# Towards Realistic Model Selection for Semi-supervised Learning

**Muyang Li** [1] **Xiaobo Xia** [1] **Runze Wu** [2] **Fengming Huang** [1] **Jun Yu** [3] **Bo Han** [4] **Tongliang Liu** [1]

## Abstract

Semi-supervised Learning (SSL) has shown remarkable success in applications with limited supervision. However, due to the scarcity of labels in the training process, SSL algorithms are known to be impaired by the lack of proper model selection, as splitting a validation set will further reduce the limited labeled data, and the size of the validation set could be too small to provide a reliable indication to the generalization error. Therefore, we seek alternatives that do not rely on validation data to probe the generalization performance of SSL models. Specifically, we find that the distinct margin distribution in SSL can be effectively utilized in conjunction with the model's spectral complexity, to provide a non-vacuous indication of the generalization error. Built upon this, we propose a novel model selection method, specifically tailored for SSL, known as **S**pectral-normalized **La**beled-margin **M**inimization (SLAM). We prove that the model selected by SLAM has upper-bounded differences w.r.t. the best model within the search space. In addition, comprehensive experiments showcase that SLAM can achieve significant improvements compared to its counterparts, verifying its efficacy from both theoretical and empirical standpoints.

## 1. Introduction

Attributed to recent advancements in deep learning methodologies, semi-supervised learning (SSL) has illustrated strong performances across multiple domains with minimal supervision (Sohn et al., 2020; Wang et al., 2022b; Huang et al., 2023). However, despite its success, model selection - one of the most fundamental problems in machine learning, has not been adequately addressed in SSL. Where existing works usually unrealistically assumed the existence of a labeled validation set that is much larger than the labeled training set itself (Rasmus et al., 2015), or simply omit the model selection process (Berthelot et al., 2019b), which are clearly problematic in real-world applications (Oliver et al., 2018). Instead, a more pragmatic strategy would be to split out a validation set from the available labeled data.

Nonetheless, this approach is fraught with difficulties, the inherent scarcity of labeled data in SSL contexts means that diverting any portion of it to a validation set will significantly hamper the performance of the SSL model. Furthermore, the limited size of the validation data could lead to ineffective model selection (Mohri et al., 2018). This conundrum underscores the urgent need for alternative methods for model selection within the SSL framework.

One potential way to address this dilemma is to directly estimate the generalization ability of the SSL model based only on the training data, which efficiently utilizes all the data at hand. In order for the estimation of generalization on training set, the notion of *hypothesis complexity*, plays a vital role in understanding generalization estimation from a philosophical perspective (Niyogi & Girosi, 1996). It is well-known that a learning model's generalization error bound can be decomposed to its empirical error and hypothesis complexity (Wei et al., 2020). When multiple models exhibit similar training errors, the one with lower complexity is anticipated to better generalize to unseen data, which aligns with the *Occam's razor* principle - the simplest hypothesis that consistent with our observations is more likely to be the correct one (MacKay, 2003; Lotfi et al., 2022).

However, as promising as it may sound, directly estimating an SSL model's generalization capability on the training sample is not easy, as two prominent obstacles arise: (1) despite the rigorous theoretical guarantee of said generalization error bounds, most existing generalization measures are not specifically designed for SSL context, as the influence of unlabeled data and noisy pseudo-labels are not accounted. Therefore, whether these measures can be effectively adapted to SSL remains a question to be explored (Mey & Loog, 2022); (2) the limited labeled sample poses an additional challenge, estimating the generalization performance of learning model heavily relies on the assumption that the observed labeled data distribution is faithful to

[1]Sydney AI Center, The University of Sydney [2]FUXI AI Lab, NetEase [3]University of Science and Technology of China [4]Hong Kong Baptist University. Correspondence to: Tongliang Liu <tongliang.liu@sydney.edu.au>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

the true distribution, which could be easily violated due tho the bias from limited observations (Wang et al., 2022a).

In response to the aforementioned challenges, in this study, we propose a novel model selection method for SSL, using training data only, known as **S**pectral-normalized **La**beled-margin **m**inimization (SLAM), which combines both the empirical error and hypothesis complexity to estimate the expected generalization error. More specifically, our investigation unveils that the classification margin of **labeled data only** can serve as an effective indicator to reflect the degree of overfitting for SSL model, where we attempt to use margin-based Probable-Approximate-Correct (PAC)-Bayes generalization metrics (McAllester, 2003; Bartlett et al., 2017; Neyshabur et al., 2017) as means to support such empirical observation with theoretical insights. Moreover, to account for the potential biases introduced by the limited volume of labeled data, we develop a local-consistency re-weighting measure to calibrate the potential bias in label data by up-weighs the representative sample, while down-weighs the bias and marginal sample.

We summarize our main contributions as follows: (1) We proposed a simple yet effective model selection method named SLAM, which is, to the best of authors' knowledge, the first model selection method that is applicable to SOTA SSL algorithms such as FixMatch (Sohn et al., 2020); (2) we prove that the model selected by SLAM has bounded differences w.r.t. optimal model within the pre-defined search space. More importantly, this difference is asymptotically governed by the SLAM metrics - which implies as SLAM is optimized, the model we end up will converge towards the optimal model; (3) through comprehensive empirical studies involving a series of relevant model selection methods, and commonly used benchmark datasets, we show that SLAM can surpass its most relevant SOTA counterparts, performs almost as good as selecting model on test data, which underscores the effectiveness of SLAM from both theoretical and empirical perspectives.

## 2. Related Work

**Model selection using unlabeled data.** Many machine learning applications involve training models using both labeled and unlabeled data, one intuitive question is whether can we use the more dispensable unlabeled data for model selection, or use them to estimate the model's generalization performances (Platanios et al., 2014; 2016; 2017). Extensive research has been conducted in this area, particularly in the field of unsupervised domain adaptation (UDA), which shares close resemblance to SSL (Morerio et al., 2018; Wei et al., 2020; Berthelot et al., 2021), the key difference is whether the unlabeled data exhibits distribution shift to the labeled data. Broadly, works in this area can be categorized into two main branches, one that leverages the labeled data loss, re-weighted by the density ratio between the marginal distribution between labeled and unlabeled, to obtain a risk-consistent estimation of the target population risk (Sugiyama et al., 2007; You et al., 2019), these approaches enjoys rigorous theoretical guarantees, but becomes trivial in the SSL setting, as the labeled data (source domain) cannot afford to split out a validation set. The second category relies solely on unlabeled data for model selection, such as considering the average prediction confidence on the unseen unlabeled data (Morerio et al., 2018), evaluating the neighborhood consistency (Saito et al., 2021; Hu et al., 2023). While these methods are more applicable to SSL, they are usually built upon heuristic intuition and do not have theoretical guarantees. Using a more concrete example to illustrate this weakness, as Saito et al. (2021) pointed out, those methods need the target model to be well-trained on the source data, so that they can exhibit meaningful indications on the unlabeled data, however, the validity of this assumption in SSL domain is still questionable. Specifically, we find that poorly trained SSL models can easily fall into the pitfalls of these heuristic-based methods.

Table 1: A summarization of model selection methods under SSL and UDA context.

| *Methods* | *Designed for SSL* | *Applicable to DNN* | *Require Valid. Split* |
|---|---|---|---|
| Stability (Lange et al., 2002) | ✓ | ✗ | ✗ |
| Co-Validation (Madani et al., 2004) | ✓ | ✗ | ✗ |
| EB-criterion (Mahsereci et al., 2017) | ✗ | ✓ | ✗ |
| DEV (You et al., 2019) | ✗ | ✓ | ✓ |
| SND (Saito et al., 2021) | ✗ | ✓ | ✓ |
| QLDS (Feofanov et al., 2023) | ✓ | ✗ | ✗ |
| MixVal (Hu et al., 2023) | ✗ | ✓ | ✓ |
| SLAM | ✓ | ✓ | ✗ |

**Model selection without validation data.** In cases where we do not even wish to dispense unlabeled data for validation purposes, some methods do not require validation data at all, where the model selection is purely built upon training data. Specifically, Lange et al. (2002) explored the possibility of model selection using training data only, via the notion of uniform-stability (Bousquet & Elisseeff, 2002). For learning with label noise, (Yuan et al., 2024) proposed early-stopping without validation set by tracing the learning stages. Under Tsybakov Margin condition (Tsybakov, 2004), Feofanov et al. (2023) developed QLDS, a margin separation approach inspired by the random matrix theory for model selection. However, those methods are either not specifically tailored for deep learning context, making them computationally inefficient, and sometimes even infeasible to compute, or does not reconciles with the Semi-supervised Learning setting without further modifications.

**Semi-supervised learning.** As one of the most fundamen-

tal research areas in the field of machine learning, SSL has continuously attracted notable attention for decades. Predominantly, these advancements leverage self-training (Wei et al., 2021; Chen et al., 2022) and pseudo-labeling techniques (Lee et al., 2013; Oh et al., 2022; Wang et al., 2022a), where models begin by learning from a modest amount of labeled data before extending their knowledge through pseudo-labels assigned to unlabeled data. Pseudo-labels deemed reliable are then added to the training dataset (Zhang et al., 2021a; Guo & Li, 2022; Wang et al., 2022b), where the criteria are usually based on prediction confidence (Berthelot et al., 2021; Xu et al., 2021; Xia et al., 2021; Li et al., 2024; Wu et al., 2024) or prediction discrepancy (Xia et al., 2023), facilitating a gradual enhancement in the model's ability to generalize to new, unseen data (Wei et al., 2020). More recently, the focus has shifted towards refining SSL performance through data augmentation and consistency regularization techniques (Xie et al., 2020; Sohn et al., 2020; Zheng et al., 2022). These methods apply varying degrees of transformation to the input data, ensuring that the model's output remains consistent across these transformations, thereby further improving the model's robustness (Xie et al., 2020).

## 3. Preliminaries

**Notations.** Under the standard setup of SSL, we have a small group of labeled instances $X_l := \{\boldsymbol{x_1}, ..., \boldsymbol{x_n}\}$ each paired with corresponding labels $Y_l := \{y_1, ..., y_n\}$, we also have a larger collection of unlabeled instances $X_u := \{\boldsymbol{x_{n+1}}, ..., \boldsymbol{x_{n+m}}\}$. Usually, we have $m >> n$, and all the data from observed dataset $D := \{X_l, Y_l, X_u\}$ are identically and independently sampled from an unknown distribution $\mathcal{D} : \mathcal{X} \times \mathcal{Y}$.

**Model Selection.** During the training process of machine learning algorithms, different models will be generated, forming a finite hypothesis set $\mathcal{F}$. One long-standing dilemma is choosing the one that will exhibit the best generalization performance. To solve this issue, we must first know what is the best possible model we can select, let's define the *Bayes error* $R^*$, which is the smallest generalization error that could be achieved from a given model family.

**Definition 3.1** (Mohri et al., 2018)**.** Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the Bayes error is the infimum of the generalization errors achieved by measurable functions $f : \mathcal{X} \to \mathcal{Y}$:

$$R^* = \inf_{f \in \mathcal{F}} R(f), \tag{1}$$

a model $f$ with $R(f) = R^*$ is known as the Bayes optimal model, which we will denote as $f^*$.

The difference between the error of our obtained model and the Bayes error is known as the *excess error*, which can be defined as $R(f) - R^*$ (Mohri et al., 2018). Our objective is therefore to find the model that minimizes excess error. However, whilst $R^*$ is often discussed in the machine learning domain, in reality, it is usually believed that $R^*$ is inaccessible, since the best model might not be appearing the pre-defined hypothesis set. Instead, we consider a simplified task, which is finding the best model within the hypothesis set $\mathcal{F}$, which can be seen as a pre-defined search space for all candidate models, a notion that is much more aligned to the task of model selection (You et al., 2019).

**Spectral-normalized margin bound.** Bounding the generalization error of the learning model has been the ultimate pursuit of learning theory, the challenging part is to find an appropriate measure to account for the complexity of the model. Recent studies have indicated that conventional complexity measures such as the Vapnik–Chervonenkis (VC) dimension and Rademacher Complexity are potentially vacuous for high capacity models such as DNNs (Zhang et al., 2021b). Instead, a more favorable notion that accounts for the classification margin and Spectral Complexity was proposed (Bartlett et al., 2017). Defining a $\rho_i$-Lipschitz continuous DNN $f_{\mathcal{A}}$ with depth $L$, parameterized by weight matrices $\{A_1, ..., A_L\}$ and non-linear activation functions $\{\sigma_1, ..., \sigma_L\}$, such that:

$$f_{\mathcal{A}}(\boldsymbol{x}) := \sigma_L(A_L \sigma_{L-1}(A_{L-1} \dots \sigma_1(A_1 \boldsymbol{x}) \dots)). \tag{2}$$

$f_{\mathcal{A}}$ is a mapping function such that $f_{\mathcal{A}} : \mathbb{R}^d \to \mathbb{R}^c$, where $c$ is the cardinality of the finite label set.

Let $\{M_1, ..., M_L\}$ be a set of reference matrices with identical dimensions to $A_1, ..., A_L$, which are used to anchor the degree of deviation of the weight matrices (Bartlett et al., 2017). Let $\|\cdot\|_\sigma$ denote the spectral norm of the weighted matrices, and $\|\cdot\|_{p,q}$ denote the $(p, q)$ matrix norm. The spectral complexity for weighted matrices $\mathcal{A}$ is:

$$\mathcal{R}_{\mathcal{A}} := \left( \prod_{i=1}^{L} \rho_i \|A_i\|_\sigma \right) \left( \sum_{i=1}^{L} \frac{\|A_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}} \right)^{3/2}. \tag{3}$$

Given the spectral complexity for $f_{\mathcal{A}}$, the following generalization error bound can be derived:

**Theorem 3.2** (Bartlett et al., 2017)**.** *For $(x, y)$ drawn i.i.d from any probability distribution over $\mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$ over $((x_i, y_i))_{i=1}^n$, every margin $\gamma > 0$ and network $f_{\mathcal{A}}$ satisfy:*

$$R(f_{\mathcal{A}}) \leq \hat{R}_D(f_{\mathcal{A}}) + \widetilde{\mathcal{O}} \left( \frac{\|X\|_2 \mathcal{R}_{\mathcal{A}}}{\gamma_f} \ln(W) + \sqrt{\frac{\ln(1/\delta)}{n}} \right),$$

*where $\hat{R}_D$ is the empirical error defined in Appendix B.1, $W$ is the maximum width of $f_{\mathcal{A}}$, and $\|X\|_2 = \sqrt{\sum_i \|x_i\|_2^2}$ is the L2 norm of the feature matrix.*

For a more precise characterization, here we define the total classification margin of labeled data:

$$\gamma_f = \sum_i^n \left( f(\boldsymbol{x}_i)_{y_i} - \max_{j \neq y_i} f(\boldsymbol{x}_i)_j \right). \tag{4}$$

where $f(\boldsymbol{x_i})_{y_i}$ refers to the $y_i$-th index of $f$'s prediction.

Essentially, Theorem 3.2 states that, the generalization error of an over-parameterized learning model such as $f_{\mathcal{A}}$ can be bounded with a composition of its classification margin and complexity, comparing with the more conventional generalization measures that only consider the classification margin (Antos et al., 2002), these Neo-generalization measures further accounts for the possibility that the learning model will continuously grow its complexity to memorize all data, hence maximizing margin while exhibits overfitting at the same time. For simplicity, with certain abuse to the notations, we simplify Theorem 3.2 by removing certain constant and low-order terms:

$$R(f_{\mathcal{A}}) \leq \hat{R}_D(f_{\mathcal{A}}) + \widetilde{\mathcal{O}} \left( \frac{\mathcal{R}_{\mathcal{A}}}{\gamma_f} + \sqrt{\frac{\ln(1/\delta)}{n}} \right). \tag{5}$$

In the following sections, to distinguish the spectral complexity for different models, $\mathcal{R}_{\mathcal{A}}$ will be denoted by $\mathcal{R}_{f_{\mathcal{A}}}$.

## 4. Method

Delving into the specifics of our proposed model selection method, SLAM, we uncover its key advantage: the elimination of the need for any validation data, whether labeled or unlabeled, to facilitate model selection. This attribute positions SLAM as a more data-efficient option compared to other relevant approaches (You et al., 2019; Saito et al., 2021; Hu et al., 2023). This is built upon the direct estimation of the model's generalization capability, a topic that remains unexplored for the SSL model. A crucial initial step in our approach is establishing the rationale behind using the classification margin of labeled training data as a reliable indicator of an SSL model's tendency to overfit. Additionally, by incorporating the concept of spectral complexity, SLAM further minimizes the risk of choosing an overfitting-prone model by favoring those with simpler structures. Finally, we draw on the principles of structural risk minimization (Koltchinskii, 2001) to validate our approach. Through this lens, we demonstrate that SLAM maintains a bounded difference from an informed oracle, providing a theoretical guarantee for its effectiveness.

### 4.1. Margin distribution in SSL

In traditional machine learning frameworks, the classification margin of training data for models with smaller capacities is often viewed as a reflection of their ability to generalize (Antos et al., 2002; McAllester, 2003). However, this

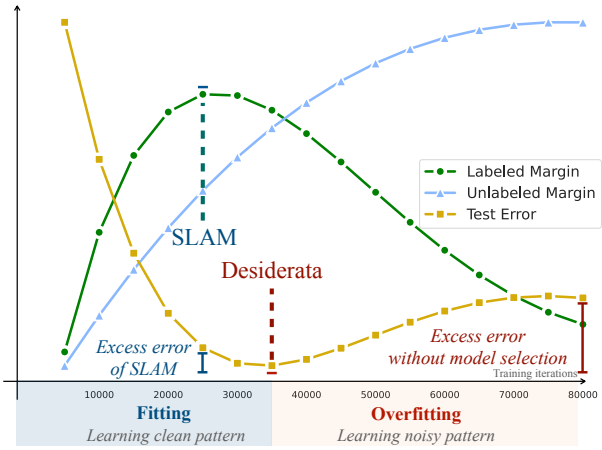



Figure 1: A visualization of the distribution of classification margin and test error is SSL training, this example uses the statistics generated from the training of MixMatch on CIFAR-10, with 40 labeled examples. In order to better observe the trend, unit scales are removed so that they can be summarized in a single figure. "Desiderata" refers to the best model defined on test set accuracy. As we can observe, SLAM aids the model selection of SSL by significantly reducing the excess error.

assumption does not hold for modern deep learning models, as these models have sufficient capacity to memorize training samples, allowing them to exhibit large margins while simultaneously overfitting. This leads to the ineffectiveness of conventional margin-based generalization measures in DNNs under the fully-supervised learning setup.

However, our primary finding in the SSL context offers a fresh perspective: the classification margin, when applied exclusively to labeled data, can often indicate the level of overfitting for the SSL model. This emerges in scenarios where the volume of unlabeled data significantly exceeds that of labeled data. In such cases, the model, while assigning pseudo-labels to unlabeled data, is prone to overfits on these noisy pseudo-labels. This overfitting is manifested as a shift in the model's outputted margin distribution, transitioning from the distribution conditioned on clean classes, $P(Y|X)$, towards that conditioned on noisy classes, $P(\hat{Y}|X)$ (Liu & Tao, 2015). By examining the labeled data $\boldsymbol{x_l}$, we can observe and track this probability distribution shift. Specifically, if the posterior probability $P(\hat{Y}|X = \boldsymbol{x_l})$ for the labeled data deviates from its expected value (which ideally should be close to 1), it indicates the extent of the model's overfitting to the noisy pseudo-labels. This insight offers a valuable method for evaluating the generalization performance of DNN models, particularly in the context of SSL.

Specifically, as shown in Figure 1, we can primarily categorize the learning process of the SSL model into two main stages. Initially, the model tends to fit the underlying true pattern, evidenced by a mutual increase in both the classifi-

cation margin of labeled and unlabeled data and a decrease in the test error. During this phase, the model's confidence in its predictions for both labeled and unlabeled data progressively strengthens. However, after a certain pivotal point, the previously observed mutual increase in margins ceases, giving way to a divergent trend: the margin of unlabeled data continues to rise monotonically, while the margin of labeled data starts to decrease, concurrently, the generalization error also start to increase around this juncture. These signs are indicative of the model beginning to overfit the noisy pseudo-labels assigned to the unlabeled data. As the model increasingly maximizes the classification margin based on these incorrect patterns, the classification margin associated with the clean labeled data correspondingly diminishes.

### 4.2. Local-consistency re-weighting

Nevertheless, using labeled margin alone in SSL still has a notable drawback, that is, due to the limited observations, the distribution of labeled data could significantly differ from the underlying true distribution, hence simply maximizing the labeled margin, without the consideration of the representativeness of those data points, can lead to the selection of the biased model. To mitigate this issue, we propose a novel approach to account for the importance of each individually labeled datum, named *local-consistency re-weighting*. Local-consistency re-weighting aims to assign higher weights to labeled data that are more representative to its class and assign lower weights to data that are potentially biased or marginal to its class distribution.

**Algorithm 1** Local-consistency re-weighting.

**Require:** SSL model $f$, labeled dataset $X_l, Y_l$, unlabeled instances $X_u$

1: **compute** the labeled margin $\gamma_f$ as defined in Equ. 4
2: **for** $i = 1, \ldots, n$ **do**
3: **sample** the KNNs of $\boldsymbol{x}_i$ from unlabeled data $X_u$, denote as $\{\boldsymbol{x}_1^i, \cdots, \boldsymbol{x}_k^i\}$
4: **compute** the pseudo-labels for unlabeled KNNs as $\{y_1^i, \cdots, y_k^i\} := f(\{\boldsymbol{x}_1^i, \cdots, \boldsymbol{x}_k^i\})$
5: **compute** the local-consistency weights for $(\boldsymbol{x}_i, y_i)$ using Equ. 7
6: **end for**
7: **return** the local-consistency weights array

For each labeled datum, we sample its K-nearest-neighbor (KNN) from the unlabeled dataset, and calculate the pseudo-label consistency, defined as follows:

$$\mathcal{W}((x,y)_i, K) = \frac{\sum_{j=1}^{K} \mathbb{1}(\hat{y}_j^i = y^i)}{K}. \tag{6}$$

Where $\mathbb{1}(\cdot)$ is an indicator function. This approach upweighs labeled data that has a neighborhood with consistent pseudo-labels, and down-weighs labeled data with inconsistent pseudo-label neighbors, as this suggests that those samples lie on the borderline of their class distribution, and are far from the class centroid, and thus should be given less weights when we wish to maximizing the margin.

Lastly, to ensure the unit scale does not change, we apply normalization to ensure the reweighted weights do not change the sum of the margin:

$$\overline{\mathcal{W}}((x,y)_i, K) = n \cdot \frac{\mathcal{W}((x,y)_i, K)}{\sum_{i=1}^{n} \mathcal{W}((x,y)_i, K)}. \tag{7}$$

### 4.3. Spectral complexity estimation

While the margin alone can be informative in SSL due to its unique distribution characteristics, it could not tell us anything about the model's complexity. If the model's complexity grows unbounded, it could eventually memorize all labeled and unlabeled data in a brute-force manner, rendering margin ineffective. Therefore, in this part, we introduce the classical complexity measure, that has been commonly employed in other literature, known as spectral norm (Bartlett et al., 2017; Miyato et al., 2018). This notion alone differs from our previously defined spectral complexity $\mathcal{R}_{f_\mathcal{A}}$, as $\mathcal{R}_{f_\mathcal{A}}$ aims to measure the model as a whole, whereas the spectral norm measures each layer individually. Nevertheless, we follow the commonly used surrogate approach, which is to use the products of the layer-wise spectral norm to approximate the spectral complexity (Jiang et al., 2019; Yang et al., 2023). We describe such measure as the empirical spectral complexity $\hat{\mathcal{R}}_{f_\mathcal{A}}$:

$$\hat{\mathcal{R}}_{f_\mathcal{A}} := \prod_{i=1}^{L} \|A_i\|_\sigma = \prod_{i=1}^{L} \max(\text{diag}(\Sigma(A_i))), \tag{8}$$

where $\Sigma(A_i)$ is the singular value of weight matrix $A_i$.

### 4.4. Spectral-normalized Labeled-margin Minimization

After knowing the classification margin of labeled data, and the empirical spectral complexity, we now present the main SLAM objective:

$$\phi_k(f) = \hat{R}_D(f) + \frac{\hat{\mathcal{R}}_{f_\mathcal{A}}}{\gamma_f} + \sqrt{\frac{\log k}{n}}. \tag{9}$$

where $\phi_k$ is the SLAM metrics for the $k$-th model from the hypothesis set. Consequently, $\hat{\mathcal{R}}_{f_\mathcal{A}}$ and $\gamma_f$ are the empirical spectral complexity and labeled margin for the $k(f)$-th model, where $k$ can be view as an index function.

Thus, the model that minimizes aforementioned objective on dataset $D$ is denoted as $f_D^\dagger$:

$$f_D^\dagger = \underset{\gamma_f, \hat{\mathcal{R}}_{f_\mathcal{A}}}{\arg\min}\, \phi_k(f) = \underset{\gamma_f, \hat{\mathcal{R}}_{f_\mathcal{A}}}{\arg\min} \left[ \hat{R}_D(f) + \frac{\hat{\mathcal{R}}_{f_\mathcal{A}}}{\gamma_f} + \sqrt{\frac{\log k}{n}} \right]. \tag{10}$$

Overall, the takeaway message that can summarize the underlying rationale of SLAM is that: *SLAM tries to find the simplest model that can best separates the labeled data, while exhibiting small training loss*.

### 4.5. Theoretical Analysis

In this section, we present a formal theoretical generalization guarantee for the model selected by SLAM. This guarantee is established through a PAC-style generalization error bound. Our proof is grounded in the principle of structural risk minimization (SRM) (Koltchinskii, 2001; 2006), which warrants and motivates model selection directly based on the training set and generalization estimation, a process that is also known as "bound-minimization", as we trying to select the model with smallest estimated generalization error bound. Specifically, SRM first assumes that the target hypothesis class $\mathcal{F}$ can be decompose into a finite set, and for every hypothesis set $\mathcal{F}_k$ within $\mathcal{F}$ is assumed to be nested, i.e. $\mathcal{F}_k \in \mathcal{F}_{k+1}$.

To demonstrate that our proposed method is provably close to the best model within the pre-defined hypothesis set, extending on Mohri et al. (2018)'s proof, we have the following inequality:

**Theorem 4.1.** *We can obtain the following inequality w.r.t the spectral-normalized margin complexity. With probability at least 1-δ:*

$$R(f_D^\dagger) \leq \inf_{f \in \mathcal{F}} \left( R(f) + 2\frac{\hat{\mathcal{R}}_{f_{\mathcal{A}}}}{\gamma_f} + \sqrt{\frac{\log k}{n}} \right) + \sqrt{\frac{2\log(3/\delta)}{n}}.$$

Theorem 4.1 demonstrates that the generalization error of the model selected by SLAM is primarily upper-bounded by the best-in-set model, with the penalization of hypothesis complexity and the sample complexity term. This can be proved using McDiarmid's inequality, which is an extension of Azuma and Hoeffding's concentration inequality, whom are commonly used to bound differences between two sequences. See Appendix B.3 for full proof.

Built upon Theorem 4.1, we can straight-forwardly arrive at the following oracle inequality:

**Corollary 4.2.** *With probability of at least 1 - δ, we have the following inequality for the model returned by our model selection function:*

$$R(f_D^\dagger) - R(f^*) \leq 2\frac{\hat{\mathcal{R}}_{f^*_{\mathcal{A}}}}{\gamma_f^*} + \sqrt{\frac{2\log(3/\delta)}{n}} + \sqrt{\frac{\log k(f^*)}{n}}.$$

*Remark* 4.3. This result directly sets the upper bound on the excess error between the model returned by SLAM and the best model within the pre-defined hypothesis set. This means, in the worst-case scenario, the SSL model selected by our method can not be worse than the optimal model up to a polynomial factor of the right-hand side (R.H.S).

## 5. Experiment

### 5.1. Setup

**Datasets.** We evaluate our methods on a series of commonly used benchmark datasets in Semi-supervised Learning (SSL): CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009). To create different level of available supervision signals, following default setting (Sohn et al., 2020; Wang et al., 2022b), for CIFAR-10, we randomly sample {4,25,400} labeled data per-class, for CIFAR-100, we randomly sample {4,25,100} labeled data per-class.

**Baselines.** While there lacks of prior methods to directly perform model selection on SOTA SSL methods, we can migrate relevant approaches from other similar domains as attempts to address this issue. First, we consider the methods that focus on model selection without a validation set, such as EB-criterion (Mahsereci et al., 2017). In addition, another highly relevant line of work is Unsupervised Domain Adaption (UDA), where methods such as density-reweighed cross-validation (Sugiyama et al., 2007; You et al., 2019) and SND (Saito et al., 2021), MixVal (Hu et al., 2023) have received notable success in perform model selection on unknown target domain.

**Target SSL algorithms.** Since the model selection on SOTA SSL baselines remains unexplored, in this study, we will conduct all model selection methods on several popular and representative SSL algorithms. Namely, in this study, we use MixMatch (Berthelot et al., 2019b), ReMixMatch (Berthelot et al., 2019a) and FixMatch (Sohn et al., 2020) as target SSL algorithms.

**Implementation details.** We follow the commonly used implementation details in SSL (Berthelot et al., 2019b; Sohn et al., 2020; Zhang et al., 2021a), where we use WideResNet-28-2 (Zagoruyko & Komodakis, 2016) for CIFAR-10. For computational efficiency reasons, we make minor modifications to use WideResNet-28-2 for CIFAR-100 instead of WideResNet-28-8. All models are trained for $2^{20}$ iterations.

### 5.2. Results

To comprehensively evaluate the capability of SLAM, we conduct several realistic model selection tasks under SSL settings, including early-stopping, hyper-parameter selection, and model selection against train/val splitting.

**Early-stopping.** For the early-stopping task, our objective is to select the model from the best epoch, where the rest of the settings are aligned with the default setup.

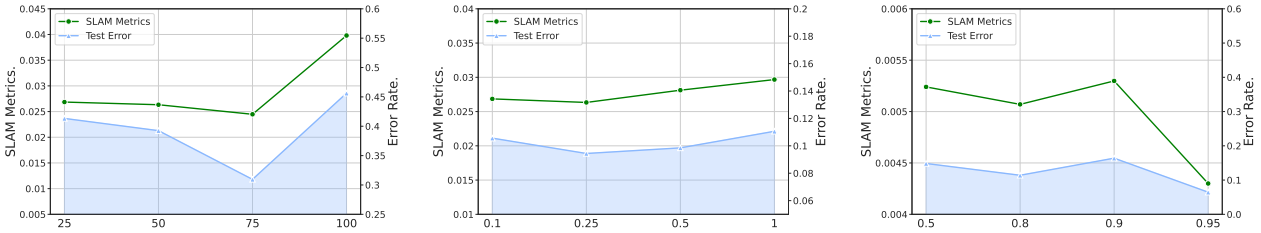
For the CIFAR-10 dataset, our first observation is that in nearly all cases, the model selected by SLAM is within one standard deviation of the optimal model. In addition, we note the gap between the model selected by SLAM and the next best baselines in MixMatch(40) and MixMatch(250),

Table 2: Early-stopping performances on CIFAR-10, where the best performing method is **bold**, and the next best method is underlined. †: Best model based on test set performance.

| METHOD | MIXMATCH | | | REMIXMATCH | | | FIXMATCH | | |
|---|---|---|---|---|---|---|---|---|---|
| # LABEL | 40 | 250 | 4000 | 40 | 250 | 4000 | 40 | 250 | 4000 |
| EB-CRITERION | $32.86 \pm 1.80$ | $61.19 \pm 5.91$ | $87.10 \pm 4.20$ | $50.57 \pm 4.14$ | $59.12 \pm 0.13$ | $91.16 \pm 0.71$ | $92.95 \pm 1.28$ | $94.40 \pm 0.18$ | $94.86 \pm 0.55$ |
| DEV | $52.65 \pm 4.30$ | $79.14 \pm 8.02$ | $92.64 \pm 0.45$ | $79.41 \pm 14.7$ | $92.28 \pm 0.28$ | $94.72 \pm 0.13$ | $93.19 \pm 0.97$ | $94.56 \pm 0.15$ | $95.25 \pm 0.07$ |
| ENTROPY | $55.47 \pm 5.98$ | $79.75 \pm 0.76$ | $88.00 \pm 0.26$ | $86.99 \pm 1.61$ | $89.96 \pm 1.15$ | $93.37 \pm 0.15$ | $72.38 \pm 7.81$ | $93.10 \pm 0.24$ | $94.29 \pm 0.37$ |
| SND | $51.40 \pm 4.08$ | $79.51 \pm 0.21$ | $91.63 \pm 0.09$ | $\mathbf{89.92 \pm 1.76}$ | $93.16 \pm 0.44$ | $94.80 \pm 0.02$ | $63.12 \pm 2.69$ | $94.63 \pm 0.12$ | $94.90 \pm 0.72$ |
| MIXVAL | $51.40 \pm 4.08$ | $79.71 \pm 2.55$ | $91.67 \pm 0.39$ | $89.79 \pm 1.38$ | $93.08 \pm 0.51$ | $94.77 \pm 0.16$ | $36.08 \pm 18.4$ | $83.32 \pm 17.9$ | $95.21 \pm 0.09$ |
| SLAM | $\mathbf{59.99 \pm 2.79}$ | $\mathbf{84.47 \pm 0.87}$ | $\mathbf{92.97 \pm 0.14}$ | $89.79 \pm 1.32$ | $\mathbf{93.43 \pm 0.15}$ | $\mathbf{94.90 \pm 0.24}$ | $\mathbf{94.43 \pm 0.73}$ | $\mathbf{94.76 \pm 0.14}$ | $\mathbf{95.51 \pm 0.20}$ |
| OPTIMAL† | $62.97 \pm 4.53$ | $86.40 \pm 0.93$ | $93.33 \pm 0.15$ | $89.95 \pm 0.91$ | $93.59 \pm 0.13$ | $95.19 \pm 0.07$ | $94.50 \pm 0.69$ | $94.95 \pm 0.08$ | $95.68 \pm 0.14$ |

Table 3: Early-stopping performances on CIFAR-100, where the best performing method is **bold**, and the next best method is underlined. †: Best model based on test set performance.

| METHOD | MIXMATCH | | | REMIXMATCH | | | FIXMATCH | | |
|---|---|---|---|---|---|---|---|---|---|
| # LABEL | 400 | 2500 | 10000 | 400 | 2500 | 10000 | 400 | 2500 | 10000 |
| EB-CRITERION | $19.91 \pm 0.84$ | $48.08 \pm 5.44$ | $66.51 \pm 0.69$ | $33.82 \pm 2.96$ | $65.51 \pm 0.43$ | $73.46 \pm 0.39$ | $15.89 \pm 8.73$ | $54.97 \pm 7.63$ | $58.48 \pm 0.16$ |
| DEV | $25.16 \pm 1.13$ | $53.12 \pm 0.87$ | $66.33 \pm 0.36$ | $43.96 \pm 2.39$ | $65.13 \pm 0.57$ | $73.24 \pm 0.37$ | $45.06 \pm 2.48$ | $64.23 \pm 0.56$ | $70.85 \pm 0.72$ |
| ENTROPY | $16.81 \pm 0.72$ | $41.55 \pm 0.34$ | $57.61 \pm 5.66$ | $37.46 \pm 2.16$ | $61.93 \pm 0.94$ | $69.67 \pm 1.52$ | $37.03 \pm 3.64$ | $58.24 \pm 1.51$ | $66.03 \pm 1.32$ |
| SND | $20.03 \pm 0.52$ | $54.07 \pm 0.59$ | $63.31 \pm 0.10$ | $36.90 \pm 5.34$ | $60.13 \pm 0.50$ | $69.46 \pm 0.58$ | $39.04 \pm 5.70$ | $59.17 \pm 2.07$ | $70.47 \pm 0.10$ |
| MIXVAL | $25.17 \pm 0.85$ | $53.49 \pm 0.30$ | $66.52 \pm 0.60$ | $27.27 \pm 2.11$ | $64.53 \pm 0.78$ | $72.76 \pm 0.54$ | $39.04 \pm 5.70$ | $64.13 \pm 2.99$ | $70.57 \pm 0.67$ |
| SLAM | $\mathbf{25.45 \pm 0.29}$ | $\mathbf{54.28 \pm 0.60}$ | $\mathbf{66.80 \pm 0.05}$ | $\mathbf{45.45 \pm 2.66}$ | $\mathbf{65.77 \pm 0.73}$ | $\mathbf{73.52 \pm 0.61}$ | $\mathbf{45.35 \pm 1.74}$ | $\mathbf{64.33 \pm 0.80}$ | $\mathbf{71.27 \pm 0.59}$ |
| OPTIMAL† | $26.26 \pm 1.14$ | $54.59 \pm 0.51$ | $67.20 \pm 0.31$ | $45.96 \pm 2.48$ | $66.05 \pm 0.50$ | $73.80 \pm 0.69$ | $45.57 \pm 1.41$ | $65.06 \pm 0.20$ | $72.01 \pm 0.16$ |




(a) MixMatch Unlabel Loss Ratio (b) ReMixMatch Rotation Loss Ratio (c) FixMatch Pseudo-label Threshold

Figure 2: Results for hyper-parameter selection on CIFAR-10(40).

are more significant than other cases, which we attributed to two main reasons: (1) for models that are more robust, such as ReMixMatch and FixMatch, or cases with more labeled data, such as MixMatch(4000), the overfitting is more benign, to wit, even overfitting has already occurred, the model can still generalize reasonably well to the unseen data (Cao et al., 2022); (2) the proposed SLAM method is less sensitive to the bias and incoherent distribution due to the limited labeled data size, which is aligned with our hypothesis and verified the efficacy of SLAM.

For the CIFAR-100 dataset, we can observe similar trends in terms of the SLAM performances against other baselines. When applied to strong base models such as ReMixMatch and FixMatch, there are baselines whose performances are reasonably close to SLAM (e.g. EB-Criterion, DEV), but no baseline method consistently exhibits robust performances. Moreover, while DEV appears to be the best alternative for SLAM, we emphasize that DEV undertook necessary modifications to adapt the SSL setting, which makes it a proxy for minimizing labeled data loss. This approach is, in principle, similar to labeled margin maximization, therefore, the advantageous performance of SLAM can be mainly attributed to the local consistency re-weighting and spectral complexity minimization.

Overall, our experiment results find that popular model selection methods defined for unsupervised model selection can indeed be borrowed to SSL to a certain extent, however, no baseline method can consistently exhibit robust performances across different target models, and different benchmark datasets, this finding underscores the cruciality of proposing model selection method specifically designed for SSL, also justified the empirical superiority of SLAM as a model selection method.

**Early-stopping on Out-of-Class data.** We also evaluate the performance of model selection on dataset that exist Out-of-Class (OOC) data (Su et al., 2021), specifically, we adapt Semi-Aves dataset as benchmark to evaluate the performance of model selection methods in OOC cases (Huang et al., 2021; Su & Maji, 2021). As shown in Table 5, we can

Table 4: Early-stopping performance against train/validation split on CIFAR-10, where the best performing method is **bold**, and the next best method is underlined. †: Best model based on test set performance.

| METHOD | MIXMATCH | | REMIXMATCH | | FIXMATCH | |
|---|---|---|---|---|---|---|
| # LABEL | 250 | 4000 | 250 | 4000 | 250 | 4000 |
| VALIDATION-(10%) | $68.64 \pm 1.56$ | $88.11 \pm 0.00$ | $90.64 \pm 1.65$ | $94.63 \pm 0.12$ | $81.25 \pm 1.24$ | $93.68 \pm 0.51$ |
| VALIDATION-(10%)† | $72.60 \pm 1.10$ | $88.91 \pm 0.71$ | $93.26 \pm 0.48$ | $95.06 \pm 0.08$ | $95.09 \pm 0.21$ | $95.57 \pm 0.19$ |
| VALIDATION-(20%) | $68.15 \pm 4.23$ | $88.29 \pm 0.14$ | $92.10 \pm 0.21$ | $94.69 \pm 0.28$ | $90.30 \pm 2.06$ | $94.91 \pm 0.55$ |
| VALIDATION-(20%)† | $71.17 \pm 1.35$ | $88.73 \pm 0.09$ | $93.61 \pm 0.25$ | $95.15 \pm 0.08$ | $95.12 \pm 0.18$ | $95.68 \pm 0.13$ |
| SLAM | $\mathbf{84.47 \pm 0.87}$ | $\mathbf{92.97 \pm 0.14}$ | $\mathbf{93.43 \pm 0.15}$ | $\mathbf{94.90 \pm 0.24}$ | $\mathbf{94.76 \pm 0.73}$ | $\mathbf{95.51 \pm 0.20}$ |

Table 5: Early-stopping performance on Semi-Aves.

| METHOD | MIXMATCH | FIXMATCH |
|---|---|---|
| # LABEL | 3959 | 3959 |
| ENTROPY | $61.56 \pm 0.20$ | $65.28 \pm 0.50$ |
| SND | $61.83 \pm 0.76$ | $67.64 \pm 0.21$ |
| MIXVAL | $58.23 \pm 0.67$ | $52.62 \pm 0.36$ |
| DEV | $60.10 \pm 0.60$ | $65.74 \pm 1.89$ |
| SLAM | $\mathbf{62.08 \pm 0.84}$ | $\mathbf{67.80 \pm 0.03}$ |
| OPTIMAL | $62.78 \pm 0.16$ | $68.03 \pm 0.03$ |

observe that SLAM consistently maintains strong performance, performing almost as good as the model with best test set performances.

**Hyper-parameter selection.** Except for early-stopping during model training, another challenge for the model selection of SSL models is the hyper-parameter selection. It is well-known that SSL algorithms are hyper-parameter-intensive, and the choice of hyper-parameter can significantly affect the final performance. In this part, we give an empirical evaluation of the effectiveness of SLAM in hyper-parameter selection. Specifically, we explore a spectrum of essential hyper-parameters across SOTA SSL algorithms, examining how the SLAM metrics correlate with actual performance outcomes. Given the wide range of hyper-parameters in the SSL model, it is impractical to enumerate all possible parameters. Instead, we can only name a few representative hyper-parameters from popular algorithms, as illustrative examples to showcase the capability of SLAM.

For instance, MixMatch's critical parameters include the weights assigned to unlabeled loss (Berthelot et al., 2019b), we define a search range for these weights set at $\{1, 25, 50, 100\}$. ReMixMatch builds upon MixMatch by introducing additional parameters, such as the weight controlling the rotation loss (Berthelot et al., 2019a), with its search range defined as $\{0.1, 0.25, 0.5, 1\}$. Similarly, for FixMatch, significant parameters include the confidence threshold (Sohn et al., 2020), with its search range specified as $\{0.5, 0.8, 0.9, 0.95\}$.

Figure 2 showcases our analysis of hyper-parameter selection for models trained on the CIFAR-10 dataset with only 40 labeled samples. This visualization reveals a pronounced correlation between the SLAM metrics and test errors, indicating that SLAM metrics can effectively guide the identification of optimal hyper-parameters as validated on the test set. This finding underscores SLAM's practical utility in refining hyper-parameter selection, further supporting its empirical success in enhancing model performance.

**Model selection against validation split.** In this part, we explore scenarios characterized by a relatively abundant supply of labeled data, so that splitting a validation set is feasible. More specifically, we consider cases where there are more than 10 labeled data per class. We test two validation split ratios, 10% and 20%.

Within the context of the CIFAR-10 dataset, as detailed in Table 4, it is evident that employing SLAM for model selection consistently yields superior results compared to the division of a validation set. More specifically, one key observation that aligns with our previous hypothesis is, that when we use a smaller portion of labeled data as a validation set (e.g. 10%), the validation set is too small to provide any meaningful indications, as we can observe, in 10% validation data cases, the model selected by validation set significantly deviates from the best model selected by the test set. Conversely, while using larger validation data, the model selected will be closer to the best model defined over the test set. Yet, it still exhibits notable differences to the model selected by SLAM.

## 6. Conclusion

In this study, we introduce a novel model selection method designed for the SSL context, an essential yet often overlooked field. This method uniquely combines empirical error and model complexity to predict the generalization ability of SSL models, leading to significant empirical enhancements. Furthermore, we demonstrate that the disparity between the model chosen by our method and the optimal model within the selection pool is upper-bounded. This provides a robust theoretical assurance, a feature rarely extended by current SOTA model selection strategies.

## Acknowledgements

Tongliang Liu is partially supported by the following Australian Research Council projects: FT220100318, DP220102121, LP220100527, LP220200949, and IC190100031. Muyang Li and Runze Wu are supported by NetEase Youling Crowdsourcing Platform. This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government. This work was supported by resources provided by the Pawsey Supercomputing Research Centre's Setonix Supercomputer, with funding from the Australian Government and the Government of Western Australia. The authors acknowledge the technical assistance provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney.

## Impact Statement

The study of model selection involves the fundamental aspect of the trustworthiness of machine learning models, previous studies in SSL often rely on empirical heuristics, and assumptions that are hard to meet in real-world applications. The scope of this study aims to bridge this significant gap, by providing model selection methods that are specifically tailored for SSL with theoretical guarantee.

## References

Antos, A., Kégl, B., Linder, T., and Lugosi, G. Data-dependent margin-based generalization bounds for classification. *Journal of Machine Learning Research*, 3(Jul): 73–98, 2002.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *NeurIPS*, 2017.

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *ICLR*, 2019a.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019b.

Berthelot, D., Roelofs, R., Sohn, K., Carlini, N., and Kurakin, A. Adamatch: A unified approach to semi-supervised learning and domain adaptation. In *ICLR*, 2021.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign overfitting in two-layer convolutional neural networks. In *NeurIPS*, pp. 25237–25250, 2022.

Chen, B., Jiang, J., Wang, X., Wan, P., Wang, J., and Long, M. Debiased self-training for semi-supervised learning. In *NeurIPS*, pp. 32424–32437, 2022.

Feofanov, V., Tiomoko, M., and Virmaux, A. Random matrix analysis to balance between supervised and unsupervised learning under the low density separation assumption. In *ICML*, pp. 10008–10033. PMLR, 2023.

Guo, L.-Z. and Li, Y.-F. Class-imbalanced semi-supervised learning with adaptive thresholding. In *ICML*, pp. 8082–8094. PMLR, 2022.

Guo, L.-Z., Zhang, Z.-Y., Jiang, Y., Li, Y.-F., and Zhou, Z.-H. Safe deep semi-supervised learning for unseen-class unlabeled data. In *ICML*, pp. 3897–3906. PMLR, 2020.

Hu, D., Liang, J., Liew, J. H., Xue, C., Bai, S., and Wang, X. Mixed samples as probes for unsupervised model selection in domain adaptation. In *NeurIPS*, 2023.

Huang, Z., Xue, C., Han, B., Yang, J., and Gong, C. Universal semi-supervised learning. In *NeurIPS*, pp. 26714–26725, 2021.

Huang, Z., Shen, L., Yu, J., Han, B., and Liu, T. Flatmatch: Bridging labeled data and unlabeled data with cross-sharpness for semi-supervised learning. In *NeurIPS*, pp. 18474–18494, 2023.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *ICLR*, 2019.

Koltchinskii, V. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

Koltchinskii, V. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, pp. 2593–2656, 2006.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lange, T., Braun, M., Roth, V., and Buhmann, J. Stability-based model selection. In *NeurIPS*, 2002.

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

Li, M., Wu, R., Liu, H., Yu, J., Yang, X., Han, B., and Liu, T. Instant: Semi-supervised learning with instance-dependent thresholds. In *NeurIPS*, 2024.

Li, Y.-F. and Liang, D.-M. Safe semi-supervised learning: a brief introduction. *Frontiers of Computer Science*, 13: 669–676, 2019.

Li, Y.-F., Zha, H.-W., and Zhou, Z.-H. Learning safe prediction for semi-supervised regression. In *AAAI*, volume 31, 2017.

Li, Y.-F., Guo, L.-Z., and Zhou, Z.-H. Towards safe weakly supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):334–346, 2019.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Lotfi, S., Izmailov, P., Benton, G., Goldblum, M., and Wilson, A. G. Bayesian model selection, the marginal likelihood, and generalization. In *ICML*, pp. 14223–14247. PMLR, 2022.

MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Madani, O., Pennock, D., and Flake, G. Co-validation: Using model disagreement on unlabeled data to validate classification algorithms. In *NeurIPS*, 2004.

Mahsereci, M., Balles, L., Lassner, C., and Hennig, P. Early stopping without a validation set. 2017.

McAllester, D. Simplified pac-bayesian margin bounds. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pp. 203–215. Springer, 2003.

Mey, A. and Loog, M. Improved generalization in semi-supervised learning: A survey of theoretical results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4747–4767, 2022.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Morerio, P., Cavazza, J., and Murino, V. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In *ICLR*, 2018.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *NeurIPS*, 2017.

Niyogi, P. and Girosi, F. On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. *Neural Computation*, 8 (4):819–842, 1996.

Oh, Y., Kim, D.-J., and Kweon, I. S. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *CVPR*, pp. 9786–9796, 2022.

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018.

Platanios, E., Poon, H., Mitchell, T. M., and Horvitz, E. J. Estimating accuracy from unlabeled data: A probabilistic logic approach. In *NeurIPS*, 2017.

Platanios, E. A., Blum, A., and Mitchell, T. Estimating accuracy from unlabeled data. In *UAI*, pp. 682–691, 2014.

Platanios, E. A., Dubey, A., and Mitchell, T. Estimating accuracy from unlabeled data: A bayesian approach. In *ICML*, pp. 1416–1425. PMLR, 2016.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. Semi-supervised learning with ladder networks. In *NeurIPS*, 2015.

Saito, K., Kim, D., Teterwak, P., Sclaroff, S., Darrell, T., and Saenko, K. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In *ICCV*, pp. 9184–9193, 2021.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, pp. 596–608, 2020.

Su, J.-C. and Maji, S. The semi-supervised inaturalist-aves challenge at fgvc7 workshop. *arXiv preprint arXiv:2103.06937*, 2021.

Su, J.-C., Cheng, Z., and Maji, S. A realistic evaluation of semi-supervised learning for fine-grained classification. In *CVPR*, pp. 12966–12975, 2021.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Tsybakov, A. B. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Wang, X., Wu, Z., Lian, L., and Yu, S. X. Debiased learning from naturally imbalanced pseudo-labels. In *CVPR*, pp. 14647–14657, 2022a.

Wang, Y., Chen, H., Heng, Q., Hou, W., Fan, Y., Wu, Z., Wang, J., Savvides, M., Shinozaki, T., Raj, B., et al. Freematch: Self-adaptive thresholding for semi-supervised learning. In *ICLR*, 2022b.

Wei, C., Shen, K., Chen, Y., and Ma, T. Theoretical analysis of self-training with deep networks on unlabeled data. In *ICLR*, 2020.

Wei, C., Sohn, K., Mellina, C., Yuille, A., and Yang, F. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *CVPR*, pp. 10857–10866, 2021.

Wu, Y., Yao, J., Xia, X., Yu, J., Wang, R., Han, B., and Liu, T. Mitigating label noise on graph via topological sample selection. *arXiv preprint arXiv:2403.01942*, 2024.

Xia, X., Liu, T., Han, B., Gong, M., Yu, J., Niu, G., and Sugiyama, M. Sample selection with uncertainty of losses for learning with noisy labels. *arXiv preprint arXiv:2106.00445*, 2021.

Xia, X., Han, B., Zhan, Y., Yu, J., Gong, M., Gong, C., and Liu, T. Combating noisy labels with sample selection by mining high-discrepancy examples. In *ICCV*, pp. 1833–1843, 2023.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. Unsupervised data augmentation for consistency training. In *NeurIPS*, pp. 6256–6268, 2020.

Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y.-F., Sun, B., Li, H., and Jin, R. Dash: Semi-supervised learning with dynamic thresholding. In *ICML*, pp. 11525–11536. PMLR, 2021.

Yang, Y., Theisen, R., Hodgkinson, L., Gonzalez, J. E., Ramchandran, K., Martin, C. H., and Mahoney, M. W. Test accuracy vs. generalization gap: Model selection in nlp without accessing training or testing data. In *KDD*, pp. 3011–3021, 2023.

You, K., Wang, X., Long, M., and Jordan, M. Towards accurate model selection in deep unsupervised domain adaptation. In *ICML*, pp. 7124–7133. PMLR, 2019.

Yuan, S., Feng, L., and Liu, T. Early stopping against label noise without validation data. In *ICLR*, 2024.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., and Shinozaki, T. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, pp. 18408–18419, 2021a.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021b.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Zheng, M., You, S., Huang, L., Wang, F., Qian, C., and Xu, C. Simmatch: Semi-supervised learning with similarity matching. In *CVPR*, pp. 14471–14481, 2022.

# A. Scope of the study and future works

**Safety of Semi-supervised Learning.** In domains where decisions have critical consequences, such as medical diagnosis and fraud detection, ensuring the reliability and safety of SSL models is paramount (Li et al., 2017; 2019; Li & Liang, 2019; Guo et al., 2020). Without robust model selection mechanisms, service providers lack a reliable method to assess a model's real-world efficacy. By introducing a model selection approach for SSL with verifiable guarantees, we contribute to bolstering the trustworthiness of models under conditions of limited labeled data. This enhancement is vital for supporting dependable decision-making in sensitive areas, ultimately safeguarding public welfare. Future work could be developed, focusing the the part model selection plays in safe SSL, i.e. whether applying model selection reconciles with the theoretical guarantee, or simply by applying appropriate model selection is sufficient for achieving safe SSL.

# B. Theoretical justification

## B.1. Assumptions

In this subsection, we give some key assumptions that are necessary for the formulation of our theoretical analysis.

**Assumption B.1.** Assume that the empirical error of function $f$ can be written as:

$$\hat{R}_D(f) \leq n^{-1} \sum_{i}^{n} \mathbb{1}\left[ f(x_i)_{y_i} \leq \gamma + \max_{j \neq y_i} f(x_i)_j \right]$$

which is an analog of the 0-1 loss under the multi-classification case, where existing surrogate loss functions such as cross-entropy loss can be proven as Bayes-consistent to the defined loss, i.e., minimizing cross-entropy loss is asymptotically equivalent to minimize the defined empirical error $\hat{R}_D$ (Zhang, 2004).

**Assumption B.2** (Decomposable Hypothesis Space)**.** For a given hypothesis set $\mathcal{F}$, it can be decomposed into countably many sub-hypothesis-set $\mathcal{F}_k$, such that $\mathcal{F} = \bigcup_{k \geq 1} \mathcal{F}_k$.

**Assumption B.3** (Nested Hypothesis Space)**.** For a given hypothesis set $\mathcal{F}$ that satisfies Assumption B.2, its decomposed hypothesis set satisfies $\mathcal{F}_k \in \mathcal{F}_{k+1}$ for all $k \geq 1$.

## B.2. Preliminaries

**Lemma B.4.** *For all $f \in \mathcal{F}$ and for all $k \in \mathbb{N}$, the following inequality holds:*

$$P\left[ \sup_{f \in \mathcal{F}} |R(f) - \phi_k(f)| > \epsilon \right] \leq 2e^{-2n\epsilon^2}. \tag{11}$$

*Proof.*

$$
\begin{aligned}
&P\left[\sup_{f\in\mathcal{F}} R(f)-\phi_k(f)>\epsilon\right] = P\left[\sup_{k\geq 1}\sup_{f\in\mathcal{F}_k} R(f)-\phi_k(f)>\epsilon\right] \\
&\leq \sum_{k=1}^{\infty} P\left[\sup_{f\in\mathcal{F}_k} R(f)-\phi_k(f)\right] \quad \text{(Boole's inequality)} \\
&= \sum_{k=1}^{\infty} P\left[\sup_{f\in\mathcal{F}_k} R(f)-\hat{R}_D(f)-\frac{\hat{\mathcal{R}}_{f_{\mathcal{A}}}}{\gamma_f}>\epsilon+\sqrt{\frac{\log k}{n}}\right] \quad \text{(Substituting equation 9)} \\
&\leq \sum_{k=1}^{\infty} \exp\left(-2n\left[\epsilon+\sqrt{\frac{\log k}{n}}\right]\right) \quad \text{(McDiarmid's inequality)} \\
&\leq \sum_{k=1}^{\infty} \exp(-2n\epsilon^2)\exp(-2\log k) \\
&= \exp(-2n\epsilon^2)\sum_{k=1}^{\infty}\frac{1}{k^2} \\
&= \frac{\pi^2}{6}\exp(-2n\epsilon^2) \quad \text{(solution to Basel's problem)} \\
&\leq 2e^{-2n\epsilon^2}.
\end{aligned}
$$

□

### B.3. Proof of Theorem 4.1

Our proof follows the one described by (Mohri et al., 2018), where we use the Spectral-Margin Complexity as the plug-in alternatives for the Rademacher Complexity, combining with Lemma B.4:

*Proof.*

$$
\begin{aligned}
&P\left[R(f_D^\dagger)-R(f)-2\frac{\hat{\mathcal{R}}_{f_{\mathcal{A}}}}{\gamma_f}-\sqrt{\frac{\log k(f)}{n}}>\epsilon\right] \\
&= P\left[R(f_D^\dagger)-\phi_{k(f_D^\dagger)}(f_D^\dagger)+\phi_{k(f_D^\dagger)}(f_D^\dagger)-R(f)-2\frac{\hat{\mathcal{R}}_{f_{\mathcal{A}}}}{\gamma_f}-\sqrt{\frac{\log k(f)}{n}}>\epsilon\right] \\
&\leq P\left[R(f_D^\dagger)-\phi_{k(f_D^\dagger)}(f_D^\dagger)>\frac{\epsilon}{2}\right]+P\left[\phi_{k(f_D^\dagger)}(f_D^\dagger)-R(f)-2\frac{\hat{\mathcal{R}}_{f_{\mathcal{A}}}}{\gamma_f}-\sqrt{\frac{\log k(f)}{n}}>\frac{\epsilon}{2}\right] \\
&\leq 2\exp(-n\epsilon^2/2)+P\left[\phi_{k(f)}(f)-R(f)-2\frac{\hat{\mathcal{R}}_{f_{\mathcal{A}}}}{\gamma_f}-\sqrt{\frac{\log k(f)}{n}}>\frac{\epsilon}{2}\right] \\
&= 2\exp(-n\epsilon^2/2)+P\left[\hat{R}_D(f)-R(f)-\frac{\hat{\mathcal{R}}_{f_{\mathcal{A}}}}{\gamma_f}>\frac{\epsilon}{2}\right] \quad \text{(Substituting equation 9)} \\
&= 2\exp(-n\epsilon^2/2)+\exp(-n\epsilon^2/2) \quad \text{(Applying Lemma B.4)}
\end{aligned}
$$

Setting $\delta := 3e^{-n\epsilon^2/2}$ hence completes the proof. □