R}_u(\ell \circ \mathcal{F}) = \frac{1}{u} \mathbb{E}[\sup_{f \in \mathcal{F}} \sum_{i=1}^u \varepsilon_i \ell_\gamma(|f(\mathbf{x}_i)|)].$$

$\ell(|x|)$ is γ^{-1} -Lipschitz function of x , hence Rademacher contraction inequality yields

$$\mathcal{R}_u(\ell \circ \mathcal{F}) \leq \gamma^{-1} \mathcal{R}_u(\mathcal{F}).$$

To proceed, applying standard generalization bound, with probability $1 - \delta/2$ over the samples, for all $f \in \mathcal{F}$, we have that

$$\mathbb{E}_{\mathcal{D}}[\ell_\gamma(|f(\mathbf{x})|)] \leq \frac{1}{u} \sum_{i=1}^u \ell_\gamma(|f(\mathbf{x}_i)|) + \frac{2}{\gamma} \mathcal{R}_u(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{u}}. \quad (\text{C.1})$$

Let $\mathcal{L}_U^* = \min \mathbb{E}_{\mathcal{D}}[\ell_\gamma(|f(\mathbf{x})|)]$ and $f^* = \arg \min \mathbb{E}_{\mathcal{D}}[\ell_\gamma(|f(\mathbf{x})|)]$. With probability $1 - \delta/2$, f^* satisfies

$$\frac{1}{u} \sum_{i=1}^u \ell_\gamma(|f^*(\mathbf{x}_i)|) \leq \mathcal{L}_U^* + \sqrt{\frac{\log(2/\delta)}{u}}.$$

Combining these two estimates and using optimality of \hat{f} , with probability at least $1 - \delta$,

$$\mathbb{E}[\ell_\gamma(|\hat{f}(\mathbf{x})|)] \leq \mathcal{L}_U^* + \frac{2}{\gamma} \mathcal{R}_u(\mathcal{F}) + 2\sqrt{\frac{\log(2/\delta)}{u}}.$$

Noticing $\mathcal{L}_U^* \leq \min_{f \in \mathcal{F}} \mathbb{P}(|f(\mathbf{x})| \leq 2\gamma)$ and $\mathbb{E}[\ell_\gamma(|\hat{f}(\mathbf{x})|)] \geq \mathbb{P}(|\hat{f}(\mathbf{x})| \leq \gamma)$ concludes the proof. \blacksquare

C.2 Proof of Theorem 5.3

C.2.1 Deterministic analysis

Following the setup of Theorem 5.3, in this section, we consider the deterministic conditions on the loss landscape that guarantees favorable properties of the constrained problem (5.8). Specifically, we make the following assumption that connects the landscape of empirical risk to the population risk.

Assumption 1 (Empirical is close to population) Fix scalars $\varepsilon > 0$, $\delta > 0$, $\bar{\Xi} \geq \tilde{\varepsilon} + \delta$. Define the sublevel set $\mathcal{F} = \mathcal{F}_{\tilde{\mathcal{L}}, \bar{\Xi} + \delta}$. The loss landscape of strong and weak supervision satisfy the following bounds.

- $\max_{f \in \mathcal{F}} |\tilde{\mathcal{L}}(f) - \tilde{\mathcal{L}}_{\mathcal{U}}(f)| \leq \delta$.
- $\max_{f \in \mathcal{F}'} |\mathcal{L}(f) - \mathcal{L}_{\mathcal{S}}(f)| \leq \varepsilon$.

Under this assumption, we have the following guarantee for the solution of the constrained empirical risk problem.

Theorem C.1 Suppose Assumption 1 holds for an (ε, δ) pair. Then, the solution to (5.8) with choice $\Xi = \bar{\Xi} + \min_{f \in \mathcal{F}} \tilde{\mathcal{L}}(f)$ with $\bar{\Xi} \geq \tilde{\varepsilon} + \delta$ satisfies

$$\mathcal{L}(\hat{f}) \leq \mathcal{L}^* + 3\varepsilon.$$

Proof Observe that, the constraint set of our problem is the sublevel set $\mathcal{F}_{\tilde{\mathcal{L}}_{\mathcal{U}}, \bar{\Xi}} = \{f \in \mathcal{F} \mid \tilde{\mathcal{L}}_{\mathcal{U}}(f) \leq \bar{\Xi}\}$. The first statement of Assumption 1 implies that the sublevel sets with respect to $\tilde{\mathcal{L}}_{\mathcal{U}}$ can be bounded via

$$\mathcal{F}_{\tilde{\mathcal{L}}, \alpha} \subseteq \mathcal{F}_{\tilde{\mathcal{L}}_{\mathcal{U}}, \alpha + \delta} \subseteq \mathcal{F}_{\tilde{\mathcal{L}}, \alpha + 2\delta} \quad \text{for all } \alpha \geq 0$$

Consequently, using $\bar{\Xi} \geq \tilde{\varepsilon} + \delta$, we find that

$$\mathcal{F}_{\tilde{\mathcal{L}}, \tilde{\varepsilon}} \subseteq \mathcal{F}_{\tilde{\mathcal{L}}_{\mathcal{U}}, \bar{\Xi}} \subseteq \mathcal{F}_{\tilde{\mathcal{L}}, \bar{\Xi} + \delta}.$$

Following the definition of $\tilde{\varepsilon}$ (i.e. (5.9)), this implies that there exists $f' \in \mathcal{F}_{\tilde{\mathcal{L}}_{\mathcal{U}}, \bar{\Xi}}$ such that $\mathcal{L}(f') \leq \mathcal{L}^* + \varepsilon$.

To proceed, the second statement of Assumption 1 guarantees that for all $f \in \mathcal{F}_{\tilde{\mathcal{L}}_{\mathcal{U}}, \bar{\Xi} + \delta}$ (thus for all feasible $f \in \mathcal{F}_{\tilde{\mathcal{L}}_{\mathcal{U}}, \bar{\Xi}}$) we have that $|\mathcal{L}(f) - \mathcal{L}_{\mathcal{S}}(f)| \leq \varepsilon$. Consequently, using the fact that \hat{f} minimizes the empirical risk over the feasible set (which includes f'), we find that

$$\mathcal{L}(\hat{f}) - \varepsilon \leq \mathcal{L}_{\mathcal{S}}(\hat{f}) \leq \mathcal{L}_{\mathcal{S}}(f') \leq \mathcal{L}(f') + \varepsilon \leq \mathcal{L}^* + 2\varepsilon,$$

concluding the proof. \blacksquare

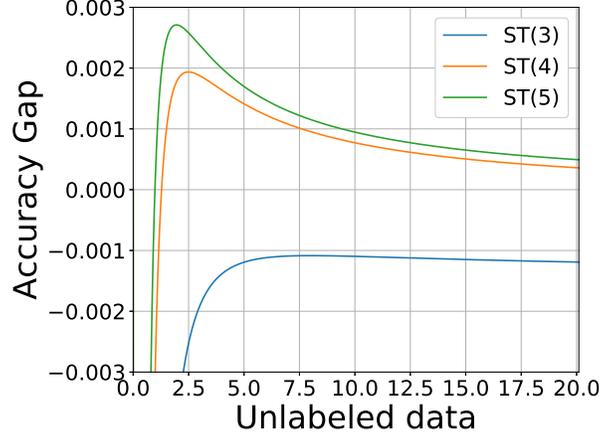


Figure 7: The gap between Fresh-ST (self-training with fresh unlabeled data) and supervised learning. Both uses u samples. This figure shows that supervised learning does not upper bound Fresh-ST and a few self-training iteration can provably go beyond supervised bound. The setup is same as in Figure 2.

C.2.2 Finishing the proof (analysis for random data)

The proof will be concluded by plugging in the necessary sample complexity bounds to guarantee that Assumption 1 holds. We pick $\delta = \tilde{\varepsilon}$ and $\bar{\Xi} \geq 2\tilde{\varepsilon}$ in our bound and in light of Theorem C.1, we would like to guarantee that

$$\max_{f \in \mathcal{F}} |\tilde{\mathcal{L}}(f) - \tilde{\mathcal{L}}_u(f)| \leq \tilde{\varepsilon} \quad \text{and} \quad \max_{f \in \mathcal{F}'} |\mathcal{L}(f) - \mathcal{L}_S(f)| \leq \varepsilon$$

where $\mathcal{F}_{\tilde{\mathcal{L}}, \bar{\Xi} + \tilde{\varepsilon}} \subseteq \mathcal{F}' = \mathcal{F}_{\tilde{\mathcal{L}}, 2\bar{\Xi}}$ (We remark that, we will ensure the desired bound holds over the larger set \mathcal{F}').

Recall that $\ell, \tilde{\ell} : \mathbb{R} \rightarrow [0, 1]$ are both L Lipschitz functions. Thus, standard Rademacher complexity based concentration bound [3] implies that, we have that

$$\sup_{f \in \mathcal{F}} |\tilde{\mathcal{L}}(f) - \tilde{\mathcal{L}}_u(f)| \leq 2L\mathcal{R}_u(\mathcal{F}) + \frac{t}{\sqrt{u}} \tag{C.2}$$

$$\sup_{f \in \mathcal{F}'} |\mathcal{L}(f) - \mathcal{L}_S(f)| \leq 2L\mathcal{R}_n(\mathcal{F}') + \frac{t}{\sqrt{n}}. \tag{C.3}$$

each with probability at least $1 - 2e^{-t^2}$. Thus, if n and u satisfies the advertised bounds, we find that $\sup_{f \in \mathcal{F}} |\tilde{\mathcal{L}}(f) - \tilde{\mathcal{L}}_u(f)| \leq \tilde{\varepsilon}$ and $\sup_{f \in \mathcal{F}'} |\mathcal{L}(f) - \mathcal{L}_S(f)| \leq \varepsilon$. Plugging these in Theorem C.1 concludes the proof.

D Self-Training with Fresh Samples Can Beat Supervised Learning

Following the setup of Figure 2, we consider the following question: Can Fresh-ST, self-training with u fresh unlabeled data at every iteration, beat supervised learning with u labels? There is no good reason for the answer to be negative however the answer is not clearly visible from Figure 2c. In this section, we zoom into Figure 2c by plotting the accuracy gap between supervised learning and Fresh-ST which is displayed in Figure 7. y-axis shows the accuracy gap $\text{acc}(\text{Fresh-ST}(\tau)) - \text{acc}(\text{supervised})$ where τ is the number of iterations. As τ increases, the accuracy gap achieves positive values proving that Fresh-ST can go beyond supervised learning. Note that this claim is already very visible for logistic regression (see Fig. 5b). This section clarifies this for averaging estimator as well.