# SemiFL: Semi-Supervised Federated Learning for Unlabeled Clients with Alternate Training

**Enmao Diao**
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27705, USA
enmao.diao@duke.edu

**Jie Ding**
School of Statistics
University of Minnesota-Twin Cities
Minneapolis, MN 55455, USA
dingj@umn.edu

**Vahid Tarokh**
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27705, USA
vahid.tarokh@duke.edu

## Abstract

Federated Learning allows the training of machine learning models by using the computation and private data resources of many distributed clients. Most existing results on Federated Learning (FL) assume the clients have ground-truth labels. However, in many practical scenarios, clients may be unable to label task-specific data due to a lack of expertise or resource. We propose SemiFL to address the problem of combining communication-efficient FL such as FedAvg with Semi-Supervised Learning (SSL). In SemiFL, clients have completely unlabeled data and can train multiple local epochs to reduce communication costs, while the server has a small amount of labeled data. We provide a theoretical understanding of the success of data augmentation-based SSL methods to illustrate the bottleneck of a vanilla combination of communication-efficient FL with SSL. To address this issue, we propose alternate training to 'fine-tune global model with labeled data' and 'generate pseudo-labels with the global model.' We conduct extensive experiments and demonstrate that our approach significantly improves the performance of a labeled server with unlabeled clients training with multiple local epochs. Moreover, our method outperforms many existing SSFL baselines and performs competitively with the state-of-the-art FL and SSL results. Our code is available here.

## 1 Introduction

For billions of users around the world, mobile devices and Internet of Things (IoT) devices are becoming common computing platforms [1]. These devices produce a large amount of data that can be used to improve a variety of existing applications [2]. Consequently, it has become increasingly appealing to process data and train models locally from privacy and economic standpoints. To address this, distributed machine learning framework of Federated Learning (FL) has been proposed [3, 4]. This method aggregates locally trained model parameters in order to produce a global inference model without sharing private local data.

Most existing works of FL focus on supervised learning tasks assuming that clients have ground-truth labels. However, in many practical scenarios, most clients may not be experts in the task of interest to label their data. In particular, the private data of each client may be completely unlabeled. For instance, a healthcare system may involve a central hub ("server") with domain experts and a

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

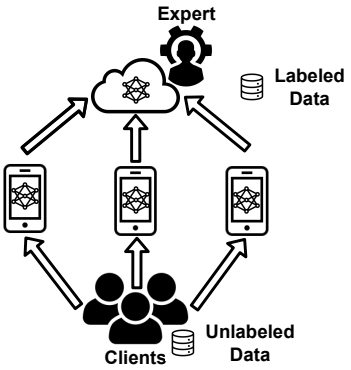


Figure 1: A resourceful server with labeled data can significantly improve its learning performance by working with distributed clients with unlabeled data without data sharing.

limited number of labeled data (such as medical records), together with many rural branches with non-experts and a massive number of unlabeled data. As another example, an autonomous driving startup ("server") may only afford beta-users assistance in labeling a road condition but desires to improve its modeling quality with the information provided by many decentralized vehicles that are not beta-users. The above scenarios naturally lead to the following important question: *How a server that hosts a labeled dataset can leverage clients with unlabeled data for a supervised learning task in the Federated Learning setting?*

We propose a new Federated Learning framework SemiFL to address the problem of Semi-Supervised Federated Learning (SSFL) as illustrated in Figure 1. We discover that it is challenging to directly combine the state-of-the-art SSL methods with the communication efficient federated learning methods such as FedAvg to allow local clients to train multiple epochs [4]. The key ingredient that enables SemiFL to allow unlabeled clients to train multiple local epochs is that we alternate the training of a labeled server and unlabeled clients to ensure that the quality of pseudo-labeling is highly maintained during training. In particular, we fine-tune the global model with labeled data and generate pseudo-labels only with the global model. We perform extensive empirical experiments to evaluate and compare our method with various baselines and state-of-the-art techniques. The results demonstrate that our method can outperform existing SSFL methods and perform close to the state-of-the-art of FL and SSL results. In particular, we contribute the following.

- We propose SemiFL in which clients have completely unlabeled data and can train multiple local epochs to reduce communication costs, while the server has a small amount of labeled data. We identify the difficulty of combining communication efficient FL method FedAvg [4] with the state-of-the-art SSL methods.
- We develop a theoretical analysis on strong data augmentation for SSL methods, the first in the literature to our best knowledge. We provide a theoretical understanding of the success of data augmentation-based SSL methods to illustrate the bottleneck of a vanilla combination of communication-efficient FL with SSL.
- To the best of our knowledge, we propose the first communication efficient SSFL method alternate training that can improve the performance of a labeled server by allowing unlabeled clients to train multiple local epochs, i.e., from $42\%$ to $88\%$ with $250$ labeled data, and from $77\%$ to $93\%$ accuracy with $4000$ labeled data on the CIFAR10 dataset. Moreover, our proposed method achieves $30\%$ improvement over the existing SSFL methods. Furthermore, SemiFL performs competitively with the state-of-the-art FL methods and SSL methods. i.e., only $1\%$ and $2\%$ away from the state-of-the-art FL and SSL results, respectively, for $4000$ labeled data on the CIFAR10 dataset.

The outline of the paper is given below. In Section 2, we review the related work. In Section 3, we identify the problem of combining SSL with communication-efficient FL methods, develop a theoretical analysis of how strong data augmentation can significantly improve the classification accuracy, and present the proposed SemiFL method with some intuitive explanations. In Section 4, we evaluate the empirical performance of the SemiFL. We make some concluding remarks in Section 5.

## 2 Related Work

**Federated Learning** The goal of Federated Learning is to scale and speed up the training of distributed models [5, 6]. FedAvg [4] allows local clients to train multiple epochs to facilitate convergence. FedProx (Li et al., 2018) performs proximal regularization against global weights. FL counterparts of Batch Normalization [7–9] are developed to further enhance the performance. The use of local momentum and global momentum [10] have been shown to facilitate faster convergence. FedOpt [11] proposes federated versions of adaptive optimizers to improve performance over FedAvg.

**Semi-Supervised Learning** Semi-Supervised Learning (SSL) refers to the general problem of learning with partially labeled data, especially when the amount of unlabeled data is much larger than that of the labeled data [12, 13]. The idea of self-training (namely to obtain artificial labels for unlabeled data from a pre-trained model) can be traced back to decades ago [14, 15]. Pseudo-labeling [16], a component of many recent SSL techniques [17], is a form of entropy minimization [18] by converting model predictions into hard labels. Consistency regularization [19] refers to training models via minimizing the distance among stochastic outputs [13, 19]. A theoretical analysis of consistency regularization was recently developed in [20]. More recently, It has been demonstrated that the technique of strong data augmentation can lead to better outcomes [21–24]. Strongly augmented examples are frequently found outside of the training data distribution, which has been shown to benefit SSL [25].

**Semi-Supervised Federated Learning (SSFL)** Most existing FL works focus on supervised learning tasks, with clients having ground-truth labels. However, in many real-world scenarios, most clients are unlikely to be experts in the task of interest, an issue raised in a recent survey paper [26]. In the research line of SSFL, the work [27] splits model parameters for labeled server and unlabeled clients separately. Another related work [28] trains and aggregates the model parameters of the labeled server and unlabeled clients in parallel with group-wise reweights. Applications of SSFL to specific applications can be found in, e.g., [29, 30].

## 3 Method

### 3.1 Problem

In a supervised learning classification task, we are given a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$, where $x_i$ is a feature vector, $y_i$ is an one-hot vector representing the class label in a $K$-class classification problem, and $N$ is the number of training examples. In a Semi-Supervised Learning classification task, we have two datasets, namely a supervised dataset $\mathcal{S}$ and an unsupervised dataset $\mathcal{U}$. Let $\mathcal{S} = \{x_s^i, y_s^i\}_{i=1}^{N_\mathcal{S}}$ be a set of $N_\mathcal{S}$ labeled data observations, and $\mathcal{U} = \{x_u^i\}_{i=1}^{N_\mathcal{U}}$ be a set of $N_\mathcal{U}$ unlabeled observations (without the corresponding true label $y_u^i$). It is often interesting to study the case where $N_\mathcal{S} \ll N_\mathcal{U}$.

In this work, we focus on Semi-Supervised Federated Learning (SSFL) with unlabeled clients, as illustrated in Figure 1. Assume $M$ clients and let $x_{u,m}$ denote the set of unsupervised data available at client $m = 1, 2, \cdots, M$. Similarly, let $(x_s, y_s)$ denote the set of labeled data available at the server. The server model is parameterized by model parameters $W_s$. The client models are parameterized respectively by model parameters $\{W_{u,1}, \ldots, W_{u,M}\}$. We assume that all models share the same model architecture, denoted by $f : (x, w) \mapsto f(x, w)$, which maps an input $x$ and parameters $W$ to a vector on the $K$-dimensional simplex, e.g., using softmax function applied to model outputs.

**Communication Efficient FL with SSL** In the standard communication efficient FL scenario where clients can train multiple local epochs before model aggregation (i.e., FedAvg [4]), existing SSFL methods have difficulty in performing close to the state-of-the-art centralized SSL methods [27,28,31]. In fact, we will demonstrate in Table 1 that existing SSFL methods cannot outperform the case of training with only labeled data. This is somewhat surprising given that their underlying methods of training unlabeled data are similar.

As shown in Figure 2, SSL methods such as FixMatch can only work with FedSGD, which requires batch-wise gradient aggregation and thus is not communication efficient. This is because SSL methods, such as FixMatch and MixMatch, sample from both labeled and unlabeled datasets for every batch of training data with a carefully tuned ratio [23, 32]. Thus, it is not straightforward how we can combine the SSL method in a communication-efficient FL scenario where we train multiple local epochs. To understand the bottleneck of this vanilla combination, we need to understand better

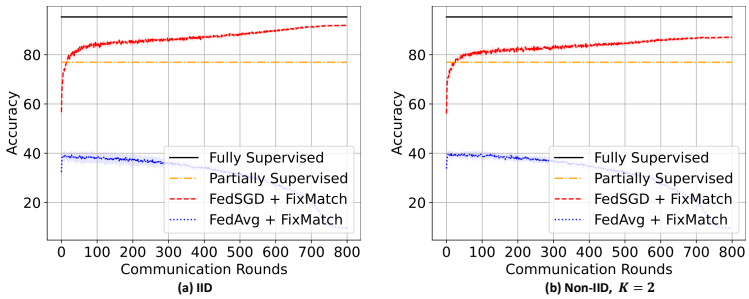


Figure 2: Results of CIFAR10 dataset with (a) IID and (b) Non-IID, $K = 2$ data partition and $N_{\mathcal{S}} = 4000$ with a vanilla combination of communication efficient FL with SSFL methods. The "Fully Supervised" and "Partially Supervised" refer to training a centralized model with full and $4000$ labeled data respectively.

why the state-of-the-art centralized SSL methods work. In section 3.2, we analyze the strong data augmentation for SSL and demonstrate that the success of FixMatch is due to using data augmentation on pseudo-labeled data of high quality.

### 3.2 Theoretical Analysis of Strong Data augmentation for SSL

Pseudo-labeling is widely used for labeling unlabeled data in SSL methods [16, 22, 32]. However, the quality (accuracy) of those pseudo-labels can be low, especially at the beginning of the training. In this light, several papers [22, 32] propose to hard-threshold or sharpen the pseudo-labels to improve the quantity of accurately labeled pseudo-labels. The problem with hard thresholding is that the data samples satisfying the confidence threshold have a small training loss. Therefore, the model cannot be significantly improved as it already performs well on the data above the threshold. To address this issue, we will use strong data augmentation [25, 32] to generate data samples that have larger training loss. The main idea is to construct a pseudo-labeling mechanism whereby our SSL method can generate more and more high-quality pseudo-labels during training. Meanwhile, the augmented data for model training can produce a more considerable drop in training loss than the original data.

To provide further insights into SSL, we develop a theoretical analysis of the strong data augmentation, which is a critical component of the state-of-the-art SSL method FixMatch [32] and SemiFL, and can be interesting in its own right. Intuitively, strong augmentation is a process that maps a data point (e.g., an image) from high quality to relatively low grade unilaterally. The low-quality data and their high-confidence pseudo-labels are then used for training so that there are sufficient "observations" in the data regime insufficiently covered by labeled data.

Our theory is based on an intuitive "adequate transmission" assumption, which means that the distribution of augmented data from high-confidence unlabeled data can adequately cover the data regime of interest during prediction. Consequently, reliable information exhibited from unlabeled data can be "transmitted" to data regimes that may have been insufficiently trained with labeled data, as illustrated in Figure 3. Instead of studying SSL in full generality, we restrict our attention to a class of nonparametric kernel-based classification learning [33–35] and derive analytically tractable statistical risk-rate analysis. More detailed background and technical details are included in the Appendix. We provide a simplified statement as follows.

**Theorem 1** *Under suitable assumptions, an SSL classifier $\hat{C}^{ssl}$ trained from $n_u$ unlabeled data and the strong data augmentation technique has a statistical risk bound at the order of $\mathcal{R}(\hat{C}^{ssl}) \sim n_u^{-q(\alpha+1)/\{q(\alpha+3+\rho)+d\}}$ where d, q, $\alpha$, $\rho$ are constants that describe the data dimension, smoothness of the conditional distribution function $(Y \mid X)$, class separability (or task difficulty), and inadequacy of transmission, respectively. The smaller $\rho$, the better risk bound. Moreover, suppose that $\hat{C}^l$ is the classifier trained from $n_l$ labeled data, where $n_l \sim n_u^{\zeta}$, $\zeta \in (0, 1)$. It can be verified that the bound of $\mathcal{R}(\hat{C}^u)$ is much smaller than that of $\mathcal{R}(\hat{C}^l)$ when $\zeta < \frac{q(\alpha+3)+d}{q(\alpha+3+\rho)+d}$. This provides an insight into the critical region of $n_u$ where significant improvement can be made from unlabeled data.*

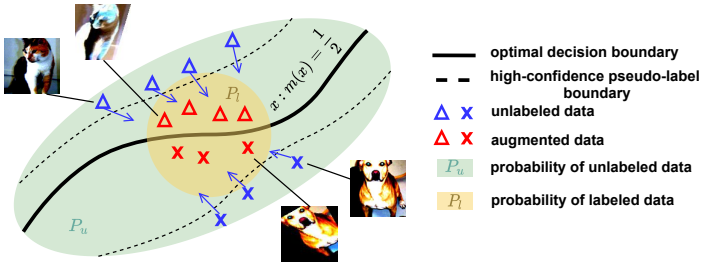


Figure 3: Illustration of the strong data augmentation-based SSL. We pick up an unlabeled point ($X \sim \mathbb{P}_{\text{u}}$) with a high-confidence pseudo-label, obtain its hard-thresholded label ($\hat{Y}$, which is believed to be close to the ground truth), maneuver $X$ into $\tilde{X}$ (which is believed to represent the test distribution $\mathbb{P}_{\text{l}}$ to some extent), and then treat $(\hat{Y}, \tilde{X})$ as labeled data for training. Consequently, reliable task-specific information exhibited from unlabeled data can be transmitted to data regimes that may have been insufficiently trained with labeled data. Note that $\mathbb{P}_{\text{l}}$ denotes the labeled data distribution as well as the out-sample test data distribution (used to evaluate the learning performance). The above ideas are theoretically formalized in Subsection 3.2 and Appendix D.

### 3.3 Alternate Training

As depicted in Figure 4(a), existing SSFL works follow the state-of-the-art SSL methods to synchronize the training of supervised and unsupervised data [23, 32]. For example, FedMatch [27] and FedRGD [28] adopt a vanilla combination of FedAvg and FixMatch. They aggregate the server model trained from labeled data and clients' models trained from unlabeled data at each communication round in parallel and generate pseudo-labels for each batch of unlabeled data with the local training model. However, existing results [27, 28] indicate that this vanilla combination has difficulty performing close to the state-of-the-art SSL methods, even if the unlabeled clients are trained with the same SSL methods.

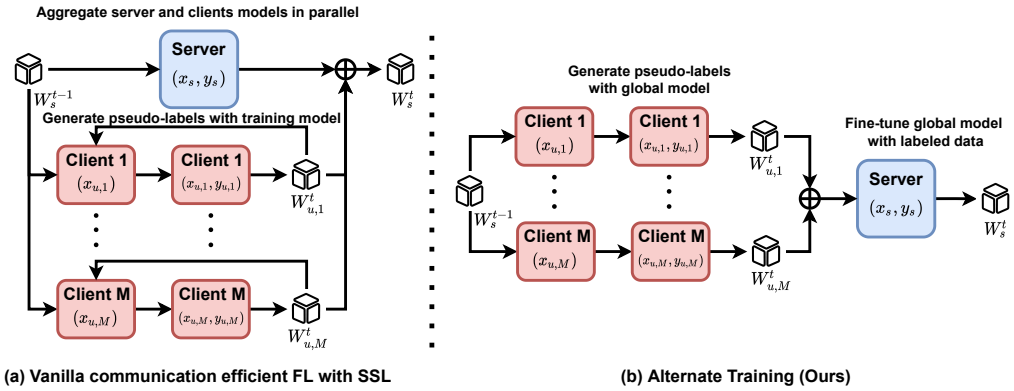


Figure 4: An illustration of (a) vanilla combination of communication efficient FL and SSL, and (b) Alternate Training (Ours). (a) The vanilla combination trains and aggregates server and client models in parallel and generates pseudo-labels with the training models for every batch of unlabeled data. (b) Alternate Training fine-tunes the aggregated global model with labeled data and generates pseudo-labels only once upon receiving the global model from the server.

In the communication efficient FL setting, we cannot guarantee an increase in the quality of pseudo-labels during training because we allow local clients to train multiple epochs, potentially deteriorating the performance (see Figure 2). Furthermore, the aggregation of a server model trained with ground-truth labels and a subset of client models trained with pseudo-labels does not constantly improve the performance of the global model over the previous communication round. A poorly aggregated global model of the last round produces worse-quality pseudo-labels. Subsequently, the performance of the aggregated model degrades in the next round.

To maintain and improve the quality of our generated pseudo-labels during training, we propose to train the labeled server and unlabeled clients in an alternate manner, as illustrated in Figure 4(b). In particular, our approach consists of two important components:

• **Fine-tune global model with labeled data** At each round, the server will retrain the global model with the labeled data. In this way, the server can provide a comparable or better model than the previous round for the active clients in the next round to generate pseudo-labels. On the contrary, the vanilla method aggregates server and client models in parallel. As a result, the quality of generated pseudo-labels will gradually degrade and thus deteriorate the performance.

• **Generate pseudo-labels with global model** We will label the unlabeled data once the active clients immediately receive the global model from the server. This way, pseudo-labels' quality will not degrade during the local training. On the contrary, the vanilla method labels every batch of data during the training of unlabeled clients. As a result, the quality of generated pseudo-labels will gradually degrade during local training, thus deteriorating performance.

Our proposed approach ensures that the clients can continually generate better quality pseudo-labels during training. Our experimental studies show that the proposed method can significantly improve the performance of the labeled server and performs competitively even with the state-of-the-art FL and centralized SSL methods. The limitation of our approach is that we need to update the aggregated client model with labeled data from the server, which will delay the computation time. We will conduct an ablation study on each component of alternative training in Table 2.

### 3.4 The SemiFL Algorithm

We summarize the pseudo-code of the proposed solution in Algorithm 1. At each iteration $t$, the server will first update the model with the standard supervised loss $L_s$ for local epochs $E$ with data batch $(x_b, y_b)$ of size $B_s$ randomly split from the supervised dataset $\mathcal{D}_s$, using

$$L_s = \ell(f(\alpha(x_b), W_s), y_b), \quad W_s = W_s - \eta \nabla_W L_s, \tag{1}$$

where $\alpha(\cdot)$ represents a weak data augmentation, such as random horizontal flipping and random cropping, that maps one image to another. Subsequently, the server updates the static Batch Normalization (sBN) statistics [9] (which is discussed in Appendix B). Next, the server distributes server model parameters $W_s$ to a subset of clients. We denote the proportion of active clients at each communication round $t$ as activity rate $C_t \in (0, 1]$. Without loss of generality, we assume that $C_t = C$ is a constant over time. After each active local client, say client $m$, receives the transmitted $W_s$, it generates pseudo-labels $y_{u,m}$ as follows:

$$W_{u,m} \leftarrow W_s, \quad y_{u,m} = f(\alpha(x_{u,m}), W_{u,m}). \tag{2}$$

Each local client will construct a high-confidence dataset $\mathcal{D}^{\text{fix}}_{u,m}$ inspired by FixMatch [32] at each iteration $t$, defined as:

$$\mathcal{D}^{\text{fix}}_{u,m} = \{(x_{u,m}, y_{u,m}) \ \texttt{with} \ \max(y_{u,m}) \geq \tau\}. \tag{3}$$

for a global confidence threshold $0 < \tau < 1$ pre-selected by all clients. If for some client $m$, we have $\mathcal{D}^{\text{fix}}_{u,m} = \emptyset$ then it will stop and refrain from transmission to the server. Otherwise, we will sample with replacement to construct a dataset inspired by MixMatch [23]. In other words,

$$\mathcal{D}^{\text{mix}}_{u,m} = \ \texttt{Sample} \ |\mathcal{D}^{\text{fix}}_{u,m}| \ \texttt{with replacement}\{(x_{u,m}, y_{u,m})\}, \tag{4}$$

where $|\mathcal{D}^{\text{fix}}_{u,m}|$ denotes the number of elements of $\mathcal{D}^{\text{fix}}_{u,m}$. Thus $|\mathcal{D}^{\text{mix}}_{u,m}| = |\mathcal{D}^{\text{fix}}_{u,m}|$. Subsequently, client $m$ trains its local model for $E$ epoch to speed up convergence [4]. For each local training epoch of the client $m$, it randomly splits local data $\mathcal{D}^{\text{fix}}_{u,m}, \mathcal{D}^{\text{mix}}_{u,m}$ into batches $\mathcal{B}^{\text{fix}}_{u,m}, \mathcal{B}^{\text{mix}}_{u,m}$ of size $B_m$. For each batch iteration, as in [36], client $m$ constructs Mixup data from one particular data batch $(x^{\text{fix}}_b, y^{\text{fix}}_b), (x^{\text{mix}}_b, y^{\text{mix}}_b)$ by

$$\lambda_{\text{mix}} \sim \text{Beta}(a, a), \quad x_{\text{mix}} \leftarrow \lambda_{\text{mix}} x^{\text{fix}}_b + (1 - \lambda_{\text{mix}}) x^{\text{mix}}_b,$$

where $a$ is the Mixup hyperparameter. Next, client $m$ defines the "fix" loss $L_{\text{fix}}$ [32] and "mix" loss $L_{\text{mix}}$ [24] by

$$\begin{aligned} L_{\text{fix}} &= \ell(f(\mathcal{A}(x^{\text{fix}}_b), W_{u,m}), y^{\text{fix}}_b), \\ L_{\text{mix}} &= \lambda_{\text{mix}} \cdot \ell(f(\alpha(x_{\text{mix}}), W_{u,m}), y^{\text{fix}}_b) + (1 - \lambda_{\text{mix}}) \cdot \ell(f(\alpha(x_{\text{mix}}), W_{u,m}), y^{\text{mix}}_b)). \end{aligned} \tag{5}$$

**Algorithm 1** Semi-Supervised Federated Learning with Alternate Training for Unlabeled Clients

**Input:** Unlabeled data $x_{u,1:M}$ distributed on $M$ local clients, activity rate $C$, the number of communication rounds $T$, the number of local training epochs $E$, server and client respective batch sizes $B_s$ and $B_m$, local learning rate $\eta$, server model parameterized by $W_s$ client models parameterized by $\{W_{u,1}, \ldots, W_{u,M}\}$, weak data augmentation function $\alpha(\cdot)$, strong data augmentation function $\mathcal{A}(\cdot)$, confidence threshold $\tau$, Mixup hyper-parameter $a$, loss hyperparameter $\lambda$, common model architecture function $f(\cdot)$

**System executes:**

```
for each communication round t = 1, 2, ... T do
    W_s^t ← ServerUpdate(x_s, y_s, W_s^t)
    Update the sBN statistics
    S_t ← max(⌊C · M⌋, 1) active clients uniformly sampled without replacement
    for each client m ∈ S_t in parallel do
        Distribute server model parameters to local client m, namely W_{u,m}^t ← W_s^t
        W_{u,m}^t ← ClientUpdate(x_{u,m}, W_{u,m}^t)
    end
    Receive model parameters from M_t clients, and calculate W_s^t = M_t^{-1} Σ_{m=1}^{M_t} W_{u,m}^t
end
W_s^T ← ServerUpdate(x_s, y_s, W_s^T)
Update the sBN statistics
```

**ServerUpdate** $(x_s, y_s, W_s)$**:**

```
Construct supervised dataset D_s = (x_s, y_s)
for each local epoch e from 1 to E do
    B_s ← Randomly split local data D_s into batches of size B_s
    for batch (x_b, y_b) ∈ B_s do
        L_s ← ℓ(f(α(x_b), W_s), y_b)
        W_s ← W_s − η∇_W L_s
    end
end
Return W_s
```

**ClientUpdate** $(x_{u,m}, W_{u,m})$**:**

```
Generate pseudo-labels with weakly augmented data α(x_{u,m}), namely y_{u,m} =
  f(α(x_{u,m}), W_{u,m})
Construct FixMatch dataset, namely D_{u,m}^fix = {(x_{u,m}, y_{u,m}) with max(y_{u,m}) ≥ τ}
If D_{u,m}^fix = ∅ then Stop. Return.
Construct an equal-size Mixup dataset, namely
D_{u,m}^mix = Sample |D_{u,m}^fix| with replacement{(x_{u,m}, y_{u,m})}
for each local epoch e from 1 to E do
    B_{u,m}^fix, B_{u,m}^mix ← Randomly split local data D_{u,m}^fix, D_{u,m}^mix into batches of size B_m^fix, B_m^mix
    for batch (x_b^fix, y_b^fix), (x_b^mix, y_b^mix) ∈ B_{u,m}^fix, B_{u,m}^mix do
        λ_mix ~ Beta(a, a)
        x_mix ← λ_mix x_b^fix + (1 − λ_mix) x_b^mix
        L_fix ← ℓ(f(A(x_b^fix), W_{u,m}), y_b^fix)
        L_mix ← λ_mix · ℓ(f(α(x_mix), W_{u,m}), y_b^fix) + (1 − λ_mix) · ℓ(f(α(x_mix), W_{u,m}), y_b^mix))
        W_{u,m} ← W_{u,m} − η∇_W (L_fix + λ · L_mix)
    end
end
Return W_{u,m} and send it to the server
```

Here, $\mathcal{A}$ represents a strong data augmentation mapping, e.g., the RandAugment [37] used in our experiments, and $\ell$ is often the cross entropy loss for classification tasks. Finally, client $m$ performs a gradient descent step with

$$W_{u,m} = W_{u,m} - \eta \nabla_W (L_{\text{fix}} + \lambda \cdot L_{\text{mix}}), \tag{6}$$

where $\lambda > 0$ is a hyperparameter set to be one in our experiments. After training for $E$ local epochs, client $m$ transmits $W_{u,m}$ to the server.

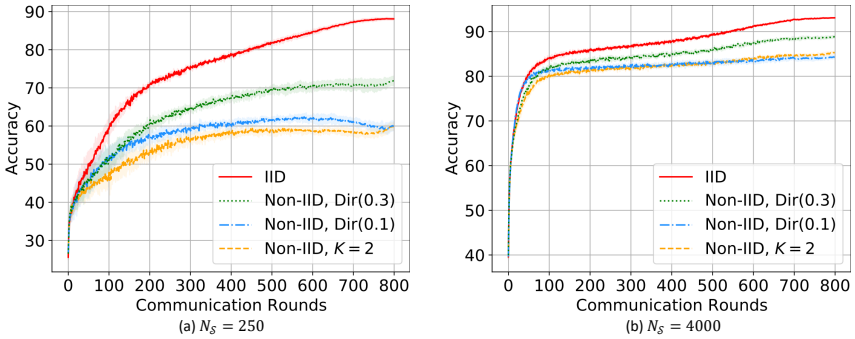


Figure 5: Results of CIFAR10 dataset with (a) $N_{\mathcal{S}} = 250$ and (b) $N_{\mathcal{S}} = 4000$.

Without loss of generality, assume that clients $1, 2, \cdots, M_t$ have sent their models to the server at time $t$. The server then aggregates client model parameters $\{W_{u,1}, \ldots, W_{u,M_t}\}$ by $W_s = M_t^{-1} \sum_{m=1}^{M_t} W_{u,m}$ [4]. This process is then repeated for multiple communication rounds $T$. After the training is finished, the server will further fine-tune the aggregated global model by additional training with the server's supervised data using its supervised loss $L_s$. Finally, it will update the sBN statistics one final time.

# 4 Experiments

**Experimental setup** To evaluate our proposed method, we conduct experiments with CIFAR10, SVHN, and CIFAR100 datasets [38,39]. To compare our method with existing FL and SSFL methods, we follow the standard communication efficient FL setting, which was originally used in FedAvg [4] and widely adopted by following works, such as [9, 40, 41]. We have 100 clients throughout our experiments, and the activity rate per communication round is $C = 0.1$. We uniformly assign the same number of data examples for IID data partition to each client. For a balanced Non-IID data partition, we ensure each client has data at most from $K$ classes and the sample size of each class is the same. We set $K = 2$ because it is the most label-skewed case for classification, and it has been evaluated in [9, 40, 41]. For unbalanced Non-IID data partition, we sample data for each client from a Dirichlet distribution $\mathrm{Dir}(\alpha)$ [41, 42]. As $\alpha \to \infty$, it reduces to IID data partition. We perform experiments with $\alpha = \{0.1, 0.3\}$.

To compare our method with the state-of-the-art SSL methods, we follow the experimental setup in [32]. We use Wide ResNet28x2 [43] as our backbone model for CIFAR10 and SVHN datasets and Wide ResNet28x8 for CIFAR100 datasets throughout our experiments. The number of labeled data at the server for CIFAR10, SVHN, and CIFAR100 datasets $N_{\mathcal{S}}$ are $\{250, 4000\}$, $\{100, 2500\}$, and $\{2500, 10000\}$ respectively. We conduct four random experiments for all the datasets with different seeds, and the standard errors are shown inside the parentheses for tables and by error bars in the figures. We demonstrate our experimental results in Table 1 and the learning curves of CIFAR10, SVNH, and CIFAR100 datasets in Figure 5, 7, and 8. Further details are included in the Appendix.

**Comparison with SSL methods** We demonstrate the results of Fully Supervised and Partially Supervised cases and existing SSL methods for comparison in Table 1. The Fully Supervised case refers to all data being labeled, while in the Partially Supervised case, we only train the model with the partially labeled $N_{\mathcal{S}}$ data. Our results significantly outperform the Partially Supervised case. In other words, SemiFL can substantially improve the performance of a labeled server with unlabeled clients in a communication-efficient scenario. Our method performs competitively with the state-of-the-art SSL methods for IID data partition. Moreover, it is foreseeable that as the clients become more label-skewed for Non-IID data partition, the performance of our method degrades. However, even the most label-skewed unlabeled clients can improve the performance of the labeled server using our approach. A limitation of our work is that as the supervised data size decreases, the performance of SemiFL degrades more than the centralized SSL methods. We believe it is because we cannot train labeled and unlabeled data simultaneously in one data batch.

**Comparison with FL and SSFL methods** We compare our results with the state-of-the-art FL and SSFL methods in Table 1. We demonstrate that SemiFL can perform competitively with the state-of-the-art FL result trained with fully supervised data. It is worth mentioning that SSFL may

Table 1: Comparison of SemiFL with the Baselines, SSL, FL, and SSFL methods. SemiFL improves the performance of the labeled server, SemiFL significantly outperforms the existing SSFL methods, and performs close to the state-of-the-art FL and SSL methods.

| Dataset | | | CIFAR10 | | SVHN | | CIFAR100 | |
|---|---|---|---|---|---|---|---|---|
| Number of Supervised | | | 250 | 4000 | 250 | 1000 | 2500 | 10000 |
| Baseline | | Fully Supervised | 95.3(0.1) | | 97.3(0.0) | | 79.3(0.1) | |
| | | Partially Supervised | 42.4(1.8) | 76.9(0.2) | 77.1(2.9) | 90.4(0.5) | 27.2(0.7) | 59.3(0.1) |
| SSL | | Π-Model [13] | 45.7(4.0) | 86.0(0.4) | 81.0(1.9) | 92.5(0.4) | 42.8(0.5) | 62.1(0.1) |
| | | Pseudo-Labeling [44] | 50.2(0.4) | 83.9(0.3) | 79.8(1.1) | 90.1(0.6) | 42.6(0.5) | 63.8(0.2) |
| | | Mean Teacher [44] | 67.7(2.3) | 90.8(0.2) | 96.4(0.1) | 96.6(0.1) | 46.1(0.6) | 64.2(0.2) |
| | | MixMatch [23] | 89.0(0.9) | 93.6(0.1) | 96.0(0.2) | 96.5(0.3) | 60.1(0.4) | 71.7(0.3) |
| | | UDA [22] | 91.2(1.1) | 95.1(0.2) | 94.3(2.8) | 97.5(0.2) | 66.9(0.2) | 75.5(0.3) |
| | | ReMixMatch [24] | 94.6(0.1) | 95.3(0.1) | 97.1(0.5) | 97.4(0.1) | 72.6(0.3) | 77.0(0.6) |
| | | FixMatch [32] | 94.9(0.7) | 95.7(0.1) | 97.5(0.4) | 97.7(0.1) | 71.7(0.1) | 77.4(0.1) |
| Non-IID, $K = 2$ | FL | HeteroFL [9] | 51.5(3.6) | | 72.3(4.4) | | 3.1(0.3) | |
| | SSFL | FedMatch [27] | 41.3(1.1) | 58.3(1.0) | 58.2(3.1) | 84.3(1.0) | 17.7(0.5) | 30.5(0.8) |
| | | FedRGD [28] | 32.7(3.6) | 48.9(1.4) | 21.2(2.2) | 21.6(2.3) | 13.8(1.4) | 26.5(3.0) |
| | | SemiFL | **60.0(0.9)** | **85.3(0.3)** | **87.5(1.1)** | **92.2(0.8)** | **35.2(0.3)** | **62.1(0.4)** |
| Non-IID, $\mathrm{Dir}(0.1)$ | FL | HeteroFL [9] | 85.0(0.6) | | 95.8(0.1) | | 74.0(0.4) | |
| | SSFL | FedMatch [27] | 41.6(1.0) | 58.9(0.7) | 58.4(3.4) | 84.3(0.6) | 17.5(0.5) | 30.8(0.6) |
| | | FedRGD [28] | 31.5(2.9) | 45.2(0.8) | 20.0(4.0) | 23.8(3.4) | 13.4(1.3) | 23.6(2.6) |
| | | SemiFL | **63.0(0.6)** | **84.5(0.4)** | **91.2(0.3)** | **93.0(0.5)** | **49.0(1.0)** | **68.0(0.2)** |
| Non-IID, $\mathrm{Dir}(0.3)$ | FL | HeteroFL [9] | 91.6(0.1) | | 96.8(0.0) | | 76.9(0.1) | |
| | SSFL | FedMatch [27] | 41.2(1.1) | 58.4(0.6) | 59.1(2.8) | 84.0(1.1) | 17.8(0.4) | 31.1(0.5) |
| | | FedRGD [28] | 32.5(3.0) | 46.9(1.6) | 24.8(5.1) | 22.0(3.9) | 13.1(2.0) | 23.8(1.9) |
| | | SemiFL | **71.9(1.2)** | **88.9(0.3)** | **94.0(0.5)** | **95.2(0.2)** | **54.9(1.4)** | **70.0(0.3)** |
| IID | FL | HeteroFL [9] | 94.3(0.1) | | 97.5(0.0) | | 77.8(0.2) | |
| | SSFL | FedMatch [27] | 41.7(1.1) | 58.6(0.5) | 58.6(3.0) | 84.3(0.9) | 17.6(0.3) | 31.3(1.0) |
| | | FedRGD [28] | 33.2(1.9) | 47.8(1.7) | 21.3(6.5) | 20.7(1.1) | 13.3(1.4) | 23.8(2.6) |
| | | SemiFL | **88.2(0.3)** | **93.1(0.1)** | **96.8(0.3)** | **96.9(0.1)** | **61.3(1.2)** | **72.1(0.2)** |

outperform FL methods in the Non-IID data partition case because the server has a small set of labeled IID data. We also demonstrate that our method significantly outperforms existing SSFL methods. Existing SSFL methods fail to perform closely to the state-of-the-art centralized SSL methods, even if their underlying SSL methods are the same. Moreover, existing SSFL methods cannot outperform the Partially Supervised case, indicating that they deteriorate the performance of the labeled server. In particular, FedMatch allocates disjoint model parameters for the server and clients, and FedRGD assigns a higher weight for the server model for aggregation. Both methods do not directly fine-tune the global model with labeled data and generate pseudo-labels with the received global model. To our best knowledge, *the proposed SemiFL is the first SSFL method that actually improves the performance of the labeled server and performs close to the state-of-the-art FL and SSL methods.*

**Ablation studies** We conduct ablation studies on SemiFL and demonstrate the results in Table 2. Based on our extensive experiments, it is evident that "Fine-tune global model with labeled data" and "Generate pseudo-labels with global model" are the critical components of the proposed 'Alternate Training' method for the success of our method. We also conduct an ablation study on static Batch Normalization(sBN), the number of local training epochs, the Mixup data augmentation, and global SGD momentum. The detailed results can be found in the Appendix.

### 4.1 Quality of Pseudo Labeling

We measure the quality of Pseudo-Labeling for Semi-Supervised Learning from three aspects, including the accuracy of pseudo-labels (Pseudo Accuracy), the accuracy of thresholded pseudo-labels (Threshold Accuracy), and the ratio of pseudo-labeled data (Label Ratio) with CIFAR10 dataset in Figure 6. We perform ablation studies on our proposed method by measuring the quality of Pseudo-Labeling. The results demonstrate that our proposed alternative training, the combination of 'Fine-tune global model with labeled data' and 'Generate pseudo-labels with global model,' can produce pseudo labels of much better quality when clients have completely unlabeled data and train multiple local epochs.

Table 2: Ablation study on each component of alternative training with CIFAR10 dataset. The combination of "Fine-tune global model with labeled data" and "Generate pseudo-labels with global model" significantly improves the performance.

| Method | Fine-tune global model with labeled data | Generate pseudo-labels with global model | Accuracy | |
|---|---|---|---|---|
| | | | Non-IID, $K = 2$ | IID |
| Fully Supervised | N/A | | 95.33 | |
| Partially Supervised | | | 76.92 | |
| FedAvg + FixMatch | ✗ | ✗ | 41.01 | 40.26 |
| SemiFL | ✗ | ✓ | 48.89 | 47.03 |
| | ✓ | ✗ | 80.42 | 81.70 |
| | ✓ | ✓ | **85.34** | **93.10** |

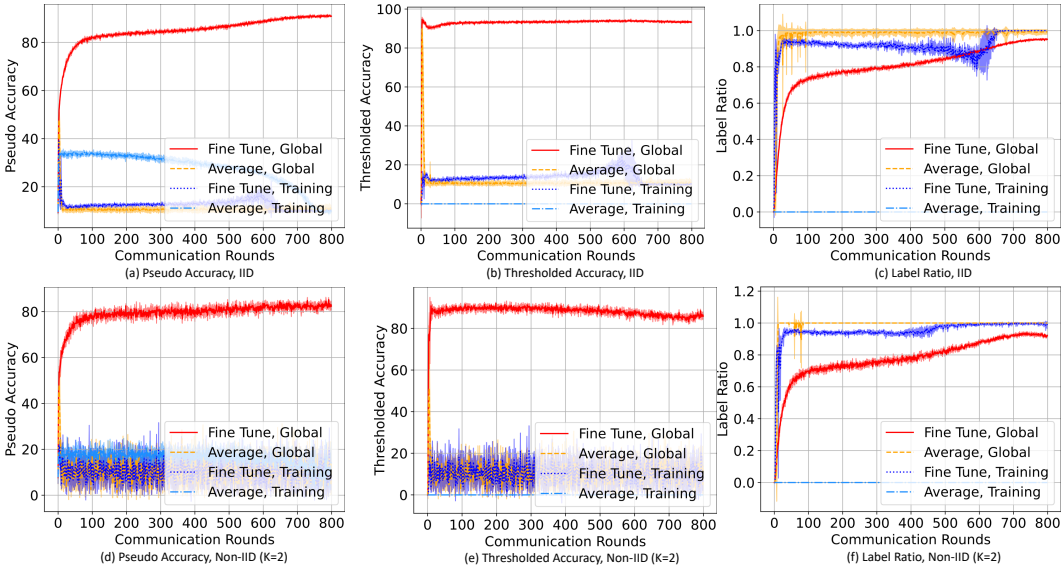


Figure 6: Ablation studies of alternative training by measuring the quality of Pseudo Labeling with CIFAR10 dataset. 'Fine Tune' and 'Global' refer to our proposed method, 'Fine-tune global model with labeled data' and 'Generate pseudo-labels with global model,' respectively. 'Average' refers to the vanilla FL method, which directly takes the average of the model parameters of the labeled server and unlabeled clients. 'Training' refers to generating pseudo-labels at each batch of local training.

## 5 Conclusion

In this work, we propose a new communication-efficient Federated Learning (FL) framework named SemiFL to address the problem of Semi-Supervised Federated Learning (SSFL) for unlabeled clients. We identify the difficulty of combining communication-efficient Federated Learning (FL) with state-of-the-art Semi-Supervised Learning (SSL). We develop a theoretical analysis of strong data augmentation for SSL, which illustrates the bottleneck of vanilla combination. We propose to train the labeled server and unlabeled clients in an alternate manner by 'fine-tune global model with labeled data' and 'generate pseudo-labels with global model.' We utilize several training techniques and establish a strong benchmark for SSFL. Extensive experimental studies demonstrate that our communication-efficient method can significantly improve the performance of a labeled server with unlabeled clients. Moreover, we show that SemiFL can perform competitively with the state-of-the-art centralized SSL and fully supervised FL methods. Our study provides a practical SSFL framework that extends the scope of FL applications.

## Acknowledgments

The work of Enmao Diao and Vahid Tarokh was supported by the Office of Naval Research (ONR) under grant number N00014-18-1-2244. The work of Jie Ding was supported by the National Science Foundation (NSF) under grant number DMS-2134148.

## References


[1] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 2020.

[2] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[3] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[4] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[5] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*, 2019.

[6] Chaoyang He, Songze Li, Jinhyun So, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, Li Shen, Peilin Zhao, Yan Kang, Yang Liu, Ramesh Raskar, Qiang Yang, Murali Annavaram, and Salman Avestimehr. Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518*, 2020.

[7] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020.

[8] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

[9] Enmao Diao, Jie Ding, and Vahid Tarokh. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. In *International Conference on Learning Representations*, 2021.

[10] Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. Slowmo: Improving communication-efficient distributed sgd with slow momentum. *arXiv preprint arXiv:1910.00643*, 2019.

[11] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

[12] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005.

[13] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder networks. *arXiv preprint arXiv:1507.02672*, 2015.

[14] Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

[15] Geoffrey J McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369, 1975.

[16] 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, 2013.

[17] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[18] Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. In *CAP*, pages 281–296, 2005.

[19] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *arXiv preprint arXiv:1412.4864*, 2014.

[20] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021.

[21] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

[22] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[23] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

[24] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.

[25] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Ruslan Salakhutdinov. Good semi-supervised learning that requires a bad gan. *arXiv preprint arXiv:1705.09783*, 2017.

[26] Yilun Jin, Xiguang Wei, Yang Liu, and Qiang Yang. Towards utilizing unlabeled data in federated learning: A survey and prospective. *arXiv e-prints*, pages arXiv–2002, 2020.

[27] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. In *International Conference on Learning Representations*, 2021.

[28] Zhengming Zhang, Yaoqing Yang, Zhewei Yao, Yujun Yan, Joseph E. Gonzalez, Kannan Ramchandran, and Michael W. Mahoney. Improving semi-supervised federated learning by reducing the gradient diversity of models. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1214–1225, 2021.

[29] Yuchen Zhao, Hanyang Liu, Honglin Li, Payam Barnaghi, and Hamed Haddadi. Semi-supervised federated learning for activity recognition. *arXiv preprint arXiv:2011.00851*, 2020.

[30] Dong Yang, Ziyue Xu, Wenqi Li, Andriy Myronenko, Holger R Roth, Stephanie Harmon, Sheng Xu, Baris Turkbey, Evrim Turkbey, Xiaosong Wang, et al. Federated semi-supervised learning for covid region segmentation in chest ct using multi-national data from china, italy, japan. *Medical image analysis*, 70:101992, 2021.

[31] Zewei Long, Liwei Che, Yaqing Wang, Muchao Ye, Junyu Luo, Jinze Wu, Houping Xiao, and Fenglong Ma. Fedsemi: An adaptive federated semi-supervised learning framework. *arXiv preprint arXiv:2012.03292*, 2020.

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

[33] Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers under the margin condition. *arXiv preprint math/0507180*, 2005.

[34] Michael Kohler and Adam Krzyzak. On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE transactions on information theory*, 53(5):1735–1742, 2007.

[35] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[36] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[37] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[38] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[39] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[40] Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

[41] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

[42] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[43] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[44] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

[45] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[46] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[47] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[48] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[49] Tehrim Yoon, Sumin Shin, Sung Ju Hwang, and Eunho Yang. Fedmix: Approximation of mixup under mean augmented federated learning. In *International Conference on Learning Representations*, 2021.

[50] 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.

[51] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[52] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[53] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See last sentences of Section 3.3 and 4.,
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not foresee any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section D.3 in Appendix.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section D.5 in Appendix.
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide source codes in the supplementary material. We use publicly available datasets.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section C.1 in Appendix.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The error bars are shown in figures and the numerical standard error are shown in brackets of tables.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] One Nvidia 1080TI is enough for one experiment run.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We cite the publicly available datasets we use.

(b) Did you mention the license of the assets? [No] They should be publicly available.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We use publicly available datasets.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We use publicly available datasets.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

# Appendix

## A Performance Goal

We outline the general performance goal of Semi-Supervised Federated Learning. The performance ceiling is obviously that of Fully Supervised Learning (FSL) (namely, assuming that all the server's and clients' data are centralized and fully labeled). For our context where clients' data are unlabeled, a vanilla approach trains the labeled data only on the server-side, referred to as Partially Supervised Learning (PSL). Clearly, the PSL performance can serve as a lower bound benchmark for other approaches that employ additional unlabeled data. When the server contains a small amount of labeled data and a substantial amount of unlabeled data (centralized), Semi-Supervised Learning (SSL) seeks to use unlabeled data to improve over the PSL. It was shown that state-of-the-art SSL methods such as FixMatch [32] could produce similar results as FSL.

Our work focuses on Semi-Supervised Federated Learning (SSFL), where the unlabeled data are distributed among many clients. The general goal of SSFL is to perform similarly to the state-of-the-art SSL and significantly outperform PSL and the existing SSFL methods. In other words, our performance goal is to achieve FSL $\gtrsim$ SSL $\gtrsim$ SSFL $\gg$ PSL.

## B Static Batch Normalization

We utilize a recently proposed adaptation of Batch Normalization (BN) named Static Batch Normalization (sBN) [9]. It was shown that this method greatly accelerates the convergence and improves the performance of FedAvg [4] compared with other forms of normalization, including InstanceNorm [45], GroupNorm (GN) [46], and LayerNorm [47]. During the training phase, sBN does not track the running statistics with momentum as in BN. Instead, it simply standardizes the data batch $x_b$ and utilizes batch-wise statistics $\mu_b$ and $\sigma_b$ in the following way.

$$\tilde{x}_b = \frac{x_b - \mu_b}{\sqrt{\sigma_b^2 + \epsilon}} \cdot \gamma + \beta, \quad \mu_b = \mathrm{E}[x_b], \quad \sigma_b^2 = \mathrm{Var}[x_b]$$

In FL training, the affine parameters $\gamma$ and $\beta$ can be aggregated as usual. We note that FedAvg with vanilla BN is not functional because the BN statistics $\mu$ and $\sigma$ used for inference are averaged from the tracked running BN statistics of local clients during training. Let $x_m$ represents the local data of client $m$ (with size $N_m$). For a total of $M$ local clients, sBN computes the global BN statistics $\mu$ and $\sigma$ for inference by querying each local client one more time after training is finished, based on

$$\mu = \frac{\sum_{m=1}^{M} N_m \mu_m}{\sum_{m=1}^{M} N_m}, \; \mu_m = \mathrm{E}[x_m], \; \sigma_m^2 = \mathrm{Var}[x_m],$$
$$\sigma^2 = \frac{\sum_{m=1}^{M} \left[(N_m - 1)\sigma_m^2 + N_m(\mu_m - \mu)^2\right]}{(\sum_{m=1}^{M} N_m) - 1}.$$

In the context of SemiFL, we need to generate pseudo-labels at every communication round. Thus, local clients need to upload BN statistics for every communication round. Fortunately, we can utilize the server data $x_s$ to update the global statistics instead of querying each local client, where $\mu = \mathrm{E}[x_s]$ and $\sigma^2 = \mathrm{Var}[x_s]$. We provide experimental results of querying the sBN statistics from all the clients and include an ablation study using only the server data in Table 3. In Table 3, we demonstrate the ablation study of the sBN statistics on the CIFAR10 dataset. Compared with updating the sBN statistics with only the server data, updating the sBN statistics with both server and clients does not provide significant improvements.

Table 3: Ablation study of sBN statistics for the CIFAR10 dataset. The alternative way of using the server data to update the global sBN statistics does not degrade the performance.

| sBN statistics | 250 | | 4000 | |
|---|---|---|---|---|
| | Non-IID, $K = 2$ | IID | Non-IID, $K = 2$ | IID |
| server only | 60.0(0.8) | 86.3(0.2) | 85.5(0.1) | 93.1(0.2) |
| server and clients | 60.0(0.9) | 88.2(0.3) | 85.3(0.3) | 93.1(0.1) |

# C Experimental Results

## C.1 Experimental setup

In Table 4, we provide the hyperparameters used in the experiments. Similar to [32], we use SGD as our optimizer and a cosine learning rate decay as our scheduler [48]. We also use the same hyperparameters as [32], where the local learning rate $\eta = 0.03$, the local momentum $\beta_l = 0.9$, and the confidence threshold $\tau = 0.95$. The Mixup hyperparameter $a$ is set to be $0.75$ as suggested by [36].

We use the standard supervised loss to train the labeled server. For training the unlabeled clients, the "fix" loss $L_{\text{fix}}$ (proposed in FixMatch [32]) leverages the techniques of consistency regularization and pseudo-labeling simultaneously. Specifically, the pseudo-labels are generated from weakly augmented data, and the model is trained with strongly augmented data. The "mix" loss (adapted from MixMatch [23, 36]) reduces the memorization of corrupted labels and increases the robustness to adversarial examples. It was also shown to benefit the SSL [24] and FL [49] methods. We have conducted an ablation study and demonstrated that the mix loss moderately improves performance.

Table 4: Hyperparameters used in our experiments.

| Dataset | | CIFAR10 | | SVHN | | CIFAR100 | |
|---|---|---|---|---|---|---|---|
| Number of Supervised | | 250 | 4000 | 250 | 1000 | 2500 | 10000 |
| Architecture | | WResNet28x2 | | | | WResNet28x8 | |
| Server | Batch size | 10 | 250 | 10 | 250 | 10 | 250 |
| | Epoch | 5 | | | | | |
| | Optimizer | SGD | | | | | |
| | Learning rate | 3.0E-02 | | | | | |
| | Weight decay | 5.0E-04 | | | | | |
| | Momentum | 0.9 | | | | | |
| | Nesterov | ✓ | | | | | |
| Client | Batch size | 10 | | | | | |
| | Epoch | 5 | | | | | |
| | Optimizer | SGD | | | | | |
| | Learning rate | 3.0E-02 | | | | | |
| | Weight decay | 5.0E-04 | | | | | |
| | Momentum | 0.9 | | | | | |
| | Nesterov | ✓ | | | | | |
| Global | Communication round | 800 | | | | | |
| | Momentum | 0.5 | | | | | |
| | Scheduler | Cosine Annealing | | | | | |

### C.2 SVHN and CIFAR100

In Figure 7 and 9, we demonstrate the results of SVHN and CIFAR100 datasets.

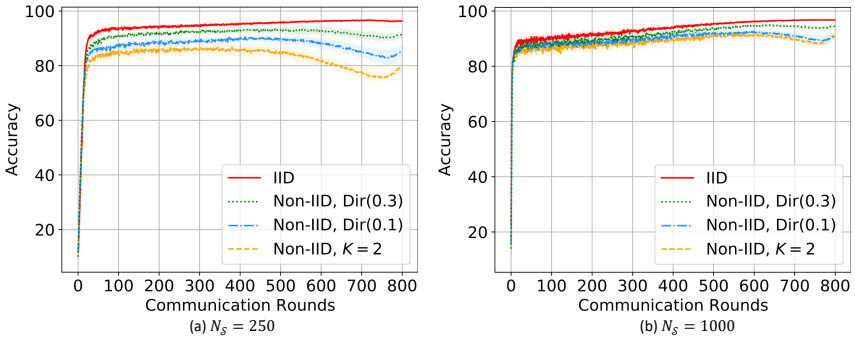


Figure 7: Experimental results for SVHN dataset with (a) $N_{\mathcal{S}} = 250$ and (b) $N_{\mathcal{S}} = 1000$.

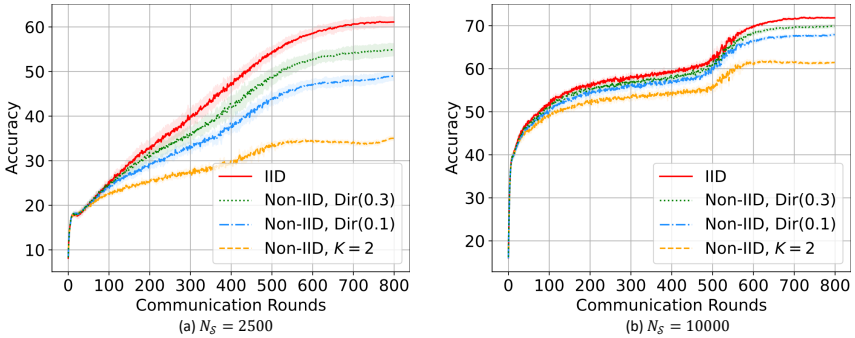


Figure 8: Experimental results for CIFAR100 dataset with (a) $N_{\mathcal{S}} = 2500$ and (b) $N_{\mathcal{S}} = 10000$.

### C.3 Ablation studies

We perform an ablation study of the training techniques adopted in our experiments. We study the efficacy of the number of local training epoch $E$, the Mixup data augmentation, and the global SGD momentum $\beta_g$ [10] as shown in Table 5. Less local training epoch significantly hurts the performance due to slow convergence. The Mixup data augmentation has around $2\%$ Accuracy improvement for the CIFAR10 dataset. It demonstrates that it is beneficial to combine strong data augmentation with Mixup data augmentation for training unlabeled data. The global momentum marginally improves the result.

Table 5: Ablation study on the CIFAR10 datasets with $4000$ labeled data at the server.

| $E$ | $\beta_g$ | Mixup | SemiFL | |
|---|---|---|---|---|
| | | | Non-IID, $K=2$ | IID |
| 1 | 0.5 | ✓ | 83.4(0.5) | 88.9(0.3) |
| 5 | 0.5 | ✗ | 84.2(0.4) | 91.3(0.2) |
| 5 | 0 | ✓ | 85.4(0.6) | 92.4(0.1) |
| 5 | 0.5 | ✓ | 85.3(0.3) | 93.1(0.1) |

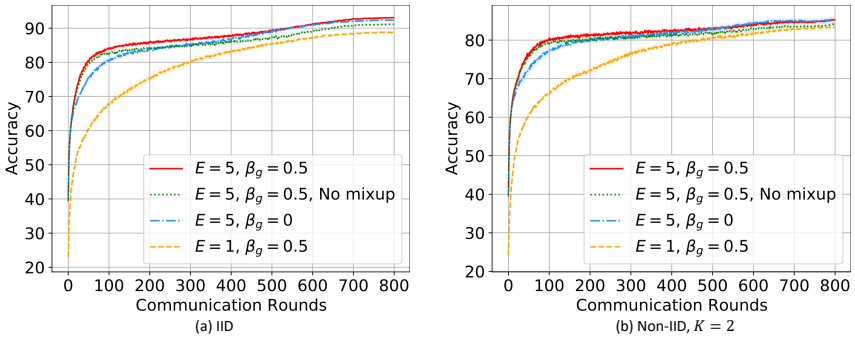


Figure 9: Ablation study of the CIFAR10 dataset with 4000 labeled data at the server for the cases of (a) IID and (b) Non-IID, $K = 2$ data partition.

# D Theoretical Analysis of Strong Data Augmentation for SSL

## D.1 Background of Classification

We take the binary classification task as an illustrating example. Let $(Y, X)$ be a random variable with values in $\mathbb{R}^d \times \{1, 0\}$. For the prediction task, we look for a classifier $C : \mathbb{R}^d \to \{1, 0\}$ such that the risk $\mathbb{P}(C(X) \neq Y)$ is small, where $\mathbb{P}$ denotes the probability measure for $(Y, X)$. Let $m(x) \triangleq \mathbb{E}(Y = 1 \mid X = x)$ denote the conditional probability of $Y$ given $X = x$. For example, the standard logistic regression model is in the form of $m(x) = 1/(1 + \exp(-\beta^{\mathrm{T}} x))$ for some $\beta \in \mathbb{R}^d$.

When the underlying $m$ is known, the risk-optimal classifier is known to be

$$C : x \mapsto \mathbb{1}\{m(x) - 1/2\} \tag{7}$$

for any given $x$. When the underlying $m$ is unknown, we need to train a classifier $\hat{C}_n$ from observed training data $(Y_i, X_i)$, $i = 1, \ldots, n$, which are often assumed to be IID random variables following the same distribution of $(Y, X)$. A general approach is to first learn $\hat{m}_n : \mathbb{R}^d \to \mathbb{R}$ and then let $\hat{C}_n(x) \triangleq \mathbb{1}\{\hat{m}_n(x) - 1/2\}$. To evaluate the prediction performance of a learned $\hat{C}_n$, we consider its gap with the optimal classifier

$$\mathcal{R}(\hat{C}_n) \triangleq \mathbb{P}(Y \neq \hat{C}_n(X)) - \mathbb{P}(Y \neq C(X)) \tag{8}$$

referred to as the classification risk of $\hat{C}_n$.

## D.2 Background of Semi-Supervised Learning

Suppose that we observe $n_{\mathrm{l}}$ IID labeled data of $(Y^{\mathrm{l}}, X^{\mathrm{l}})$, denoted by $D^{\mathrm{l}} = \{(Y_i^{\mathrm{l}}, X_i^{\mathrm{l}})\}_{i=1}^{n_{\mathrm{l}}}$, where $X^{\mathrm{l}}$ has probability distribution $\mathbb{P}_{\mathrm{l}}$ and $\mathbb{E}(Y^{\mathrm{l}} \mid X^{\mathrm{l}} = x) = m(x)$. We also observe $n_{\mathrm{u}}$ unlabeled data of $(X^{\mathrm{u}})$, denoted by $\{X_j^{\mathrm{u}}\}_{j=1}^{n_{\mathrm{u}}}$, where each $X^{\mathrm{u}}$ has probability distribution $\mathbb{P}_{\mathrm{u}}$. Here, $\mathbb{P}_{\mathrm{u}}$ may or may not be the same as $\mathbb{P}_{\mathrm{l}}$. The Semi-Supervised Learning problem of interest concerns the case $n_{\mathrm{u}} \gg n_{\mathrm{l}}$ and solutions that can properly utilize the unlabeled data to boost the performance of a classifier trained from labeled data. In other words, we look for a classifier $\hat{C}_n^{\mathrm{ssl}}(x)$ trained from observations of both $(Y^{\mathrm{l}}, X^{\mathrm{l}})$ and $X^{\mathrm{u}}$, so that its risk satisfies

$$\mathcal{R}(\hat{C}^{\mathrm{ssl}}) \ll \mathcal{R}(\hat{C}^{\mathrm{l}})$$

where $\hat{C}^{\mathrm{l}}$ is the classifier trained from observations of $(Y^{\mathrm{l}}, X^{\mathrm{l}})$ only.

## D.3 A new perspective of Semi-Supervised Learning

As we mentioned in Section 2, there has been a lot of empirical success in using new techniques such as consistency regularization and strong augmentation to improve the classification risk of classical Semi-Supervised Learning. Recently, the work of [50] provides a theoretical understanding of the

consistency regularization in reducing classification risk. Its analysis is based on an "expansion" assumption that a low-probability subset of data must expand to a large-probability neighborhood, and there is little overlap between neighborhoods of different classes. To the best of our knowledge, the existing theories do not explain why the strong augmentation technique works so well (to achieve state-of-the-art performance) for Semi-Supervised Learning. Intuitively, strong augmentation is a process that maps a data point (e.g., an image) from high quality to relatively low quality in a unilateral manner (illustrated in Figure 10). Strong augmentation such as RandAugment [37] consists of a set of data augmentation strategies, e.g., rotating the image, shearing the image, translating the image, adjusting the color balance, and modifying the brightness. The low-quality data and their high-confidence pseudo-labels are then used for training so that there are sufficient "observations" near the difficult data regimes (e.g., near the decision boundary).

In line with the above intuition, we develop a theoretical understanding of how and when using strong augmentation can significantly reduce the classification risk obtained from only labeled data. Instead of studying Semi-Supervised Learning in full generality, we restrict our attention to a class of nonparametric kernel-based classification learning and derive analytically tractable statistical risk-rate analysis. Our theory is based on an intuitive "*adequate transmission*" assumption, which means that the distribution of augmented data from high-confidence unlabeled data can adequately cover the data regime of interest during the test. Consequently, reliable information exhibited from unlabeled data can be "transmitted" to data regimes that may have been insufficiently trained with labeled data.

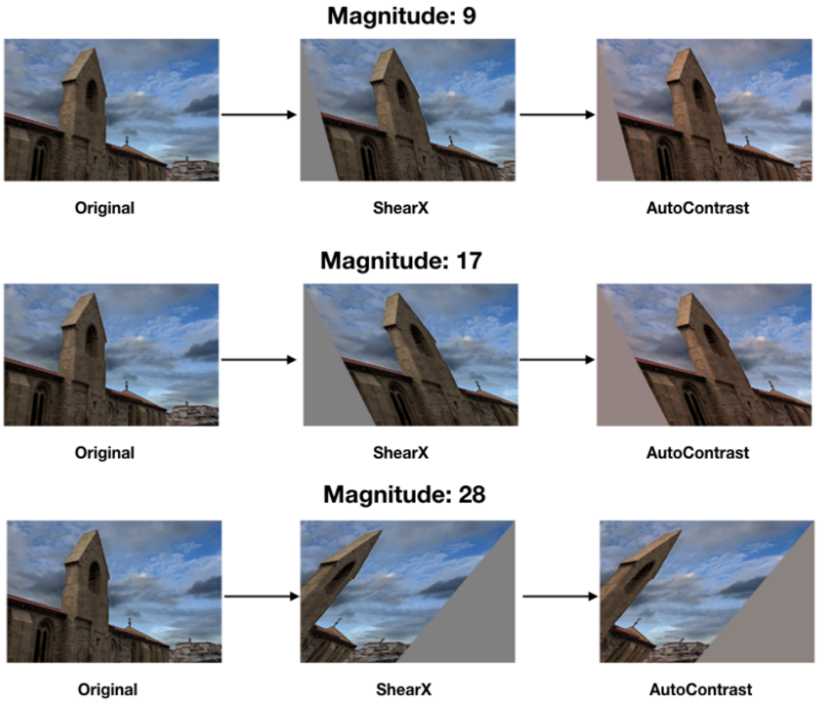


Figure 10: Examples of strong data augmentations based on the RandAugment technique [37]. As the distortion magnitude increases, the strength of the augmentation increases. Here, "ShearX" means shearing the image along the horizontal axis, and "AutoConstrast" means maximizing the image contrast by setting the darkest (respectively lightest) pixel to black (respectively white).

In addition to the notations made in Subsections D.1 and D.2, we will let $\tilde{X}$ denote strongly-augmented data from $X^{\mathrm{u}}$, and $\tilde{Y}$ its corresponding label that follows the same conditional distribution, namely $\mathbb{P}(\tilde{Y} = 1 \mid \tilde{X}) = m(\tilde{X})$. Recall that $\mathbb{P}_{\mathrm{u}}$ and $\mathbb{P}_{\mathrm{l}}$ are the probability measures of unlabeled $X^{\mathrm{u}}$ and labeled $X^{\mathrm{l}}$, respectively. We suppose that the test data distribution for evaluating the classification performance also follows $\mathbb{P}_{\mathrm{l}}$. In other words, the probability measure in (8) is the product of $\mathbb{P}_{Y|X}$ or $\mathbb{P}_{\tilde{Y}|\tilde{X}}$ (as determined by $m(\cdot)$) and $\mathbb{P}_{\mathrm{l}}$. Let $\hat{m}_0$ denote an initial estimate of $m$. For generality, we will assume $\hat{m}_0$ is learned from all or only part of the available labeled data. To develop theoretical analyses, we consider the following generic SSL classifier with strong augmentation.

**Generic Semi-Supervised Learning with Strong Data Augmentation**

• *Step 1*. From $\{X_i^{\mathrm{u}}\}_{i=1}^{n_{\mathrm{u}}}$, we pick up those "high-confidence" $x$ satisfying

$$\min\{1-\hat{m}_0(x), \hat{m}_0(x)\} \le \delta \tag{9}$$

for some $\delta$ (to be quantified), and denote the set as $\mathcal{X}^{\mathrm{aug}}$.

• *Step 2*. For each $X \in \mathcal{X}^{\mathrm{aug}}$, we calculate the pseudo-label $\hat{Y} = \mathbb{1}\{\hat{m}_0(X) - 1/2\}$; meanwhile, we generate the strongly augmented data $\tilde{X}$. Consequently, we obtain a set of data $(\hat{Y}, \tilde{X})$ and denote that set as $D^{\mathrm{aug}}$.

• *Step 3*. Train an estimate of $m$, denoted by $m^{\mathrm{ssl}}$, and the associated classifier $C^{\mathrm{ssl}}$ using the labeled and augmented data $D^{\mathrm{ssl}} \triangleq D^{\mathrm{l}} \cup D^{\mathrm{aug}}$.

Note that if $\hat{m}_0$ is learned from data independent with $D^{\mathrm{l}}$, the data in $D^{\mathrm{ssl}}$ are independent but not necessarily identically distributed (since $\mathbb{P}_{\mathrm{l}}$ and $\mathbb{P}_{\mathrm{u}}$ may not be the same).

To show how SSL with strong augmentation can potentially enhance classification learning, we consider a classical nonparametric classifier $\hat{C}$ defined in the following way. Let $K : \mathbb{R}^d \to \mathbb{R}^+$ denote the box kernel function that maps $u$ to $\mathbb{1}\{\|u\| \le 1\}$, where $\mathbb{1}\{\cdot\}$ denotes the indicator function. With $n$ labeled data $(Y_i, X_i)$, similarly to (7), we define

$$\hat{C}_n : x \mapsto \mathbb{1}\{\hat{m}_n(x) - 1/2\}, \quad \text{where } \hat{m}_n(x) = \frac{\sum_{i=1}^n K(h_n^{-1}(x - X_i)) \cdot Y_i}{\sum_{i=1}^n K(h_n^{-1}(x - X_i))} \tag{10}$$

if $\sum_{i=1}^n K(h_n^{-1}(x - X_i)) \neq 0$, and $\hat{m}_n(x) = 0$ otherwise. Here, $\hat{m}_n$ is known as the Nadaraya-Watson kernel estimate [51, 52] of the underlying $m$, and $h_n > 0$ is the bandwidth.

In our setting, we suppose that $n_0 > 0$ labeled data are used to learn $\hat{m}_0$, and another $n_{\mathrm{l}} \ge 0$ labeled data along with $n_{\mathrm{u}} > 0$ unlabeled data to learn $\hat{m}^{\mathrm{ssl}}$ and thus the subsequent classifier $\hat{C}^{\mathrm{ssl}}$. Note that the $n_{\mathrm{l}}$ is introduced only for generality. Our technical analysis includes $n_{\mathrm{l}} = 0$ as a special case. In the main result to be introduced, the risk bound will only involve $n_{\mathrm{u}}$ but eliminate $n_{\mathrm{l}}$ during technical derivations since we are interested in the regime of $n_{\mathrm{u}} \gg n_0 + n_{\mathrm{l}}$.

Before starting the main result, we make the following additional technical assumptions and provide the intuitions.

(A1) There exists positive constants $c_1$ and $s$ such that $\mathbb{P}_{\mathrm{u}}(\min\{1 - m(X), m(X)\} \le \delta) \ge g_s(\delta)$ for all sufficiently small $\delta > 0$, where $g_s(\delta) \triangleq c_1\delta^s$.

*Explanation of (A1)*: Recall that $\mathbb{P}_{\mathrm{u}}$ is the probability measure of unlabeled data. This condition requires a nontrivial amount of unlabeled data with high confidence (or large margin) in the sense that $m(X)$ is close to either zero or one. The function $g_s$ quantifies the "sufficiency" of data at the tail part of $X$. Take logistic regression $m(x) = 1/(1 + \exp(-\beta^{\mathrm{T}} x))$ as an example. It can be easily verified that

$$\mathbb{P}_{\mathrm{u}}(1 - m(X) \le \delta) \ge \mathbb{P}_{\mathrm{u}}(\beta^{\mathrm{T}} X \ge -\log\delta), \quad \mathbb{P}_{\mathrm{u}}(m(X) \le \delta) \ge \mathbb{P}_{\mathrm{u}}(\beta^{\mathrm{T}} X \le \log\delta),$$

so $\mathbb{P}_{\mathrm{u}}(\min\{1-m(X), m(X)\} \le \delta) = \mathbb{P}_{\mathrm{u}}(1-m(X) \le \delta) + \mathbb{P}_{\mathrm{u}}(m(X) \le \delta) \ge \mathbb{P}_{\mathrm{u}}(|\beta^{\mathrm{T}} X| \ge -\log\delta)$ for all $\delta \in (0, 1/2)$. For example, if $|\beta^{\mathrm{T}} X|$ follows standard Exponential, we let $g_s : \delta \mapsto \delta$.

(A2) There exists a constant $c_3 \in (0, 1/2)$ such that the strong augmentation $X^{\mathrm{u}} \to \tilde{X}$ satisfies $\mathbb{P}(\tilde{Y} = 1 \mid \tilde{X} = \tilde{x}, X^{\mathrm{u}} = x) = m(x)$ for all $x$ such that $\min\{1 - \hat{m}_0(x), \hat{m}_0(x)\} \le c_3$.

*Explanation of (A2)*: Let us think $X^{\mathrm{u}}$ as a high-confidence image, with $m(X^{\mathrm{u}})$ close to either zero or one. Meanwhile, $\tilde{X}$ is a strongly augmented version of $X^{\mathrm{u}}$, e.g., by random masking or noise injection, so $m(\tilde{X})$ is closer to $1/2$ than $m(X^{\mathrm{u}})$. The condition of (A2) means that if conditioning on both images, the label $\tilde{Y}$ has a distribution that is only determined by the higher-quality image, which is quite intuitive. A mathematically equivalent way to describe (A2) is that $\tilde{X} \to X^{\mathrm{u}} \to \tilde{Y}$ follows a Markov chain.

(A3) There exist positive constants $c_2$, $c_4$, and a non-negative $v$ such that for every $\mathbb{P}_{\mathrm{l}}$-measurable ball $B \subseteq \mathbb{R}^d$ with $\mathbb{P}_{\mathrm{l}}(B) \le c_4$, for the strong augmentation $X^{\mathrm{u}} \to \tilde{X}$, we have $\mathbb{P}_{\mathrm{u}}(\tilde{X} \in B \mid \min\{1 - \hat{m}_0(X^{\mathrm{u}}), \hat{m}_0(X^{\mathrm{u}})\} \le \delta)/\mathbb{P}_{\mathrm{l}}(B) \ge g_v(\delta)$ for all sufficiently small $\delta > 0$, where $g_v(\delta) \triangleq c_2\delta^v$.

*Explanation of (A3)*: The above numerator is the probability of the augmented data $\tilde{X}$ falling into $B$ conditional on the original unlabeled data (with probability $\mathbb{P}_{\mathrm{u}}$) having high confidence. This assumption ensures that for every regime of significant interest in evaluating the prediction performance (since $\mathbb{P}_{\mathrm{l}}$ is the measure for test data), there will be a sufficient probability coverage of the augmented data. This is an intuitive condition since otherwise, the augmented data cannot represent the test data of interest to boost the test performance. In this assumption, the function $g_v$ determines the coverage as a function of tail probability $\delta$. For example, if $v = 0$, a sufficiently small $\delta$ (or higher confidence) gives a non-vanishing coverage. The combination of (A2) and (A3) can be interpreted as an "adequate transmission" condition, under which a small amount of high-confidence unlabeled data can induce augmented data that can accurately represent the test data regime of interest. Such transmitted data can be basically approximated as labeled data for supervised training.

(A4) There exist positive constants $c_6$ and $\alpha$ such that $\mathbb{P}_{\mathrm{l}}(|m(X^{\mathrm{l}}) - 1/2| \leq t) \leq c_6 t^{\alpha}$ for all $t > 0$. Moreover, $X^{\mathrm{l}} \in [0, 1]^d$.

*Explanation of (A4)*: The inequality is a margin condition that has been used in the classical learning literature (see, e.g., [34, 35] and the references therein). It determines the difficulty of the underlying classification task. Intuitively speaking, a larger $\alpha$ means more separability of the two classes under the probability $\mathbb{P}_{\mathrm{l}}$. The boundedness of $X^{\mathrm{l}}$ is for technical convenience.

(A5) There exist positive constants $q$ and $c_7$ such that $|m(x) - m(x')| \leq c_7 \|x - x'\|^q$ for all $x, x' \in [0, 1]^d$, where $\|\cdot\|$ denotes the Euclidean norm.

*Explanation of (A5)*: This condition assumes a Lipschitz-type condition of $m(\cdot)$, where $q$ is allowed to be different from one. Intuitively, it assumes the underlying classifier to learn cannot be too bumpy. For $q \in (0, 1]$, a larger $q$ means more smoothness of $m(\cdot)$.

(A6) There exist positive constants $r$, $c_8$, and $\Delta$ such that $|\hat{m}_0(x) - m(x)| \leq c_8 n_0^{-r}$ for all $x$ satisfying $\min\{1 - \hat{m}_0(x), \hat{m}_0(x)\} \leq \Delta$.

*Explanation of (A6)*: This assumption requires that conditional on $X$ falls into a large-margin area, the estimation error of the initial function $\hat{m}_0$ is not too large.

(A7) For the constants $s$, $v$, $\alpha$, $q$, and $r$ defined in the above assumptions, we have

$$\frac{q \cdot s}{q \cdot (\alpha + 3 + v + s) + d} < \frac{1}{2}, \tag{11}$$

$$\frac{n_0^{-r}}{n_{\mathrm{u}}^{-q/\{q(\alpha+3+v+s)+d\}}} \to 0, \text{ as } \min\{n_0, n_{\mathrm{u}}\} \to \infty. \tag{12}$$

*Explanation of (A7)*: The two inequalities will be technical conditions used in the proof. A sufficient condition for (11) to hold is that $\alpha \geq s$. Intuitively, this requires that $\alpha$, which describes the separability of the decision boundary (the larger, the better), is not smaller than $s$, which quantifies the sufficiency of tail samples (the smaller, the better). The inequality (12) means that the initial classifier $\hat{m}_0$ cannot perform too poorly. This matches our empirical observations that the SSL training in each round has to immediately follow a preceding round that uses some labeled data. Also, the denominator in (12) favors relatively small $s, d$ compared with $\alpha, v, q$.

### D.4 Main result

Our **main result** is provided below.

**Theorem 1**: Under Assumptions (A1)-(A7), the generic SSL classifier with strong augmentation (namely the above Steps 1-3) satisfies

$$\mathcal{R}(\hat{C}^{\mathrm{ssl}}) \leq C n_{\mathrm{u}}^{-q(\alpha+1)/\{q(\alpha+3+v+s)+d\}} \tag{13}$$

for some constant $C$ that does not depend on the sample size.

*Explanation of Theorem 1*: The theorem gives an explicit rate of convergence for the SSL classification risk using unlabeled data of size $n_{\mathrm{u}}$. It is the informal statement made in the main paper with $\rho \triangleq v + s$. We interpret the power

$$\frac{q(\alpha + 1)}{q(\alpha + 3 + v + s) + d}$$

as follows. If the margin parameter $\alpha$ is large, the classification is relatively easy, and the ratio can go up to one, namely $\mathcal{R}(\hat{C}^{\text{ssl}}) \sim n_{\text{u}}^{-1}$. This is reminiscent of an existing result that uses labeled data and a large margin to achieve the $n_{\text{l}}^{-1}$ rate [33]. If the tail sufficiency parameter $s$ or the coverage parameter $v$ is large, the ratio becomes approximately $(\alpha+1)/(v+s)$. Intuitively, a larger $s$ or $v$ indicates that there will be fewer high-confidence unlabeled data to be transmitted to benefit the classification learning (on the evaluation measure $\mathbb{P}_{\text{l}}$ of interest), which is in line with a slower rate of convergence $n_{\text{u}}^{-(\alpha+1)/(v+s)}$.

On the contrary, consider the other extreme that $v = s = 0$. Then, the ratio becomes $q(\alpha+1)/\{q(\alpha+3)+d\}$, which matches an existing result in classification learning [34]. For comparison, we define the baseline classifier that only uses $n_{\text{l}}$ labeled data based on the kernel estimation in (10). We denote that classifier as $\hat{C}^{\text{l}}$. The risk would be $\mathcal{R}(\hat{C}^{\text{l}}) \leq C' n_{\text{l}}^{-q(\alpha+1)/\{q(\alpha+3)+d\}}$ for some constant $C'$. Comparing this with (13), we can determine the region where employing SSL can significantly improve supervised learning. To illustrate this point, let us suppose that

$$n_{\text{l}} \sim n_{\text{u}}^{\zeta}$$

for some constant $\zeta \in (0,1)$. It can be verified that the bound of $\mathcal{R}(\hat{C}^{\text{l}})$ is much larger than that of $\mathcal{R}(\hat{C}^{\text{ssl}})$ when

$$\frac{q(\alpha+1)}{q(\alpha+3+v+s)+d} > \frac{\zeta q(\alpha+1)}{q(\alpha+3)+d},$$

or equivalently,

$$\zeta < \frac{q(\alpha+3)+d}{q(\alpha+3+v+s)+d}. \tag{14}$$

The inequality (14) provides an insight into the *critical region of* $n_u$ where significant improvement can be made from unlabeled data, as dependent on constants that describe the underlying function smoothness ($q$), data dimension ($d$), task difficulty ($\alpha$), and "adequate transmission" parameters ($s, v$).

### D.5 Proof of Theorem 1

We first give a sketch of the proof. We first relate the risk bound of $\mathcal{R}(\hat{C}^{\text{ssl}})$ to the estimation error of $\hat{m}^{\text{ssl}}$, and then decompose the error into a bias term and a variance term. Each term is then bounded using concentration inequalities, in a way similar to the techniques used in [53, Ch. 5] and [34]. Different from the standard nonparametric analysis of classification learning with IID data, we will use the aforementioned "adequate transmission" conditions to derive the rate of convergence from data that are contributed from both labeled and pseudo-labeled data. The analysis involves a careful choice of the tuning parameters, e.g., the $\delta$ in Assumption (A1) and the kernel bandwidth, so that the biases introduced from pseudo-labeled data have a diminishing influence on the risk rate. Next, we provide detailed proof.

We let $n = n_{\text{l}} + n_{\text{u}}$ denote the total size of labeled and unlabeled data available to the SSL training. For notational clarity, we sometimes put subscript $n$, e.g., $\delta_n$ instead of $\delta$ (in Step 1), to highlight a quantity that is designed to vanish at some rate as $n$ becomes large. Recall that $D^{\text{ssl}} = D^{\text{l}} \cup D^{\text{aug}}$. Let $n_{\text{l}}$ and $n_{\text{u}}^{\text{aug}}$ denote the sample sizes of $D^{\text{l}}$ and $D^{\text{aug}}$, respectively. Note that $n_{\text{u}}^{\text{aug}}$ is random since the Step 1 depends on $n_0$ labeled data. We first consider the risk conditional on a fixed $n_{\text{u}}^{\text{aug}}$, denoted by $\mathcal{R}_{n_{\text{u}}^{\text{aug}}}(\hat{C}^{\text{ssl}})$.

Direct calculations show that

$$\mathcal{R}_{n_{\text{u}}^{\text{aug}}}(\hat{C}^{\text{ssl}}) = \mathbb{E}_{\text{l}}\Big(|2m(X)-1| \cdot \mathbb{1}\{\hat{C}^{\text{ssl}}(X) \neq C(X)\}\Big) = T_1 + T_2, \text{ where} \tag{15}$$

$$T_1 = 2\mathbb{E}_{\text{l}}\Big(|m(X)-1/2| \cdot \mathbb{1}\Big\{|m(X)-1/2| \leq t_n, \hat{C}^{\text{ssl}}(X) \neq C(X)\Big\}\Big)$$

$$T_2 = 2\mathbb{E}_{\text{l}}\Big(|m(X)-1/2| \cdot \mathbb{1}\Big\{|m(X)-1/2| > t_n, \hat{C}^{\text{ssl}}(X) \neq C(X)\Big\}\Big)$$

for an arbitrary $t_n > 0$ to be selected. From Assumption (A4), $|m(X)-1/2| \leq 1/2$, and $\mathbb{1}\{|m(X)-1/2| > t_n, \hat{C}^{\text{ssl}}(X) \neq C(X)\} \leq \mathbb{1}\{|m(X)-\hat{m}(X)| > t_n\}$, we have

$$T_1 \leq 2t_n \cdot \mathbb{P}_{\text{l}}\big(|m(X)-1/2| \leq t_n\big) \leq 2c_6 t_n^{1+\alpha}, \quad T_2 \leq \mathbb{P}_{\text{l}}\big(|m(X)-\hat{m}(X)| > t_n\big). \tag{16}$$

Moreover, by the triangle inequality, we have

$$T_2 \leq \mathbb{P}_{\mathrm{l}}\big(|m(X) - \bar{m}(X)| > t_n/2\big) + \mathbb{P}_{\mathrm{l}}\big(|\bar{m}(X) - \hat{m}(X)| > t_n/2\big), \tag{17}$$

where we define the function $\bar{m}$ by

$$\bar{m}(x) = \frac{\sum_{X \in D^{\mathrm{ssl}}} K(h_n^{-1}(x - X))m(X)}{\sum_{X \in D^{\mathrm{ssl}}} K(h_n^{-1}(x - X))}$$

if the denominator is nonzero, and $\bar{m}(x) = 0$ otherwise.

In the sequel, we bound each term in (17). First, we rewrite

$$\mathbb{P}_{\mathrm{l}}\big(|m(X) - \bar{m}(X)| > t_n/2\big) = \int_{x \in [0,1]^d} \mathbb{P}\big(|m(x) - \bar{m}(x)| > t_n/2\big) d\mathbb{P}_{\mathrm{l}}(x), \tag{18}$$

where $\mathbb{P}$ denotes the probability measure induced by $D^{\mathrm{ssl}}$ (which is implicitly used to define $\bar{m}$). For each $x$, we define the event

$$E_x = \left\{\omega : \sum_{X \in D^{\mathrm{ssl}}} K(h_n^{-1}(x - X))\right\}.$$

Then, from Assumption (A5) and the definition that $K(u) = \mathbb{1}\{\|u\| \leq 1\}$, we have

$$\begin{aligned}
|m(x) - \bar{m}(x)| &= \frac{|\sum_{X \in D^{\mathrm{ssl}}} K(h_n^{-1}(x - X))(m(x) - m(X))|}{\sum_{X \in D^{\mathrm{ssl}}} K(h_n^{-1}(x - X))} \cdot \mathbb{1}\{E_x\} + m(x)(1 - \mathbb{1}\{E_x\}) \\
&\leq \frac{\sum_{X \in D^{\mathrm{ssl}}} K(h_n^{-1}(x - X))|x - X|^q}{\sum_{X \in D^{\mathrm{ssl}}} K(h_n^{-1}(x - X))} \cdot \mathbb{1}\{E_x\} + m(x)(1 - \mathbb{1}\{E_x\}) \\
&\leq c_7 h_n^q + m(x)(1 - \mathbb{1}\{E_x\}).
\end{aligned} \tag{19}$$

Let $B_{x,h} \triangleq \{u \in \mathbb{R}^d : \|u - x\| \leq h\}$ denote the Euclidean ball of center $x$ and radius $h$. If we choose

$$t_n/2 > c_7 h_n^q, \tag{20}$$

the above inequality (19) implies that

$$\begin{aligned}
\mathbb{P}(|m(x) - \bar{m}(x)| \geq t_n/2) &\leq \mathbb{P}\bigg(m(x)(1 - \mathbb{1}\{E_x\}) \geq t_n/2 - c_7 h_n^q\bigg) \\
&\leq \mathbb{P}\bigg\{\sum_{X \in D^{\mathrm{ssl}}} K(h_n^{-1}(x - X)) = 0\bigg\} \\
&= \mathbb{P}\bigg\{\|x - X\| > h_n, \forall X \in D^{\mathrm{ssl}}\bigg\} \\
&= (1 - \mathbb{P}_{\mathrm{l}}(B_{x,h_n}))^{n_{\mathrm{l}}} \cdot (1 - \mathbb{P}_{\mathrm{u}}(B_{x,h_n}))^{n_{\mathrm{u}}^{\mathrm{aug}}} && (21) \\
&\leq \exp\{-n_{\mathrm{l}}\mathbb{P}_{\mathrm{l}}(B_{x,h_n})\} \cdot \exp\{-n_{\mathrm{u}}^{\mathrm{aug}}\mathbb{P}_{\mathrm{u}}(B_{x,h_n})\} && (22)
\end{aligned}$$

Let $c_9 \triangleq \max_{v>0} ve^v$. Let $\{z_i\}_{i=1}^{M_n}$ be a set of points in $\mathbb{R}^d$ such that $[0,1]^d \subseteq \cup_{i=1}^{M_n} B_{z_i,h_n/2}$, with $M_n = c_{10} h_n^{-d}$ for some $c_{10}$. Taking (22) into (18), and invoking Assumption (A3), we obtain

$$\begin{aligned}
&\mathbb{P}_{\mathrm{l}}\big(|m(X) - \bar{m}(X)| > t_n/2\big) \\
&= \int_{x \in [0,1]^d} \exp\{-n_{\mathrm{l}}\mathbb{P}_{\mathrm{l}}(B_{x,h_n})\} \cdot \exp\{-n_{\mathrm{u}}^{\mathrm{aug}}\mathbb{P}_{\mathrm{u}}(\tilde{X} \in B_{x,h_n} \mid \tilde{X} \in D^{\mathrm{aug}})\} d\mathbb{P}_{\mathrm{l}}(x) \\
&\leq \int_{x \in [0,1]^d} \exp\{-n_{\mathrm{l}}\mathbb{P}_{\mathrm{l}}(B_{x,h_n}) - g_v(\delta_n) n_{\mathrm{u}}^{\mathrm{aug}}\mathbb{P}_{\mathrm{l}}(B_{x,h_n})\} d\mathbb{P}_{\mathrm{l}}(x) \\
&= \int_{x \in [0,1]^d} \exp\{-\tilde{n}\mathbb{P}_{\mathrm{l}}(B_{x,h_n})\} d\mathbb{P}_{\mathrm{l}}(x) \\
&\leq c_9 \int_{x \in [0,1]^d} \frac{1}{\tilde{n}\mathbb{P}_{\mathrm{l}}(B_{x,h_n})} d\mathbb{P}_{\mathrm{l}}(x) \\
&\leq c_9 \sum_{i=1}^{M_n} \int_{x \in [0,1]^d} \frac{\mathbb{1}\{x \in B_{z_i,h_n/2}\}}{\tilde{n}\mathbb{P}_{\mathrm{l}}(B_{x,h_n})} d\mathbb{P}_{\mathrm{l}}(x) \\
&\leq c_9 \tilde{n}^{-1} M_n = c_9 c_{10} \tilde{n}^{-1} h_n^{-d}
\end{aligned} \tag{23}$$

where we let $\tilde{n} \triangleq n_{\mathrm{l}} + g_v(\delta_n) n_{\mathrm{u}}^{\mathrm{aug}}$. The technique of covering used in the last two inequalities was from [53, Eq. 5.1].

To bound the second term in (17), we write

$$\hat{m}(x) - \bar{m}(x) = \sum_{(Y,X)\in D^{\mathrm{ssl}}} \frac{K(h_n^{-1}(x-X))}{\sum_{(Y,X)\in D^{\mathrm{ssl}}} K(h_n^{-1}(x-X))}(Y - m(X)). \tag{24}$$

Recall that $D^{\mathrm{ssl}} = D^{\mathrm{l}} \cup D^{\mathrm{aug}}$. For every $(Y^{\mathrm{l}}, X^{\mathrm{l}}) \in D^{\mathrm{l}}$, we have $\mathbb{E}(Y^{\mathrm{l}} \mid X^{\mathrm{l}}) = m(X)$.

For any $\delta_n$ that satisfies $\delta_n \leq \min\{c_3, \Delta, 1/4\}$, where $c_3$ was introduced in Assumption (A2) and $\Delta$ was introduced in Assumption (A6), we have

$$\begin{aligned}
&\mathbb{P}(\hat{Y} = 1, \tilde{Y} = 0 \mid \tilde{X}, X^{\mathrm{u}}) \\
&= \mathbb{P}(\hat{Y} = 1, \tilde{Y} = 0, \hat{m}_0(X^{\mathrm{u}}) \geq 1 - \delta_n \mid \tilde{X}, X^{\mathrm{u}}) + \mathbb{P}(\hat{Y} = 1, \tilde{Y} = 0, \hat{m}_0(X^{\mathrm{u}}) \leq \delta_n \mid \tilde{X}, X^{\mathrm{u}}) \\
&= \mathbb{P}(\hat{Y} = 1, \tilde{Y} = 0, \hat{m}_0(X^{\mathrm{u}}) \geq 1 - \delta_n \mid \tilde{X}, X^{\mathrm{u}}) \\
&\leq \mathbb{P}(\tilde{Y} = 0, \hat{m}_0(X^{\mathrm{u}}) \geq 1 - \delta_n, m(X^{\mathrm{u}}) \geq 1 - \delta_n - c_8 n_0^{-r} \mid \tilde{X}, X^{\mathrm{u}}) \\
&\quad + \mathbb{P}(\hat{m}_0(X^{\mathrm{u}}) \geq 1 - \delta_n, m(X^{\mathrm{u}}) \leq 1 - \delta_n - c_8 n_0^{-r} \mid \tilde{X}, X^{\mathrm{u}}) \\
&\leq \mathbb{P}(\tilde{Y} = 0, m(X^{\mathrm{u}}) \geq 1 - \delta_n - c_8 n_0^{-r}) + 0 \\
&\leq \delta_n + c_8 n_0^{-r},
\end{aligned}$$

and similarly, $\mathbb{P}(\hat{Y} = 0, \tilde{Y} = 1 \mid \tilde{X}, X^{\mathrm{u}}) \leq \delta_n + c_8 n_0^{-r}$. Thus,

$$\mathbb{E}(|\hat{Y} - \tilde{Y}| \mid \tilde{X}) = \mathbb{E}\{\mathbb{E}(|\hat{Y} - \tilde{Y}| \mid \tilde{X}, X^{\mathrm{u}}) \mid \tilde{X}\} \leq 2\delta_n + 2c_8 n_0^{-r}.$$

Consequently, for every $(\hat{Y}, \tilde{X}) \in D^{\mathrm{aug}}$, we have

$$\mathbb{E}(\hat{Y} \mid \tilde{X}) = \mathbb{E}(\tilde{Y} \mid \tilde{X}) + \kappa(\tilde{X}) = m(\tilde{X}) + \kappa(\tilde{X}) \tag{25}$$

where $\kappa(\tilde{X}) \triangleq \mathbb{E}(\hat{Y} - \tilde{Y} \mid \tilde{X}) \leq 2\delta_n + 2c_8 n_0^{-r}$.

Back in (24), let $u(Y) = Y$ if $(Y, X) \in D^{\mathrm{l}}$ and $u(Y) = \tilde{Y}$ if $(Y, X) \in D^{\mathrm{aug}}$, where $\tilde{Y}$ is the pseudo-label random variable as in Assumption (A2) and equality (25). In this way, we have $\mathbb{E}(u(Y) \mid X) = m(X)$. We rewrite (24) as

$$\begin{aligned}
&\hat{m}(x) - \bar{m}(x) = T_3(x) + T_4(x), \text{ where} \\
&T_3(x) \triangleq \sum_{(Y,X)\in D^{\mathrm{ssl}}} \frac{K(h_n^{-1}(x-X))}{\sum_{(Y,X)\in D^{\mathrm{ssl}}} K(h_n^{-1}(x-X))}(u(Y) - m(X)) \\
&T_4(x) \triangleq \sum_{(\hat{Y},\tilde{X})\in D^{\mathrm{aug}}} \frac{K(h_n^{-1}(x-X))}{\sum_{(Y,X)\in D^{\mathrm{ssl}}} K(h_n^{-1}(x-X))}(\hat{Y} - \tilde{Y}) \\
&\qquad \leq \sum_{(\hat{Y},\tilde{X})\in D^{\mathrm{aug}}} \frac{K(h_n^{-1}(x-X))}{\sum_{(\hat{Y},\tilde{X})\in D^{\mathrm{aug}}} K(h_n^{-1}(x-X))}(\hat{Y} - \tilde{Y}).
\end{aligned}$$

Let $\mathcal{X}^{\text{ssl}} \triangleq \{X : (\cdot, X) \in D^{\text{ssl}}\}$ and $\mathcal{X}^{\text{aug}} \triangleq \{X : (\cdot, X) \in D^{\text{aug}}\}$. Then, we can bound

$$\mathbb{P}\big(|\bar{m}(x) - \hat{m}(x)| > t_n/2 \mid \mathcal{X}^{\text{ssl}}\big) \tag{26}$$

$$\leq \mathbb{P}\big(|T_3(x)| > t_n/4 \mid \mathcal{X}^{\text{ssl}}\big) + \mathbb{P}\big(|T_4(x)| > t_n/4 \mid \mathcal{X}^{\text{ssl}}\big)$$

$$\leq 2\exp\bigg\{-\frac{2(t_n/4)^2}{\sum_{X\in\mathcal{X}^{\text{ssl}}} K^2(h_n^{-1}(x-X))/\{\sum_{X'} K(h_n^{-1}(x-X'))\}^2}\bigg\} + \tag{27}$$

$$+ \mathbb{P}\bigg(\bigg|\sum_{(\hat{Y},\tilde{X})\in D^{\text{aug}}} \frac{K(h_n^{-1}(x-X))}{\sum_{X'\in\mathcal{X}^{\text{aug}}} K(h_n^{-1}(x-X'))}(\hat{Y} - \tilde{Y} - \mathbb{E}(\hat{Y} - \tilde{Y} \mid \tilde{X}))\bigg| > t_n/8 \mid \mathcal{X}^{\text{ssl}}\bigg) +$$

$$+ \mathbb{P}\bigg(\bigg|\sum_{\tilde{X}\in\mathcal{X}^{\text{aug}}} \frac{K(h_n^{-1}(x-\tilde{X}))}{\sum_{X'\in\mathcal{X}^{\text{aug}}} K(h_n^{-1}(x-X'))}\kappa(\tilde{X}))\bigg| > t_n/8 \mid \mathcal{X}^{\text{aug}}\bigg)$$

$$\leq 2\exp\bigg\{-\frac{1}{8}t_n^2 \sum_{X\in\mathcal{X}^{\text{ssl}}} K(h_n^{-1}(x-X))\bigg\} + 2\exp\bigg\{-\frac{1}{128}t_n^2 \sum_{X\in\mathcal{X}^{\text{aug}}} K(h_n^{-1}(x-X))\bigg\} + \tag{28}$$

$$+ \mathbb{P}\bigg(2\delta_n + 2c_8 n_0^{-r} > t_n/8\bigg) \tag{29}$$

$$\leq 4\exp\bigg\{-\frac{1}{128}t_n^2 \sum_{X\in\mathcal{X}^{\text{aug}}} K(h_n^{-1}(x-X))\bigg\} \tag{30}$$

$$\leq 4\mathbb{1}\bigg\{\sum_{X\in\mathcal{X}^{\text{aug}}} K(h_n^{-1}(x-X)) < \frac{1}{2}n_{\text{u}}^{\text{aug}}\mathbb{P}_{\text{u}}(B_{x,h_n}) - \log^2 n_{\text{u}}^{\text{aug}}\bigg\} +$$

$$4\exp\bigg\{-\frac{1}{256}t_n^2 n_{\text{u}}^{\text{aug}}\mathbb{P}_{\text{u}}(B_{x,h_n}) + \frac{1}{128}t_n^2\log^2 n_{\text{u}}^{\text{aug}}\bigg\} \tag{31}$$

provided that

$$2\delta_n + 2c_8 n_0^{-r} \leq t_n/8. \tag{32}$$

In the above derivation, (27) uses the Hoeffding's inequality, the fact that $K^2(\cdot) = K(\cdot)$, and the triangle inequality, (28) uses the Hoeffding's inequality again, (29) follows from (25), (30) is from $\mathcal{X}^{\text{aug}} \subseteq \mathcal{X}^{\text{ssl}}$, and (31) is by the definition of the indicator function. Consequently, with the choice of

$$t_n \log n_{\text{u}}^{\text{aug}} \leq 1, \tag{33}$$

we have

$$\mathbb{P}\big(|\bar{m}(x) - \hat{m}(x)| > t_n/2\big) \tag{34}$$

$$\leq 4\mathbb{P}_{\text{u}}\bigg\{\sum_{X\in\mathcal{X}^{\text{aug}}} K(h_n^{-1}(x-X)) < \frac{1}{2}n_{\text{u}}^{\text{aug}}\mathbb{P}_{\text{u}}(B_{x,h_n}) - \log^2 n_{\text{u}}^{\text{aug}}\bigg\}$$

$$+ 8\exp\bigg\{-\frac{1}{256}t_n^2 n_{\text{u}}^{\text{aug}}\mathbb{P}_{\text{u}}(B_{x,h_n})\bigg\}. \tag{35}$$

The first term in (35), according to the Bernstein inequality, can be upper bounded by

$$4\exp\bigg\{-\frac{1}{2}\frac{(n_{\text{u}}^{\text{aug}}\mathbb{P}_{\text{l}}(B_{x,h_n})/2 + \log^2 n_{\text{u}}^{\text{aug}})^2}{n_{\text{u}}^{\text{aug}}\mathbb{P}_{\text{l}}(B_{x,h_n}) + (n_{\text{u}}^{\text{aug}}\mathbb{P}_{\text{l}}(B_{x,h_n})/2 + \log^2 n_{\text{u}}^{\text{aug}})/3}\bigg\}\bigg\}$$

$$\leq 4\exp\bigg\{-\frac{3}{14}(n_{\text{u}}^{\text{aug}}\mathbb{P}_{\text{l}}(B_{x,h_n})/2 + \log^2 n_{\text{u}}^{\text{aug}})\bigg\} \leq 4\exp\bigg\{-\frac{3}{14}\log^2 n_{\text{u}}^{\text{aug}})\bigg\}.$$

Therefore, we can bound the second term in (17) by

$$\mathbb{P}_{\text{l}}\big(|\bar{m}(X) - \hat{m}(X)| > t_n/2\big)$$

$$\leq \int_{x\in[0,1]^d} \mathbb{P}\big(|\bar{m}(x) - \hat{m}(x)| > t_n/2\big)d\mathbb{P}_{\text{l}}(x)$$

$$\leq 4\exp\bigg\{-\frac{3}{14}\log^2 n_{\text{u}}^{\text{aug}})\bigg\} + 8\int_{x\in[0,1]^d} \exp\bigg\{-\frac{1}{256}t_n^2 n_{\text{u}}^{\text{aug}}\mathbb{P}_{\text{u}}(B_{x,h_n})\bigg\}d\mathbb{P}_{\text{l}}(x).$$

The second term in (35), according to the same arguments as in (23), can be upper bounded by $8 \cdot 256 \cdot c_9 c_{10}/(g_v(\delta_n) t_n^2 n_{\mathrm{u}}^{\mathrm{aug}} h_n^d)$. Therefore, we have

$$\mathbb{P}_{\mathrm{l}}\big(|\bar{m}(X) - \hat{m}(X)| > t_n/2\big) \leq 4 \exp\bigg\{-\frac{3}{14} \log^2 n_{\mathrm{u}}^{\mathrm{aug}})\bigg\} + \frac{2^{11} c_9 c_{10}}{g_v(\delta_n) t_n^2 n_{\mathrm{u}}^{\mathrm{aug}} h_n^d}.$$

Combining inequalities (15), (16), (17), and (23), we obtain

$$\mathcal{R}_{n_{\mathrm{u}}^{\mathrm{aug}}}(\hat{C}^{\mathrm{ssl}}) \leq 2c_6 t_n^{1+\alpha} + \frac{c_9 c_{10}}{(n_{\mathrm{l}} + g_v(\delta_n) n_{\mathrm{u}}^{\mathrm{aug}}) h_n^d} + 4 \exp\bigg\{-\frac{3}{14} \log^2 n_{\mathrm{u}}^{\mathrm{aug}})\bigg\} + \frac{2^{11} c_9 c_{10}}{g_v(\delta_n) t_n^2 n_{\mathrm{u}}^{\mathrm{aug}} h_n^d}.$$

Finally, we use a probabilistic lower bound of $n_{\mathrm{u}}^{\mathrm{aug}}$ to obtain the risk bound. Let $E$ denote the event $\min\{1 - \hat{m}_0(X), \hat{m}_0(X)\} \leq \delta_n$. By the triangle inequality, assumptions (A1) and (A6), we have

$$\begin{aligned}
&\mathbb{P}_{\mathrm{u}}(\min\{1 - \hat{m}_0(X), \hat{m}_0(X)\} \leq \delta_n) \\
&\geq \mathbb{P}_{\mathrm{u}}(\min\{1 - m(X), m(X)\} \leq \delta_n - c_8 n_0^{-r}) - \mathbb{P}_{\mathrm{u}}(|m(X) - \hat{m}_0(X)| > c_8 n_0^{-r}, E) \\
&\geq g_s(\delta_n - c_8 n_0^{-r})
\end{aligned}$$

Note that $n_{\mathrm{u}}^{\mathrm{aug}}$ is a sum of $n_{\mathrm{u}}$ IID Bernoulli random variables $Z$ with probability $\mathbb{P}(Z = 1) = \mathbb{P}_{\mathrm{u}}(\min\{1 - \hat{m}_0(X), \hat{m}_0(X)\} \leq \delta_n)$. By the Hoeffding's inequality, with probability at least $1 - 2\exp\{-n_{\mathrm{u}}(\tilde{n}_{\mathrm{u}}/n_{\mathrm{u}})^2/2\}$, we have

$$\frac{3\tilde{n}_{\mathrm{u}}}{2} \geq n_{\mathrm{u}}^{\mathrm{aug}} \geq \frac{\tilde{n}_{\mathrm{u}}}{2}, \quad \text{where } \tilde{n}_{\mathrm{u}} \triangleq g_s(\delta_n - c_8 n_0^{-r}) \cdot n_{\mathrm{u}}.$$

Therefore, we have

$$\begin{aligned}
\mathcal{R}(\hat{C}^{\mathrm{ssl}}) &= \mathbb{E}\mathcal{R}_{n_{\mathrm{u}}^{\mathrm{aug}}}(\hat{C}^{\mathrm{ssl}}) \\
&\leq 2c_6 t_n^{1+\alpha} + \frac{c_9 c_{10}}{(n_{\mathrm{l}} + g_v(\delta_n)\tilde{n}_{\mathrm{u}}/2) h_n^d} + 4\exp\bigg\{-\frac{3}{14}(\log \tilde{n}_{\mathrm{u}} - \log 2)^2)\bigg\} + \\
&\quad \frac{2^{11} c_9 c_{10}}{g_v(\delta_n) t_n^2 \tilde{n}_{\mathrm{u}} h_n^d/2} + \exp\bigg\{-\frac{n_{\mathrm{u}}}{2}\Big(g_s(\delta_n - c_8 n_0^{-r})\Big)^2\bigg\},
\end{aligned} \tag{36}$$

provided that the choices of (20), (32), and (33) are made, namely

$$t_n/2 > c_7 h_n^q, \quad 2\delta_n + 2c_8 n_0^{-r} \leq t_n/8, \quad t_n \log(3\tilde{n}_{\mathrm{u}}/2) \leq 1.$$

Choosing $h_n$, $t_n$, and $\delta_n$ at the rate of

$$h_n \sim n_{\mathrm{u}}^{-1/\{q(\alpha+3+v+s)+d\}}, \quad t_n \sim h_n^q, \quad \delta_n \sim h_n^q,$$

and invoking the assumption (A7), we can verify that the rate of convergence in (36) is at the order of

$$\mathcal{R}(\hat{C}^{\mathrm{ssl}}) \sim n_{\mathrm{u}}^{-q(\alpha+1)/\{q(\alpha+3+v+s)+d\}},$$

which concludes the proof.