WHY PSEUDO LABEL BASED ALGORITHM IS EFFECTIVE? –FROM THE PERSPECTIVE OF PSEUDO LABELED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, pseudo label based semi-supervised learning has achieved great success in many fields. The core idea of the pseudo label based semi-supervised learning algorithm is to use the model trained on the labeled data to generate pseudo labels on the unlabeled data, and then train a model to fit the previously generated pseudo labels. We give a theory analysis for why pseudo label based semi-supervised learning is effective in this paper. We mainly compare the generalization error of the model trained under two settings: (1) There are N labeled data. (2) There are N unlabeled data and a suitable initial model. Our analysis shows that, firstly, when the amount of unlabeled data tends to infinity, the pseudo label based semi-supervised learning algorithm can obtain model which have the same generalization error upper bound as model obtained by normally training in the condition of the amount of labeled data tends to infinity. More importantly, we prove that when the amount of unlabeled data is large enough, the generalization error upper bound of the model obtained by pseudo label based semi-supervised learning algorithm can converge to the optimal upper bound with linear convergence rate. We also give the lower bound on sampling complexity to achieve linear convergence rate. Our analysis contributes to understanding the empirical successes of pseudo label-based semi-supervised learning.

1 INTRODUCTION

Neural networks often require a large amount of labeled data to train, but in practice labeled data is usually very time-consuming and labor-intensive to obtain. However, unlabeled data is often less expensive to obtain. Therefore, semi-supervised learning has become popular in the field of deep learning. The key to the success of semi-supervised learning is to effectively use unlabeled data to help us obtain better models. (Kingma et al., 2014), (Laine & Aila, 2016), (Sohn et al., 2020), (Xie et al., 2020), (Shu et al., 2018), (Zhang et al., 2019) and (Laine & Aila, 2016) have put a lot of effort into using unlabeled data.

1.1 PSEUDO LABEL BASED SEMI-SUPERVISED LEARNING ALGORITHM

In general, pretraining and generating pseudo label are two main ways to use unlabeled data. Famous works of pretrain models include (Devlin et al., 2018), (Brown et al., 2020), (Baevski et al., 2020) and (Liu et al., 2019). In this paper, we focus on pseudo label based semi-supervised learning algorithm (Grandvalet & Bengio, 2004) and (Lee et al., 2013). The core idea of the pseudo label based semi-supervised learning algorithm is to use the model trained on the labeled data to generate pseudo labels on the unlabeled data, and then train a model to fit the previously generated pseudo labels. The sketch of pseudo label based semi-supervised learning algorithm is shown in Figure 1.

Algorithm 1 Pseudo label based semi-supervised learning algorithm

1: Input: proper initial model f_0, unlabeled data $\mathcal{T}$, $i = 0$, iteration number I.
2: **repeat**
3: Generate pseudo label on $\mathcal{T}$ using f_i.
4: Train on pseudo labeled $\mathcal{T}$ and get f_{i+1}.
5: $i \leftarrow i + 1$
6: **until** $i = I$

Output: Model f_I

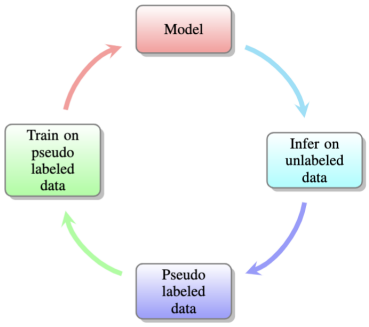

Figure 1: Pseudo label based semi-supervised learning algorithm sketch

1.2 INSIGHT

We provide a novel theory analysis of pseudo label based semi-supervised learning algorithm. Under a simple and realistic assumption on model, we show that when the amount of unlabeled data tends to infinity, the pseudo label based semi-supervised learning algorithm can obtain model which have the same generalization error upper bound as when the amount of labeled data tends to infinity. And when the amount of unlabeled data is large enough, the generalization error upper bound of the model obtained by pseudo label based semi-supervised learning algorithm can convergence to the optimal upper bound with linear convergence rate. Here the optimal upper bound represents the generalization error upper bound of the model trained on equal amount of labeled data.

Let us first introduce our assumption . Our assumption about the model is that if we train the model on dataset which part of them are randomly labeled, when the proportion of randomly labeled data is low, the model we get can have a lower empirical error on the correct labeled data, and a higher empirical error on the randomly labeled data. This is reasonable especially when the model is under parameterized. Though in our analysis the data is pseudo labeled and we always assume that we have large amount of unlabeled data, the assumption is reasonable. It is worth mentioning that even for when the DNN models are over parameterized, the DNN models still tends to fitting correct data before mislabeled data.(Liu et al., 2020) and (Arora et al., 2019). We also show the reasonable of it in Figure 2 by a toy example. In this toy setting, the dataset consists of $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ (in color green) and small part of random mislabeled of them (in color orange). So when we train the DNN models on the dataset as in the Figure 2, the model we get will ignore the mislabeled data.

Then, as our assumption, our insight is that when we train model on the dataset with small part of them are random labeled, the model tends to fit the correct data first and ignore the small random data especially if the model is under parameterized. So if we have a proper inital model and use it to generate pseudo labels on the unlabeled dataset, there will be small part of the pseudo labels are wrong labeled. And if we use the generated pseudo label to train a new model, the model will try to fit the correct labels and will be well-trained since the absolute quantity of the pseudo labels is often large. What's more, if the model trained on the pseudo labels is good enough, we can reuse it to generate new pseudo labels on the unlabeled dataset. We hope that the pseudo labels generated by previous model trained on the pseudo labels will have less wrong labels. So we can get better model

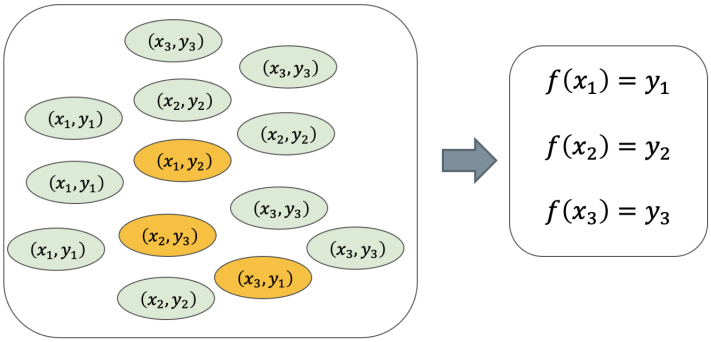

Figure 2: A toy example for our assumption

by them and the iteration can continue. Thus, our analysis contributes to understand the empirical successes of pseudo label-based semi-supervised learning (Laine & Aila, 2016), (Tarvainen & Valpola, 2017), (Lee et al., 2013), (Li et al., 2019), (Graves, 2012), (Chiu et al., 2018) and (Bengio et al., 2015).

In summary, our contributions include:

- We give a therory analysis of the pseudo label based algorithm from the perspective of pseduo labeled data and contribute to understand the empirical successes of pseudo label-based semi-supervised learning.
- We show that, if we have a proper initial model, when the amount of unlabeled data tends to infinity, the pseudo label based semi-supervised learning algorithm can obtain model which have the same generalization error upper bound as model obtained by noramlly training in the condition of the amount of labeled data tends to infinity.
- We prove that when the amount of unlabeled data is large enough (not need to be infinite), the generalization error upper bound of the model obtained by pseudo label based semi-supervised learning algorithm can convergence to the optimal upper bound with linear convergence rate. We also give the lower bound on sampling complexity to achieve linear convergence rate.

2 Related work

2.1 Theory about the pseudo label based semi-supervised learning

In the early stage of machine learning, (Sain, 1996) proposes transductive SVM which tried to utilize the unlabeled data. Then (Derbeko et al., 2003) estimates error bounds for transduction learning. Later, (Oymak & Gulcu, 2020) shows pseudo label based semi-supervised learning iterations improve model accuracy even though the model may be plagued by suboptimal fixed points. (Chen et al., 2020) shows that, for a certain class of distributions, entropy minimization on unlabeled target data will reduce the interference of fake features. However, the analysis in (Oymak & Gulcu, 2020) and (Chen et al., 2020) mainly focus on linear models and DNN models are not analyzed. For DNN models, (Wei et al., 2020) shows that pseudo label based semi-supervised learning method is indeed beneficial to improve the performance of the DNN models, and gives a sample complexity.

2.2 Random labeled data and population risk

There are lots of methods to estimate the population risk of the DNN models. One of the most important methods is estimating the upper bound on the generalization error of DNN models by estimating the complexity of the hypothesis classes (Neyshabur et al., 2015), (Neyshabur et al., 2017), (Ma et al., 2018) and (Weinan et al., 2019). However, this method often is strict to a specific

model and it is hard to use it to create a unified analysis to illustrate the advantage of pseudo label based semi-supervised learning for DNN models. Recently, (Garg et al., 2021) established a method to estimate the population risk of the DNN models via the model performance on random labeled data. In the method of (Garg et al., 2021), we will not need to estimate the complexity of the hypothesis classes. So it can help us create a unified analysis for DNN models. And it is clearly using this method is very convenient to show the benefit of pseudo label based semi-supervised learning since there often will be some mislabeled data in the pseudo labels.

3 PRELIMINARY

3.1 NOTATION

To be clearly, we first show the notation in our paper. We mainly focus on k classification problem. Using $\mathcal{S}$ represents the labeled data, n represents the amount of dataset $\mathcal{S}$, $\tilde{\mathcal{S}}$ represents the randomly labeled data and m represents the amount of dataset in $\tilde{\mathcal{S}}$. Using $\mathcal{E}_{\mathcal{S}}$ represents 0-1 loss on $\mathcal{S}$, $\mathcal{E}_{\tilde{\mathcal{S}}}$ represents 0-1 loss on $\tilde{\mathcal{S}}$, $\mathcal{E}_{\mathcal{D}}$ represents popluation 0-1 loss.

3.2 POPULATION RISK UPPER BOUND BASED ON RANDOM LABELED DATA

As described in Section 2.2, estimating population risk upper bound based on random labeled data is convenient to show the benefit of pseudo label based semi-supervised learning since there often will be some mislabeled data in the pseudo labels. What's more, it can help us to make an unified analysis for DNN models. Now we describe the theorem. This is obviously crucial for our following analysis.

Assumption 1 *Let $\hat{f}$ be a model obtained by training with an algorithm $\mathcal{A}$ on a mixture of clean data S and randomly labeled data $\tilde{\mathcal{S}}$. Then with probability $1-\delta$ over the (uniform but without the correct label) mislabeled data $\tilde{\mathcal{S}}_M$, we assume that the following condition holds:*

$$\mathcal{E}_{\tilde{\mathcal{S}}_M}(\widehat{f}) \leq \mathcal{E}_{\mathcal{D}'}(\widehat{f}) + c\sqrt{\frac{\log(1/\delta)}{2m}} \tag{1}$$

for a fixed constant $c > 0$. Where the $\mathcal{E}_{\mathcal{D}'}(\widehat{f})$ represents the population loss of $\widehat{f}$ on (uniform but without the correct label) mislabeded data. (Garg et al., 2021)

Theorem 1 *Under the Assumption 1, then for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$\mathcal{E}_{\mathcal{D}}(\widehat{f}) \leq \mathcal{E}_{\mathcal{S}}(\widehat{f}) + (k-1)\left(1 - \frac{k}{k-1}\mathcal{E}_{\tilde{\mathcal{S}}}(\widehat{f})\right) + c\sqrt{\frac{\log\left(\frac{4}{\delta}\right)}{2m}} \tag{2}$$

for some constantc satisfy

$$c \leq \left(2k + \sqrt{k} + \frac{m}{n\sqrt{k}}\right) \tag{3}$$

Where m represents the amount of dataset $\tilde{\mathcal{S}}$ and n represents the amount of dataset $\mathcal{S}$. (Garg et al., 2021)

Remark 1 *The randomly labeled data is not equal to mislabeled data. For k classification problem, the randomly labeled data means for any x, its label is uniformly randomly selected from k labels $y_1, y_2, y_3, ..., y_k$. However, the mislabeled data means for any x, its label is uniformly randomly selected from all k labels except its ground truth label. For example, for x_1, and we suppose its ground truth label is y_1, then mislabeled data of x_1 is uniformly randomly selected from $k-1$ labels $y_2, y_3, ..., y_k$.*

Remark 2 *The Assumption 1 holds in almost all scenarios. Since when we train DNN models, they always tend to overfit the training data. In practice, we often need to take steps to prevent the overfitting.*

3.3 ASSUMPTION ON THE MODEL AND THE DATA STRUCTURE

As we describe in Section 1.2, our assumption about the model is that if we train the model on the dataset, which parts of it are randomly labeled. When the proportion of randomly labeled data is low, we can obtain models with low empirical error on correctly labeled data and high empirical error on randomly labeled data. Here we give a mathematical formula that describes this assumption in Assumption 2.

Assumption 2 $\exists\varepsilon,\ \exists\tilde{\delta}$, *if the training dataset* $\mathcal{S}\cup\widetilde{\mathcal{S}}$ *satisfy*

$$\frac{m}{n}\leq\widetilde{\delta}<1 \tag{4}$$

we can get $\hat{f}$ *that satisfy*

$$\mathcal{E}_{\mathcal{S}}(\hat{f})\leq\varepsilon \tag{5}$$

$$\mathcal{E}_{\tilde{\mathcal{S}}}(\hat{f})\geq 1-\frac{1}{k}-\varepsilon \tag{6}$$

At the following section, we will go further under the condition of Assumption 2. Since the ε and $\tilde{\delta}$ change as the architecture of the model and training data change, exploring how the ε and $\tilde{\delta}$ change as the architecture of the model and training data change is still an important work, and we think we will do it in the future. And we will discuss under the fixed ε and $\tilde{\delta}$ in this paper.

In Section 4, we discuss the population risk under the condition that we have N labeled data as the normal training setting. In Section 5, we discuss the population risk under the condition that we have N unlabeled data and a proper initial model as the pseudo label based algorithm setting. And we compare the population risk in Section 4 and Section 5 as amount of (unlabeled) data tends to infinite. This is to show the power of pseudo label based algorithm in ONE iteration as the amount of unlabeled data tend infinite. In Section 6, We focus on whether the population risk by pseudo label based algorithm can approximated to the normally trained population risk $\mathcal{E}_D^*$ as the pseudo label based algorithm iteration progresses and the convergence rate of it. Besides, we provide a precise mathematical characterization of the required initialization model. This is to show the great power of pseudo label based algorithm when we have enough but limited unlabeled data as the iteration progresses.

4 TRAINING ON THE LABELED DATA

Firstly, according to Assumption 2, we have

$$\frac{m}{n}\leq\widetilde{\delta}<1 \tag{7}$$

and we can give a more relax upper bound in Theorem 1 to simplify our analysis and notation latter.

Theorem 2 *Under the Assumption 1, and if* $\frac{m}{n}<1$*, then for any* $\delta>0$*, with probability at least* $1-\delta$*, we have*

$$\begin{aligned}\mathcal{E}_{\mathcal{D}}(\widehat{f})\leq\mathcal{E}_{\mathcal{S}}(\widehat{f})+(k-1)\left(1-\frac{k}{k-1}\mathcal{E}_{\tilde{\mathcal{S}}}(\widehat{f})\right)\\ +4k\sqrt{\frac{\log\left(\frac{4}{\delta}\right)}{m}}\end{aligned} \tag{8}$$

Where m *represents the amount of dataset* $\tilde{\mathcal{S}}$ *and* n *represents the amount of dataset* $\mathcal{S}$*. (Garg et al., 2021)*

Now, we want to estimate the population risk upper bound of the models normal training on N labeled data by Theorem 2. On the one hand, we observe that the second term in Theorem 2 is related to performance on random labeled data $\tilde{\mathcal{S}}$. And the third term $4k\sqrt{\frac{\log\left(\frac{4}{\delta}\right)}{m}}$ decreases as m

increases. On the other hand, in Assumption 2, the ε show the model performance on both correct labeled data $\mathcal{S}$ and random labeled data $\tilde{\mathcal{S}}$. The $\widetilde{\delta}$ limit the upper bound of amount of random labeled data $\tilde{\mathcal{S}}$.

Thus, we can randomly labeled small part of N labeled data and we can use the Theorem 2 to estimate the population risk and the randomly labeled small part won't affect much compared with training purely on N labeled data. To satisfy Assumption 2, we have

$$\begin{cases} \dfrac{m}{n} = \tilde{\delta} \\ m + n = N \end{cases} \tag{9}$$

So we have

$$\begin{cases} m = \dfrac{\tilde{\delta}}{1+\tilde{\delta}} N \\ n = \dfrac{1}{1+\tilde{\delta}} N. \end{cases} \tag{10}$$

According to Assumption 2 and Theorem 2, we have f^* which satisfy

$$\mathcal{E}_D(f^*) \leq (k+1)\varepsilon + 4k\sqrt{\log\left(\frac{4}{\delta}\right)} \frac{1}{\sqrt{\frac{\tilde{\delta}}{1+\tilde{\delta}} N}} \tag{11}$$

We denote the population risk upper bound of f^* as

$$\mathcal{E}_D^* := (k+1)\varepsilon + 4k\sqrt{\log\left(\frac{4}{\delta}\right)} \frac{1}{\sqrt{\frac{\tilde{\delta}}{1+\tilde{\delta}} N}} \tag{12}$$

which reflect the population risk upper bound under the normal setting which we have N labeled training dataset.

Remark 3 *We observe that the $\mathcal{E}_D^*$ is nearly optimal since it has the error rate order of $O(\frac{1}{\sqrt{N}})$ which is equal to Monte Carlo estimation error rate order. This also implies the reasonableness of our assumption.*

5 AMOUNT OF DATA TENDS TO INFINITY

In this section, we firstly discuss the population risk under the condition that we have N unlabeled data and a proper initial model as the pseudo label based algorithm setting. Then we compare the population risk in Section 4 and Section 5 as amount of (unlabeled) data tends to infinite. We show that when the amount of unlabeled data tends to infinity, the pseudo label based semi-supervised learning algorithm can obtain model which have the same generalization error upper bound as model obtained by noramlly training in the condition of the amount of labeled data tends to infinity. We show the limitation of pseudo label based algorithm in ONE iteration as the amount of unlabeled data tend infinite. We reemphasize here that we discuss under the fixed ε and $\tilde{\delta}$ in this paper as in Section 3.3.

5.1 TRAINING ON PSEUDO LABELS

Now we talk about when we only have N unlabeled data and a initial f_0 as the pseudo label based semi-supervised learning algorithm setting as described in Section 1.1.

We need to generate the pseudo labels by the f_0 then train model by pseudo labels. We denote the $\mathcal{E}_D(f_0)$ as γ_0. Obviously there are about $(1-\gamma_0)N$ correct labels and $\gamma_0 N$ wrong labels in the generated pseudo labels. However, we can not view wrong labels in the generated pseudo labels as random label. A direct evidence is there are no correct labels in the wrong labels in the generated pseudo labels, but there are around $\frac{1}{k}$ correct labels in the random labels. Since both Assumption

2 and Theorem 2 connect to the model performance in random labeled data. A correct way is that we can select small part of generated pseudo labels then random labeled them. Then we can use Assumption 2 and Theorem 2 to estimate the population risk. To be more clearly, the modified pseudo label based algorithm, which is used to analyze are shown in Algorithm 2. The step 5 in Algorithm 2, which is mainly modified compared with Algorithm 1, has little effect on pseudo label based algorithm in Algorithm 1, because in practice the $m << N$ and we will show how to determine m below.

Algorithm 2 Modified pseudo label based algorithm

1: Input: initial model f_0, N unlabeled data $\mathcal{T}$, test data $\mathcal{T}_{test}$, $\varepsilon, \widetilde{\delta}$ that satisfy Assumption 2, iteration number I, $i = 0$.
2: **repeat**
3: Estimate the f_i population risk γ_i on $\mathcal{T}_{test}$
4: Generate pseudo label on $\mathcal{T}$ using f_i.
5: Random select proper $m(m << N)$ part of data from pseudo labeled $\mathcal{T}$ and random labeled them
6: Update the pseudo labeled $\mathcal{T}$.
7: Train on pseudo label $\mathcal{T}$.
8: $i \leftarrow i + 1$
9: **until** $i = I$
Output: Model f_I

To satisfy Assumption 2, the amount of data selected to random label m in Algorithm 2 has the following restriction.

$$\frac{m + \gamma_0(N - m)}{(1 - \gamma_0)(N - m)} \leq \widetilde{\delta} \tag{13}$$

So, we have

$$m \leq \frac{\widetilde{\delta}(1 - \gamma_0) - \gamma_0}{(1 + \widetilde{\delta})(1 - \gamma_0)} N \tag{14}$$

$$\gamma_0 \leq \frac{\widetilde{\delta}}{1 + \widetilde{\delta}} \tag{15}$$

The equation 14 shows that we can select at most $\frac{\widetilde{\delta}(1-\gamma_0)-\gamma_0}{(1+\widetilde{\delta})(1-\gamma_0)} N$ data from N generated pseudo labels then random labeled them when the γ_0 satisfy equation 21. According to Assumption 2 and Theorem 2, we can get f_1 and with at least $(1 - \delta)^{O(1)}$ probability we have

$$\mathcal{E}_D(f_1) \leq (k + 1)\varepsilon + 4k\sqrt{\log\left(\frac{4}{\delta}\right)}\sqrt{\frac{(1 + \widetilde{\delta})(1 - \gamma_0)}{\widetilde{\delta}(1 - \gamma_0) - \gamma_0}}\frac{1}{\sqrt{N}} \tag{16}$$

Compare $\mathcal{E}_D(f_1)$ with $\mathcal{E}_D^* := (k + 1)\varepsilon + 4k\sqrt{\log\left(\frac{4}{\delta}\right)}\frac{1}{\sqrt{\frac{\widetilde{\delta}}{1+\widetilde{\delta}} N}}$ in equation 12, we can easily show that

$$\lim_{N \to +\infty} \frac{\mathcal{E}_D(f_1)}{\mathcal{E}_D^*} \leq 1. \tag{17}$$

In summary, we have

Theorem 3 *Under the condition of Assumption 2, for fixed ε, $\widetilde{\delta}$ and $\forall \delta \in (0, 1)$, if we have f_0 with $\gamma_0 := \mathcal{E}_D(f_0) < \frac{\widetilde{\delta}}{1+\widetilde{\delta}}$ and N unlabeled data, then by Algorithm 2, with at least $(1 - \delta)^{O(1)}$ probability, we can get f_1 that satisfies*

$$\lim_{N \to +\infty} \frac{\mathcal{E}_D(f_1)}{\mathcal{E}_D^*} = 1. \tag{18}$$

This result implies that if we have a proper f_0 and the amount of input unlabeled data tends to infinite, pseudo label based semi-supervised algorithm can obtain a model which generalization error upper bound equal to model trained in condition of amount of labeled data tends to infinite, and this can achieve even in ONE iteration. This actually shows the power of the pseudo label based semi-supervised algorithm.

6 CONVERGENCE RATE OF PSEUDO LABEL BASED SEMI-SUPERVISED LEARNING

In this section, contrast to Section 5, we show the limitation of pseudo label based algorithm when we have enough but limited unlabeled data as the iteration progresses. We focus on whether the population risk by pseudo label based algorithm can approximate to the normally trained population risk $\mathcal{E}_D^*$ as the pseudo label based algorithm iteration progresses.

More importantly, we show that, when amount of unlabeled data N is big enough, a linear convergence rate can be achieved as the pseudo label based algorithm iteration progresses. We also give the lower bound on sampling complexity to achieve linear convergence rate. This is an extremely exciting finding that theoretically demonstrates the effectiveness of pseudo label based algorithms. We believe this finding contributes to understanding the empirical successes of pseudo label-based semi-supervised learning. We reemphasize here that we discuss under the fixed ε and $\tilde{\delta}$ in this paper as in Section 3.3.

6.1 ITERATION ON PSEUDO LABELS

As analysis in Section 5.1, if we use an initial model with population risk γ_0, the population risk uppper bound of f_1 by Algorithm 2 with at least $(1-\delta)^{O(1)}$ probability we have

$$\mathcal{E}_D(f_1) \leq (k+1)\varepsilon + 4k\sqrt{\log\left(\frac{4}{\delta}\right)}\sqrt{\frac{(1+\widetilde{\delta})(1-\gamma_0)}{\tilde{\delta}(1-\gamma_0)-\gamma_0}}\frac{1}{\sqrt{N}} \tag{19}$$

If the model f_1 trained on the pseudo labels is good enough, we can reuse it to generate new pseudo labels on the unlabeled dataset then obtained the new model f_2. We hope that the pseudo labels generated by f_1 model trained on the pseudo labels will have less wrong labels than f_0. So we can get better model by them and the iteration can continue. Here, we are interested in if the population risk upper bound can approximate the $\mathcal{E}_D^*$ and how fast it is. So we should care if we can achieve

$$\frac{\mathcal{E}_D(f_{i+1}) - \mathcal{E}_D^*}{\mathcal{E}_D(f_i) - \mathcal{E}_D^*} \leq p \tag{20}$$

where f_i, f_{i+1} denote the output of ith and $i+1$th iteration in Algorithm 2, p is a constant in $(0,1)$.We denote the population risk of f_i as γ_i. Then according to the analysis in Section 5.1, if

$$\gamma_i \leq \frac{\widetilde{\delta}}{1+\tilde{\delta}} \tag{21}$$

we can get f_{i+1} that with at least $(1-\delta)^{O(1)}$ probability we have

$$\mathcal{E}_D(f_{i+1}) \leq (k+1)\varepsilon + 4k\sqrt{\log\left(\frac{4}{\delta}\right)}\sqrt{\frac{(1+\widetilde{\delta})(1-\gamma_i)}{\tilde{\delta}(1-\gamma_i)-\gamma_i}}\frac{1}{\sqrt{N}} \tag{22}$$

Without loss of generality, we can further assume the γ_i satisfies

$$\mathcal{E}_D^* + c_1 \leq \mathcal{E}_D(f_i) \leq \mathcal{E}_D^* + c_2 \tag{23}$$

where c_1 and c_2 are two positive constant and c_1 can be arbitrarily small. Then solve the equation 20 we can get the lower bound of amount of data N.

$$N \geq \left(\frac{4k}{pc_1}\right)^2 \left[\sqrt{\frac{\tilde{\delta}+1}{\tilde{\delta}-\frac{\mathcal{E}_D^*+c_2}{1-\mathcal{E}_D^*-c_2}}}-\sqrt{\frac{\tilde{\delta}+1}{\tilde{\delta}}}\right]^2 \log\left(\frac{4}{\delta}\right) \tag{24}$$

In summary, we have

Theorem 4 *Under the condition of Assumption 2, for fixed ε, $\widetilde{\delta}$ satisfy Assumption 2 and $\forall \delta, c_1, c_2, p \in (0,1)$ and $c_1 < c_2$, we define $\mathcal{E}_D^*$ as equation 12. If the number of unlabeled data N satisfy*

$$N \geq \left(\frac{4k}{pc_1}\right)^2 \left[\sqrt{\frac{\tilde{\delta}+1}{\tilde{\delta}-\frac{\mathcal{E}_D^*+c_2}{1-\mathcal{E}_D^*-c_2}}}-\sqrt{\frac{\tilde{\delta}+1}{\tilde{\delta}}}\right]^2 \log\left(\frac{4}{\delta}\right) \tag{25}$$

and

$$\mathcal{E}_D^* + c_1 \leq \mathcal{E}_D(f_i) \leq \mathcal{E}_D^* + c_2 \tag{26}$$

$$\mathcal{E}_D(f_i) \leq \frac{\tilde{\delta}}{1+\tilde{\delta}} \tag{27}$$

then with at least $O(1-\delta)^{O(1)}$ probability we have

$$\frac{\mathcal{E}_D(f_{i+1}) - \mathcal{E}_D^*}{\mathcal{E}_D(f_i) - \mathcal{E}_D^*} \leq p \tag{28}$$

where f_i, f_{i+1} denote the output of ith and $i+1$th iteration in Algorithm 2.

Since c_1 could be arbitrarily small, this result indeed shows that the the population risk by pseudo label based algorithm can approximated to the normally trained population risk $\mathcal{E}_D^*$ as the pseudo label based algorithm iteration progresses. What's more, we show that a linear convergence rate can be achieved and we give the lower bound on sampling complexity to achieve linear convergence rate.

7 CONCLUTION AND FUTURE WORK

We conduct a theoretical analysis of pseudo-label based algorithm. We show that when the amount of (unlabeled) data tends to infinity, if we have a suitable initial model, pseudo label based semi-supervised learning algorithm can obtain models with the same upper bound on generalization error as we usually train models. We also demonstrate that the generalization error upper bound of the model obtained by the pseudo label based semi-supervised learning algorithm can converge to the optimal upper bound with a linear convergence rate. Our Assumption 2 is important to our analysis and we have explained the reasonableness of the assumption in Section 1.2. But how the ε and $\tilde{\delta}$ change as the architecture of the model and training data change is still mysterious to us, and we will explore it in the future. We hope that our analysis helps to understand the empirical success and reveal the potential of pseudo label based semi-supervised learning algorithm, and facilitate its application in wider scenarios.

REFERENCES

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. *Advances in Neural Information Processing Systems*, 33:21061–21071, 2020.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4774–4778. IEEE, 2018.

Philip Derbeko, Ran El-Yaniv, and Ron Meir. Error bounds for transductive learning via compression and clustering. *Advances in Neural Information Processing Systems*, 16, 2003.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Saurabh Garg, Sivaraman Balakrishnan, Zico Kolter, and Zachary Lipton. Ratt: Leveraging unlabeled data to guarantee generalization. In *International Conference on Machine Learning*, pp. 3598–3609. PMLR, 2021.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

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

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013.

Jie Li, Xiaorui Wang, Yan Li, et al. The speechtransformer for large-scale mandarin chinese speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7095–7099. IEEE, 2019.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Chao Ma, Lei Wu, et al. A priori estimates of the population risk for two-layer neural networks. *arXiv preprint arXiv:1810.06397*, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pp. 1376–1401. PMLR, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

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

Stephan R Sain. The nature of statistical learning theory, 1996.

Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

E Weinan, Chao Ma, and Qingcan Wang. A priori estimates of the population risk for residual networks. *arXiv preprint arXiv:1903.02154*, 1(7), 2019.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pp. 7404–7413. PMLR, 2019.