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 μ_b and σ_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 γ and β can be aggregated as usual. We note that FedAvg with vanilla BN is not functional because the BN statistics μ and σ 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 μ and σ 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_{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.

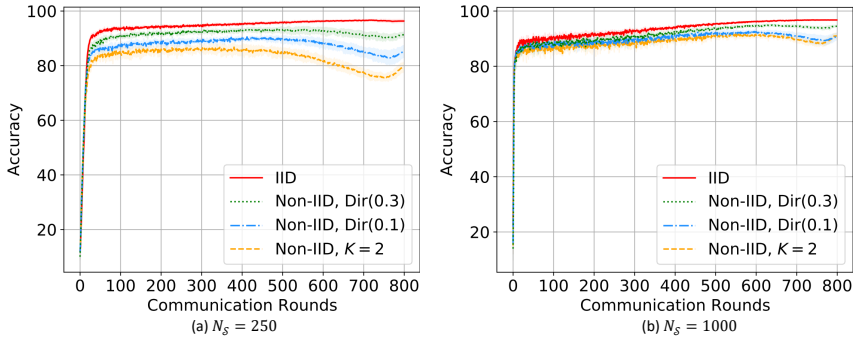

Figure 7: Experimental results for SVHN dataset with (a) $N_{\mathcal{S}} = 250$ and (b) $N_{\mathcal{S}} = 1000$.

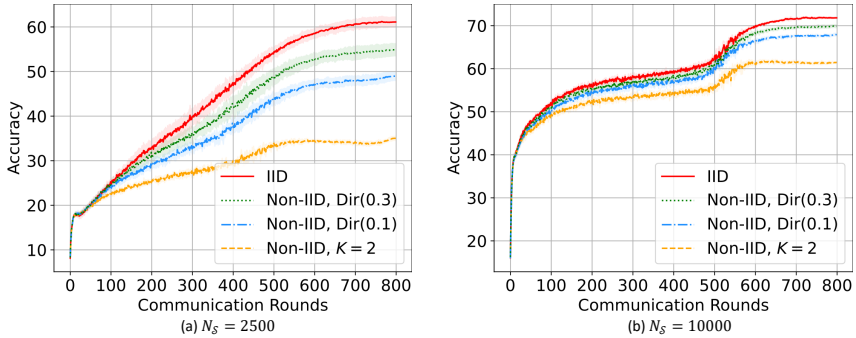

Figure 8: Experimental results for CIFAR100 dataset with (a) $N_{\mathcal{S}} = 2500$ and (b) $N_{\mathcal{S}} = 10000$.

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 β_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	β_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)

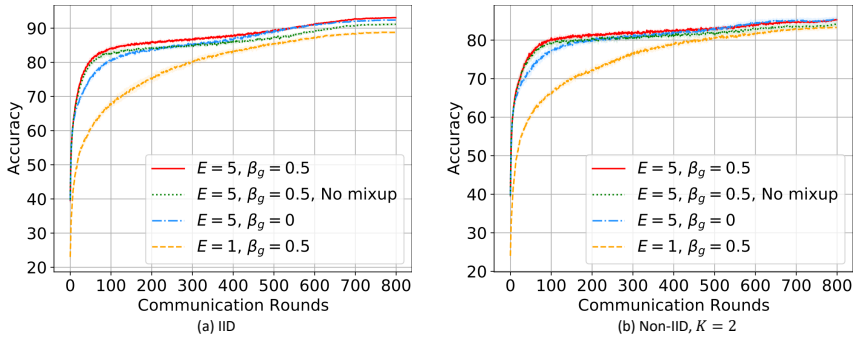

Figure 9: Ablation study of the CIFAR10 dataset with 4000 labeled data at the server for the cases of (a) IID and (b) Non-IID, $K = 2$ data partition.

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_{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^{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_{u} unlabeled data of (X^{u}), denoted by $\{X_j^{\mathrm{u}}\}_{j=1}^{n_{\mathrm{u}}}$, where each X^{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^{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.

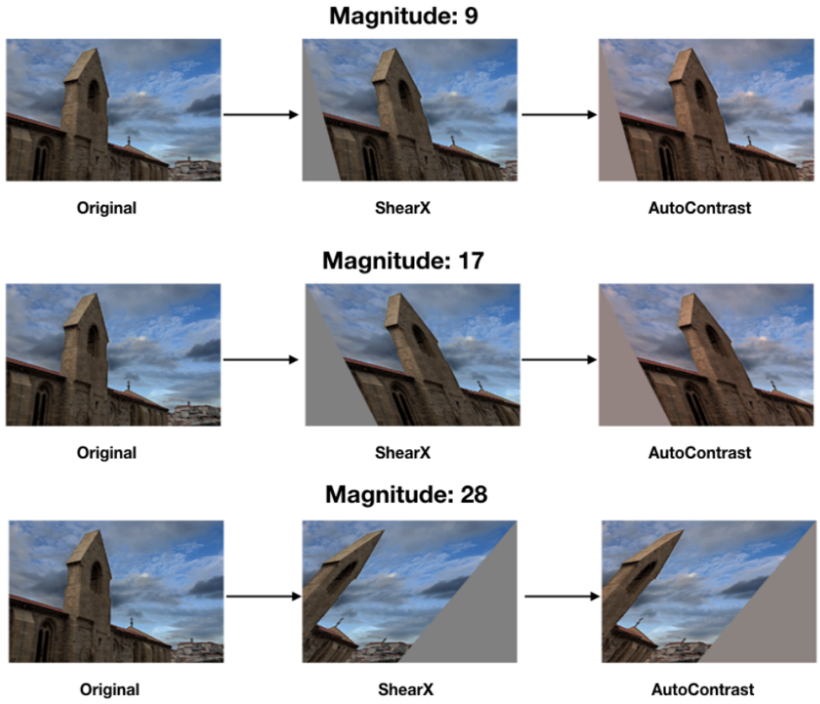

Figure 10: Examples of strong data augmentations based on the RandAugment technique [37]. As the distortion magnitude increases, the strength of the augmentation increases. Here, "ShearX" means shearing the image along the horizontal axis, and "AutoConstrast" means maximizing the image contrast by setting 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^{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^{u} and labeled X^{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 δ (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^{aug}.

• *Step 3*. Train an estimate of m, denoted by m^{ssl}, and the associated classifier C^{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^{l}, the data in D^{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_{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_{u} but eliminate n_{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^{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^{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 δ. For example, if $v = 0$, a sufficiently small δ (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 α 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 α means more separability of the two classes under the probability $\mathbb{P}_{\mathrm{l}}$. The boundedness of X^{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 Δ 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, α, 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 α, 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 α, 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_{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 α 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_{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_{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 (α), 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 δ 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., δ_n instead of δ (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_{l} and $n_{\text{u}}^{\text{aug}}$ denote the sample sizes of D^{l} and D^{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^{ssl} (which is implicitly used to define $\bar{m}$). For each x, we define the event

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

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

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

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

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

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

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

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 δ_n that satisfies $\delta_n \leq \min\{c_3, \Delta, 1/4\}$, where c_3 was introduced in Assumption (A2) and Δ 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_{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

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

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 δ_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.