Journal of Machine Learning Research 25 (2024) 1-47 Submitted 1/23; Revised 8/23; Published 3/24

Multi-class Probabilistic Bounds for Majority Vote Classifiers with Partially Labeled Data

Vasilii Feofanov VASILII.FEOFANOV@HUAWEI.COM
Univ. Grenoble Alpes / Huawei Noah's Ark Lab
92100 Boulogne-Billancourt, France

Emilie Devijver EMILIE.DEVIJVER@UNIV-GRENOBLE-ALPES.FR
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
38000 Grenoble, France

Massih-Reza Amini MASSIH-REZA.AMINI@UNIV-GRENOBLE-ALPES.FR
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
38000 Grenoble, France

Editor: Sanmi Koyejo

Abstract

In this paper, we propose a probabilistic framework for analyzing a multi-class majority vote classifier in the case where training data is partially labeled. First, we derive a multi-class transductive bound over the risk of the majority vote classifier, which is based on the classifier's vote distribution over each class. Then, we introduce a mislabeling error model to analyze the error of the majority vote classifier in the case of the pseudo-labeled training data. We derive a generalization bound over the majority vote error when imperfect labels are given, taking into account the mean and the variance of the prediction margin. Finally, we demonstrate an application of the derived transductive bound for self-training to find automatically the confidence threshold used to determine unlabeled examples for pseudo-labeling. Empirical results on different data sets show the effectiveness of our framework compared to several state-of-the-art semi-supervised approaches.

Keywords: semi-supervised learning, learning theory, multi-class classification, transductive inference, self-training

1. Introduction

We consider multi-class classification problems where the scarce labeled training set comes along with a huge number of unlabeled training examples. This is for example the case in web-oriented applications where a huge number of unlabeled observations arrive sequentially, and there is not enough time to manually label them all. In this context, the use of traditional supervised approaches trained on available labeled data usually leads to poor learning performance. In semi-supervised learning (Chapelle et al., 2010), it is generally assumed that unlabeled training examples contain valuable information about the prediction problem, so the aim is to exploit both available labeled and unlabeled training observations in order to provide an improved solution.

A common approach when working with partially-labeled data is to pseudo-label unlabeled data using the associated predictions of a classification model and treat them like

labeled training examples (Tür et al., 2005; Lee, 2013). This paper theoretically studies the pseudo-labeling approach with a focus on error estimation of majority vote classifiers such as Random Forest (Lorenzen et al., 2019), AdaBoost (Germain et al., 2015) and SVM (Fakeri-Tabrizi et al., 2015). The majority vote classifier is well studied in the binary supervised case, where the majority vote risk is usually bounded indirectly by twice the risk of related stochastic Gibbs classifier (Langford and Shawe-Taylor, 2003). This is used to derive tight PAC-Bayesian guarantees (McAllester, 1999), but not relevant for a more profound analysis of the risk behavior as the voters may compensate the errors of each other, so the majority vote risk can be much smaller than the Gibbs risk. This is why some works are focused on deriving direct upper-bounds on the majority vote error (Lacasse et al., 2007), and the results were recently extended to the multi-class case (Laviolette et al., 2017; Masegosa et al., 2020). However, when the multi-class framework meets semi-supervised learning, the theoretical analysis is not straightforward as it is not clear how unlabeled data may be integrated while some results for binary classifiers (Lacasse et al., 2007; Amini et al., 2008) do not hold in this case.

Semi-supervised majority vote classifiers can be studied in two settings: the classical inductive case and the transductive one (Vapnik, 1998, p. 339), which aims for correctly classifying the given unlabeled training examples. In the latter case, Feofanov et al. (2019) derived a bound for the multi-class majority vote classifier by analyzing distribution of the class vote, focusing on the class confusion matrix as an error indicator as proposed by Morvant et al. (2012). The proposed transductive bound is applied for a self-training algorithm (Amini et al., 2023) to automatically find the subset of examples for pseudo-labeling in the multi-class case. Our paper may be seen of as an expanded version of Feofanov et al. (2019) with two distinct contributions. In order to improve upon earlier work, we first generalize the transductive bound to the probabilistic framework, where soft labels may be applied to the unlabeled set. Secondly, we propose a new direction for theoretical analysis of the majority vote classifier in the inductive case when it is trained on pseudo-labeled training examples, which inevitably contain label noise. For this, we take explicitly into account possible mislabeling following the model introduced by Chittineni (1980). First, we derive the connection between the classification error of the true and the imperfect label. Based on this, we propose a new probabilistic generalization bound over the error of the multi-class majority vote classifier in the presence of imperfect labels. This bound is based on the mean and the variance of the prediction margin (Lacasse et al., 2007), so it reflects both the individual strength of voters and their correlation in prediction.

The rest of this paper is organized as follows. In Section 2, we overview the related work. Section 3 introduces the problem statement and the proposed framework. In Section 4, we present a transductive error bound for the multi-class majority vote classifier; this section is an extension of Feofanov et al. (2019) to the probabilistic case. Section 5 introduces a mislabeling error model and derives a probabilistic bound taking into account mislabeling errors. In Section 6, we describe the extended self-training algorithm that learns the threshold using the proposed transductive bound, present empirical evidence that the proposed self-training strategy is effective compared to several state-of-the-art approaches, and illustrate the behavior of the new generalization bound on real data sets. Finally, in Section 7 we summarize the outcome of this study and discuss the future work.

2. Related Work

Generalization guarantees of majority vote classifiers are well studied in the binary supervised setting. Many works are focused on deriving tight PAC guarantees for the Gibbs classifier in the inductive case (McAllester, 2003; Maurer, 2004; Catoni, 2007) as well as in the transductive one (Derbeko et al., 2004; Bégin et al., 2014), and applying these results for optimization (Thiemann et al., 2017), linear classifiers (Germain et al., 2009), random forests (Lorenzen et al., 2019), neural networks (Letarte et al., 2019). While this bound can be tight, it reflects only the individual strength of voters, so using it as a minimization criterion often leads to an increase in the test error (Masegosa et al., 2020). This motivates to opt for bounds that directly upper bound the majority vote error. In the transductive setting, Amini et al. (2008) derive a risk upper bound based on how voters agree on every unlabeled example. In the inductive setting, Lacasse et al. (2007) propose an upper bound (later called *C-bound*) for the generalization error that is based on the first and the second statistical moments of the margin of the majority vote classifier. While the first moment reflects the individual errors of voters (the Gibbs risk), the second moment takes into account the error correlation between them.

In the binary case, the majority vote classifier is usually defined through the sign function (Germain et al., 2009, p. 790), which implies that the mathematical derivations are not directly extendable to the multi-class case. In fact, most of the results hold for the binary classification only, and only few results exist for the multi-class majority vote classifier. In the supervised setting, Morvant et al. (2012) derive generalization guarantees on the confusion matrix norm, whereas Laviolette et al. (2017) extend the C-bound of Lacasse et al. (2007) to the multi-class case. Masegosa et al. (2020) study tight estimations from data by deriving a relaxed version of Laviolette et al. (2017). In the transductive setting, Feofanov et al. (2019) extend the bound of Amini et al. (2008) to the multi-class case.

However, the aforementioned studies are limited by assuming that all training examples are perfectly labeled. Learning with an imperfect supervisor, in which training data contains an unknown portion of imperfect labels, has been considered in both supervised (Natarajan et al., 2013; Scott, 2015; Xia et al., 2019) and semi-supervised settings (Amini and Gallinari, 2003). In most cases, methods focus on the estimation of the mislabeling errors like the anchor points approach either to train a classifier with a corrected loss (Patrini et al., 2017; Xia et al., 2019) or to correct the classifier's output after training (Zhang et al., 2021). From the theoretical point of view, one can highlight the method of unbiased estimators studied both in the binary (Natarajan et al., 2013; Scott, 2015) and the multi-class case (van Rooyen and Williamson, 2018). Chittineni (1980) analyzes the connection between the true and the imperfect label in the multi-class case but only for the maximum a posteriori classifier. We extend the latter result to an arbitrary classifier and use it to derive a new C-bound with imperfect labels. To the best of our knowledge, this is the first attempt to extend the theoretical analysis of the majority vote classifier to the imperfect label scenario.

In this paper, our theoretical development has a particular focus on semi-supervised learning. The question of learning on labeled and unlabeled examples is usually studied under three related yet different assumptions (Chapelle et al., 2010). While the approaches based on data *clustering* (Peikari et al., 2018; Maximov et al., 2018) suggest that the labeled and unlabeled examples are divided into informative clusters, the graph-based approaches

(Zhou et al., 2004; Chong et al., 2020) assume that the data lies on a low-dimensional *manifold*. Then, pseudo-labeling approaches like self-training (Amini et al., 2023) and unlabeled margin maximization (Feofanov et al., 2023) are based on an assumption that the decision boundary passes through *low-density regions* of unlabeled data (Chapelle and Zien, 2005), which is implemented by using pseudo-labeled unlabeled examples for classifier's training.

The self-training algorithm has been present in the literature since the late 1960s (Scudder, 1965; Fralick, 1967) and is still widely used in practice (Amini et al., 2023). Starting from a supervised classifier initially trained on the labeled data only, the algorithm iteratively re-trains the classifier by assigning pseudo-labels to unlabeled examples with the confidence score above a certain threshold and including them to the training set. To the best of our knowledge, the theoretical analysis of self-training is limited to the binary classification setting (Amini et al., 2008; Frei et al., 2022). Recently, Frei et al. (2022) derived guarantees for the self-training with a binary linear classifier considering a specific class of mixture models, and Zhang et al. (2022) proved the convergence rate for one-hidden-layer self-training neural networks in the case of regression and isotropic Gaussian distribution. In practice, self-training is usually performed with a fixed threshold; another method consists in controlling the number of pseudo-labeled examples by curriculum learning (Cascante-Bonilla et al., 2021). We show that this threshold can be effectively found at every iteration as a trade-off between the number of pseudo-labeled examples and the bounded transductive error evaluated on them. The proposed policy allows us to partially address the confirmation bias which underlies the self-training and consists in including wrongly pseudo-labeled data into the training set thereby increasing the bias of the model towards its initial belief. In the context of deep learning, there are also attempts to overcome this issue, we refer to Arazo et al. (2020), Li et al. (2021) and Radhakrishnan et al. (2024) for more details.

We would also like to point out that the self-training approach has similarities with the abstention paradigm (Freund, 1995; Bartlett and Wegkamp, 2008), whose theoretical studies for the multi-class classification (Ramaswamy et al., 2018) may interest the reader. However, the two main differences are that the abstention is studied in the supervised setting, and a classifier guided by self-training is iteratively trained on pseudo-labeled examples, which can be erroneous and for which the true labels are unknown.

3. Framework and Definitions

We consider multi-class classification problems with an input space $\mathcal{X} \subset \mathbb{R}^d$ and an output space $\{1, \dots, K\}$, $K \geq 2$. We denote by $\mathbf{X} = (X_1, \dots, X_d) \in \mathcal{X}$ (resp. $Y \in \{1, \dots, K\}$) an input (resp. output) random variable. Considering the semi-supervised framework, we assume an available set of labeled training examples $\mathrm{Z}_{\mathcal{L}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l} \in (\mathcal{X} \times \{1, \dots, K\})^l$, where $\mathbf{x}_i$ is identically and independently distributed (i.i.d.) with respect to a fixed yet unknown marginal density $f_{\mathbf{X}}$ over $\mathcal{X}$ while y_i is sampled from the true label generator $P(Y|\mathbf{X} = \mathbf{x}_i)$ defined over $\{1, \dots, K\}$, and an available set of unlabeled training examples $\mathrm{X}_{\mathcal{U}} = \{\mathbf{x}_i\}_{i=l+1}^{l+u} \in \mathcal{X}^u$ drawn i.i.d. from the density $f_{\mathbf{X}}$ over the domain $\mathcal{X}$.

Further, we denote by $\mathbf{0}_K$ the zero vector of size K, $\mathbf{0}_{K,K}$ is the zero matrix of size $K \times K$ and we set $n := l + u$. In this work, a fixed class of classifiers $\mathcal{H} = \{h : \mathcal{X} \to \{1, \dots, K\}\}$, called the *hypothesis space*, is considered and defined without reference to the training set.

Notation	Description
$\mathcal{X} \in \mathbb{R}^d$	input space
$\{1, \ldots, K\}$	output space
$l,\ u$	the number of labeled and unlabeled examples
$\mathrm{Z}_{\mathcal{L}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$	labeled set of training examples
$\mathrm{X}_{\mathcal{U}} = \{\mathbf{x}_i\}_{i=l+1}^{l+u}$	unlabeled set of training examples
$\mathcal{H} = \{h : \mathcal{X} \to \{1, \ldots, K\}\}$	hypothesis space, supposed to be fixed
Q_0 and Q	the prior and the posterior distribution over $\mathcal{H}$
$B_Q(\mathbf{x})$	Q-weighted majority vote (Bayes) classifier, Eq. (1)
$G_Q(\mathbf{x})$	stochastic Gibbs classifier
$v_Q(\mathbf{x}, y)$	class vote, Eq. (2)
$m_Q(\mathbf{x}, y)$	class margin, Eq. (3)

Table 1: List of notations used in this paper.

Over $\mathcal{H}$, two probability distributions are introduced: the prior Q_0 and the posterior Q that are defined respectively before and after observing the training set. It can be useful when some knowledge is given with the data, if not, one can use the uniform distribution. We focus on two classifiers: the *Q-weighted majority vote classifier* (also called the Bayes classifier[1], which is reflected in our notations) defined for all $\mathbf{x} \in \mathcal{X}$ as

$$B_Q(\mathbf{x}) := \underset{\hat{y} \in \{1,\ldots,K\}}{\operatorname{argmax}} \left[\mathbb{E}_{h \sim Q} \mathbb{I}(h(\mathbf{x}) = \hat{y})\right], \tag{1}$$

and, the stochastic *Gibbs classifier* G_Q that for every $\mathbf{x} \in \mathcal{X}$ predicts the label using a randomly chosen classifier $h \in \mathcal{H}$ according to Q. The former one represents a class of learning methods, where the predictions of hypotheses are aggregated using the majority vote rule scheme, while the latter one is often used to analyze the behavior of the Q-weighted majority vote classifier.

The goal of learning is formulated as to choose a posterior distribution Q over $\mathcal{H}$ based on the training set $\mathrm{Z}_{\mathcal{L}} \cup \mathrm{X}_{\mathcal{U}}$ such that the classifier B_Q will have the smallest possible error value. In contrast to the studies of Derbeko et al. (2004), Bégin et al. (2014) and Feofanov et al. (2019), which considered the deterministic case where there is one and only one possible label for each unlabeled example, in this study, we investigate a more generic scenario with probabilistic labels assuming the possibility of different outcomes.

To measure confidence of the majority vote classifier in its prediction, the notions of class votes and margin are further considered. Given an observation $\mathbf{x}$, we define a vector of *class votes* $\mathbf{v}_{\mathbf{x}} := (v_Q(\mathbf{x}, \hat{y}))_{\hat{y}=1}^{K}$ where the $\hat{y}$-th component corresponds to the total vote given to the class $\hat{y}$:

$$v_Q(\mathbf{x}, \hat{y}) := \mathbb{E}_{h \sim Q} \mathbb{I}(h(\mathbf{x}) = \hat{y}) = \sum_{h : h(\mathbf{x}) = \hat{y}} Q(h). \tag{2}$$

In practice, $v_Q(\mathbf{x}, \hat{y})$ can be regarded as an approximation of the posterior probability $P(Y = \hat{y} | \mathbf{X} = \mathbf{x})$; a large value indicates a high confidence that the true label of $\mathbf{x}$ is $\hat{y}$.

1. This should not be confused with other learning paradigms based on the Bayesian inference, e.g., the Bayesian statistics or the oracle Bayes classifier of the form $B_O(\mathbf{x}) := \max_{\hat{y} \in \{1,\ldots,K\}} P(Y = \hat{y} | \mathbf{X} = \mathbf{x})$.

Given an observation $\mathbf{x}$ and its true class y, its *margin* is defined in the following way:

$$m_Q(\mathbf{x}, y) := \mathbb{E}_{h\sim Q}\mathbb{I}(h(\mathbf{x}) = y) - \max_{\substack{\hat{y}\in\{1,\dots,K\}\\ \hat{y}\neq y}} \mathbb{E}_{h\sim Q}\mathbb{I}(h(\mathbf{x}) = \hat{y}) = v_Q(\mathbf{x}, y) - \max_{\substack{\hat{y}\in\{1,\dots,K\}\\ \hat{y}\neq y}} v_Q(\mathbf{x}, \hat{y}). \tag{3}$$

The margin measures a gap between the vote of the true class and the maximal vote among all other classes. If the value is strictly positive for an example $\mathbf{x}$, then y will be the output of the majority vote, so the example will be correctly classified.

4. Probabilistic Transductive Bounds

In this section, we derive guarantees for the multi-class majority vote classifier in the transductive setting (Vapnik, 1982, 1998), i.e., where the error is evaluated on the unlabeled set $\mathrm{X}_{\mathcal{U}}$ only. The proposed bound assumes that the majority vote classifier makes mistakes on low class votes and thereby uses votes as an indicator of confidence. This section is an extension of Feofanov et al. (2019) to the setting with probabilistic labels. The derivations of the bounds (established in Theorem 4.1 and Corollary 4.2) follow the similar steps, while the analysis of the bound's tightness (Proposition 4.3) differs in order to take into account the probabilistic nature of the errors. All proofs are deferred to Appendix A.

4.1 Transductive Conditional Risk

At first, we show how to upper bound the risk evaluated conditionally to the values of the true and the predicted class. Given a classifier h, for each class pair $(\hat{y}, y) \in \{1, \dots, K\}^2$ such that $\hat{y} \neq y$, the *transductive conditional risk* is defined as follows:

$$R_{\mathcal{U}}(h, y, \hat{y}) := \frac{1}{u_y} \sum_{\mathbf{x}\in\mathrm{X}_{\mathcal{U}}} P(Y = y|X = \mathbf{x})\mathbb{I}(h(\mathbf{x}) = \hat{y}), \tag{4}$$

where $u_y = \sum_{\mathbf{x}\in\mathrm{X}_{\mathcal{U}}} P(Y = y|X = \mathbf{x})$ is the expected number of unlabeled observations from the class $y \in \{1, \dots, K\}$. The value of $R_{\mathcal{U}}(h, y, \hat{y})$ indicates the expected proportion of unlabeled examples that are classified to the class $\hat{y}$ being from the class y. Particularly, $R_{\mathcal{U}}(B_Q, y, \hat{y})$ is called the *transductive Q-weighted majority vote conditional risk*. In the similar way, the *transductive Gibbs conditional risk* is defined for all $(y, \hat{y}) \in \{1, \dots, K\}^2,\ y \neq \hat{y}$ by

$$R_{\mathcal{U}}(G_Q, y, \hat{y}) := \mathbb{E}_{h\sim Q} R_{\mathcal{U}}(h, y, \hat{y}).$$

Although the Gibbs classifier is stochastic, its error is defined in expectation over Q. In other words, the Gibbs conditional risk represents the Q-weighted average conditional risk of hypotheses $h \in \mathcal{H}$.

In addition, we define the transductive *joint Q-weighted majority vote conditional risk* for a threshold vector $\boldsymbol{\theta} \in [0, 1]^K$, for $(y, \hat{y}) \in \{1, \dots, K\}^2,\ y \neq \hat{y}$, as follows:

$$R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, y, \hat{y}) := \frac{1}{u_y} \sum_{\mathbf{x}\in\mathrm{X}_{\mathcal{U}}} P(Y = y|X = \mathbf{x})\mathbb{I}(B_Q(\mathbf{x}) = \hat{y})\mathbb{I}(v_Q(\mathbf{x}, \hat{y}) \geq \theta_{\hat{y}}). \tag{5}$$

If the Q-weighted majority vote classifier makes mistakes, i.e., outputs the class $\hat{y}$ when the true class is y, on the examples with low values of $v_Q(\mathbf{x}, \hat{y})$, then the joint risk computes the

probability to make the conditional error on confident observations when a large enough $\theta_{\hat{y}}$ is set with respect to the distribution of $v_Q(\mathbf{x}, \hat{y})$. In Section 6.1, it will be seen that the joint risk can be interpreted as the error evaluated on those unlabeled examples that are going to be pseudo-labeled by the self-training algorithm, and the threshold is set to $\boldsymbol{\theta}$. The following Theorem 4.1 derives a transductive bound over the joint Q-weighted majority vote conditional risk.

Theorem 4.1 *Let $\mathbf{X} \in \mathcal{X}$ and $Y \in \{1, \ldots, K\}$ be the input and the output random variables, and $\mathrm{X}_{\mathcal{U}}$ be a set of unlabeled examples. Let $\mathcal{H}$ be the fixed hypothesis space, Q be the posterior distribution over $\mathcal{H}$, and B_Q be the Q-weighted majority vote classifier defined by Eq. (1). Denote by $v_Q(\mathbf{x}, \hat{y})$ the assigned class vote of a given observation $\mathbf{x}$ for the class $\hat{y}$. Let u_y be the expected number of unlabeled observations from the class $y \in \{1, \ldots, K\}$. Then, for any set $\mathrm{X}_{\mathcal{U}}$, for any given $\boldsymbol{\theta} \in [0,1]^K$, for all $(y, \hat{y}) \in \{1, \ldots, K\}^2$ we have*

$$R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q, y, \hat{y}) \leq \inf_{\gamma \in [\theta_{\hat{y}}, 1]} \left\{ I^{(\leq, <)}_{y,\hat{y}}(\theta_{\hat{y}}, \gamma) + \frac{1}{\gamma} \Big[K_{y,\hat{y}} - V^{(\leq, <)}_{y,\hat{y}}(\theta_{\hat{y}}, \gamma) \Big]_+ \right\}, \qquad (\mathrm{TB}_{y,\hat{y}})$$

where

- $K_{y,\hat{y}} = \frac{1}{u_y} \sum_{\mathbf{x} \in X_{\mathcal{U}}} P(Y = y | X = \mathbf{x}) v_Q(\mathbf{x}, \hat{y}) \mathbb{I}(B_Q(\mathbf{x}) = \hat{y})$ *is the transductive Gibbs conditional risk evaluated on the examples for which the majority vote class is $\hat{y}$,*
- $I^{(\leq, <)}_{y,\hat{y}}(\theta_{\hat{y}}, \gamma) = \frac{1}{u_y} \sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} P(Y = y | X = \mathbf{x}) \mathbb{I}(\theta_{\hat{y}} \leq v_Q(\mathbf{x}, \hat{y}) < \gamma)$ *is the expected proportion of unlabeled examples in the class y satisfying $\theta_{\hat{y}} \leq v_Q(\mathbf{x}, \hat{y}) < \gamma$,*
- $V^{(\leq, <)}_{y,\hat{y}}(\theta_{\hat{y}}, \gamma) = \frac{1}{u_y} \sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} P(Y = y | X = \mathbf{x}) v_Q(\mathbf{x}, \hat{y}) \mathbb{I}(\theta_{\hat{y}} \leq v_Q(\mathbf{x}, \hat{y}) < \gamma)$ *is the average of $\hat{y}$-votes in the class y satisfying $\theta_{\hat{y}} \leq v_Q(\mathbf{x}, \hat{y}) < \gamma$.*

The proof stands in Appendix A.1. By sorting the prediction votes over classes in the ascending order, the transductive bound ($\mathrm{TB}_{y,\hat{y}}$) is derived as a solution of a linear program, where the risk is maximized while the link with the Gibbs classifier is employed as a linear constraint. The solution of the linear program can be formulated as the greatest feasible solution in the lexicographic order. In other words, starting from the minimal class vote value in the ascending order, we assign the maximal possible error on each unique class vote value γ (i.e., the proportion of unlabeled examples with the class vote γ) until we reach the equality for the imposed linear constraints. In the proof, we also show that the bound can be computed[2] without explicitly solving the linear program as its solution is the infimum of the following function:

$$U_{y,\hat{y}} : \gamma \mapsto I^{(\leq, <)}_{y,\hat{y}}(\theta_{\hat{y}}, \gamma) + \frac{1}{\gamma} \Big[K_{y,\hat{y}} - V^{(\leq, <)}_{y,\hat{y}}(\theta_{\hat{y}}, \gamma) \Big]_+$$

on the interval $[\theta_{\hat{y}}, 1]$. As illustrated in Figure 1, the optimal value $\gamma^* = \operatorname{argmin}_\gamma U_{y,\hat{y}}(\gamma)$ is found from minimization of the sum of the first term, an increasing function of γ, and the second term that decreases with the increase of γ.

2. An implementation in python is publicly available for research purposes: `https://github.com/vfeofanov/trans-bounds-maj-vote`.

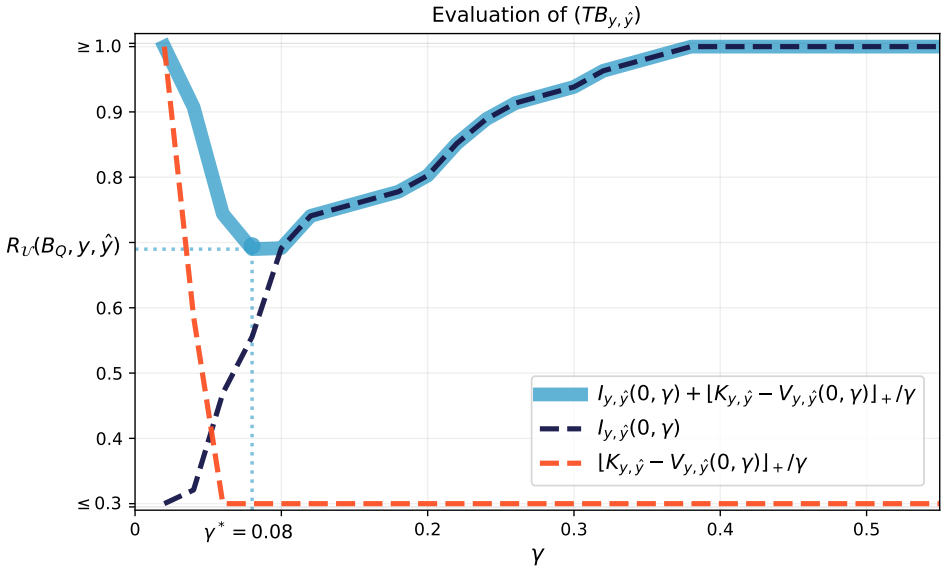

Figure 1: Illustration of how the transductive bound is evaluated on the `Vowel` data set (presented in Table 2).

When $\theta_{\hat{y}} = 0$, a bound over the transductive Q-weighted majority vote conditional risk is directly obtained from $(\mathrm{TB}_{y,\hat{y}})$:

$$R_{\mathcal{U}}(B_Q, y, \hat{y}) \leq \inf_{\gamma \in [0,1]} \left\{ I^{(\leq,<)}_{y,\hat{y}}(0,\gamma) + \frac{1}{\gamma} \left\lfloor K_{y,\hat{y}} - V^{(\leq,<)}_{y,\hat{y}}(0,\gamma) \right\rfloor_+ \right\}. \tag{6}$$

We note that for the transductive bound obtained in the binary case by Amini et al. (2008), the transductive Gibbs risk used inside the linear program can be bounded either by the PAC-Bayesian bound (Derbeko et al., 2004; Bégin et al., 2014) or by 1/2 (the worst possible error of the binary classifier), which allows to compute the transductive bound. In the multi-class case, all the terms must be approximated. To achieve it, we approximate the posterior probabilities by the predicted probabilities of a supervised classifier. We discuss this choice in more details in Section 6.1 and Section C.1. Thus, we use this approach to estimate all the terms of $(\mathrm{TB}_{y,\hat{y}})$, so we directly approximate the transductive conditional Gibbs risk and do not upper bound it.

4.2 Transductive Confusion Matrix and Transductive Error Rate

Based on Theorem 4.1, we derive bounds for two other error measures: the *error rate* and the *confusion matrix* (Morvant et al., 2012). We define the transductive error rate and the *transductive joint error rate* of the Q-weighted majority vote classifier B_Q over the

unlabeled set $\mathrm{X}_{\mathcal{U}}$ given a vector $\boldsymbol{\theta} = (\theta_y)_{y=1}^K \in [0,1]^K$, as

$$R_{\mathcal{U}}(B_Q) := \frac{1}{u} \sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} \sum_{\substack{y \in \{1,\ldots,K\} \\ y \neq B_Q(\mathbf{x})}} P(Y = y | \mathbf{X} = \mathbf{x}),$$

$$R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q) := \frac{1}{u} \sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} \sum_{\substack{y \in \{1,\ldots,K\} \\ y \neq B_Q(\mathbf{x})}} P(Y = y | X = \mathbf{x}) \mathbb{I}(v_Q(\mathbf{x}, B_Q(\mathbf{x})) \geq \theta_{B_Q(\mathbf{x})}). \tag{7}$$

Then, we define the *transductive joint Q-weighted majority vote confusion matrix* for $\boldsymbol{\theta} \in [0,1]^K$, and $(y, \hat{y}) \in \{1, \ldots, K\}^2$, as follows:

$$\left[\mathbf{C}_{B_Q}^{\mathcal{U} \wedge \boldsymbol{\theta}}\right]_{y,\hat{y}} := \begin{cases} 0 & y = \hat{y}, \\ R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q, y, \hat{y}) & y \neq \hat{y}. \end{cases} \tag{8}$$

From Theorem 4.1, we derive corresponding transductive bounds for the confusion matrix norm and the error rate of the Q-weighted majority vote classifier. To simplify notations, we introduce a matrix $\mathbf{U}_{\boldsymbol{\theta}}$ of size $K \times K$ with zeros on the main diagonal and the following $(y, \hat{y})$-entries, $y \neq \hat{y}$:

$$[\mathbf{U}_{\boldsymbol{\theta}}]_{y,\hat{y}} := \inf_{\gamma \in [\theta_{\hat{y}}, 1]} \left\{ I_{y,\hat{y}}^{(\leq,<)}(\theta_{\hat{y}}, \gamma) + \frac{1}{\gamma} \Big| K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(\theta_{\hat{y}}, \gamma) \Big|_+ \right\}, \tag{9}$$

which corresponds to the transductive bound proposed in Theorem 4.1.

Corollary 4.2 *Let $\mathbf{X} \in \mathcal{X}$ and $Y \in \{1, \ldots, K\}$ be the input and the output random variables, and $\mathrm{X}_{\mathcal{U}}$ be a set of unlabeled examples. Let $\mathcal{H}$ be the fixed hypothesis space, Q be the posterior distribution over $\mathcal{H}$, and B_Q be the Q-weighted majority vote classifier defined by Eq.* (1)*. Let u_y be the expected number of unlabeled observations from the class $y \in \{1, \ldots, K\}$. Denote $\|.\|_2$ the spectral norm of degree 2.*
Then, for any set $\mathrm{X}_{\mathcal{U}}$, for any given $\boldsymbol{\theta} \in [0,1]^K$, the spectral norm of the confusion matrix $\mathbf{C}_{B_Q}^{\mathcal{U} \wedge \boldsymbol{\theta}}$ is bounded as

$$\|\mathbf{C}_{B_Q}^{\mathcal{U} \wedge \boldsymbol{\theta}}\|_2 \leq \|\mathbf{U}_{\boldsymbol{\theta}}\|_2. \tag{10}$$

where $\mathbf{U}_{\boldsymbol{\theta}}$ is defined by Eq. (9)*. Moreover, we have the following bound for the transductive joint error rate $R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q)$:*

$$R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q) \leq \left\| \mathbf{U}_{\boldsymbol{\theta}}^{\intercal} \, \mathbf{p} \right\|_1. \tag{11}$$

where $\mathbf{p} = \{u_y / u\}_{y=1}^K$.

Note that the transductive bound of the Q-weighted majority vote error rate is obtained from Eq. (11) by taking $\boldsymbol{\theta} = \mathbf{0}_K$:

$$R_{\mathcal{U}}(B_Q) \leq \left\| \mathbf{U}_{\mathbf{0}_K}^{\intercal} \, \mathbf{p} \right\|_1. \tag{TB}$$

The proof stands in Appendix A.2.

4.3 Tightness Guarantees

In this section, we assume that the Q-weighted majority vote classifier makes most of its error on unlabeled examples with a low prediction vote, i.e., class votes can be considered as indicators of confidence. In the following proposition, we show that the bound becomes tight under certain conditions. As the value of the conditional error may be never zero in the case of probabilistic labels for any confidence value, we introduce a hyperparameter τ that is used to decide from which value the error is considered to be non-negligible.

Proposition 4.3 *Let $\mathbf{X} \in \mathcal{X}$ and $Y \in \{1, \dots, K\}$ be the input and the output random variables, and $\mathrm{X}_{\mathcal{U}}$ be a set of unlabeled examples. Let $\mathcal{H}$ be the fixed hypothesis space, Q be the posterior distribution over $\mathcal{H}$, and B_Q be the Q-weighted majority vote classifier defined by Eq. (1). Denote by $v_Q(\mathbf{x}, \hat{y})$ the assigned class vote of a given observation $\mathbf{x}$ for the class $\hat{y}$. Let u_y be the expected number of unlabeled observations from the class $y \in \{1, \dots, K\}$. For $1 \leq \hat{y} \leq K, \tau \in [0, 1]$, let consider the set $\Gamma^{\tau}_{y,\hat{y}}$ of unique votes with conditional error larger than a threshold τ:*

$$\Gamma^{\tau}_{y,\hat{y}} := \left\{ \gamma \in [0,1] \,\middle|\, \frac{1}{u_y} \sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} P(Y = y | X = \mathbf{x}) \mathbb{I}(B_Q(\mathbf{x}) = \hat{y}) \mathbb{I}(v_Q(\mathbf{x}, \hat{y}) = \gamma) \geq \tau \right\}.$$

Assume there exists a lower bound $C \in [0, 1]$ such that for all $\gamma \in \Gamma^{\tau}_{y,\hat{y}}$:

$$\sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} P(Y = y | X = \mathbf{x}) \mathbb{I}(B_Q(\mathbf{x}) = \hat{y}) \mathbb{I}(v_Q(\mathbf{x}, \hat{y}) < \gamma) \geq C \sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} P(Y = y | X = \mathbf{x}) \mathbb{I}(v_Q(\mathbf{x}, \hat{y}) < \gamma). \tag{12}$$

Then, for any set $\mathrm{X}_{\mathcal{U}}$, for all $(y, \hat{y}) \in \{1, \dots, K\}^2$, the following inequality holds:

$$[\mathbf{U}_{\mathbf{0}_K}]_{y,\hat{y}} - R_{\mathcal{U}}(B_Q, y, \hat{y}) \leq \frac{1 - C}{C} R_{\mathcal{U}}(B_Q, y, \hat{y}) + r_{y,\hat{y}} \left(\frac{1}{\gamma^*_{y,\hat{y}}} - 1 \right),$$

where

- $\gamma^*_{y,\hat{y}} := \max \Gamma^{\tau}_{y,\hat{y}}$ *is the highest vote on which the conditional error is larger than τ, and*
- $r_{y,\hat{y}} := \sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} P(Y = y | X = \mathbf{x}) v_Q(\mathbf{x}, \hat{y}) \mathbb{I}(B_Q(\mathbf{x}) = \hat{y}) \mathbb{I}(v_Q(\mathbf{x}, \hat{y}) > \gamma^*_{y,\hat{y}}) / u_y$ *corresponds to the average of $\hat{y}$-votes in the class y that greater than $\gamma^*_{y,\hat{y}}$ and on which the Q-weighted majority vote classifier makes the conditional mistake.*

The proof stands in Appendix A.3. This proposition states that if Eq. (12) holds, the difference between the transductive Q-weighted majority vote conditional risk and its upper bound does not exceed an expression that depends on a constant C and a threshold τ. When the majority vote classifier makes most of its mistake for the class $\hat{y}$ on observations with a low value of $v_Q(\mathbf{x}, \hat{y})$, $r_{y,\hat{y}}$ and $\gamma^*_{y,\hat{y}}$ are decreasing with a reasonable choice of τ. This also implies that Eq. (12) accepts a high value C (close to 1) and the bound will be tighter. The closer our framework to the deterministic one, the closer $r_{y,\hat{y}}$ will be to 0 (in

the deterministic case, τ can be set to 0, so $r_{y,\hat{y}}$ will be 0), so the bound becomes tight. Although our bound is tight only under the condition of making mistakes on low prediction votes, the assumption is reasonable from the theoretical point of view, since if, for some observation, the Q-weighted majority vote classifier gives a relatively high vote to the class $\hat{y}$, we expect that the observation is most probably from this class and not from the class y. From the practical point of view, this assumption requires the learning model to be well calibrated (Gebel, 2009).

5. Probabilistic C-Bound with Imperfect Labels

In this section, we consider another setup: the inductive *generalization error* is taken as the learning objective, which is defined for any $h : \mathcal{X} \to \{1, \ldots, K\}$ in the case of probabilistic labels as follows:

$$R(h) := \mathbb{E}_{\mathbf{X}}\left[r(h, \mathbf{x})\right], \tag{13}$$
$$\text{where } r(h, \mathbf{x}) := \sum_{\substack{y \in \{1,\ldots,K\} \\ y \neq h(\mathbf{x})}} P(Y = y | \mathbf{X} = \mathbf{x}) = 1 - P(Y = h(\mathbf{x}) | \mathbf{X} = \mathbf{x}).$$

In addition, we consider that pseudo-labels have been inferred by a teacher model that is trained independently, either by using a hold-out set or pre-trained on a similar benchmark. We propose a way to evaluate the error of the classifier that is trained on both labeled and pseudo-labeled data, which implies that the training example come with the label noise. For this, we derive a new generalization bound in the presence of imperfect labels.

5.1 C-Bound in the Probabilistic Setting

Lacasse et al. (2007) proposed the C-bound that upper bounds the Q-weighted majority vote error by taking into account the mean and the variance of the prediction margin, which, we recall Eq. (3), is defined as $v_Q(\mathbf{x}, y) - \max_{\hat{y} \in \{1,\ldots,K\} \setminus \{y\}} v_Q(\mathbf{x}, \hat{y})$ with y denoting the true class for $\mathbf{x}$. A similar result was obtained in a different context by Breiman (2001). Laviolette et al. (2017) extended the C-bound to the multi-class case. Below, we derive their C-bound in the probabilistic setting.

Theorem 5.1 *Let $\mathbf{X} \in \mathcal{X}$ and $Y \in \{1, \ldots, K\}$ be the input and the output random variables, respectively. Let $\mathcal{H}$ be the fixed hypothesis space, and Q be the posterior distribution over $\mathcal{H}$. Let B_Q and m_Q be the majority vote classifier and the associated margin defined by Eq.* (1) *and Eq.* (3)*, respectively. Assuming m_Q is measurable, let M_Q be a random variable defined as $M_Q := m_Q(\mathbf{X}, Y)$ with its first and second statistical moments denoted by $\mu_1^{M_Q}$ and $\mu_2^{M_Q}$, respectively. Then, for all choices of Q on a fixed hypothesis space $\mathcal{H}$, and for any density $f_{\mathbf{X}}$ over $\mathcal{X}$ and any distribution $P(Y|\mathbf{X})$ over $\{1, \ldots, K\}$ such that $\mu_1^{M_Q} > 0$, we can upper-bound the generalization error $R(B_Q)$, defined in Eq.* (13)*, as follows:*

$$R(B_Q) \leq 1 - \frac{\left(\mu_1^{M_Q}\right)^2}{\mu_2^{M_Q}}. \tag{CB}$$

Proof At first, we show that $R(B_Q) = P(M_Q \leq 0)$. We notice that

$$P(M_Q \leq 0|\mathbf{X}=\mathbf{x}) = \sum_{y=1}^{K} P(Y=y|\mathbf{X}=\mathbf{x})\mathbb{I}(m_Q(\mathbf{x},y)\leq 0) = \sum_{\substack{y\in\{1,\dots,K\}\\ y\neq B_Q(\mathbf{x})}} P(Y=y|\mathbf{X}=\mathbf{x}).$$

Then, we obtain that

$$P(M_Q \leq 0) = \int_{\mathcal{X}} P(M_Q \leq 0|\mathbf{X}=\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} = \mathbb{E}_{\mathbf{X}} P(M_Q \leq 0|\mathbf{X}=\mathbf{x}) = R(B_Q). \quad (14)$$

By applying the Cantelli-Chebyshev inequality (Lemma B.1, Appendix B), we deduce:

$$P(M_Q \leq 0) \leq \frac{\mu_2^{M_Q} - \left(\mu_1^{M_Q}\right)^2}{\mu_2^{M_Q} - \left(\mu_1^{M_Q}\right)^2 + \left(\mu_1^{M_Q}\right)^2} = 1 - \frac{\left(\mu_1^{M_Q}\right)^2}{\mu_2^{M_Q}}. \quad (15)$$

The bound is obtained by combining Eq. (14) and Eq. (15). ■

Thus, the probabilistic C-bound allows to bound the generalization error of the Q-weighted majority vote classifier when examples are provided with probabilistic labels. Note that when only one label is possible for every example, the bound comes back to the usual deterministic case. The main advantage of C-bound is the involvement of the second margin moment, which can be related to correlation between hypotheses' predictions (Lacasse et al., 2007).

5.2 Mislabeling Error Model

In this section, we further assume that pseudo-labeling of unlabeled training examples is performed by a teacher classifier that did not have access to the training labeled and unlabeled data during its training. We model the pseudo-labels by considering a random variable $\hat{Y}$ that may differ from the true label Y in its distribution, thereby containing label noise. From the point of view of the training of the student classifier, which error we want to evaluate, pseudo-labeled unlabeled training data $\{(\mathbf{x}_i, \hat{y}_i)\}_{i=l+1}^{l+u}$ may be thought as identically and independently distributed with the marginal density $f_{\mathbf{X}}$ but another label generator $P(\hat{Y}|\mathbf{X}=\mathbf{x}_i)$, compared to $P(Y|\mathbf{X}=\mathbf{x})$ used for the labeled data.

The goal of introducing the random variable $\hat{Y}$ is to understand the difference between the risk of a classifier $h : \mathcal{X} \to \{1,\dots,K\}$, when it is evaluated on the true label Y, $R(h)$, and on the imperfect label $\hat{Y}$, defined in the following way:

$$\hat{R}(h) := \mathbb{E}_{\mathbf{X}}\left[\hat{r}(\mathbf{x})\right]$$
$$\text{where } \hat{r}(h,\mathbf{x}) := \sum_{\substack{y\in\{1,\dots,K\}\\ y\neq h(\mathbf{x})}} P(\hat{Y}=y|\mathbf{X}=\mathbf{x}).$$

One can notice that $P(\hat{Y}=\hat{y}|\mathbf{X}=\mathbf{x}) = \sum_{y=1}^{K} P(\hat{Y}=\hat{y}|Y=y,\mathbf{X}=\mathbf{x})P(Y=y|\mathbf{X}=\mathbf{x})$ for all $(\hat{y},y)\in\{1,\dots,K\}^2$, $\mathbf{x}\in\mathcal{X}$. Probabilities $P(\hat{Y}=\hat{y}|Y=y,\mathbf{X}=\mathbf{x})$ are called

the mislabeling probabilities, and they allow us to explicitly model imperfection of labels. However, their estimation is very challenging as they depend on $\mathbf{x}$. A common approach to overcome this is to assume that the mislabeling probabilities are class-related and instance-independent (Chittineni, 1980; Amini and Gallinari, 2003; Scott, 2015).

Assumption 1 (Class-related Mislabeling Model) *We assume that the imperfect label $\hat{Y}$ does not influence the true class distribution: $P(\mathbf{X}|Y,\hat{Y}) = P(\mathbf{X},Y)$, and the label imperfection is summarized through the* mislabeling matrix $\mathbf{P} = (p_{\hat{y},y})_{1\leq\hat{y},y\leq K}$*, defined by*

$$p_{\hat{y},y} := P(\hat{Y} = \hat{y}|Y = y) \quad \forall(\hat{y},y) \in \{1,\ldots,K\}^2, \tag{16}$$

where $\sum_{\hat{y}=1}^{K} p_{\hat{y},y} = 1$.

Particularly, this assumption implies that the posterior distribution of $\hat{Y}$ is decomposed for any $\hat{y} \in \{1,\ldots,K\}$ as follows:

$$P(\hat{Y} = \hat{y}|\mathbf{X} = \mathbf{x}) = \sum_{y=1}^{K} p_{\hat{y},y} P(Y = y|\mathbf{X} = \mathbf{x}). \tag{17}$$

Assumption 1 can be regarded as realistic or unrealistic depending on the application. For example, in the case of MNIST classification the class Y generates features $\mathbf{X}$, so we generally expect that mislabeling of $\hat{Y}$ would depend mostly on how digits differ fundamentally from each other. It is well known that in the MNIST data set digits 4 and 9 can be confused between each other, so we anticipate $P(\hat{Y} = 4|Y = 9)$ and $P(\hat{Y} = 9|Y = 4)$ to be high even if we do not know where exactly the classifier is going to mistake.

Further, we derive several results assuming the class-related model described in Eq. (17). Nevertheless, Theorem 5.2 and Theorem 5.3, which will be given later, hold also for a more general case when mislabeling probabilities are instance-dependent. In the following theorem, we derive a bound that connects the error of the true and the imperfect label in misclassifying a particular example $\mathbf{x} \in \mathcal{X}$.

Theorem 5.2 *Let $\mathbf{X} \in \mathcal{X}$, $Y \in \{1,\ldots,K\}$ and $\hat{Y} \in \{1,\ldots,K\}$ be the input, the true output and the imperfect output random variables, respectively. Let $\mathcal{H}$ be the fixed hypothesis space, Q be the posterior distribution over $\mathcal{H}$, and B_Q be the majority vote classifier defined by Eq.* (1)*. Following Assumption 1, let $\mathbf{P}$ be the mislabeling matrix defined in Eq.* (16)*, and assume that $p_{\hat{y},\hat{y}} > p_{\hat{y},y}$, for all $\hat{y},y \in \{1,\ldots,K\}^2$. Then, for any $\mathbf{x} \in \mathcal{X}$,*

$$r(B_Q,\mathbf{x}) \leq \frac{\hat{r}(B_Q,\mathbf{x})}{\delta(\mathbf{x})} - \frac{1-\alpha(\mathbf{x})}{\delta(\mathbf{x})}, \tag{18}$$

with

- $\delta(\mathbf{x}) := p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})} - \max_{y\in\{1,\ldots,K\}\setminus\{B_Q(\mathbf{x})\}} p_{B_Q(\mathbf{x}),y}$,
- $\alpha(\mathbf{x}) := p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})}$.

Proof First, from the definition of $\hat{r}(B_Q, \mathbf{x})$ and applying Eq. (17) we obtain that

$$\hat{r}(B_Q, \mathbf{x}) = 1 - P(\hat{Y} = B_Q(\mathbf{x})|\mathbf{X} = \mathbf{x}) = 1 - \sum_{y=1}^{K} p_{B_Q(\mathbf{x}),y} P(Y = y|\mathbf{X} = \mathbf{x})$$

$$= 1 - p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})} P(Y = B_Q(\mathbf{x})|\mathbf{X} = \mathbf{x}) - \sum_{\substack{y=1 \\ y \neq B_Q(\mathbf{x})}}^{K} p_{B_Q(\mathbf{x}),y} P(Y = y|\mathbf{X} = \mathbf{x})$$

One can notice that

$$\sum_{\substack{y=1 \\ y \neq B_Q(\mathbf{x})}}^{K} p_{B_Q(\mathbf{x}),y} P(Y = y|\mathbf{X} = \mathbf{x}) \leq \left(\max_{y \in \{1,\dots,K\} \setminus \{B_Q(\mathbf{x})\}} p_{B_Q(\mathbf{x}),y} \right) \sum_{\substack{y=1 \\ y \neq B_Q(\mathbf{x})}}^{K} P(Y = y|\mathbf{X} = \mathbf{x})$$

$$= \left(\max_{y \in \{1,\dots,K\} \setminus \{B_Q(\mathbf{x})\}} p_{B_Q(\mathbf{x}),y} \right) (1 - P(Y = B_Q(\mathbf{x})|\mathbf{X} = \mathbf{x})).$$

Finally, we infer the following inequality:

$$\hat{r}(B_Q, \mathbf{x}) \geq (p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})} - \max_{y \in \{1,\dots,K\} \setminus \{B_Q(\mathbf{x})\}} p_{B_Q(\mathbf{x}),y})(1 - P(Y = B_Q(\mathbf{x})|\mathbf{X} = \mathbf{x}))$$

$$+ 1 - p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})} = \delta(\mathbf{x}) r(B_Q, \mathbf{x}) + 1 - \alpha(\mathbf{x}).$$

Taking into account the assumption that $p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})} > p_{B_Q(\mathbf{x}),y}$, $\forall B_Q(\mathbf{x}) \in \{1, \dots, K\}$, $y \in \{1, \dots, K\} \setminus \{B_Q(\mathbf{x})\}$, we deduce that $\delta(\mathbf{x}) > 0$, which concludes the proof. ■

This theorem gives us insights on how the true error rate can be bounded given the error rate of the imperfect label and the mislabeling matrix. With the quantities $\delta(\mathbf{x})$ and $\alpha(\mathbf{x})$, we perform a correction of $\hat{r}(B_Q, \mathbf{x})$. Note that when there is no mislabeling, the left and right sides of Eq. (18) are equal, since $\alpha(\mathbf{x}) = 1$ and $\delta(\mathbf{x}) = 1$ in this case.

In the theorem, the mislabeling matrix is assumed given, while in practice it has to be estimated. Since the number of matrix entries grows quadratically with the increase of K, a direct estimation of the true posterior probabilities from Eq. (17) may be more affected by the estimation error than the bound itself as the latter needs to know only $2K$ entries. We give more details about estimation of the mislabeling matrix in Section 7.

The bound can be compared with a bound derived by Chittineni (1980, Eq. (3.14), p. 284) for the oracle Bayes classifier $B_O(\mathbf{x}) := \max_{\hat{y} \in \{1,\dots,K\}} P(Y = \hat{y}|\mathbf{X} = \mathbf{x})$. It is shown that $r(B_O, \mathbf{x}) \leq 1 - \frac{1 - \hat{r}(B_O, \mathbf{x})}{\beta}$, where $\beta = \max_{\hat{y}=1,\dots,K} \left(\sum_{y=1}^{K} p_{\hat{y},y} \right)$. One can notice that the regularizer β is constant with respect to $\mathbf{x}$, so the penalization of the error rate $\hat{r}(B_O, \mathbf{x})$ does not depend on the label the classifier predicts. Another limitation is that the bound holds for the oracle Bayes classifier only, while Theorem 5.2 holds for any classifier.

The assumption of Theorem 5.2 requires that the diagonal entries of the mislabeling matrix are the largest elements in their corresponding columns, which means that the imperfect label is reasonably correlated with the true label. However, in practice, the assumption may not hold (at least, for some of the classes), so the theorem is not applicable.

To overcome this, it can be relaxed by considering $\lambda > 0$ such that $\lambda + \delta(\mathbf{x}) > 0$, and so we obtain the following bound:

$$r(B_Q, \mathbf{x}) \leq \frac{\hat{r}(B_Q, \mathbf{x})}{\lambda + \delta(\mathbf{x})} - \frac{1 - \lambda - \alpha(\mathbf{x})}{\lambda + \delta(\mathbf{x})}. \tag{19}$$

When $\delta(\mathbf{x})$ is close to 0, it also avoids the bound to become arbitrarily large. The use of this bound is illustrated in Section C.3 of Appendix.

5.3 C-Bounds with Imperfect Labels

Based on Theorem 5.2, we bound the generalization error of the majority vote classifier $R(B_Q)$, defined as the expectation of $r(B_Q, \mathbf{x})$. By taking expectation in Eq. (18), we obtain that

$$R(B_Q) = \mathbb{E}_{\mathbf{X}} r(B_Q, \mathbf{x}) \leq \mathbb{E}_{\mathbf{X}} \frac{\hat{r}(B_Q, \mathbf{x})}{\delta(\mathbf{x})} - \mathbb{E}_{\mathbf{X}} \frac{1 - \alpha(\mathbf{x})}{\delta(\mathbf{x})}. \tag{20}$$

One can see that for every $\mathbf{x}$, $\hat{r}(B_Q, \mathbf{x})$ is multiplied by a positive weight $1/\delta(\mathbf{X}) > 0$, so the first term of the right-hand side is a weighted generalization error of the imperfect label. In the following theorem, we show how to derive a C-bound in this scenario.

Theorem 5.3 *Let $\mathbf{X} \in \mathcal{X}$, $Y \in \{1, \ldots, K\}$ and $\hat{Y} \in \{1, \ldots, K\}$ be the input, the true output and the imperfect output random variables, respectively. Let $\mathcal{H}$ be the fixed hypothesis space, and Q be the posterior distribution over $\mathcal{H}$. Let B_Q and m_Q be the majority vote classifier and the associated margin defined by Eq. (1) and Eq. (3), respectively. Assuming m_Q is measurable, let $\hat{M}_Q$ be a random variable defined as $\hat{M}_Q := m_Q(\mathbf{X}, \hat{Y})$ with its first and second statistical moments denoted by $\mu_1^{\hat{M}_Q}$ and $\mu_2^{\hat{M}_Q}$, respectively. Following Assumption 1, let $\mathbf{P}$ be the mislabeling matrix defined in Eq. (16), and assume that every diagonal entry of $\mathbf{P}$ is the largest element in the corresponding column, i.e., $p_{\hat{y},\hat{y}} > p_{\hat{y},y}$, for all $(\hat{y}, y) \in \{1, \ldots, K\}^2$. Then, for all choice of Q on a fixed hypothesis space $\mathcal{H}$, and for any density $f_{\mathbf{X}}$ over $\mathcal{X}$ and all distributions $P(Y|\mathbf{X})$ and $P(\hat{Y}|\mathbf{X})$ over $\{1, \ldots, K\}$, we can upper-bound the generalization error $R(B_Q)$, defined in Eq. (13), as follows:*

$$R(B_Q) \leq \psi_{\mathbf{P}} - \frac{\left(\mu_1^{\hat{M}_Q, \mathbf{P}}\right)^2}{\mu_2^{\hat{M}_Q, \mathbf{P}}}, \tag{CBIL}$$

if $\mu_1^{\hat{M}_Q, \mathbf{P}} > 0$, where

- $\psi_{\mathbf{P}} := \mathbb{E}_{\mathbf{X}} \frac{\alpha(\mathbf{x})}{\delta(\mathbf{x})}$ *with δ and α defined as in Theorem 5.2,*
- $\mu_1^{\hat{M}_Q, \mathbf{P}} := \int_{\mathbb{R}^{d+1}} \frac{m}{\delta(\mathbf{x})} f_{\hat{M}_Q, \mathbf{X}}(m, \mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}m$ *is the weighted 1st margin moment, where $f_{\hat{M}_Q, \mathbf{X}}$ denotes the joint density of $\hat{M}_Q$ and $\mathbf{X}$,*
- $\mu_2^{\hat{M}_Q, \mathbf{P}} := \int_{\mathbb{R}^{d+1}} \frac{m^2}{\delta(\mathbf{x})} f_{\hat{M}_Q, \mathbf{X}}(m, \mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}m$ *is the weighted 2nd margin moment.*

Proof At first, let us introduce a normalization factor $\omega_{\mathbf{P}}$ defined as follows:

$$\omega_{\mathbf{P}} := \mathbb{E}_{\mathbf{X}} \frac{1}{\delta(\mathbf{x})} = \int_{\mathbb{R}^{d+1}} \frac{f_{\hat{M}_Q,\mathbf{X}}(m,\mathbf{x})}{\delta(\mathbf{x})} \mathrm{d}\mathbf{x}\mathrm{d}m.$$

Remind that $\hat{r}(h,\mathbf{x}) = P(\hat{M}_Q \leq 0 | \mathbf{X} = \mathbf{x})$. Then, we can write:

$$\begin{aligned}\mathbb{E}_{\mathbf{X}} \frac{\hat{r}(B_Q,\mathbf{x})}{\delta(\mathbf{x})} &= \int_{\mathbb{R}^d} \frac{1}{\delta(\mathbf{x})} P(\hat{M}_Q \leq 0 | \mathbf{X} = \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_{-\infty}^{0} \int_{\mathbb{R}^d} \frac{f_{\hat{M}_Q,\mathbf{X}}(m,\mathbf{x})}{\delta(\mathbf{x})} \mathrm{d}\mathbf{x}\mathrm{d}m \\ &= \omega_{\mathbf{P}} \int_{-\infty}^{0} \frac{\int_{\mathbb{R}^d} f_{\hat{M}_Q,\mathbf{X}}(m,\mathbf{x})/\delta(\mathbf{x})\mathrm{d}\mathbf{x}}{\int_{\mathbb{R}^{d+1}} f_{\hat{M}_Q,\mathbf{X}}(m,\mathbf{x})/\delta(\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}m} \mathrm{d}m = \omega_{\mathbf{P}} P(\hat{M}_\omega < 0), \end{aligned} \tag{21}$$

where the last equality is given by a random variable $\hat{M}_\omega$ coming from the density f_ω defined as the expression inside the integral in Eq. (21). We further notice that the weighted first and second moments can be represented as

$$\begin{aligned}\mu_1^{\hat{M}_Q,\mathbf{P}} &= \int_{\mathbb{R}^{d+1}} \frac{m}{\delta(\mathbf{x})} f_{\hat{M}_Q,\mathbf{X}}(m,\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}m = \omega_{\mathbf{P}} \mu_1^{\hat{M}_\omega}, \\ \mu_2^{\hat{M}_Q,\mathbf{P}} &= \int_{\mathbb{R}^{d+1}} \frac{m^2}{\delta(\mathbf{x})} f_{\hat{M}_Q,\mathbf{X}}(m,\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}m = \omega_{\mathbf{P}} \mu_2^{\hat{M}_\omega}. \end{aligned}$$

From this, we also obtain that $var(M_\omega) = \left(\mu_2^{\hat{M}_Q,\mathbf{P}}/\omega_{\mathbf{P}}\right) - \left(\mu_1^{\hat{M}_Q,\mathbf{P}}/\omega_{\mathbf{P}}\right)^2$. Then, using the Cantelli-Chebyshev inequality (Lemma B.1) with $\lambda = \mu_1^{\hat{M}_\omega} = \mu_1^{\hat{M}_Q,\mathbf{P}}/\omega_{\mathbf{P}}$ we deduce the following inequality:

$$P(\hat{M}_\omega < 0) \leq \frac{\left(\mu_2^{\hat{M}_Q,\mathbf{P}}/\omega_{\mathbf{P}}\right) - \left(\mu_1^{\hat{M}_Q,\mathbf{P}}/\omega_{\mathbf{P}}\right)^2}{\left(\mu_2^{\hat{M}_Q,\mathbf{P}}/\omega_{\mathbf{P}}\right) - \left(\mu_1^{\hat{M}_Q,\mathbf{P}}/\omega_{\mathbf{P}}\right)^2 + \left(\mu_1^{\hat{M}_Q,\mathbf{P}}/\omega_{\mathbf{P}}\right)^2} = 1 - \frac{\left(\mu_1^{\hat{M}_Q,\mathbf{P}}\right)^2}{\omega_{\mathbf{P}} \mu_2^{\hat{M}_Q,\mathbf{P}}}. \tag{22}$$

Combining Eq. (22) and Eq. (20) we infer (CBIL):

$$R(B_Q) \leq \mathbb{E}_{\mathbf{X}} \frac{\hat{r}(B_Q,\mathbf{x})}{\delta(\mathbf{x})} - \mathbb{E}_{\mathbf{X}} \frac{1-\alpha(\mathbf{x})}{\delta(\mathbf{x})} = \omega_{\mathbf{P}} P(\hat{M}_\omega < 0) - \omega_{\mathbf{P}} + \psi_{\mathbf{P}} \leq \psi_{\mathbf{P}} - \frac{\left(\mu_1^{\hat{M}_Q,\mathbf{P}}\right)^2}{\mu_2^{\hat{M}_Q,\mathbf{P}}}.$$

■

Given data with imperfect labels, the direct evaluation of the generalization error rate may be biased, leading to an overly optimistic evaluation. Using the mislabeling matrix $\mathbf{P}$ we derive a more conservative C-bound, where the error of $\mathbf{x}$ is penalized by the factor $1/\delta(\mathbf{x})$. When there is no mislabeling, $\psi_{\mathbf{P}} = 1$, $\mu_1^{\hat{M}_Q,\mathbf{P}}$ and $\mu_2^{\hat{M}_Q,\mathbf{P}}$ are equivalent to $\mu_1^{\hat{M}_Q}$ and $\mu_2^{\hat{M}_Q}$, so we obtain the regular C-bound (CB).

In particular, we can use this general result to evaluate the error rate when mislabeling is caused by pseudo-labeling of unlabeled data. Note that Lacasse et al. (2007) and Roy et al.

(2011) proposed another way to evaluate the C-bound in the semi-supervised setting: they use unlabeled data to estimate the second margin moment by expressing it via disagreement of hypotheses. However, this is only possible in the binary classification case.

While we have combined the mislabeling bound given by Eq. (18) with the supervised C-bound (Laviolette et al., 2017), the bound based on the second-order Markov's inequality could be an alternative. As pointed out by Masegosa et al. (2020), the latter can be regarded as a relaxation of the C-bound but is easier to estimate from data in some cases.

5.4 PAC-Bayesian Theorem for C-Bound Estimation

When the margin mean, the margin variance and the mislabeling matrix are empirically estimated from data, evaluation of (CBIL) may be optimistically biased. In this section, we analyze the behavior of the estimate with respect to the sample size. To achieve that, we use the PAC-Bayesian theory initiated by McAllester (1999, 2003) to derive a Probably Approximately Correct bound defined below.

Theorem 5.4 *Under the notations of Theorem 5.3, for any fixed set of classifiers $\mathcal{H}$, for any prior distribution Q_0 on $\mathcal{H}$ and any $\epsilon \in (0,1]$, with a probability at least $1-\epsilon$ over the choice of the sample $\{\mathbf{x}_i, y_i\}_{i=1}^{l} \cup \{\mathbf{x}_i\}_{i=l+1}^{l+u}$, for every posterior distribution Q over $\mathcal{H}$, if $\mu_1^{\hat{M}_Q} > 0$ and $\tilde{\delta}(\mathbf{x}) > 0$, we have:*

$$R(B_Q) \leq \tilde{\psi} - \frac{\tilde{\mu}_1^2}{\tilde{\mu}_2}, \tag{23}$$

where

$$\tilde{\mu}_1 := \frac{1}{u}\sum_{i=l+1}^{l+u} \frac{1}{\tilde{\delta}(\mathbf{x}_i)} \sum_{k=1}^{K} m_Q(\mathbf{x}_i, k) P(\hat{Y}=k|\mathbf{X}=\mathbf{x}_i) - J_1 \sqrt{\frac{2}{u}\left[KL(Q\,\|\,Q_0) + \ln\frac{2\sqrt{u}}{\epsilon/\rho}\right]}$$

$$\tilde{\mu}_2 := \frac{1}{u}\sum_{i=l+1}^{l+u} \frac{1}{\tilde{\delta}(\mathbf{x}_i)} \sum_{k=1}^{K} (m_Q(\mathbf{x}_i, k))^2 P(\hat{Y}=k|\mathbf{X}=\mathbf{x}_i) + J_2 \sqrt{\frac{2}{u}\left[2KL(Q\,\|\,Q_0) + \ln\frac{2\sqrt{u}}{\epsilon/\rho}\right]}$$

$$\tilde{\psi} := \frac{1}{u}\sum_{i=l+1}^{l+u} \frac{\tilde{\alpha}(\mathbf{x}_i)}{\tilde{\delta}(\mathbf{x}_i)} + J_3 \sqrt{\frac{2}{u}\ln\frac{2\sqrt{u}}{\epsilon/\rho}}$$

$$\tilde{\delta}(\mathbf{x}) := \hat{\delta}(\mathbf{x}) - \sqrt{\frac{1}{2l_{k_\mathbf{x}}}\ln\frac{2\sqrt{l_{k_\mathbf{x}}}}{\epsilon/\rho}} - \sqrt{\frac{1}{2l_{j_\mathbf{x}}}\ln\frac{2\sqrt{l_{j_\mathbf{x}}}}{\epsilon/\rho}}, \text{ with } k_\mathbf{x} := B_Q(\mathbf{x}), j_\mathbf{x} := \operatorname*{argmin}_{j \neq k_\mathbf{x}} l_j,$$

$$\tilde{\alpha}(\mathbf{x}) := \hat{\alpha}(\mathbf{x}) + \sqrt{\frac{1}{2l_{k_\mathbf{x}}}\ln\frac{2\sqrt{l_{k_\mathbf{x}}}}{\epsilon/\rho}}, \text{ for all } \mathbf{x} \in \mathcal{X}$$

$$J_1 := \max_{\mathbf{x}} |(1/\tilde{\delta}(\mathbf{x})) \sum_{\hat{y}=1}^{K} m_Q(\mathbf{x}, \hat{y}) P(\hat{Y}=\hat{y}|\mathbf{X}=\mathbf{x})|$$

$$J_2 := \max_{\mathbf{x}} |1/\tilde{\delta}(\mathbf{x}) \sum_{\hat{y}=1}^{K} (m_Q(\mathbf{x}, \hat{y}))^2 P(\hat{Y}=\hat{y}|\mathbf{X}=\mathbf{x})|$$

$$J_3 := \max_{x \in \mathcal{X}} [\hat{\alpha}(\mathbf{x}) + \iota(l_{k_\mathbf{x}})] / [\hat{\delta}(\mathbf{x}) - \iota(l_{k_\mathbf{x}}) - \iota(l_{j_\mathbf{x}})]$$

$$\iota(l_y) := \sqrt{\frac{1}{2l_y} \ln \frac{2\sqrt{l_y}}{\epsilon}},$$

and where $\hat{\delta}(\mathbf{x})$ *and* $\hat{\alpha}(\mathbf{x})$ *are empirical estimates respectively of* $\delta(\mathbf{x})$ *and* $\alpha(\mathbf{x})$ *based on the available labeled set,* $KL(Q \,\|\, Q_0)$ *is the Kullback-Leibler divergence between* Q *and* Q_0*,* $l_y = \sum_{i=1}^{l} \mathbb{I}(y_i = y)/l$ *is the proportion of the labeled training examples from the true class* y*, and* $\rho := 2K + 3$ *comes from applying a union bound.*

The proof is a combination of Propositions B.5, B.7 and B.9 deferred to Appendix B.

Thus, by using Eq. (23) we additionally penalize the C-bound by the sample size and the divergence between Q and Q_0. As u grows, the penalization becomes less severe, so $\tilde{\mu}_1$ and $\tilde{\mu}_2$ are close to $\mu_1^{\hat{M}_Q}$ and $\mu_2^{\hat{M}_Q}$. Similarly, $\tilde{\delta}(\mathbf{x})$ and $\tilde{\alpha}(\mathbf{x})$ are closer to $\hat{\delta}(\mathbf{x})$ and $\hat{\alpha}(\mathbf{x})$ with the increase of the number examples used to estimate the mislabeling matrix, which we take l for the sake of simplicity. Note that, in contrast to the supervised case (Laviolette et al., 2017, Theorem 3), J_1 and J_2 can have a drastic influence on the bound's value, when $\tilde{\delta}(\mathbf{x})$ is close to 0, which motivates in practice to use the λ-relaxation given by Eq. (19).

The obtained bound may be used to estimate the error of the majority vote from data, with the pseudo-labeled unlabeled examples serving as a hold-out set for estimating the margin moments, and the labeled examples are used to estimate the mislabeling matrix. In the case of classical ensembles, it can be performed in the out-of-bag fashion following Thiemann et al. (2017) and Lorenzen et al. (2019). However, the bound does not appear tighter in practice compared to the supervised case (Laviolette et al., 2017) due to the additional penalization on estimation of the mislabeling matrix. Making this bound tighter could be a good direction for future work. Nevertheless, when the focus is set on model selection, a common choice is to simply use an empirical estimate of the C-bound as an optimization criterion (Bauvin et al., 2020).

6. Algorithm and Experimental Results

In this section, we show that the proposed bound on the transductive conditional risk found in Theorem 4.1 can be applied for developing a new self-training technique for multi-class classification. Then, in order to support our suggested framework, we carried out several numerical experiments in real-world scenarios and compared with other semi-supervised classification algorithms. Finally, we illustrate the proposed (CBIL) on real data sets and analyze its behavior.

6.1 Multi-class Self-training Algorithm

In this section, we consider the classical setting where the self-training algorithm is initialized by a supervised base classifier that has been trained first on available labeled training data. Then, at each iteration, the predictions of the base classifier, called pseudo-labels, are assigned to those unlabeled examples that have a confidence score above a certain threshold. The pseudo-labeled examples are then included in the training set, and the base classifier is retrained. The process is repeated until no examples for pseudo-labeling are left.

The central question of applying the self-training algorithm is how to choose the confidence threshold. While setting the threshold to a low value would imply a lot of label noise, setting it to a very high value would put excessive trust in the confidence score initially biased by the small labeled set. Considering the prediction vote of the majority vote classifier as an indicator of confidence, we propose the strategy to automatically select the threshold by minimizing the following criterion $R_{\mathcal{U}|\boldsymbol{\theta}}(B_Q)$ defined as

$$R_{\mathcal{U}|\boldsymbol{\theta}}(B_Q) := \frac{R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q)}{\frac{1}{u}\sum_{\mathbf{x}\in \mathrm{X}_{\mathcal{U}}} \mathbb{I}(v_Q(\mathbf{x}, B_Q(\mathbf{x})) \geq \theta_{B_Q(\mathbf{x})})}. \tag{24}$$

Thus, the threshold is found by making a trade-off between the error we induce by pseudo-labeling and the number of pseudo-labeled examples. Algorithm 1 summarizes mains steps of our method that we further call `MSTA`[3].

To evaluate $R_{\mathcal{U}|\boldsymbol{\theta}}(B_Q)$, we bound the numerator of Eq. (24) by Corollary 4.2. However, the bound can practically be computed only with assumptions, since the posterior probabilities $P(Y=y|X=\mathbf{x})$ for unlabeled examples are not known. In this work, we approximate the posterior $P(Y=y|X=\mathbf{x})$ by $v_Q(\mathbf{x}, y)$ of the base classifier trained on labeled examples only (the initial step of `MSTA`). Although this approximation is optimistic, by formulating the bound as probabilistic we keep some chances for other classes so the error of the supervised classifier can be smoothed. Nevertheless, it must be borne in mind that the hypothesis space should be diverse enough so that the entropy of $(v_Q(\mathbf{x}, y))_{y=1}^K$ would not be always zero, and the errors are made mostly on low prediction votes. In our experiments, as the base classifier, we use the random forest (Breiman, 2001) that aggregates predictions from trees learned on different bootstrap samples. In Appendix C.1, we validate the proposed approximation by empirically comparing it with the case when the posterior probabilities are set to $1/K$, i.e., when we treat all classes as equally probable.

To find an optimal $\boldsymbol{\theta}^*$ we perform a grid search over the hypercube $(0, 1]^K$. The same algorithm is used for computing the optimal $\gamma^*_{y,\hat{y}}$ that provides the value of an upper bound for the conditional risk (see Theorem 4.1). As the direct grid search in the multi-class setting costs $O\left(R^K\right)$, where R is the sampling rate of the grid, we notice that

$$R_{\mathcal{U}|\boldsymbol{\theta}}(B_Q) \leq \sum_{\hat{y}=1}^{K} \frac{R^{(\hat{y})}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q)}{\frac{1}{u}\sum_{\mathbf{x}\in \mathrm{X}_{\mathcal{U}}} \mathbb{I}(v_Q(\mathbf{x}, \hat{y}) \geq \theta_{\hat{y}})\mathbb{I}(B_Q(\mathbf{x}) = \hat{y})}, \tag{$*$}$$

where $R^{(\hat{y})}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q) = \sum_{y=1}^{K} u_y R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, y, \hat{y})/u$. Thus, $(*)$ might be minimized term by term, tuning independently each component of $\boldsymbol{\theta}$. This replaces the K-dimensional minimization task by K tasks of 1-dimensional minimization.

6.2 Experimental Setup

All experiments were performed on a cluster with an `Intel(R) Xeon(R) CPU E5-2640 v3` at `2.60GHz`, `32` cores, `256GB` of RAM, the `Debian 4.9.110-3 x86_64` OS. Experiments are conducted on publicly available data sets (Dua and Graff, 2017; Chang and Lin, 2011; Xiao et al., 2017). Since we are interested in the practical use of our approach in the

3. The source code of `MSTA` can be found at `https://github.com/vfeofanov/trans-bounds-maj-vote`.

Algorithm 1 Multi-class Self-training algorithm (MSTA)

Input:
Labeled observations $\mathrm{Z}_{\mathcal{L}}$
Unlabeled observations $\mathrm{X}_{\mathcal{U}}$
Initialisation:.
A set of pseudo-labeled instances, $\mathrm{Z}_{\mathcal{P}} \leftarrow \emptyset$
A classifier B_Q trained on $\mathrm{Z}_{\mathcal{L}}$
repeat

1. Compute the vote threshold $\boldsymbol{\theta}^*$ that minimizes the conditional Q-weighted majority vote error rate:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in (0,1]^K}{\operatorname{argmin}} R_{\mathcal{U}|\boldsymbol{\theta}}(B_Q). \qquad (\star)$$

2. $S \leftarrow \{(\mathbf{x}, y') | \mathbf{x} \in \mathrm{X}_{\mathcal{U}}; [v_Q(\mathbf{x}, y') \geq \theta^*_{y'}] \wedge [y' = \operatorname{argmax}_{k \in \{1,\dots,K\}} v_Q(\mathbf{x}, k)]\}$
3. $\mathrm{Z}_{\mathcal{P}} \leftarrow \mathrm{Z}_{\mathcal{P}} \cup S$, $\mathrm{X}_{\mathcal{U}} \leftarrow \mathrm{X}_{\mathcal{U}} \setminus S$
4. Learn a classifier B_Q with the following loss function:

$$\mathcal{L}(B_Q, Z_{\mathcal{L}}, \mathrm{Z}_{\mathcal{P}}) = \frac{l + |\mathrm{Z}_{\mathcal{P}}|}{l} \mathcal{L}(B_Q, Z_{\mathcal{L}}) + \frac{l + |\mathrm{Z}_{\mathcal{P}}|}{|\mathrm{Z}_{\mathcal{P}}|} \mathcal{L}(B_Q, \mathrm{Z}_{\mathcal{P}})$$

until $\mathrm{X}_{\mathcal{U}} = \emptyset$ or $S = \emptyset$
Output: The final classifier B_Q

semi-supervised context, we would like to see if it has good performance when $l \ll u$. Therefore, we do not use the train/test splits that are proposed by data sources. Instead, we propose our own splits that makes a situation closer to the semi-supervised context. Each experiment is conducted 20 times, by randomly splitting an original data set on a labeled and an unlabeled parts keeping fixed their respective size at each iteration. The reported performance results are averaged over the 20 trials. We evaluate the performance as the accuracy score over the unlabeled training set (`ACC-U`).

In all our experiments, we consider the Random Forest algorithm (Breiman, 2001) (denoted by `RF`) with 200 trees and the maximal depth of trees as the majority vote classifier with the uniform prior and posterior distributions. For an observation $\mathbf{x}$, we evaluate the vector of class votes $\{v(\mathbf{x}, i)\}_{i=1}^K$ by averaging over the trees the vote given to each class by the tree. A tree computes a class vote as the fraction of training examples in a leaf belonging to a class.

Experiments are conducted on 11 real data sets. The associated applications are image classification with the `Fashion` data set, the `Pendigits` and the `MNIST` databases of handwritten digits; a signal processing application with the `SensIT` data set for vehicle type classification and the human activity recognition `HAR` database; speech recognition using the `Vowel`, the `Isolet` and the `Letter` data sets; document recognition using the `Page Blocks` database; and finally applications to bioinformatics with the `Protein` and `DNA` data sets. The main characteristics of these data sets are summarized in Table 2.

The proposed `MSTA` that automatically finds the threshold by minimizing the criterion given by Eq. (24), is compared with the following baselines:

Data set	# of labeled examples, l	# of unlabeled examples, u	Dimension, d	# of classes, K
Vowel	99	891	10	11
Protein	129	951	77	8
DNA	31	3155	180	3
PageBlocks	1094	4379	10	5
Isolet	389	7408	617	26
HAR	102	10197	561	6
Pendigits	109	10883	16	10
Letter	400	19600	16	26
Fashion	175	69825	784	10
MNIST	175	69825	784	10
SensIT	49	98479	100	3

Table 2: Characteristics of data sets used in our experiments ordered by the size of the training set ($n = l + u$).

- a fully supervised RF trained using only labeled examples. The approach is obtained at the initialization step of MSTA and once learned it is directly applied to predict the class labels of the whole unlabeled set;
- the scikit-learn implementation (Pedregosa et al., 2011) of the graph-based, label spreading algorithm (Zhou et al., 2004) denoted by LS;
- the one-versus-all extension of a transductive support vector machine Joachims (1999) using the Quasi-Newton scheme. The approach was proposed by Gieseke et al. (2014) ans is further denoted as QN-S3VM[4];
- a semi-supervised extension of the linear discriminant analysis Semi-LDA, which is based on the contrastive pessimistic likelihood estimation proposed by Loog (2015);
- a semi-supervised extension of the random forest DAS-RF proposed by Leistner et al. (2009) where the classifier is repeatedly re-trained on the labeled and all the unlabeled examples with pseudo-labels optimized via deterministic annealing;
- the multi-class extension of the classical self-training approach (denoted by FSTA, Tür et al., 2005) with a fixed prediction vote threshold $\boldsymbol{\theta}$;
- a self-training approach (denoted by CSTA) where the threshold is defined via curriculum learning by taking it as the $(1 - t \cdot \Delta)$-th percentile of the prediction vote distribution at the step $t = 1, 2, \ldots$ (Cascante-Bonilla et al., 2021).

As the size of the labeled training examples $|\mathrm{Z}_{\mathcal{L}}|$ is small, the hyperparameter tuning can not be performed properly. At the same time, the performance of baselines may be sensitive to some of their hyperparameters. For this reason, we compute LS, QN-S3VM, Semi-LDA, DAS-RF on a grid of hyperparameters' values, and then choose the

4. The source code for the binary QN-S3VM is available at http://www.fabiangieseke.de/index.php/code/qns3vm.

Data set	RF	LS	QN-S3VM	Semi-LDA	DAS-RF	FSTA $_{\theta=0.7}$	CSTA $_{\Delta=1/3}$	MSTA
Vowel	$.586 \pm .028$	**.602** $\pm .026$	$.208^{\downarrow} \pm .029$	$.432^{\downarrow} \pm .029$	$.587 \pm .028$	$.531^{\downarrow} \pm .034$	$.576^{\downarrow} \pm .031$	$.586 \pm .026$
Protein	$.764^{\downarrow} \pm .032$	$.825 \pm .028$	$.72^{\downarrow} \pm .034$	**.842** $\pm .029$	$.768^{\downarrow} \pm .036$	$.687^{\downarrow} \pm .036$	$.771^{\downarrow} \pm .035$	$.781^{\downarrow} \pm .034$
DNA	$.693^{\downarrow} \pm .074$	$.584^{\downarrow} \pm .038$	**.815** $\pm .025$	$.573^{\downarrow} \pm .037$	$.693^{\downarrow} \pm .083$	$.521^{\downarrow} \pm .095$	$.671^{\downarrow} \pm .112$	$.702^{\downarrow} \pm .082$
PageBlocks	$.965 \pm .003$	$.905^{\downarrow} \pm .004$	$.931^{\downarrow} \pm .003$	$.935^{\downarrow} \pm .009$	$.965 \pm .003$	$.964 \pm .004$	$.965 \pm .003$	**.966** $\pm .002$
Isolet	$.854^{\downarrow} \pm .016$	$.727^{\downarrow} \pm .01$	$.652^{\downarrow} \pm .016$	$.787^{\downarrow} \pm .019$	$.859^{\downarrow} \pm .018$	$.7^{\downarrow} \pm .04$	$.843^{\downarrow} \pm .021$	**.875** $\pm .014$
HAR	$.851 \pm .024$	$.215^{\downarrow} \pm .05$	$.78^{\downarrow} \pm .02$	$.743^{\downarrow} \pm .043$	$.852 \pm .024$	$.81^{\downarrow} \pm .041$	$.841 \pm .029$	**.854** $\pm .026$
Pendigits	$.863^{\downarrow} \pm .022$	**.916** $\pm .013$	$.675^{\downarrow} \pm .022$	$.824^{\downarrow} \pm .012$	$.872^{\downarrow} \pm .023$	$.839^{\downarrow} \pm .036$	$.871^{\downarrow} \pm .029$	$.884^{\downarrow} \pm .022$
Letter	$.711 \pm .011$	$.664^{\downarrow} \pm .01$	$.064^{\downarrow} \pm .013$	$.589^{\downarrow} \pm .016$	$.718 \pm .012$	$.651^{\downarrow} \pm .015$	**.72** $\pm .013$	$.717 \pm .013$
Fashion	$.718 \pm .022$	NA	NA	$.537^{\downarrow} \pm .027$	$.722 \pm .023$	$.64^{\downarrow} \pm .04$	$.713 \pm .026$	**.723** $\pm .023$
MNIST	$.798^{\downarrow} \pm .015$	NA	NA	$.423^{\downarrow} \pm .029$	$.822^{\downarrow} \pm .017$	$.705^{\downarrow} \pm .055$	$.829^{\downarrow} \pm .02$	**.857** $\pm .013$
SensIT	**.723** $\pm .022$	NA	NA	$.647^{\downarrow} \pm .042$	**.723** $\pm .022$	$.692^{\downarrow} \pm .023$	$.713 \pm .024$	$.722 \pm .021$

Table 3: Classification performance on different data sets described in Table 2. The performance is computed using the accuracy score on the unlabeled training examples (ACC-U). The sign $^{\downarrow}$ shows if the performance is statistically worse than the best result on the level 0.01 of significance. NA indicates the case when the time limit was exceeded.

value for which the performance is the best in average on 20 trials. We tune the RBF kernel parameter $\sigma \in \{10, 1.5, 0.5, 10^{-1}, 10^{-2}, 10^{-3}\}$ for LS, the regularization parameters $(\lambda, \lambda') \in \{10^{-1}, 10^{-2}, 10^{-3}\}^2$ for QN-S3VM, the learning rate $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ for Semi-LDA, the initial temperature $T_0 \in \{10^{-3}, 5 \cdot 10^{-3}, 10^{-2}\}$ for DAS-RF. Other hyperparameters for these algorithms are left to their default values. Particularly, in DAS-RF the strength parameter and the number of iterations are respectively set to 0.1 and 10.

While the aforementioned parameters are rather data-dependent, the choice of θ for FSTA and Δ for CSTA depend more on what prediction vote distribution the base classifier outputs. After manually testing different values, we have found that FSTA $_{\theta=0.7}$ and CSTA $_{\Delta} = 1/3$ are good choices for the random forest. For FSTA, we terminate the learning procedure as soon as the algorithm makes 10 iterations, which reduces the computation time and may also improve the performance, since, in this case, the algorithm is less affected by noise. Cascante-Bonilla et al. (2021) used for CSTA a slightly other architecture for self-training, where the set of selected pseudo-labeled examples included just for one iteration (like if in Algorithm 1 Step 3 would be replaced by $\mathrm{Z}_{\mathcal{P}} \leftarrow S$). In our context, we have found that the performance of CSTA is identical for both architectures.

6.3 Illustration of MSTA

In our setup, a time deadline is set: we stop computation for an algorithm if one trial takes more than 4 hours. Table 3 summarizes results obtained by RF, LP, QN-S3VM, Semi-LDA, DAS-RF, FSTA, CSTA and MSTA. We used bold face to indicate the highest performance rates and the symbol $\downarrow$ indicates that the performance is significantly worse than the best result, according to Mann-Whitney U test (Mann and Whitney, 1947) used at the p-value threshold of 0.01.

From these results it comes out that

- in 5 of 11 cases, the `MSTA` performs better than its opponents. On data sets `Isolet` and `MNIST` it significantly outperforms all the others, and it significantly outperforms the baseline `RF` on `Isolet`, `Pendigits` and `MNIST` (6% improvement);
- the `LS` and the `QN-S3VM` did not pass the scale over larger data sets (`Fashion`, `MNIST` and `SensIT`), while the `MSTA` did not exceeded 2 minutes per trial on these data sets (see Table 5);
- the performance of `LS` and `Semi-LDA` performance varies greatly on different data sets, which may be caused by the topology of data. In contrast, `MSTA` has more stable results over all data sets as it is based on the predictive score, and the `RF` is used as the base classifier;
- since the `QN-S3VM` is a binary classifier by nature, its one-versus-all extension is not robust with respect to the number of classes. This can be observed on `Vowel`, `Isolet` and `Letter`, where the number of classes is high;
- from our observation, both `LS` and `QN-S3VM` are highly sensitive to the choice of the hyperparameters. However, it is not very clear whether these hyperparameters can be properly tuned given a insufficient number of labeled examples. The same concern is applied to all the other semi-supervised baselines, while `MSTA` does not require any particular tuning since it finds automatically the threshold $\boldsymbol{\theta}$;
- while the approach proposed by Loog (2015) always guarantees an improvement of the likelihood compared to the supervised case, we have observed that the classification accuracy is not always improved for `Semi-LDA` and may even degrade over the supervised linear discriminant analysis;
- compared to the fully supervised approach, `RF`, the use of pseudo-labeled unlabeled training data (in `DAS-RF`, `FSTA`, `CSTA` or `MSTA`) may generally give no benefit or even degrade performance in some cases (`Vowel`, `PageBlocks`, `SensIT`). This may be due to the fact that the learning hypotheses are not met regarding the data sets where this effect is observed;
- although for `DAS-RF` the performance is usually not degraded when T_0 is properly chosen, it has rather little improvement compared to `RF`. The performance of `FSTA` degrades most of the time, while degradation for `CSTA` is observed on 6 data sets. The latter suggests that the choice of the threshold for pseudo-labeling is crucial and challenging in the multi-class framework. Using the proposed criterion based on Eq. (24), we can find the threshold efficiently;
- from the results it can be seen that self-training is also sensitive to the choice of the initial classifier. On some data sets, the number of labeled examples might be too small leading to a bad initialization of the first classifier trained over the labeled set. This implies that the initial votes are biased, so even with a well picked threshold we do not expect a great increase in performance (see Appendix C.1 for more details).

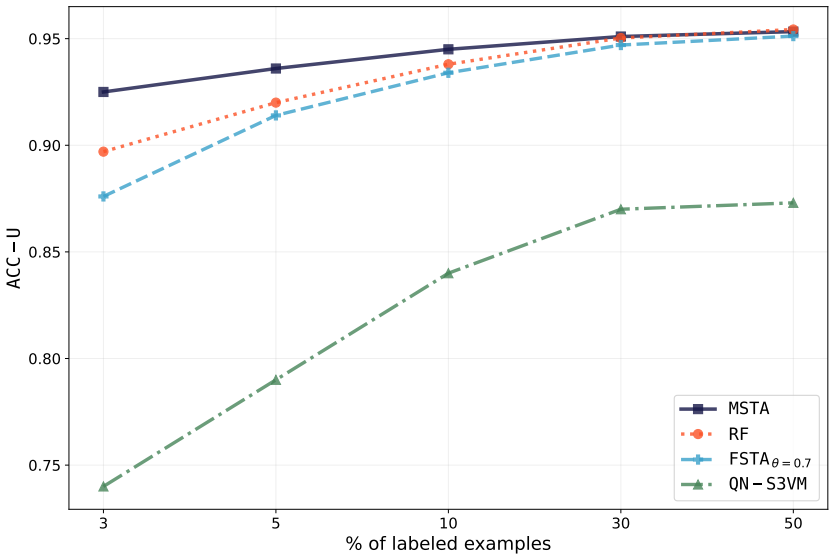

Figure 2: Classification accuracy with respect to the proportion of unlabeled examples for the `MNIST` data set (a subsample of 3500 examples). On the graph, dots represent the average performance on the unlabeled examples over 20 random splits. For simplicity of illustration, the other considered algorithms are not displayed.

We also analyze the behavior of the various algorithms with an increasing number of labeled examples in the training set. Figure 2 illustrates this by showing the accuracy on a subsample of 3500 observations from `MNIST` of `RF`, `QN-S3VM`, $\texttt{FSTA}_{\theta=0.7}$ and `MSTA` with respect to the percentage of the labeled training examples. As expected, all performance curves increase monotonically with respect to the additional labeled data. When there are sufficient labeled training examples, `MSTA`, `FSTA` and `RF` actually converge to the same accuracy score, suggesting that the labeled data carries enough information, so it can not be extracted further from the unlabeled data.

Now, we present a comparison of the learning algorithms under consideration by analyzing their complexity. The time complexity of the random forest `RF` is $O(Td\tilde{l}\log^2\tilde{l})$ (Louppe, 2014), where T is the number of decision trees in the forest and $\tilde{l} \approx 0.632 \cdot l$ is the number of training examples used for each tree. Since `RF` is employed in `DAS-RF` and self-training, the time complexity of `DAS-RF`, `FSTA` and `CSTA` is $O(CTd\tilde{n}\log^2\tilde{n})$, where C is the number of times `RF` has been learned, $\tilde{n} \approx 0.632 \cdot n$. In our experimental setup, $C = 11$ for `FSTA` and `DAS-RF`, and $C = 1/\Delta + 1 = 4$ for `CSTA`.

The time required for finding the optimal threshold at every iteration of the `MSTA` is $O(K^2R^2n)$, where R is the sampling rate of the grid. From this we deduce that the complexity of `MSTA` is $O(C\max(Tdn\log^2 n, K^2R^2n))$. As n grows, the complexity is written as $O(dn\log^2 n)$, since C, T, R are constant. This indicates a good scalability of all consid-

ered pseudo-labeling methods for large-scale data as they also have a memory consumption proportional to nd, so the computation can be performed on a regular PC even for the large-scale applications. In the label spreading algorithm, an iterative procedure is performed, where at every step the affinity matrix is computed. Hence, the time complexity of the `LS` is $O(Mn^2d)$, where M is the maximal number of iterations. From our observation, the convergence of `LS` is highly influenced by the value of σ and the data topology. The time complexity of the `QN-S3VM` is $O(n^2d)$ (Gieseke et al., 2014). Both algorithms suffer from high run-time for large-scale applications. Since `LS` and `QN-S3VM` evaluate respectively the affinity matrix and the kernel matrix of size n by n, these algorithms have also large space complexity proportional to n^2. From our observation, for the large-scale data (`Fashion`, `MNIST`, `SensIT`) the maximal resident set size[5] of `LS` and `QN-S3VM` may reach up to 200GB of RAM, which is practically infeasible with the lack of resources.

Finally, the time complexity of `Semi-LDA` is $O(M\max(nd^2, d^3))$, where M is the maximal number of iterations and $O(\max(nd^2, d^3))$ is the complexity of the linear discriminant analysis assuming $n > d$ (Cai et al., 2008), and the space complexity is $O(nd)$. The approach passes the scale well with respect to the sample size, but may significantly slow down in the case of very large dimension. In Section C.2, we further analyze the time complexity empirically for all the methods under consideration.

6.4 Illustration of (CBIL)

In this section, we empirically illustrate the value of (CBIL) evaluated in the following way. We split the labeled data into two separate sets, where one is used to train a teacher model, and another is for training a student model. Then, the performance of the student model is evaluated on the unlabeled examples pseudo-labeled by the teacher model, which ensures the i.i.d. assumption imposed on the pseudo-labeled data. We empirically compare (CBIL) with the oracle C-bound (CB) evaluated as if the labels for the considered unlabeled data would be known.

To do so, we compute the value of the two bounds varying the number of examples used for evaluation with respect to the prediction confidence: the pseudo-labeled examples are sorted by the value of the student's prediction vote in the descending order, and we keep only the first $\rho\%$ of the examples for $\rho \in \{20, 40, 60, 80, 100\}$.

We use the votes of the current classifier and expect that with increase of ρ we have more mislabels, so the (CBIL) is more penalized. In (CBIL), we use the true value of the mislabeling matrix (i.e., evaluated using the labels of unlabeled data) for clear illustration of the C-bound's penalization. In Section 7, we discuss the possible estimations of the mislabeling matrix. The experimental results on 4 data sets `HAR`, `Isolet`, `Letter` and `MNIST` are illustrated in Figure 3. As expected, the classifier makes mistakes mostly on low class votes, so the error increases when ρ grows. One can see that when ρ is small and the majority of pseudo-labels are true, (CBIL) appears to be conservative giving a pessimistic result. This may be due to the fact that even if each example is subject to a small penalty, the value of (CBIL) will accumulate these penalties. When more noisy pseudo-labels are included, the difference between the two bounds becomes small, and it

5. Maximal resident set size (maxRSS) is the peak portion of memory that was occupied in RAM during the run.

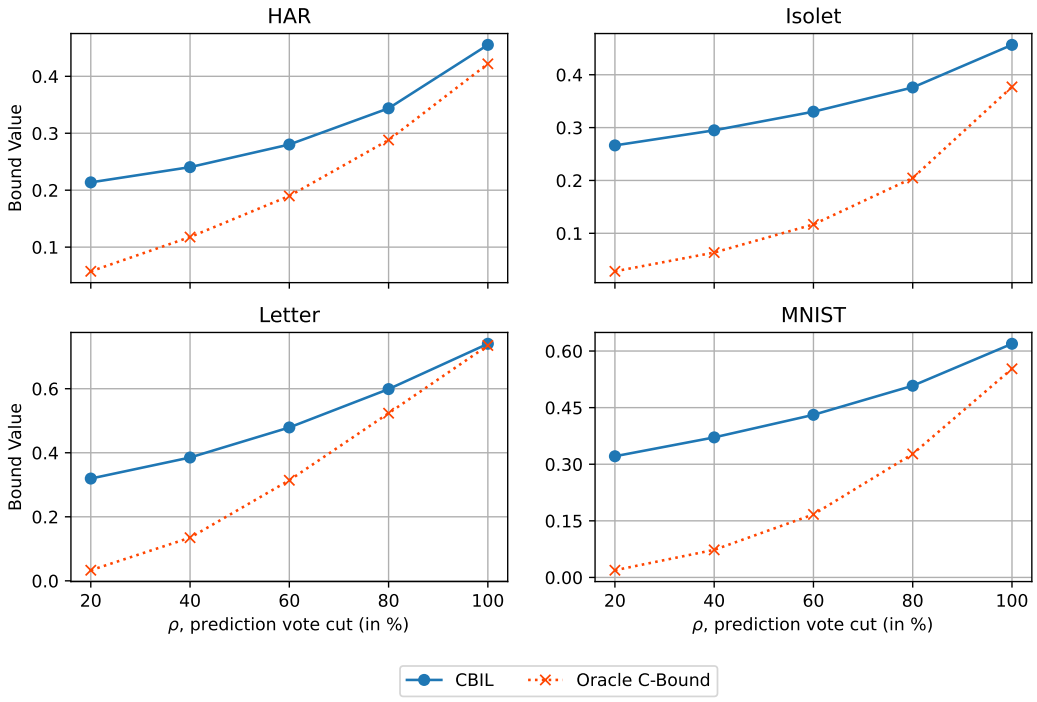

Figure 3: (CBIL) and Oracle C-Bound when varying the number of pseudo-labels on 4 data sets. We keep the most confident one (with respect to prediction vote) from 20% to 100%.

becomes more evident that a mislabeling error model needs to be considered, since otherwise the pseudo-label-estimated bound would be overly optimistic.

7. Conclusion and Future Work

In this paper, we proposed a new probabilistic framework for analysis of the majority vote classifier in the case of semi-supervised multi-class classification. At first, we derived a bound for the transductive conditional risk of the majority vote classifier. This probabilistic bound is based on the distribution of the class vote over unlabeled examples for a predicted class. We deduced corresponding bounds on the confusion matrix norm and the error rate as a corollary and determined when the bounds are tight. Based on this result, we proposed a multi-class self-training algorithm where the threshold for selecting unlabeled data to pseudo-label is automatically found from minimization of the transductive bound on the majority vote error rate. From the numerical results, it came out that the self-training algorithm is sensitive to the supervised performance of the base classifier, but it passes well the scale and significantly outperforms the case when the threshold is manually fixed. However, when the classifier is trained on the pseudo-labeled examples, the training labels can be erroneous, so we proposed a mislabeling error model to take explicitly into

account these mislabeling errors. We established the connection between the true and the imperfect output and consequently extended the C-bound to the case of imperfect labeling, and derived a PAC-Bayesian Theorem for controlling the sample effect. We illustrated the influence of the mislabeling error model on the bound's value on several real data sets. With the proposed bound we initiated a new direction to study the effect of pseudo-labeled data on the error of the majority vote classifier.

This work raises several open practical and theoretical questions. Firstly, the proposed self-training policy has been experimentally validated when it is coupled with the random forest, but it would be interesting to test with other classifiers, e.g., with deep learning methods. However, we should note that modern neural networks are not always well calibrated, and examples can be misclassified with a high prediction vote (Guo et al., 2017). This is a significant limitation in our case, since we make the assumption that the classifier makes its mistakes on examples with low prediction votes, which is used for the bound's approximation. Possible solutions to overcome this issue could be negative entropy regularization (Zou et al., 2019) or ensemble diversification (Odonnat et al., 2023), but a theoretical study of the problem is also valuable.

Secondly, further analysis of the learning model learned on pseudo-labels is perplexing due to the so-called confirmation bias, which arises from classifier's overconfidence to its initial decisions that could be erroneous. When self-training assigns highly confident but wrong predictions to unlabeled examples, the hypotheses tend to have a small disagreement on the unlabeled set, so the votes are no more adequate for measuring prediction confidence. A correct estimation of mislabeling probabilities or changing the self-training procedure are possible solutions.

Thirdly, (CBIL) requires in practice the estimation of the mislabeling matrix, which is a complex problem, but an active field of study (Natarajan et al., 2013). Most of these studies tackle this problem from an algorithmic point of view: for example, in the semi-supervised setting, Krithara et al. (2008) learn the mislabeling matrix together with the classifier parameters through the classifier likelihood maximization for document classification; in the supervised setting, a common approach is to detect anchor points whose labels are surely true (Scott, 2015). A potential idea would be to transfer this idea to the semi-supervised case in order to detect the anchor points in the unlabeled set and use them together with the labeled set for correct estimation of the noise in pseudo-labels; this may require additional assumptions such as the existence of clusters (Rigollet, 2007; Maximov et al., 2018) or manifold structure (Belkin and Niyogi, 2004).

Fourthly, in this paper, we derived (CBIL) assuming that the pseudo-labels come independently from the training data, which requires us to have a hold-out set or a pre-trained model. Then, a promising direction is to extend this result to a broader scenario when the noisy training examples are interdependent, which would fit the generic self-training framework and give the light on the learnability of this algorithm. One way to achieve it would be to model the dependency using β-mixing processes following the work on learning with interdependent data (Mohri and Rostamizadeh, 2010; Amini and Usunier, 2015).

Finally, we also point out possible applications of (CBIL). At first, the bound can be used for model selection tasks such as semi-supervised wrapper feature selection (Sheikhpour et al., 2017; Feofanov et al., 2022). Since minimization of the C-bound implies simultaneously maximization of the margin mean and minimization of the margin variance, (CBIL)

would guide a feature selection algorithm to choose an optimal feature subset based on the labeled and the pseudo-labeled sets. Next, (CBIL) can be used as a criterion to learn the posterior Q in the semi-supervised setting. This issue is actively studied in the supervised context, e.g., Roy et al. (2016); Bauvin et al. (2020) have been developed the boosting-based C-bound optimization algorithms. It should be noticed that for these two applications, the main objective is to rank models so that the best model has the minimal error on the unlabeled set. Hence, the bound analysis goes beyond the classical question of tightness: the tightest bound does not always imply the minimal error, and a bound relaxation can have a positive effect (see Appendix C.3).

Acknowledgments

We acknowledge support for the project IRIS from the IRS Grant (Grenoble INP, University Grenoble Alpes).

Appendix A. Tools for Section 4

In this section, we provide complete proofs of all theoretical results presented in Section 4.

A.1 Tools for Theorem 4.1

The proof of Theorem 4.1 relies on Lemma A.1 and Lemma A.2. First, we prove Lemma A.1 that establishes a connection between the joint majority vote and the Gibbs conditional risks.

Lemma A.1 *For $\hat{y} \in \{1, \dots, K\}$, let $\Gamma_{\hat{y}} = \{\gamma_{\hat{y}} \in [0,1] | \ \exists\ \mathbf{x} \in \mathrm{X}_{\mathcal{U}} : \gamma_{\hat{y}} = v_Q(\mathbf{x}, \hat{y})\}$ be the set of unique votes for the unlabeled examples to the class $\hat{y}$. Let enumerate its elements such that they form an ascending order:*

$$\gamma_{\hat{y}}^{(1)} \leq \gamma_{\hat{y}}^{(2)} \leq \cdots \leq \gamma_{\hat{y}}^{(N_{\hat{y}})},$$

where $N_{\hat{y}} := |\Gamma_{\hat{y}}|$. Denote $b_{y,\hat{y}}^{(t)} := \frac{1}{u_y} \sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} P(Y = y | X = \mathbf{x}) \mathbb{I}(B_Q(\mathbf{x}) = \hat{y}) \mathbb{I}(v_Q(\mathbf{x}, \hat{y}) = \gamma_{\hat{y}}^{(t)})$. Then, for all $(y, \hat{y}) \in \{1, \dots, K\}^2$:

$$R_{\mathcal{U}}(G_Q, y, \hat{y}) \geq K_{y,\hat{y}} := \sum_{t=1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)} \gamma_{\hat{y}}^{(t)}, \tag{25}$$

$$R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q, y, \hat{y}) = \sum_{t=m_{\hat{y}}+1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)}, \tag{26}$$

where $m_{\hat{y}} = \max\{t | \gamma_{\hat{y}}^{(t)} < \theta_{\hat{y}}\}$ with $\max(\emptyset) = 0$ by convention.

Proof First, we obtain Eq. (25):

$$\begin{aligned}
R_{\mathcal{U}}(G_Q, y, \hat{y}) &= \frac{1}{u_y}\mathbb{E}_{h\sim Q}\sum_{\mathbf{x}\in X_{\mathcal{U}}} P(Y=y|X=\mathbf{x})\mathbb{I}(h(\mathbf{x})=\hat{y}) \\
&= \frac{1}{u_y}\sum_{\mathbf{x}\in X_{\mathcal{U}}} P(Y=y|X=\mathbf{x})v_Q(\mathbf{x},\hat{y}) \\
&\geq \frac{1}{u_y}\sum_{\mathbf{x}\in X_{\mathcal{U}}} P(Y=y|X=\mathbf{x})v_Q(\mathbf{x},\hat{y})\mathbb{I}(B_Q(\mathbf{x})=\hat{y}) \\
&= \frac{1}{u_y}\sum_{t=1}^{N_{\hat{y}}}\sum_{\mathbf{x}\in X_{\mathcal{U}}}\left(P(Y=y|X=\mathbf{x})\mathbb{I}(B_Q(\mathbf{x})=\hat{y})\mathbb{I}(v_Q(\mathbf{x},\hat{y})=\gamma_{\hat{y}}^{(t)})\right)\gamma_{\hat{y}}^{(t)} \\
&= \sum_{t=1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)}\gamma_{\hat{y}}^{(t)}.
\end{aligned}$$

Then, we deduce Eq. (26):

$$\begin{aligned}
R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, y, \hat{y}) &= \frac{1}{u_y}\sum_{\mathbf{x}\in X_{\mathcal{U}}} P(Y=y|X=\mathbf{x})\mathbb{I}(B_Q(\mathbf{x})=\hat{y})\mathbb{I}(v_Q(\mathbf{x},\hat{y})\geq\theta_{\hat{y}}) \\
&= \frac{1}{u_y}\sum_{t=1}^{N_{\hat{y}}}\sum_{\mathbf{x}\in X_{\mathcal{U}}} P(Y=y|X=\mathbf{x})\mathbb{I}(B_Q(\mathbf{x})=\hat{y})\mathbb{I}(v_Q(\mathbf{x},\hat{y})=\gamma_{\hat{y}}^{(t)})\mathbb{I}(\gamma_{\hat{y}}^{(t)}\geq\theta_{\hat{y}}) \\
&= \frac{1}{u_y}\sum_{t=m_{\hat{y}}+1}^{N_{\hat{y}}}\sum_{\mathbf{x}\in X_{\mathcal{U}}} P(Y=y|X=\mathbf{x})\mathbb{I}(B_Q(\mathbf{x})=\hat{y})\mathbb{I}(v_Q(\mathbf{x},\hat{y})=\gamma_{\hat{y}}^{(t)}) \\
&= \sum_{t=m_{\hat{y}}+1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)}.
\end{aligned}$$

■

The next lemma formulates a linear program and shows that its solution is the greatest feasible one in the lexicographic order.

Lemma A.2 (Amini et al., 2008, Lemma 4) *Let $(g_i)_{i\in\{1,\ldots,N\}}$ be such that $0 < g_1 < \cdots < g_N \leq 1$. Consider also $(p_i)_{i\in\{1,\ldots,N\}}$ with $p_i \geq 0$, $Z \geq 0$, $m \in \{1,\ldots,N\}$. Then, the optimal solution of the linear program:*

$$\begin{cases}
\max_{\mathbf{q}:=(q_1,\ldots,q_N)} F(\mathbf{q}) := \max_{q_1,\ldots,q_N}\sum_{i=m+1}^{N} q_i \\
0 \leq q_i \leq p_i \quad \forall i \in \{1,\ldots,N\} \\
\sum_{i=1}^{N} q_i g_i \leq Z
\end{cases}$$

will be $\mathbf{q}^* := (q_1^*,\ldots,q_N^*)$ defined as $q_i^* = \min\left(p_i, \left\lfloor\frac{Z-\sum_{j<i} q_j^* g_j}{g_i}\right\rfloor_+\right)\mathbb{I}(i > m)$ for all $i \in \{1,\ldots,N\}$, where, the sign $\lfloor\cdot\rfloor_+$ denotes the positive part of a number, $\lfloor x\rfloor_+ = x\cdot\mathbb{I}(x>0)$.

Proof of Theorem 4.1 We would like to find an upper bound for the joint Q-weighted majority vote conditional risk. Hence, for all $(y, \hat{y}) \in \{1, \dots, K\}^2$, for all $\boldsymbol{\theta} \in [0,1]^K$, we consider the case when the mistake is maximized. Then, using Lemma A.1:

$$R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q, y, \hat{y}) = \sum_{t=m_{\hat{y}}+1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)} \leq \max_{b_{y,\hat{y}}^{(1)}, \dots, b_{y,\hat{y}}^{(N_{\hat{y}})}} \sum_{t=m_{\hat{y}}+1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)}, \tag{27}$$

with $m_{\hat{y}} = \max\{t | \gamma_{\hat{y}}^{(t)} < \theta_{\hat{y}}\} \mathbb{I}(\{t | \gamma_{\hat{y}}^{(t)} < \theta_{\hat{y}}\} \neq \emptyset)$.

Let $\bar{b}_{y,\hat{y}}^{(t)} = \sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} P(Y = y | X = \mathbf{x}) \mathbb{I}(v_Q(\mathbf{x}, \hat{y}) = \gamma_{\hat{y}}^{(t)}) / u_y$. Then, it can be noticed that $0 \leq b_{y,\hat{y}}^{(t)} \leq \bar{b}_{y,\hat{y}}^{(t)}$. Remember that $K_{y,\hat{y}}$ can also be written as $\sum_{t=1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)} \gamma_{\hat{y}}^{(t)}$. Hence the bound defined by Eq. (27) should satisfy the following linear program :

$$\begin{aligned} &\max_{b_{y,\hat{y}}^{(1)}, \dots, b_{y,\hat{y}}^{(N_{\hat{y}})}} \sum_{t=m_{\hat{y}}+1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)} \\ &\text{s.t.} \quad \forall t,\ 0 \leq b_{y,\hat{y}}^{(t)} \leq \bar{b}_{y,\hat{y}}^{(t)} \text{ and } \sum_{t=1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)} \gamma_{\hat{y}}^{(t)} = K_{y,\hat{y}}. \end{aligned} \tag{28}$$

Applying Lemma A.2, the solution of (28) is found as

$$b_{y,\hat{y}}^{(t)} = \min \left(\bar{b}_{y,\hat{y}}^{(t)}, \left\lfloor \frac{1}{\gamma_{\hat{y}}^{(t)}} (K_{y,\hat{y}} - \sum_{m_{\hat{y}} < w < t} \gamma_{\hat{y}}^{(w)} \bar{b}_{y,\hat{y}}^{(w)}) \right\rfloor_+ \right) \mathbb{I}(t \leq m_{\hat{y}}). \tag{29}$$

Further, we can notice that, for all $(y, \hat{y}) \in \{1, \dots, K\}^2$,

$$\sum_{m_{\hat{y}} < w < t} \gamma_{\hat{y}}^{(w)} \bar{b}_{y,\hat{y}}^{(w)} = V_{y,\hat{y}}^{(\leq, <)}(\theta_{\hat{y}}, \gamma_{\hat{y}}^{(t)}).$$

Let $p = \max\{t | K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(\theta_{\hat{y}}, \gamma_{\hat{y}}^{(t)}) > 0\}$. Then, Eq. (29) can be re-written as follows:

$$b_{y,\hat{y}}^{(t)} = \begin{cases} 0 & t \leq m_{\hat{y}} \\ \bar{b}_{y,\hat{y}}^{(t)} & m_{\hat{y}} + 1 \leq t < p \\ \frac{1}{\gamma_{\hat{y}}^{(p)}} \left(K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(\theta_{\hat{y}}, \gamma_{\hat{y}}^{(p)}) \right) & t = p \\ 0 & t > p. \end{cases} \tag{30}$$

Notice that $\sum_{t=m_{\hat{y}}+1}^{p-1} \bar{b}_{y,\hat{y}}^{(t)} = I_{y,\hat{y}}^{(\leq,<)}(\theta_{\hat{y}}, \gamma_{\hat{y}}^{(p)})$. Using this fact as well as Eq. (30), we infer:

$$R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q, y, \hat{y}) \leq I_{y,\hat{y}}^{(\leq,<)}(\theta_{\hat{y}}, \gamma_{\hat{y}}^{(p)}) + \frac{1}{\gamma_{\hat{y}}^{(p)}} \left(K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(\theta_{\hat{y}}, \gamma_{\hat{y}}^{(p)}) \right).$$

Consider the following function:

$$\gamma \mapsto U_{y,\hat{y}}(\gamma) := I_{y,\hat{y}}^{(\leq,<)}(\theta_{\hat{y}}, \gamma) + \frac{1}{\gamma} \left\lfloor K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(\theta_{\hat{y}}, \gamma) \right\rfloor_+.$$

To prove the theorem, it remains to verify that, for all $(y, \hat{y}) \in \{1, \ldots, K\}^2$, for all $\gamma \in [\theta_{\hat{y}}, 1]$, $U_{y,\hat{y}}(\gamma_{\hat{y}}^{(p)}) \leq U_{y,\hat{y}}(\gamma)$. For this, consider $\gamma_{\hat{y}}^{(w)}$ with $w \in \{1, \ldots, N_{\hat{y}}\}$.

If $w > p$, then $U_{y,\hat{y}}(\gamma_{\hat{y}}^{(p)}) \leq I_{y,\hat{y}}^{(\leq,<)}(\theta_{\hat{y}}, \gamma_{\hat{y}}^{(p+1)}) \leq U_{y,\hat{y}}(\gamma_{\hat{y}}^{(w)})$.

If $w < p$, then

$$\begin{aligned} U_{y,\hat{y}}(\gamma_{\hat{y}}^{(p)}) - U_{y,\hat{y}}(\gamma_{\hat{y}}^{(w)}) &= \sum_{t=w}^{p} b_{y,\hat{y}}^{(t)} - \frac{1}{\gamma_{\hat{y}}^{(w)}} \left(K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(\theta_{\hat{y}}, \gamma_{\hat{y}}^{(w)}) \right) \\ &= \sum_{t=w}^{p} b_{y,\hat{y}}^{(t)} - \frac{1}{\gamma_{\hat{y}}^{(w)}} \left(\sum_{t=m_{\hat{y}}+1}^{p} b_{y,\hat{y}}^{(t)} \gamma_{\hat{y}}^{(t)} - \sum_{t=m_{\hat{y}}+1}^{w-1} \gamma_{\hat{y}}^{(t)} b_{y,\hat{y}}^{(t)} \right) \\ &= \frac{1}{\gamma_{\hat{y}}^{(w)}} \left(\sum_{t=w}^{p} b_{y,\hat{y}}^{(t)} \gamma_{\hat{y}}^{(w)} - \sum_{t=w}^{p} b_{y,\hat{y}}^{(t)} \gamma_{\hat{y}}^{(t)} \right) \leq 0, \end{aligned}$$

which completes the proof of Theorem 4.1. ■

A.2 Tools for Corollary 4.2

The following lemma establishes a connection between the joint error rate and the joint confusion matrix of the Q-weighted majority vote classifier.

Lemma A.3 *Let B_Q be the majority vote classifier. Given a vector $\boldsymbol{\theta} \in [0,1]^K$, for $\mathbf{p} := \{u_y/u\}_{y=1}^K$, where $u_y = \sum_{\mathbf{x} \in \mathrm{X}_\mathcal{U}} P(Y = y | X = \mathbf{x})$, we have:*

$$R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q) = \left\| \left(\mathbf{C}_{B_Q}^{\mathcal{U} \wedge \boldsymbol{\theta}} \right)^\intercal \mathbf{p} \right\|_1 . \tag{31}$$

Proof To prove Eq. (31), combine the definition of transductive joint Q-weighted majority vote conditional risk given in Eq. (5) and Eq. (7) as follows:

$$\begin{aligned} R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q) &= \frac{1}{u} \sum_{y=1}^{K} \sum_{\substack{\hat{y}=1 \\ \hat{y} \neq y}}^{K} \sum_{\mathbf{x} \in \mathrm{X}_\mathcal{U}} P(Y = y | X = \mathbf{x}) \mathbb{I}(B_Q(\mathbf{x}) = \hat{y}) \mathbb{I}(v_Q(\mathbf{x}, \hat{y}) \geq \theta_{\hat{y}}) \\ &= \sum_{y=1}^{K} \frac{u_y}{u} \sum_{\substack{\hat{y}=1 \\ \hat{y} \neq y}}^{K} R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q, y, \hat{y}) = \left\| \left(\mathbf{C}_{B_Q}^{\mathcal{U} \wedge \boldsymbol{\theta}} \right)^\intercal \mathbf{p} \right\|_1 . \end{aligned}$$

■

Proof of Corollary 4.2 The confusion matrix $\mathbf{C}_{B_Q}^{\mathcal{U} \wedge \boldsymbol{\theta}}$ is always non-negative, and from Theorem 4.1, each of its entries is smaller than the corresponding entry of $\mathbf{U}_{\boldsymbol{\theta}}$. Hence, from the property of spectral norm for two positive matrices $\mathbf{A}$ and $\mathbf{B}$:

$$\mathbf{0}_{K,K} \preceq \mathbf{A} \preceq \mathbf{B} \Rightarrow \|\mathbf{A}\| \leq \|\mathbf{B}\|,$$

where $\mathbf{A} \preceq \mathbf{B}$ denotes that each element of $\mathbf{A}$ is smaller than the corresponding element of $\mathbf{B}$, we deduce Eq. (10).

With the same computations, we observe the following inequality:

$$\left(\mathbf{C}_{B_Q}^{\mathcal{U}\wedge\boldsymbol{\theta}}\right)^\intercal \mathbf{p} \leq \mathbf{U}_{\boldsymbol{\theta}}^\intercal \mathbf{p}.$$

Elements of the left vector are non-negative. Hence the inequality holds for the ℓ_1-norm, and taking into account Lemma A.3 we infer:

$$R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q) = \left\| \left(\mathbf{C}_{B_Q}^{\mathcal{U}\wedge\boldsymbol{\theta}}\right)^\intercal \mathbf{p} \right\|_1 \leq \left\| \mathbf{U}_{\boldsymbol{\theta}}^\intercal \mathbf{p} \right\|_1 .$$

■

A.3 Tools for Proposition 4.3

Before proving Proposition 4.3, we formulate the following lemma.

Lemma A.4 *For all $\mathbf{x} \in \mathrm{X}_{\mathcal{U}}$, for all $(y, \hat{y}) \in \{1, \dots, K\}^2$, the following inequality holds:*

$$R_{\mathcal{U}}(B_Q, y, \hat{y}) \geq \frac{1}{u_y} \sum_{\mathbf{x}\in \mathrm{X}_{\mathcal{U}}} P(Y=y|X=\mathbf{x}) \mathbb{I}(B_Q(\mathbf{x})=\hat{y}) \mathbb{I}(v_Q(\mathbf{x},\hat{y}) < \gamma^*_{y,\hat{y}}) + \frac{1}{\gamma^*_{y,\hat{y}}} \left\lfloor \lfloor K_{y,\hat{y}} - V^{(\leq,<)}_{y,\hat{y}}(0, \gamma^*_{y,\hat{y}}) \rfloor_+ - r_{y,\hat{y}} \right\rfloor_+ + r_{y,\hat{y}}, \quad (32)$$

where $\gamma^*_{y,\hat{y}} := \max \Gamma^\tau_{y,\hat{y}}$.

Proof Following notations of Section A.1, we denote $\gamma^*_{y,\hat{y}}$ as $\gamma^{(p)}_{\hat{y}}$. According to Lemma A.1, $K_{y,\hat{y}} = \sum_{t=1}^{N_{\hat{y}}} b^{(t)}_{y,\hat{y}} \gamma^{(t)}_{\hat{y}}$, where $b^{(t)}_{y,\hat{y}} := \frac{1}{u_y} \sum_{\mathbf{x}\in \mathrm{X}_{\mathcal{U}}} P(Y=y|X=\mathbf{x}) \mathbb{I}(B_Q(\mathbf{x})=\hat{y}) \mathbb{I}(v_Q(\mathbf{x},\hat{y})=\gamma^{(t)}_{\hat{y}})$. We can express $b^{(p)}_{y,\hat{y}}$ in the following way:

$$b^{(p)}_{y,\hat{y}} = \frac{K_{y,\hat{y}} - \sum_{t=1}^{p-1} b^{(t)}_{y,\hat{y}} \gamma^{(t)}_{\hat{y}} - \sum_{t=p+1}^{N_{\hat{y}}} b^{(t)}_{y,\hat{y}} \gamma^{(t)}_{\hat{y}}}{\gamma^{(p)}_{\hat{y}}} = \frac{K_{y,\hat{y}} - \sum_{t=1}^{p-1} b^{(t)}_{y,\hat{y}} \gamma^{(t)}_{\hat{y}} - r_{y,\hat{y}}}{\gamma^{(p)}_{\hat{y}}}.$$

Remind $\bar{b}^{(t)}_{y,\hat{y}} = \frac{1}{u_y} \sum_{\mathbf{x}\in \mathrm{X}_{\mathcal{U}}} P(Y=y|X=\mathbf{x}) \mathbb{I}(v_Q(\mathbf{x},\hat{y})=\gamma^{(t)}_{\hat{y}})$. From this we derive the following:

$$-\sum_{t=1}^{p-1} b^{(t)}_{y,\hat{y}} \gamma^{(t)}_{\hat{y}} \geq -\sum_{t=1}^{p-1} \bar{b}^{(t)}_{y,\hat{y}} \gamma^{(t)}_{\hat{y}} = -V^{(\leq,<)}_{y,\hat{y}}(0, \gamma^{(p)}_{\hat{y}}) = -V^{(\leq,<)}_{y,\hat{y}}(0, \gamma^*_{y,\hat{y}}).$$

Taking into account this as well as $b^{(p)}_{y,\hat{y}} \geq 0$, we deduce a lower bound for $b^{(p)}_{y,\hat{y}}$:

$$b^{(p)}_{y,\hat{y}} \geq \frac{1}{\gamma^*_{y,\hat{y}}} \lfloor K_{y,\hat{y}} - V^{(\leq,<)}_{y,\hat{y}}(0, \gamma^*_{y,\hat{y}}) - r_{y,\hat{y}} \rfloor_+ = \frac{1}{\gamma^*_{y,\hat{y}}} \left\lfloor \lfloor K_{y,\hat{y}} - V^{(\leq,<)}_{y,\hat{y}}(0, \gamma^*_{y,\hat{y}}) \rfloor_+ - r_{y,\hat{y}} \right\rfloor_+ . \quad (33)$$

Also, taking into account Lemma A.1, one can notice that:

$$\begin{aligned} R_{\mathcal{U}}(B_Q, y, \hat{y}) &= \sum_{t=1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)} = \sum_{t=1}^{p-1} b_{y,\hat{y}}^{(t)} + b_{y,\hat{y}}^{(p)} + \sum_{t=p+1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)} \\ &\geq \frac{1}{u_y} \sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} P(Y = y | X = \mathbf{x}) \mathbb{I}(B_Q(\mathbf{x}) = \hat{y}) \mathbb{I}(v_Q(\mathbf{x}, \hat{y}) < \gamma_{y,\hat{y}}^*) + b_{y,\hat{y}}^{(p)} + r_{y,\hat{y}}, \end{aligned} \tag{34}$$

since $\sum_{t=p+1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)} \geq \sum_{t=p+1}^{N_{\hat{y}}} b_{y,\hat{y}}^{(t)} \gamma_{\hat{y}}^{(t)}$. Combining Eq. (33) and Eq. (34) we infer Eq. (32):

$$\begin{aligned} R_{\mathcal{U}}(B_Q, y, \hat{y}) \geq \frac{1}{u_y} \sum_{\mathbf{x} \in \mathrm{X}_{\mathcal{U}}} & P(Y = y | X = \mathbf{x}) \mathbb{I}(B_Q(\mathbf{x}) = \hat{y}) \mathbb{I}(v_Q(\mathbf{x}, \hat{y}) < \gamma_{y,\hat{y}}^*) \\ & + \frac{1}{\gamma_{y,\hat{y}}^*} \left\lfloor \lfloor K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(0, \gamma_{y,\hat{y}}^*) \rfloor_+ - r_{y,\hat{y}} \right\rfloor_+ + r_{y,\hat{y}}. \end{aligned}$$

■

Proof of Proposition 4.3 Taking into account Eq. (32) and Eq. (12) we deduce the following:

$$R_{\mathcal{U}}(B_Q, y, \hat{y}) \geq C\, I_{y,\hat{y}}^{(\leq,<)}(0, \gamma_{y,\hat{y}}^*) + \frac{1}{\gamma_{y,\hat{y}}^*} \left\lfloor \lfloor K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(0, \gamma_{y,\hat{y}}^*) \rfloor_+ - r_{y,\hat{y}} \right\rfloor_+ + r_{y,\hat{y}}. \tag{35}$$

By definition of $\mathbf{U}_{\mathbf{0}_K}$ we have, for all $(y, \hat{y}) \in \{1, \ldots, K\}^2$,

$$[\mathbf{U}_{\mathbf{0}_K}]_{y,\hat{y}} \leq I_{y,\hat{y}}^{(\leq,<)}(0, \gamma_{y,\hat{y}}^*) + \frac{1}{\gamma_{y,\hat{y}}^*} \left\lfloor K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(0, \gamma_{y,\hat{y}}^*) \right\rfloor_+. \tag{36}$$

Subtracting Eq. (35) from Eq. (36) we obtain:

$$\begin{aligned} [\mathbf{U}_{\mathbf{0}_K}]_{y,\hat{y}} - R_{\mathcal{U}}(B_Q, y, \hat{y}) \leq & (1 - C) I_{y,\hat{y}}^{(\leq,<)}(0, \gamma_{y,\hat{y}}^*) \\ & + \frac{1}{\gamma_{y,\hat{y}}^*} \left(\left\lfloor K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(0, \gamma_{y,\hat{y}}^*) \right\rfloor_+ - \left\lfloor \lfloor K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(0, \gamma_{y,\hat{y}}^*) \rfloor_+ - r_{y,\hat{y}} \right\rfloor_+ \right) - r_{y,\hat{y}}. \end{aligned}$$

We can notice that for all $a, b \in \mathbb{R}^+ :\ b - \lfloor b - a \rfloor_+ \leq a$. Then, we have:

$$[\mathbf{U}_{\mathbf{0}_K}]_{y,\hat{y}} - R_{\mathcal{U}}(B_Q, y, \hat{y}) \leq (1 - C) I_{y,\hat{y}}^{(\leq,<)}(0, \gamma_{y,\hat{y}}^*) + r_{y,\hat{y}} \left(\frac{1}{\gamma_{y,\hat{y}}^*} - 1 \right). \tag{37}$$

Also, from Eq. (35) one can derive:

$$\begin{aligned} I_{y,\hat{y}}^{(\leq,<)}(0, \gamma_{y,\hat{y}}^*) &\leq \frac{1}{C} \left(R_{\mathcal{U}}(B_Q, y, \hat{y}) - \frac{1}{\gamma_{y,\hat{y}}^*} \left\lfloor \lfloor K_{y,\hat{y}} - V_{y,\hat{y}}^{(\leq,<)}(0, \gamma_{y,\hat{y}}^*) \rfloor_+ - r_{y,\hat{y}} \right\rfloor_+ - r_{y,\hat{y}} \right) \\ &\leq \frac{R_{\mathcal{U}}(B_Q, y, \hat{y})}{C}. \end{aligned} \tag{38}$$

Taking into account Eq. (37) and Eq. (38), we infer:

$$[\mathbf{U}\mathbf{0}_K]_{y,\hat{y}} - R_{\mathcal{U}}(B_Q, y, \hat{y}) \leq \frac{1-C}{C} R_{\mathcal{U}}(B_Q, y, \hat{y}) + r_{y,\hat{y}} \left(\frac{1}{\gamma^*_{y,\hat{y}}} - 1 \right).$$

■

Appendix B. Tools for Section 5

In this section, we detail the proofs of Theorem 5.1 and Theorem 5.4.

B.1 Tools for Theorem 5.1

Theorem 5.1 is based on the well-known Cantelli-Chebyshev inequality (for example, see Boucheron et al., 2013, Ex. 2.3).

Lemma B.1 (Cantelli-Chebyshev inequality) *Let Z be a random variable with the mean μ and the variance σ^2. Then, for every $a > 0$, we have:*

$$P(Z \leq \mu - a) \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

B.2 Tools for Theorem 5.4

We divide the proof of Theorem 5.4 into several parts formulated as Proposition B.4, B.5, B.7 and B.9. The proofs of all propositions are based on the following two lemmas.

Lemma B.2 (Pinsker's Inequality, Boucheron et al., 2013, Theorem 4.19) *For all $p_1, p_2 \in [0,1]^2$,*

$$\begin{aligned} &2(p_2 - p_1)^2 \leq kl(p_2||p_1) \\ &kl(p_2||p_1) := p_2 \ln \frac{p_2}{p_1} + (1-p_2) \ln \frac{1-p_2}{1-p_1} = KL(P_2 \,\|\, P_1), \end{aligned}$$

where P_2 and P_1 are Bernoulli distributions with parameters p_2 and p_1 respectively.

Lemma B.3 (Maurer, 2004, Theorem 1, and Germain et al., 2015, Lemma 19) *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector, whose components $X_i \in [0,1]$ are i.i.d. with an expectation μ. Let $\mathbf{X}' = (X'_1, \ldots, X'_n)$ denotes a random vector, where each X'_i is the unique Bernoulli random variable of the corresponding X_i: $P(X'_i = 1) = \mathbb{E}X'_i = \mathbb{E}X_i = \mu$, for all $i \in \{1, \ldots, n\}$. Then,*

$$\mathbb{E}\left[e^{nKL(\bar{X} \,\|\, \mu)} \right] \leq \mathbb{E}\left[e^{nKL(\bar{X}' \,\|\, \mu)} \right] \leq 2\sqrt{n},$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and $\bar{X}' = \frac{1}{n} \sum_{i=1}^n X'_i$.

B.2.1 Bounds for the Mislabeling Matrix' Entries

We remind that the imperfection is summarized through the mislabeling matrix $\mathbf{P} = (p_{\hat{y},y})_{1\leq\hat{y},y\leq K}$ with

$$p_{\hat{y},y} := P(\hat{Y} = \hat{y}|Y = y) \quad \text{for all } (\hat{y}, y) \in \{1, \ldots, K\}^2$$

such that $\sum_{\hat{y}=1}^{K} p_{\hat{y},y} = 1$. Also, recall that $\alpha(\mathbf{x}) = p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})}$ and $\delta(\mathbf{x}) := p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})} - \max_{y\in\{1,\ldots,K\}\setminus\{B_Q(\mathbf{x})\}} p_{B_Q(\mathbf{x}),y}$.

Proposition B.4 *Let $\mathbf{P}$ be the mislabeling matrix, and assume that $p_{\hat{y},\hat{y}} > p_{\hat{y},y},\ \forall(\hat{y}, y) \in \{1, \ldots, K\}^2$. For any $\epsilon \in (0, 1]$, with probability $1 - \epsilon$ over the choice of the sample $S = \{\mathbf{x}_i\}_{i=1}^{n_S}$ with the true class labels $\{y_i\}_{i=1}^{n_S}$, for all $(\hat{y}, y) \in \{1, \ldots, K\}^2$, for all $\mathbf{x} \in \mathcal{X}$,*

$$\hat{p}_{\hat{y},y} - \iota(n_{S,y}) \leq p_{\hat{y},y} \leq \hat{p}_{\hat{y},y} + \iota(n_{S,y}), \tag{39}$$

$$\alpha(\mathbf{x}) \leq \tilde{\alpha}(\mathbf{x}) := \hat{\alpha}(\mathbf{x}) + \iota(n_{S,k_\mathbf{x}}), \tag{40}$$

$$\frac{1}{\delta(\mathbf{x})} \leq \frac{1}{\tilde{\delta}(\mathbf{x})} := \frac{1}{\hat{\delta}(\mathbf{x}) - \iota(n_{S,k_\mathbf{x}}) - \iota(n_{S,j_\mathbf{x}})}, \quad \text{if } \hat{\delta}(\mathbf{x}) \geq \iota(n_{S,k_\mathbf{x}}) + \iota(n_{S,j_\mathbf{x}}), \tag{41}$$

where

- $\iota(n_{S,y}) = \sqrt{\frac{1}{2n_{S,y}} \ln \frac{2\sqrt{n_{S,y}}}{\epsilon}}$,
- $n_{S,y} = \sum_{i=1}^{n_S} \mathbb{I}(y_i = y)/n_S$ *is the proportion of the examples from S with the true class label y,*
- $k_\mathbf{x} := B_Q(\mathbf{x})$, $j_\mathbf{x} := \operatorname{argmin}_{\hat{y}\in\{1,\ldots,K\}\setminus\{k_\mathbf{x}\}} n_{S,\hat{y}}$,
- $\hat{p}_{\hat{y},y}$, $\hat{\alpha}(\mathbf{x})$ *and* $\hat{\delta}(\mathbf{x})$ *are empirical estimates respectively of* $p_{\hat{y},y}$, $\alpha(\mathbf{x})$ *and* $\delta(\mathbf{x})$ *based on the available sample S.*

Proof Let $S_y = \{\mathbf{x}_i \in S|y = y_i\}$ be the subset of the available examples for which the true class is y. Consider the non-negative random variable $\exp\left\{2n_{S,y}(\hat{p}_{\hat{y},y} - p_{\hat{y},y})^2\right\}$. From the Markov's inequality we obtain that the following holds with probability at least $1 - \epsilon$ over $S_y \sim P(\mathbf{X}|Y = y)^{n_{S,y}}$:

$$\exp\left\{2n_{S,y}(\hat{p}_{\hat{y},y} - p_{\hat{y},y})^2\right\} \leq \frac{1}{\epsilon}\mathbb{E}_{S_y} \exp\left\{2n_{S,y}(\hat{p}_{\hat{y},y} - p_{\hat{y},y})^2\right\}. \tag{42}$$

By successively applying Lemma B.2 and Lemma B.3, we deduce that

$$\mathbb{E}_{S_y} \exp\left\{2n_{S,y}(\hat{p}_{\hat{y},y} - p_{\hat{y},y})^2\right\} \leq \mathbb{E}_{S_y} \exp\left\{n_{S,y} \cdot kl(\hat{p}_{\hat{y},y}||p_{\hat{y},y})\right\} \leq 2\sqrt{n_{S,y}}. \tag{43}$$

Combining Eq. (42) and Eq. (43), we infer $2n_{S,y}(\hat{p}_{\hat{y},y} - p_{\hat{y},y})^2 \leq \ln\left(2\sqrt{n_{S,y}}/\epsilon\right)$. Eq. (39) is directly obtained from the last inequality, and hence, we derive also Eq. (40). To prove Eq. (41), let us define

$$q_\mathbf{x} := \operatorname*{argmax}_{y\in\{1,\ldots,K\}\setminus\{B_Q(\mathbf{x})\}} p_{k_\mathbf{x},y}, \qquad \hat{q}_\mathbf{x} := \operatorname*{argmax}_{y\in\{1,\ldots,K\}\setminus\{B_Q(\mathbf{x})\}} \hat{p}_{k_\mathbf{x},y}.$$

Then, we write:

$$\frac{1}{\delta(\mathbf{x})} = \frac{1}{p_{k_\mathbf{x},k_\mathbf{x}} - p_{k_\mathbf{x},q_\mathbf{x}}} \leq \frac{1}{\hat{p}_{k_\mathbf{x},k_\mathbf{x}} - \hat{p}_{k_\mathbf{x},q_\mathbf{x}} - \iota(n_{S,k_\mathbf{x}}) - \iota(n_{S,q_\mathbf{x}})}$$
$$\leq \frac{1}{\hat{p}_{k_\mathbf{x},k_\mathbf{x}} - \hat{p}_{k_\mathbf{x},\hat{q}_\mathbf{x}} - \iota(n_{S,k_\mathbf{x}}) - \iota(n_{S,j_\mathbf{x}})} = \frac{1}{\hat{\delta}(\mathbf{x}) - \iota(n_{S,k_\mathbf{x}}) - \iota(n_{S,j_\mathbf{x}})}.$$

These transitions hold only when the denominator is positive, which is ensured if $\hat{\delta}(\mathbf{x}) \geq \iota(n_{S,k_\mathbf{x}}) + \iota(n_{S,j_\mathbf{x}})$. ■

B.2.2 Lower Bound of the First Moment of the Margin

Proposition B.5 *Given the input $\mathbf{X}$ and the imperfect output Y, let $\hat{M}_Q$ be a random variable defined as $\hat{M}_Q := m_Q(\mathbf{X}, \hat{Y})$ with its first statistical moment denoted by $\mu_1^{\hat{M}_Q}$. Given the result of Proposition B.4, for any set of classifiers $\mathcal{H}$, for any prior distribution Q_0 on $\mathcal{H}$ and any $\epsilon \in (0, 1]$, with a probability at least $1 - \epsilon$ over the choice of the sample $S = \{\mathbf{x}_i\}_{i=1}^{n_S}$, for every posterior distribution Q over $\mathcal{H}$, we have:*

$$\mu_1^{\hat{M}_Q} \geq \bar{\mu}_1^S - J_1 \sqrt{\frac{2}{n_S}\left[KL(Q \,\|\, Q_0) + \ln \frac{2\sqrt{n_S}}{\epsilon}\right]},$$

where

- $\bar{\mu}_1^S = \frac{1}{n_S}\sum_{i=1}^{n_S}(1/\tilde{\delta}(\mathbf{x}_i))\sum_{\hat{y}=1}^{K} m_Q(\mathbf{x}_i, \hat{y})P(\hat{Y} = \hat{y}|\mathbf{X} = \mathbf{x}_i)$ *is the empirical weighted margin mean based on the available sample S,*
- $\tilde{\delta}(\mathbf{x})$ *is defined as in Eq.* (41)*,*
- $J_1 := \max_\mathbf{x} |(1/\tilde{\delta}(\mathbf{x}))\sum_{\hat{y}=1}^{K} m_Q(\mathbf{x}, \hat{y})P(\hat{Y}=\hat{y}|\mathbf{X}=\mathbf{x})|$,
- KL *denotes the Kullback–Leibler divergence.*

Proof Let μ_1^h and $\bar{\mu}_1^{S,h}$ be the random variables such that $\mu_1^{\hat{M}_Q} = \mathbb{E}_{h\sim Q}\mu_1^h$ and $\bar{\mu}_1^S = \mathbb{E}_{h\sim Q}\bar{\mu}_1^{S,h}$. We apply the Markov's inequality to $\mathbb{E}_{h\sim Q_0}\exp\left\{\frac{n_S}{2J_1^2}(\bar{\mu}_1^{S,h} - \mu_1^h)^2\right\}$, that is a non-negative random variable, and obtain that with probability at least $1 - \epsilon$ over $S \sim P(\mathbf{X}, \hat{Y})^{n_S}$:

$$\mathbb{E}_{h\sim Q_0}\exp\left\{\frac{n_S}{2J_1^2}(\bar{\mu}_1^{S,h} - \mu_1^h)^2\right\} \leq \frac{1}{\epsilon}\mathbb{E}_S\mathbb{E}_{h\sim Q_0}\exp\left\{\frac{n_S}{2J_1^2}(\bar{\mu}_1^{S,h} - \mu_1^h)^2\right\}. \tag{44}$$

Since the prior distribution Q_0 over $\mathcal{H}$ is independent on S, we can swap $\mathbb{E}_S$ and $\mathbb{E}_{h\sim Q_0}$. One can notice that

$$\frac{1}{2J_1^2}\left(\bar{\mu}_1^{S,h} - \mu_1^h\right)^2 = 2\left[\frac{1}{2}\left(1 - \frac{\bar{\mu}_1^{S,h}}{J_1}\right) - \frac{1}{2}\left(1 - \frac{\mu_1^h}{J_1}\right)\right]^2,$$

which is the squared of the difference of two random variables that are both $\in [0,1]$. Then, we successively apply Lemma B.2 and Lemma B.3 deriving that:

$$\mathbb{E}_{h\sim Q_0}\mathbb{E}_S \exp\left\{2n_S\left[\frac{1}{2}\left(1-\frac{\bar{\mu}_1^{S,h}}{J_1}\right)-\frac{1}{2}\left(1-\frac{\mu_1^h}{J_1}\right)\right]^2\right\}$$
$$\leq \mathbb{E}_{h\sim Q_0}\mathbb{E}_S \exp\left\{n_S\cdot kl\left[\frac{1}{2}\left(1-\frac{\bar{\mu}_1^{S,h}}{J_1}\right)\middle\|\frac{1}{2}\left(1-\frac{\mu_1^h}{J_1}\right)\right]\right\} \leq \mathbb{E}_{h\sim Q_0} 2\sqrt{n_S} = 2\sqrt{n_S}.$$

We apply this result for Eq. (44), and by taking the natural logarithm from the both sides we obtain that:

$$\ln\left(\mathbb{E}_{h\sim Q_0}\exp\left\{\frac{n_S}{2J_1^2}(\bar{\mu}_1^{S,h}-\mu_1^h)^2\right\}\right) \leq \ln\left(\frac{2\sqrt{n_S}}{\epsilon}\right). \tag{45}$$

Using the change of measure (Lemma B.6) and the Jensen's inequalities, we derive that:

$$\ln\left(\mathbb{E}_{h\sim Q_0}\exp\left\{\frac{n_S}{2J_1^2}(\bar{\mu}_1^{S,h}-\mu_1^h)^2\right\}\right) \geq \mathbb{E}_{h\sim Q}\frac{n_S}{2J_1^2}\left(\bar{\mu}_1^{S,h}-\mu_1^h\right)^2 - KL(Q\,\|\,Q_0)$$
$$\geq \frac{n_S}{2J_1^2}\left(\mathbb{E}_{h\sim Q}[\bar{\mu}_1^{S,h}-\mu_1^h]\right)^2 - KL(Q\,\|\,Q_0).$$

Combining with Eq. (45), we derive:

$$\frac{n_S}{2J_1^2}\left(\bar{\mu}_1^S-\mu_1^{\hat{M}_Q}\right)^2 \leq \ln\left(\frac{2\sqrt{n_S}}{\epsilon}\right) + KL(Q\,\|\,Q_0). \tag{46}$$

The final inequality is directly inferred from Eq. (46).

Lemma B.6 (Change of Measure Inequality, Donsker and Varadhan, 1975) *For any measurable function ϕ defined on the hypothesis space $\mathcal{H}$ and all distributions Q_0, Q on $\mathcal{H}$, the following inequality holds:*

$$\mathbb{E}_{h\sim Q}\phi(h) \leq KL(Q\,\|\,Q_0) + \ln \mathbb{E}_{h\sim Q_0} e^{\phi(h)}.$$

■

B.2.3 Other Required Bounds

Proposition B.7 *Given the input $\mathbf{X}$ and the imperfect output Y, let $\hat{M}_Q$ be a random variable defined as $\hat{M}_Q := m_Q(\mathbf{X}, \hat{Y})$ with its first statistical moment denoted by $\mu_1^{\hat{M}_Q}$. Given the result of Proposition B.4, for any set of classifiers $\mathcal{H}$, for any prior distribution Q_0 on $\mathcal{H}$ and any $\epsilon \in (0,1]$, with a probability at least $1-\epsilon$ over the choice of the sample $S = \{\mathbf{x}_i\}_{i=1}^{n_S}$, for every posterior distribution Q over $\mathcal{H}$*

$$\mu_2^{\hat{M}_Q} \leq \bar{\mu}_2^S + J_2\sqrt{\frac{2}{n_S}\left[2KL(Q\,\|\,Q_0) + \ln\frac{2\sqrt{n_S}}{\epsilon}\right]},$$

where

- $\bar{\mu}_2^S = \frac{1}{n_S}\sum_{i=1}^{n_S}(1/\tilde{\delta}(\mathbf{x}_i))\sum_{\hat{y}=1}^{K}(m_Q(\mathbf{x}_i,\hat{y}))^2 P(\hat{Y}=\hat{y}|\mathbf{X}=\mathbf{x}_i)$ *is the empirical weighted 2nd margin moment based on the available sample* S,
- $\tilde{\delta}(\mathbf{x})$ *is defined as in Eq.* (41),
- $J_2 := \max_{\mathbf{x}} |1/\tilde{\delta}(\mathbf{x})\sum_{\hat{y}=1}^{K}(m_Q(\mathbf{x},\hat{y}))^2 P(\hat{Y}=\hat{y}|\mathbf{X}=\mathbf{x})|$,
- KL *denotes the Kullback–Leibler divergence.*

Proof The proof is similar to the one given for Proposition B.5, but relies on the extension of the change of measure inequality (Lemma B.8).

Lemma B.8 (Laviolette et al., 2017, Lemma 1) *For any set of voters* $\mathcal{H}$*, for any distributions* Q_0, Q *on* $\mathcal{H}$*, and for any measurable function* $\phi : \mathcal{H}\times\mathcal{H} \to \mathbb{R}$*, the following inequality holds:*

$$\mathbb{E}_{(h,h')\sim Q^2}\phi(h,h') \le 2KL(Q\,\|\,Q_0) + \ln \mathbb{E}_{(h,h')\sim Q_0^2} e^{\phi(h,h')}.$$

■

Proposition B.9 *Given the result of Proposition B.4, for any* $\epsilon \in (0,1]$*, with a probability at least* $1-\epsilon$ *over the choice of the sample* $S = \{\mathbf{x}_i\}_{i=1}^{n_S}$*,*

$$\psi_{\mathbf{P}} \le \frac{1}{n_S}\sum_{i=1}^{n_S}\frac{\tilde{\alpha}(\mathbf{x}_i)}{\tilde{\delta}(\mathbf{x}_i)} + J_3\sqrt{\frac{2}{n_S}\ln\frac{2\sqrt{n_S}}{\epsilon}},$$

where $J_3 := \max_{x\in\mathcal{X}}[\tilde{\alpha}(\mathbf{x})]/[\tilde{\delta}(\mathbf{x})]$*, and* $\tilde{\alpha}(\mathbf{x})$ *and* $\tilde{\delta}(\mathbf{x})$ *are defined in Eq.* (40) *and Eq.* (41)*, respectively.*

Proof First, we take into consideration the result of Proposition B.4 and deduce that $\psi_{\mathbf{P}} \le \mathbb{E}_{\mathbf{X}}[\tilde{\alpha}(\mathbf{x}_i)/\tilde{\delta}(\mathbf{x}_i)]$. The rest of proof is similar to those are given for Proposition B.4 and Proposition B.5. ■

Appendix C. Additional Experiments

In this section, we provide some additional experimental results. First, we give more details on estimation of the transductive bound ($\mathrm{TB}_{y,\hat{y}}$) in practice. Then, we compare the runtime of `MSTA` and its competitors. Finally, we empirically analyze the behavior of (CBIL) in the case of relaxation of one of the assumptions.

C.1 Approximation of the Posterior Probabilities for Self-training

In this section, we analyze the behavior of `MSTA` depending on how the transductive bound given by Eq. ($\mathrm{TB}_{y,\hat{y}}$) is evaluated. Since the posterior probabilities for unlabeled data are not known, we have proposed to estimate them as the votes of the base supervised

classifier learned using the labeled data only (Sup. Estimation). This approach has been used in Section 6 for running MSTA. We compare it with another strategy that is to assign $P(Y = y|\mathbf{X} = \mathbf{x}) = 1/K,\ \forall \mathbf{x} \in \mathrm{X}_{\mathcal{U}},\ \forall y \in \{1, \ldots, K\}$. In this case, we consider the worst case when every class is equally probable for each example (Unif. Estimation). Finally, we provide the performance of MSTA when the labels of unlabeled data are given, which means that the transductive bound is truly estimated (Oracle). Table 4 illustrates the performance results. As we can see, the supervised approximation generally outperforms the uniform one (significantly on MNIST). This might be explained by the fact that the supervised votes may give some additional information on the most probable labels for each example. In addition, we have observed that on the last iterations the votes of MSTA tend to be biased, so such posteriors can play a role of regularization. The performance results of the oracle show that better estimation of the posteriors can give an improvement, though not significantly on most of data sets. Note that the performance of the oracle is not perfect, because the true labels are used only for the bound estimation, and the votes are used for pseudo-labeling.

Data set	MSTA		
	Unif. Estimation	Sup. Estimation	Oracle
Vowel	.586 ± .029	.586 ± .026	.599 ± .028
Protein	.773 ± .034	.781 ± .034	.805 ± .036
DNA	.697 ± .079	.702 ± .082	.721 ± .09
Page Blocks	.965 ± .002	.966 ± .002	.966 ± .002
Isolet	.869 ± .015	.875 ± .014	.885 ± .012
HAR	.852 ± .025	.854 ± .026	.856 ± .022
Pendigits	.873 ± .024	.884 ± .022	.892 ± .016
Letter	.716 ± .013	.717 ± .013	.723 ± .012
Fashion	.722 ± .022	.723 ± .023	.728 ± .024
MNIST	.834 ± .016	.857 ± .013	.87 ± .012
SensIT	.722 ± .021	.722 ± .021	.722 ± .021

Table 4: The performance comparison of MSTA depending on how the posterior probabilities are estimated in the evaluation of the transductive bound (Eq. ($\mathrm{TB}_{y,\hat{y}}$)).

C.2 Time

In this section, we present the run-time of all the algorithms empirically compared in Section 6.3. The results are depicted in Table 5. In general, the obtained run-time is coherent with the complexity analysis presented in Section 6.3. LS and QN-S3VM have a very large run-time when they converge slowly, and they are generally slower than the other algorithms. Semi-LDA is fast on the considered data sets, though it may slow down on data of large dimension not considered in this paper.

It can be seen that DAS-RF is slower than the self-training algorithms, which is due to the fact that the classifier is trained on all labeled and unlabeled examples at each iteration. CSTA is the fastest approach since it re-trains the base classifier only 3 times compared to 10 times for FSTA. From our observation, MSTA needs usually around 3-5 iterations to pseudo-label the whole unlabeled set, but it takes more time than CSTA, since it searches at each iteration the threshold by minimizing the conditional Bayes error. We have implemented the search in a single core, but it can be potentially parallelized. Nevertheless, the MSTA still runs fast.

Data set	RF	LS	QN-S3VM	Semi-LDA	DAS-RF	$\text{FSTA}_{\theta=0.7}$	$\text{CSTA}_{\Delta=1/3}$	MSTA
Vowel	1 s	6 s	2 s	3 s	7 s	11 s	2 s	5 s
Protein	1 s	22 s	4 m	5 s	6 s	10 s	2 s	4 s
DNA	1 s	1 m	26 s	1 s	9 s	7 s	3 s	4 s
PageBlocks	1 s	2 m	2 m	14 s	9 s	12 s	3 s	6 s
Isolet	1 s	1 m	1 h	10 s	38 s	16 s	5 s	28 s
HAR	1 s	18 m	32 m	3 s	42 s	23 s	6 s	13 s
Pendigits	1 s	30 m	10 m	37 s	13 s	13 s	3 s	14 s
Letter	1 s	3 h	40 m	1 m	20 s	16 s	5 s	1 m
Fashion	1 s	>4 h	>4 h	1 m	2 m	1 m	29 s	1 m
MNIST	1 s	>4 h	>4 h	1 m	2 m	1 m	29 s	1 m
SensIT	1 s	>4 h	>4 h	2 m	3 m	2 m	30 s	1 m

Table 5: The average run-time of the learning algorithms under consideration on the data sets described in Table 2. s stands for seconds, m for minutes and h for hours.

C.3 Relaxation of CBIL

The proposed (CBIL) is based on Eq. (18), which holds only when $\delta(\mathbf{x}) \geq 0$. As it was discussed in Section 5.2, Eq. (18) can be relaxed by adding some $\lambda > 0$ leading to Eq. (19). In practice, it not only can make the bound computable, but also make it smoother, since arbitrarily small values of $\delta(\mathbf{x})$ implies arbitrarily large values of $\hat{r}(B_Q, \mathbf{x})/\delta(\mathbf{x})$. The latter should be avoided if (CBIL) is used as some optimization or selection criterion.

In this section, we study the impact of λ on the bound's value on different data sets. In Figure 4, we display the results of all 20 experimental trials for HAR, Isolet, Letter, MNIST and Fashion when $\lambda \in [0.1, 0.2, \ldots, 1]$. One can observe that when the bound is not penalized much (i.e., $\delta(\mathbf{x})$ is far from 0), then the increase of λ makes the bound looser, so $\lambda = 0.1$ is the tightest choice. Exactly the opposite situation is observed when $\delta(\mathbf{x})$ is small (trials 4 and 14 for Letter, most of trials for Fashion): higher values of λ diminish the influence of hyperbolic weights $1/\delta(\mathbf{x})$, so $\lambda = 1$ leads to the tightest bound.

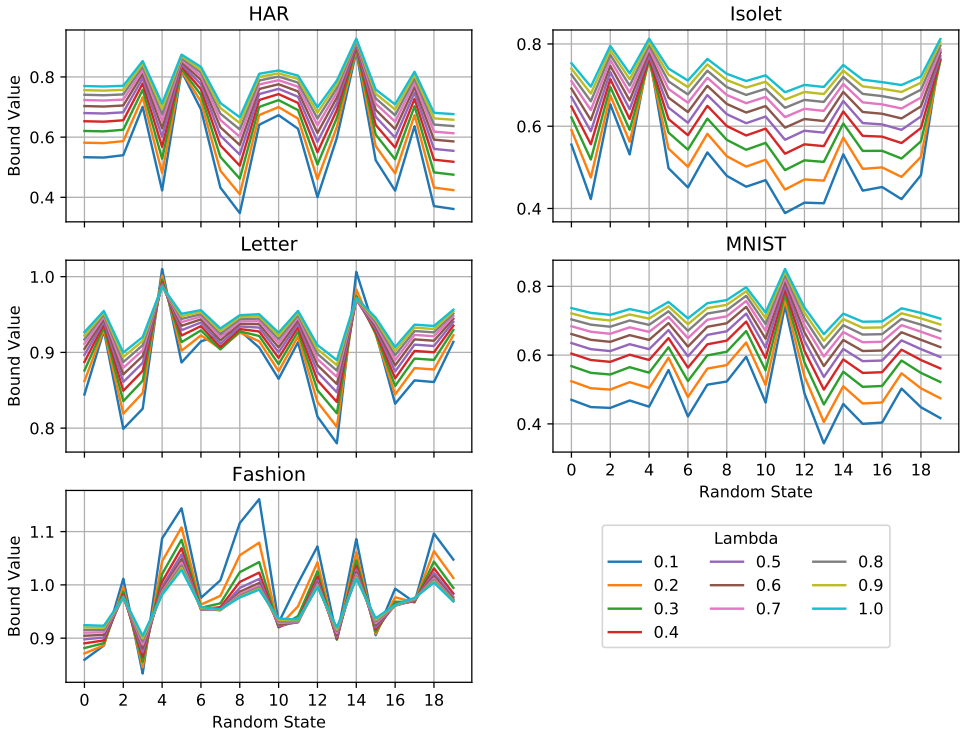

Figure 4: The value of (CBIL) with different λ over 20 different labeled/unlabeled splits of 5 data sets.

References

Massih-Reza Amini and Patrick Gallinari. Semi-supervised learning with explicit misclassification modeling. In *International Joint Conference on Artificial Intelligence*, pages 555–560. Morgan Kaufmann, 2003.

Massih-Reza Amini and Nicolas Usunier. *Learning with Partially Labeled and Interdependent Data*. Springer, 2015.

Massih-Reza Amini, François Laviolette, and Nicolas Usunier. A transductive bound for the voted classifier with an application to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 65–72, 2008.

Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Lies Hadjadj, Emilie Devijver, and Yury Maximov. Self-training: A survey. *arXiv preprint arXiv:2202.12040*, 2023.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2020.

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

Baptiste Bauvin, Cécile Capponi, Jean-Francis Roy, and François Laviolette. Fast greedy C-bound minimization with guarantees. *Machine Learning*, 109(9):1945–1986, 2020.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian theory for transductive learning. In *International Conference on Artificial Intelligence and Statistics*, volume 33, pages 105–113. PMLR, 2014.

Mikhail Belkin and Partha Niyogi. Semi-supervised learning on riemannian manifolds. *Machine Learning*, 56(1-3):209–239, 2004.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Deng Cai, Xiaofei He, and Jiawei Han. Training linear discriminant analysis in linear time. In *2008 IEEE 24th International Conference on Data Engineering*, pages 209–217. IEEE, 2008.

Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6912–6920, 2021.

Olivier Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *International Workshop on Artificial Intelligence and Statistics*, volume R5 of *Proceedings of Machine Learning Research*, pages 57–64. PMLR, 2005.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

CB Chittineni. Learning with imperfectly labeled patterns. *Pattern Recognition*, 12(5): 281–291, 1980.

Yanwen Chong, Yun Ding, Qing Yan, and Shaoming Pan. Graph-based semi-supervised learning: A review. *Neurocomputing*, 408:216–230, 2020.

Philip Derbeko, Ran El-Yaniv, and Ron Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, 22(1):117–142, 2004.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28 (1):1–47, 1975.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Ali Fakeri-Tabrizi, Massih-Reza Amini, Cyril Goutte, and Nicolas Usunier. Multiview self-learning. *Neurocomputing*, 155(C):117–127, 2015.

Vasilii Feofanov, Emilie Devijver, and Massih-Reza Amini. Transductive bounds for the multi-class majority vote classifier. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3566–3573, 2019.

Vasilii Feofanov, Emilie Devijver, and Massih-Reza Amini. Wrapper feature selection with partially labeled data. *Applied Intelligence*, 52(11):12316–12329, 2022.

Vasilii Feofanov, Malik Tiomoko, and Aladin Virmaux. Random matrix analysis to balance between supervised and unsupervised learning under the low density separation assumption. In *International Conference on Machine Learning*, pages 10008–10033. PMLR, 2023.

S Fralick. Learning to recognize patterns without a teacher. *IEEE Transactions on Information Theory*, 13(1):57–64, 1967.

Spencer Frei, Difan Zou, Zixiang Chen, and Quanquan Gu. Self-training converts weak learners to strong learners in mixture models. In *International Conference on Artificial Intelligence and Statistics*, pages 8003–8021. PMLR, 2022.

Yoav Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, 1995.

Martin Gebel. *Multivariate calibration of classifier scores into the probability space*. PhD thesis, The University of Dortmund, 2009.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning*, pages 353–360, 2009.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario March, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26):787–860, 2015.

Fabian Gieseke, Antti Airola, Tapio Pahikkala, and Oliver Kramer. Fast and simple gradient-based optimization for semi-supervised support vector machines. *Neurocomputing*, 123:23–32, 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning*, pages 200–209. Morgan Kaufmann Publishers Inc., 1999.

Anastasia Krithara, Massih-Reza Amini, Jean-Michel Renders, and Cyril Goutte. Semi-supervised document classification with a mislabeling error model. In *Advances in Information Retrieval*, pages 370–381, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *Advances in Neural Information Processing Systems*, pages 769–776, 2007.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003.

François Laviolette, Emilie Morvant, Liva Ralaivola, and Jean-Francis Roy. Risk upper bounds for general ensemble methods with an application to multiclass classification. *Neurocomputing*, 219:15–25, 2017.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on Challenges in Representation Learning, ICML*, 3 (2):896, 2013.

Christian Leistner, Amir Saffari, Jakob Santner, and Horst Bischof. Semi-supervised random forests. In *2009 IEEE 12th International Conference on Computer Vision*, pages 506–513. IEEE, 2009.

Gaël Letarte, Pascal Germain, Benjamin Guedj, and François Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In *Advances in Neural Information Processing Systems*, pages 6872–6882, 2019.

Y Li, L Liu, and RT Tan. Certainty-driven consistency loss for semi-supervised learning. *arXiv preprint arXiv:1901.05657*, 2021.

Marco Loog. Contrastive pessimistic likelihood estimation for semi-supervised classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):462–475, 2015.

Stephan S Lorenzen, Christian Igel, and Yevgeny Seldin. On PAC-Bayesian bounds for random forests. *Machine Learning*, 108(8-9):1503–1522, 2019.

Gilles Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

Andres Masegosa, Stephan Lorenzen, Christian Igel, and Yevgeny Seldin. Second order PAC-Bayesian bounds for the weighted majority vote. *Advances in Neural Information Processing Systems*, 33, 2020.

Andreas Maurer. A note on the PAC Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

Yury Maximov, Massih-Reza Amini, and Zaid Harchaoui. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm. *Journal of Artificial Intelligence Research*, 61(1):761–786, 2018.

David McAllester. Simplified PAC-Bayesian margin bounds. In *Learning Theory and Kernel Machines*, pages 203–215. Springer Berlin Heidelberg, 2003.

David A McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(26):789–814, 2010.

Emilie Morvant, Sokol Koço, and Liva Ralaivola. PAC-Bayesian generalization bound on confusion matrix for multi-class classification. In *International Conference on Machine Learning*, 2012.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204, 2013.

Ambroise Odonnat, Vasilii Feofanov, and Ievgen Redko. Leveraging ensemble diversity for robust self-training in the presence of sample selection bias. *arXiv preprint arXiv:2310.14814*, 2023.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Mohammad Peikari, Sherine Salama, Sharon Nofech-Mozes, and Anne L Martel. A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific Reports*, 8(1):1–13, 2018.

Aswathnarayan Radhakrishnan, Jim Davis, Zachary Rabin, Benjamin Lewis, Matthew Scherreik, and Roman Ilin. Design choices for enhancing noisy student self-training. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1926–1935, 2024.

Harish G Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12:530–554, 2018.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(49):1369–1392, 2007.

Jean-Francis Roy, François Laviolette, and Mario Marchand. From PAC-Bayes bounds to quadratic programs for majority votes. In *International Conference on Machine Learning*, pages 649–656. Omnipress, 2011.

Jean-Francis Roy, Mario Marchand, and François Laviolette. A column generation bound minimization approach with PAC-Bayesian generalization guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 1241–1249, 2016.

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *International Conference on Artificial Intelligence and Statistics*, pages 838–846, 2015.

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

Razieh Sheikhpour, Mehdi Agha Sarram, Sajjad Gharaghani, and Mohammad Ali Zare Chahooki. A survey on semi-supervised feature selection methods. *Pattern Recognition*, 64(C):141–158, 2017.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *International Conference on Algorithmic Learning Theory*, pages 466–492. PMLR, 2017.

Gökhan Tür, Dilek Z. Hakkani-Tür, and Robert E. Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45:171–186, 2005.

Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

Vladimir Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Advances in Neural Information Processing Systems*, pages 6838–6849, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Mingyuan Zhang, Jane Lee, and Shivani Agarwal. Learning from noisy labels with no change to the training process. In *International Conference on Machine Learning*, pages 12468–12478. PMLR, 2021.

Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. How unlabeled data improve generalization in self-training? A one-hidden-layer theoretical analysis. In *International Conference on Learning Representations*, 2022.

Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, pages 321–328, 2004.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.