An Unbiased Risk Estimator for Learning with Augmented Classes

Yu-Jie Zhang ZHANGYJ@LAMDA.NJU.EDU.CN
Peng Zhao ZHAOP@LAMDA.NJU.EDU.CN
Zhi-Hua Zhou ZHOUZH@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China

Abstract

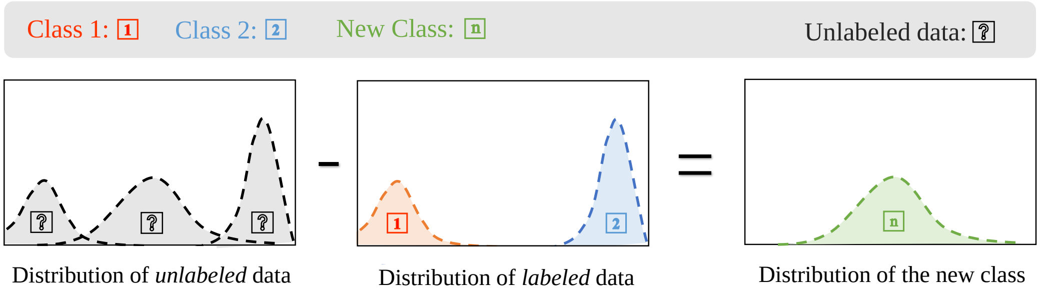
In this paper, we study the problem of learning with augmented classes (LAC), where new classes that do not appear in the training dataset might emerge in the testing phase. The mixture of known classes and new classes in the testing distribution makes the LAC problem quite challenging. Our discovery is that by exploiting cheap and vast unlabeled data, the testing distribution can be estimated in the training stage, which paves us a way to develop algorithms with nice statistical properties. Specifically, we propose an unbiased risk estimator over the testing distribution for the LAC problem, and further develop an efficient algorithm to perform the empirical risk minimization. Both asymptotic and non-asymptotic analyses are provided as theoretical guarantees. The efficacy of the proposed algorithm is also confirmed by experiments.

1. Introduction

Recent advances in machine learning encourage its application in high-stack scenarios, where robustness is the central requirement (Dietterich, 2017, 2019). The robustness of a learning system relies on its adaptability to open and dynamic environments, in which the distribution, the label space and the feature space of a task could change. Under such a circumstance, classical supervised learning approaches might fail since they do not take the non-stationarity into consideration. Consequently, it is of great importance to design more robust and reliable algorithms to adapt to these unreliable and changing factors in environments.

This paper investigates the class-incremental learning (Zhou and Chen, 2002), where new classes may emerge in the learning process. Specifically, we concern about the problem of learning with augmented classes (LAC) (Da et al., 2014), one of the core tasks of class-incremental learning. In the LAC problem, classes that do not appear in the training dataset might emerge in the testing stage. These new classes, if simply neglected, will seriously deteriorate learning performance due to the misclassification of instances therein. Thus, a desired learning system should be able to identify these unknown classes and retain good generalization ability over the testing distribution.

The fundamental obstacle of the LAC problem is how to model relationships between known and new classes. In the literature, researchers propose various assumptions to model such relationships, including low-density separation (Da et al., 2014) or open space property (Scheirer et al., 2013), etc. A variety of algorithms are proposed with satisfactory empirical performance. Nevertheless, their theoretical properties are generally unclear. To the best of our knowledge, there is no method that provides theoretical investigations on classifiers' generalization ability over the testing distribution, where both known and new classes exist.



Distribution of *unlabeled* data

Distribution of *labeled* data

Distribution of the new class

Figure 1: Approximate the distribution of new classes by distributions of labeled data and unlabeled data in the training dataset.

In this paper, we discover that by exploiting the cheap and vast unlabeled data, an *unbiased* risk estimator over the *testing* distribution for the LAC problem can be developed. This paves a way to design algorithms with nice statistical properties. Our intuition is that, though labeled instances from new classes are absent from labeled data, their distribution information is contained in unlabeled data, though mixed with the distribution of known classes. Therefore, we can access the distribution of new classes by separating the distribution of labeled data from that of unlabeled data, based on which an unbiased risk estimator can be established. Figure 1 illustrates our idea.

More precisely, we propose the *class shift condition* to model the connection between training and testing distributions for the LAC problem. Under such a condition, the distribution of new classes can be directly reduced to the difference between the distribution of unlabeled data and that of labeled data (from known classes). Based on the reduction, we can evaluate the risk of classifiers over testing distribution in the training stage, where minimizing its empirical estimator finally gives our EULAC algorithm, short for Exploiting Unlabeled data for Learning with Augmented Classes.

Since the empirical risk estimator is evaluated directly over the testing distribution, the EULAC algorithm enjoys several favorable properties. Theoretically, our algorithm enjoys both asymptotic (consistency) and non-asymptotic (generalization error bound) guarantees. Notably, the non-asymptotic analysis further justifies the capability of our algorithm in utilizing unlabeled data, since the generalization error becomes smaller with an increasing number of unlabeled data. Moreover, extensive experiments further validate the effectiveness of our approach. It is noteworthy to mention that our approach can now perform the standard cross validation procedure to select parameters, while most geometric-based algorithms cannot and thus heavily rely on experience to tune parameters, since the cross validation of these algorithms is biased due to unavailability of the testing distribution.

The main contributions of this paper are as follows.

(1) We propose the *class shift* condition for the LAC problem, which models the connection between training and testing distributions.

(2) Based on the class shift condition, we establish an unbiased risk estimator over the testing distribution, by exploiting unlabeled data.

(3) We propose the EULAC algorithm to empirically minimize the unbiased estimator. We further prove its theoretical effectiveness by both asymptotic and non-asymptotic analysis, and validate its empirical superiority by extensive experiments.

This paper is organized as follows. First, we will discuss the relationship between our method and the related works of LAC problem. Meanwhile, some other relevant topics regarding reliable machine learning algorithms will also be included (Section 2). Then, we formally describe the LAC problem and introduce notations used in the rest of this paper (Section 3). Next, we propose the class shift condition and show how to establish an unbiased risk estimator with labeled and unlabeled training data, which finally gives our algorithm (Section 4). Afterwards, we provide the theoretical properties of our algorithm containing both asymptotic and non-asymptotic analyses(Section 5). Finally, we validate the effectiveness of our method by extensive experiments (Section 6).

2. Related Work

Class-incremental learning (C-IL) (Zhou and Chen, 2002) aims to facilitate the learning system with capability of handling new classes that appear in the learning process, which is a fundamental task for robust and reliable learning in open and dynamic environments. Learning with augmented classes (LAC), the focus of this paper, is one of the core tasks of C-IL, where classes that do not appear in training data might emerge in the testing stage. The pioneering work of Da et al. (2014) proposes to exploit unlabeled data for the LAC problem, and authors design a novel algorithm by tuning the decision boundary to pass through low-density regions. While we share the same problem setup, our approach differs from theirs in various aspects. We propose the class shift condition to model connections between known and new classes, whereas theirs use the geometric assumption instead. Moreover, our approach enjoys stronger theoretical guarantees and superior empirical performance.

Another related topic is the open set recognition (Scheirer et al., 2013), which is an alternative terminology mainly used in the pattern recognition community. Scheirer et al. (2013) introduce the concept of “open space risk” to penalize predictions outside the support of training data, based on which many approaches are proposed (Scheirer et al., 2013, 2014). Besides, there are also works based on the nearest neighbor method (Mendes-Junior et al., 2017) or the extreme value theory (Rudd et al., 2018). Although these algorithms achieve satisfactory empirical behavior, they generally lack theoretical guarantees.

Two exceptions are works of Scott and Blanchard (2009) and Liu et al. (2018). Authors focus on the Neyman-Pearson classification problem, where false positive predictions on known classes are minimized with the constraint of desired novelty detection ratio, or vice. Liu et al. (2018) design a meta-algorithm to take the existing novelty detection algorithm as a subroutine to recognize new classes. Although the meta-algorithm enjoys PAC-style guarantees, statistical properties of subroutines are unclear. Moreover, these works mainly contribute to the analysis for novelty detection, while the predictive error over the testing distribution, or the generalization ability, is not investigated.

Apart from the batch setting, researchers also consider the scenario of emerging new classes in the streaming data (Fink et al., 2006; Muhlbaier et al., 2009), where a few labeled instances from new classes are available and the learner requires to update model incrementally to adapt to emerging classes. Subsequently, Mu et al. (2017) and Cai et al. (2019) study the problem of streaming classification with emerging new classes, which is more challenging since the learner requires to detect new classes, update model, and predict with unlabeled streaming data.

Learning with rejection also concerns about reliability of predictions, where the classifier is provided with an option to reject an instance instead of producing a low confidence prediction (Chow, 1970). Plenty of works are proposed with effective implementation (Yuan and Wegkamp, 2010; Cortes et al., 2016b; Wang and Qiao, 2018; Geifman and El-Yaniv, 2019) or theoretical foundations (Herbei and Wegkamp, 2006; Bartlett and Wegkamp, 2008; Cortes et al., 2016a). However, algorithms for learning with rejection are not applicable to the LAC problem, since new classes do not necessarily locate in the low confidence region.

3. Problem Statement

In this section, we provide formal descriptions of the learning with augmented classes (LAC) problem and introduce notations used throughout the paper.

LAC Problem. In the training stage, the learner receives a labeled dataset $D_L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_l}$, sampled from the training distribution P_{tr} over $\mathcal{X} \times \mathcal{Y}'$, where $\mathcal{X}$ denotes the feature space and $\mathcal{Y}' = \{1, \ldots, K\}$ is the label space of K known classes.

In the testing stage, the learner requires to predict instances from the testing distribution P_{te}, where new classes might emerge. Since the specific partition of new classes is unobserved, the learner would predict all of them as a single new class nc. So the testing distribution is defined over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{1, \ldots, K, nc\}$ is the augmented label space.

The goal of the learner is to train a classifier $g : \mathcal{X} \mapsto \mathcal{Y}$ minimizing the following expected risk with respect to 0-1 loss in order to retain a good generalization ability over the testing distribution,

$$R(g) = \mathbb{E}_{(\mathbf{x},y)\sim P_{te}} \left[\mathbb{1}(g(\mathbf{x}) \neq y)\right], \tag{1}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function.

In our setup, the learner can additionally receive an unlabeled dataset $D_U = \{\mathbf{x}_i\}_{i=1}^{n_u}$ sampled from the testing distribution apart from labeled data.

Notations. We use the uppercase P to denote distributions, where training, testing and new classes distributions are denoted by P_{tr}, P_{te} and P_{new}, respectively. Besides, the lowercase p is the density function, where the joint, conditional and marginal density functions are indicated by the subscripts XY, $X|Y$ $(Y|X)$ and X (Y), respectively. For instance, we denote by $p_X^{te}(\mathbf{x})$ the marginal density function of the testing distribution over feature space.

Although the original training distribution is defined over label space $\mathcal{X} \times \mathcal{Y}'$, the learning is implemented on the augmented label space $\mathcal{Y}$. Thus, we redefine all the distributions over the space $\mathcal{X} \times \mathcal{Y}$, where $p_{XY}^{tr}(\mathbf{x}, y) = 0$ holds for all $\mathbf{x} \in \mathcal{X}$ and $y = nc$. Meanwhile, $p_{XY}^{new}(\mathbf{x}, y) = 0$ holds for all $\mathbf{x} \in \mathcal{X}$ and $y \neq nc$.

4. Our Proposal

In this section, we propose an unbiased risk estimator over the testing distribution, where the *class shift* condition is introduced first to bridge training and testing distributions. Based on this condition, we then study an ideal situation where the testing distribution is accessible. Next, we show how to approximate the ideal situation with empirical training data, which finally gives our EULAC algorithm.

4.1 Problem Refinement

The LAC problem studies the situation where new classes appear in the testing stage. Although not explicitly stated, previous works (Scheirer et al., 2013; Da et al., 2014; Mendes-Junior et al., 2017) actually rely on the essential assumption that the distribution of known classes remains unchanged with the augmentation of new classes. Following the same spirit, we introduce the following *class shift* condition to rigorously depict the connection between training and testing distributions of the LAC problem.

Definition 1 (Class Shift Condition)**.** We claim the training distribution P_{tr}, the testing distribution P_{te}, and distribution of new classes P_{new} are under the *class shift* condition when

$$P_{te} = \theta \cdot P_{tr} + (1 - \theta) \cdot P_{new}, \tag{2}$$

where $0 < \theta \leq 1$ is a certain mixture proportion.

Class shift condition essentially states that the testing distribution can be regarded as a mixture of those of known and new classes with a certain proportion θ, where P_{tr} is actually the distribution of known classes.

4.2 A Practical Unbiased Risk Estimator

We now turn to develop the unbiased risk estimator for the LAC problem, where the major obstacle is how to approximate the testing distribution with training data. Instead of imposing assumptions on new classes, we find that the information contained in vast and cheap unlabeled data collected from the environments is actually a good advisor. Before describing the way to estimate the testing distribution via unlabeled data, we first consider an ideal situation where the testing distribution is available.

An Ideal Case. When the testing distribution is available, the LAC problem degenerates to a standard multi-class classification problem, which can be addressed by many established algorithms. Among those approaches, we adopt the one-versus-rest (OVR) strategy, which enjoys sound theoretical foundations (Zhang, 2004) and nice practical performance (Rifkin and Klautau, 2004). The risk minimization problem for the OVR strategy is formulated as,

$$\min_{f_1,\dots,f_{K+1}} R_\psi(f_1, \dots, f_{K+1}) = \mathbb{E}_{(\mathbf{x},y)\sim P_{te}} \Big[\psi(f_y(\mathbf{x})) + \sum_{k=1,k\neq y}^{K+1} \psi(-f_k(\mathbf{x}))\Big], \tag{3}$$

where f_i is the classifier trained for the i-th class, $i = 1, \dots, K$; and f_{nc} denotes the classifier for the new class. For simplicity, we substitute f_{nc} with f_{K+1} in the formulation. $\psi : \mathbb{R} \mapsto [0, +\infty]$ is a certain binary surrogate loss function such as hinge loss, exponential loss, etc. After obtaining classifiers by minimizing the empirical version of (3), the learner can predict an instance to the index of the classifier with maximum output, namely, $\arg\max_{i=1,\dots,K,nc} f_i(\mathbf{x})$.

Approximating Testing Distribution. Unfortunately, the testing distribution is unavailable in the training stage, due to the existence of new class. Thus, the risk minimization problem (3) is far from practice. We now proceed to reduce the OVR risk R_ψ to a more operational one established over distributions of labeled and unlabeled data.

First, we note that, according to the definition of class shift condition, the joint probability density of the testing distribution can be decomposed into two parts,

$$p_{XY}^{te}(\mathbf{x}, y) \overset{(2)}{=} \theta \cdot p_{XY}^{tr}(\mathbf{x}, y) + (1 - \theta) \cdot p_{XY}^{new}(\mathbf{x}, y)$$

$$= \theta \cdot p_{XY}^{tr}(\mathbf{x}, y) + (1-\theta) \cdot \mathbb{1}(y = nc) \cdot p_X^{new}(\mathbf{x}),$$

where the last equality follows from the fact that $p_{XY}^{new}(\mathbf{x}, y) = 0$ holds for all $\mathbf{x} \in \mathcal{X}$ and $y \neq nc$. The first part is the joint probability density function of the training distribution, which can be accessed by the labeled data. The only unknown term is the second part, namely, the marginal density function of the new class. So, to access the testing distribution, it is sufficient to estimate the distribution of new classes.

The estimation can be achieved by exploiting the labeled and unlabeled data. The basic observation is that, under the class shift condition, by summing over the label space $\mathcal{Y}$ we have

$$(1-\theta) \cdot p_X^{new}(\mathbf{x}) = p_X^{te}(\mathbf{x}) - \theta \cdot p_X^{tr}(\mathbf{x}), \tag{4}$$

which shows that the marginal density function of the new class $p_X^{new}(\mathbf{x})$ can be calculated by the difference between those of testing and training distributions. Although the calculation still requires the knowledge of testing distribution, what we really demand is only its marginal distribution over the feature space, which can be estimated effectively by unlabeled data.

Consequently, we can evaluate the OVR risk R_ψ in the training stage through an equivalent risk R'_{LAC}.

Proposition 1. *Under the class shift condition, the following equality holds for all measurable functions* $f_1, \ldots, f_K, f_{nc}$*,*

$$R_\psi(f_1, \ldots, f_K, f_{nc}) = R'_{LAC}(f_1, \ldots, f_K, f_{nc}),$$

where R'_{LAC} *is defined as,*

$$\begin{aligned} R'_{LAC} =& \theta \cdot \mathbb{E}_{(\mathbf{x},y)\sim P_{tr}} \left[\psi(f_y(\mathbf{x})) - \psi(-f_y(\mathbf{x})) + \psi(-f_{nc}(\mathbf{x})) - \psi(f_{nc}(\mathbf{x}))\right] \\ &+ \mathbb{E}_{\mathbf{x}\sim p_X^{te}(\mathbf{x})} \left[\psi(f_{nc}(\mathbf{x})) + \sum_{k=1}^{K} \psi(-f_k(\mathbf{x}))\right]. \end{aligned}$$

Remark 1. The risk R'_{LAC} can be evaluated while training since the first term is the expectation over P_{tr} which can be effectively approximated by the labeled data, and the second term is established on the marginal distribution of testing data, which can be approached by the unlabeled data.

One last problem regarding the risk R'_{LAC} is that even with convex binary surrogate loss functions, the corresponding risk minimization problem is non-convex, and the non-convexity comes from the terms $-\psi(-f_y(\mathbf{x}))$ and $-\psi(f_{nc}(\mathbf{x}))$. Inspired by the work of du Plessis et al. (2014), we eliminate these non-convex terms by carefully choosing surrogate loss functions, as stated in the following proposition.

Proposition 2. *Under the class shift condition, the following equality holds for all measurable functions* $f_1, \ldots, f_K, f_{nc}$ *and when the surrogate loss function satisfies* $\psi(z) - \psi(-z) = -z$ *for all* $z \in \mathbb{R}$*,*

$$R_\psi(f_1, \ldots, f_K, f_{nc}) = R_{LAC}(f_1, \ldots, f_K, f_{nc}),$$

where the LAC risk R_{LAC} *is defined as,*

$$R_{LAC} = \theta \cdot \mathbb{E}_{(\mathbf{x},y)\sim P_{tr}} \left[f_{nc}(\mathbf{x}) - f_y(\mathbf{x})\right] + \mathbb{E}_{\mathbf{x}\sim p_X^{te}(\mathbf{x})} \left[\psi(f_{nc}(\mathbf{x})) + \sum_{k=1}^{K} \psi(-f_k(\mathbf{x}))\right].$$

Proposition 2 is a direct consequence of Proposition 1 with desired surrogate loss functions. Many loss functions satisfy the condition, such as logistic loss $\psi(z) = \log(1 + \exp(-z))$, square loss $\psi(z) = (1 - z)^2/4$ and double hinge loss$\psi(z) = \max(-z, \max(0, 1/2 - z/2))$.

After acquiring the LAC risk R_{LAC}, the classifiers can be trained via minimizing the corresponding empirical risk estimator $\widehat{R}_{LAC}$. Since LAC risk R_{LAC} is equivalent to the ideal OVR risk R_ψ, its empirical estimator $\widehat{R}_{LAC}$ is unbiased over the testing distribution. Such a property guarantees performance of our algorithm in the testing stage.

4.3 Empirical Implementation

Specifically, we consider minimizing the empirical LAC risk $\widehat{R}_{LAC}$ in a reproducing kernel Hilbert space (RKHS), formulated as

$$\min_{f_1,\ldots,f_k,f_{nc}\in\mathbb{F}} \widehat{R}_{LAC} + \lambda\Big(\sum_{i=1}^{K} \|f_i\|_{\mathbb{F}}^2 + \|f_{nc}\|_{\mathbb{F}}^2\Big), \tag{5}$$

where $\mathbb{F}$ is the RKHS associated to a PDS kernel $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ and $\|\cdot\|_{\mathbb{F}}$ is its norm. The empirical LAC risk $\widehat{R}_{LAC}$ is the empirical version of R_{LAC} defined by

$$\widehat{R}_{LAC} = \frac{\theta}{n_l}\sum_{i=1}^{n_l}(f_{nc}(\mathbf{x}_i) - f_{y_i}(\mathbf{x}_i)) + \frac{1}{n_u}\sum_{i=1}^{n_u}\Big(\psi(f_{nc}(\mathbf{x}_i)) + \sum_{k=1}^{K}\psi(-f_k(\mathbf{x}_i))\Big), \tag{6}$$

where ψ can be any surrogate loss functions satisfying the condition in Proposition 2.

According to the representer theorem (Scholkopf and Smola, 2001), the optimal solution of the optimization problem (5) is provably in the form of

$$f_k(\cdot) = \sum_{\mathbf{x}_i\in D_L} \alpha_i^k K(\cdot, \mathbf{x}_i) + \sum_{x_j\in D_U} \alpha_j^k K(\cdot, \mathbf{x}_j), \tag{7}$$

where α_i^k is the i-th coefficient of the k-th classifier.

Plugging (7) into (5), we get a convex optimization problem with respect to $\boldsymbol{\alpha}$, which can be solved efficiently. After obtaining the classifiers $f_1, \ldots, f_K, f_{nc}$, the learner can just predict as $g(\mathbf{x}) = \arg\max_{i=1,\ldots,K,nc} f_i(\mathbf{x})$.

Notice that the implementation of our algorithm requires the knowledge of the mixture proportion θ, where plenty of works (Kawakubo et al., 2016; Ramaswamy et al., 2016) have explored to estimate θ from the labeled dataset and the unlabeled data. We adopt the method of Ramaswamy et al. (2016). Algorithm 1 summarizes main procedures.

The last issue regarding the implementation is the parameters selection. Since the risk estimator $\widehat{R}_{LAC}$ is established on the testing distribution directly, we can perform an unbiased cross validation procedure to select the parameters. On the contrary, the cross validation process could be biased for geometric-based algorithms since the distribution of new classes is unknown, and thus the setting of parameters heavily relies on experience for these approaches.

5. Theoretical Analysis

In this section, we establish both asymptotic and non-asymptotic analysis for our approach. Specifically, we first show the infinite-sample consistency of the LAC risk R_{LAC}. Then, we investigate the generalization property via generalization error analysis. All proofs can be found in the appendix.

Algorithm 1 EULAC Algorithm

```
Require: labeled dataset $D_L$, unlabeled dataset $D_U$, kernel function $K(\cdot,\cdot)$, regularization parameter $\lambda$.
Ensure: Classifiers' coefficients $\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^K, \boldsymbol{\alpha}^{nc}$
1: Estimate the mixture ratio $\theta$ of the observed classes in the testing distribution.
2: Solve the convex optimization problem (5).
```

5.1 Infinite-sample Consistency

At first, we show that the LAC risk R_{LAC} is infinite-sample consistent with the risk over the testing distribution with respect to 0-1 loss. Namely, by minimizing the expected risk on R_{LAC}, we can get classifiers achieving Bayes rule over the testing distribution, which is the ultimate goal of the LAC problem. The formal statement is as follows.

Theorem 1. *Under the class shift condition, suppose the surrogate loss function ψ is convex, bounded below, differential, satisfying $\psi(z) - \psi(-z) = -z$ and $\psi(z) < \psi(-z)$ when $z > 0$, then for any $\epsilon_1 > 0$, there exists $\epsilon_2 > 0$ such that for all measurable functions $f_1, \ldots, f_K, f_{nc}$,*

$$R_{LAC}(f_1, \ldots, f_K, f_{nc}) \leq R^*_{LAC} + \epsilon_2$$

implies

$$R\Big(\underset{i=1,\ldots,K,nc}{\arg\max}\ f_i\Big) \leq R^* + \epsilon_1,$$

where $R^*_{LAC} = \min_{f_1,\ldots,f_K,f_{nc}} R_{LAC}(f_1, \ldots, f_K, f_{nc})$ and R^* is the Bayes error over the testing distribution P_{te}.

Theorem 1 follows from Proposition 2 and analysis in the seminal work of Zhang (2004), where the consistency property of OVR risk R_ψ is investigated. Since the LAC risk R_{LAC} is equivalent to the OVR risk R_ψ, it is naturally infinite-sample consistent.

Many loss functions satisfy assumptions in Theorem 1 such as logistic loss $\psi(z) = \log(1 + \exp(-z))$ and square loss $\psi(z) = (1-z)^2/4$. Further, a more quantitative result can be obtained for the square loss.

Theorem 2. *Under the class shift condition, when using $\psi(z) = (1-z)^2/4$ as the surrogate loss function, for all measurable functions $f_1, \ldots, f_K, f_{nc}$, we have,*

$$R\Big(\underset{i=1,\ldots,K,nc}{\arg\max}\ f_i\Big) - R^* \leq \sqrt{2\big(R_{LAC}(f_1, \ldots, f_K, f_{nc}) - R^*_{LAC}\big)}.$$

Theorem 2 shows that the excess risk of R_{LAC} is actually the upper bound of the excess risk of 0-1 loss. Thus, by minimizing the LAC risk R_{LAC}, we can obtain classifiers performing well on the testing distribution with respect to 0-1 loss.

5.2 Finite-sample Convergence

We establish the generalization bound for the proposed algorithm in this part. Since the algorithm actually minimizes the empirical risk estimator $\widehat{R}_{LAC}$ with a regularization term of the RKHS $\mathbb{F}$, it is equivalent to investigate the generalization ability of classifiers $f_1, \ldots, f_K, f_{nc}$ in the kernel-based hypothesis set $\mathcal{F}$,

$$\mathcal{F} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \Phi(\mathbf{x})\rangle \mid \|\mathbf{w}\|_{\mathbb{F}} \leq \Lambda\}, \tag{8}$$

where $\Phi : \mathbf{x} \mapsto \mathbb{F}$ is a feature mapping associated with the PDS kernel K, and $\mathbf{w}$ is an element in RKHS $\mathbb{F}$. We have the following generalization error bound.

Theorem 3. *Assume that $K(\mathbf{x}, \mathbf{x}) \leq r^2$ holds for all $\mathbf{x} \in \mathcal{X}$ and the surrogate loss function ψ is bounded by $B_\psi \geq 0$ and is L-Lipschitz continuous*[1]*. Let $\mathcal{F} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \Phi(\mathbf{x})\rangle \mid \|\mathbf{w}\|_{\mathbb{F}} \leq \Lambda\}$ be the kernel-based hypothesis set. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of labeled samples D_L of size n_l from P_{tr} and unlabeled samples D_U of size n_u from P_{te}, the following holds for all $f_1, \ldots, f_K, f_{nc} \in \mathcal{F}$,*

$$\begin{aligned} & R_{LAC}(f_1, \ldots, f_K, f_{nc}) - \widehat{R}_{LAC}(f_1, \ldots, f_K, f_{nc}) \\ \leq & \frac{2(K+1)\Lambda r}{\sqrt{n_l}} + 6\Lambda r\sqrt{\frac{2\log(4/\delta)}{n_l}} + \frac{2(K+1)L\Lambda r}{\sqrt{n_u}} + 3(K+1)B_\psi\sqrt{\frac{\log(4/\delta)}{n_u}}. \end{aligned}$$

Based on Theorem 3, by the standard argument (Bousquet et al., 2003; Mohri et al., 2012), we can obtain the estimation error bound.

Theorem 4. *Under the assumptions of Theorem 3 and let $\widehat{f}_1, \ldots, \widehat{f}_K, \widehat{f}_{nc}$ be the optimal solution of the optimization problem (5) with certain $\lambda > 0$, we have*

$$R_{LAC}(\widehat{f}_1, \ldots, \widehat{f}_K, \widehat{f}_{nc}) - \inf_{\boldsymbol{f} \in \mathscr{F}} R_{LAC}(f_1, \ldots, f_K, f_{nc}) \leq \mathcal{O}_p\left(\frac{K+1}{\sqrt{n_l}} + \frac{K+1}{\sqrt{n_u}}\right),$$

where $\boldsymbol{f}$ denotes $(f_1, \ldots, f_K, f_{nc})$ and $\mathscr{F} = \{\boldsymbol{f} \mid f_1, \ldots, f_K, f_{nc} \in \mathbb{F}, \sum_{k=1}^K \|f_k\|_{\mathbb{F}}^2 + \|f_{nc}\|_{\mathbb{F}}^2 \leq c_\lambda^2\}$. The parameter $c_\lambda > 0$ is a constant related to λ in (5). For a better presentation, we use the $\mathcal{O}_p$-notation to keep the dependence on n_u, n_l and K only.

Remark 2. Theorem 3 and Corollary 4 show that, the estimation error of the trained classifiers decreases with a growing number of labeled and *unlabeled* data. An important message delivered here is that our algorithm can achieve better performance by collecting more unlabeled data, which theoretically justifies its effectiveness in exploiting unlabeled data. Experiments also validate the same tendency.

5.3 Overview of Theoretical Results

Recall that the ultimate goal of the LAC problem is to obtain classifiers that approach Bayes rule over the testing distribution, and thus we need to minimize the excess risk $R\left(\operatorname{argmax}_{i=1,\ldots,K,nc} f_i\right) - R^*$. According to the the consistency guarantee presented in Section 5.1, it suffices to minimize the excess risk $R_{LAC}(\boldsymbol{f}) - R_{LAC}^*$, which can be further decomposed into the estimation error and the approximation error as follows,

$$R_{LAC}(\boldsymbol{f}) - R_{LAC}^* = \underbrace{R_{LAC}(\boldsymbol{f}) - \inf_{\boldsymbol{f} \in \mathscr{F}} R_{LAC}(\boldsymbol{f})}_{\texttt{estimation error}} + \underbrace{\inf_{\boldsymbol{f} \in \mathscr{F}} R_{LAC}(\boldsymbol{f}) - R_{LAC}^*}_{\texttt{approximation error}}.$$

[1] Common surrogate loss functions including logistic loss, exponential loss and square loss satisfy these conditions, since $K(\mathbf{x}, \mathbf{x}) \leq r^2$ and $\|\mathbf{w}\|_{\mathbb{F}} \leq \Lambda$.

Theorem 4 demonstrates that the estimation error converges to zero with an increasing number of labeled and unlabeled data. Meanwhile, the term of approximation error measures how well the hypothesis set is in approximating Bayes risk, which is not accessible in general (Mohri et al., 2012).

To conclude, the consistency guarantee (in Section 5.1) and estimation error bound (in Section 5.2) theoretically justify the effectiveness of our algorithm.

6. Experiment

In this section, we conduct experiments to examine performance of the proposed Eulac algorithm from the following three aspects.

(i) **Comparisons on benchmark datasets:** we compare various algorithms on benchmark datasets, to validate the efficacy of our approach.

(ii) **Comparisons with an increasing number of unlabeled data:** we examine empirical behavior with an increasing number of unlabeled data, to demonstrate effectiveness of our approach in utilizing unlabeled data.

(iii) **Performance in various environments:** we report performance of our algorithm in various environments, where the mixture ratio θ varies.

In all experiments, we randomly generate 10 class configurations for each dataset to simulate the augmentation of classes unless otherwise specified, where half of the total classes are chosen as new classes. For datasets whose class number is less than 4, we generate the maximum number of class configurations it can produce. In each class configuration, 500 instances are randomly selected as training data from known classes. The testing and unlabeled datasets contain 1000 instances sampled from the whole dataset. The instances sampling procedure also repeats 10 times.

6.1 Comparisons on Benchmark Datasets

In this part, we conduct experiments on benchmark datasets.

Datasets. We perform the empirical studies over 10 benchmark datasets, including 9 datasets: usps, segment, satimage, optdigits, pendigits, SenseVeh, landset, mnist and shuttle. The brief statistics of these datasets are listed in Table 1.

Contenders. We compare Eulac with six methods, including four without exploiting unlabeled data and two utilizing them. The four algorithms are,

- **OVR-SVM** is a powerful strategy for the multi-class classification problem (Rifkin and Klautau, 2004). In order to adapt OVR-SVM to the LAC problem, the algorithm predicts an instance as new when $\max_{k\in[K]} f_k < 0$, otherwise it predicts as the classical OVR-SVM.
- **W-SVM** (Scheirer et al., 2014) is an SVM-based algorithm, where both one-class SVM and binary SVM incorporating with extreme value theory (EVT) are used to predict for the new class.
- **OSNN** (Mendes-Junior et al., 2017) is a nearest neighbor-based algorithm, which predicts an instance as new class if it shares similar distances with two nearest neighbors from different classes.
- **EVM** (Rudd et al., 2018) is also based on the extreme value theory, and it uses non-linear radial basis functions.

Table 1: Statistics of datasets used in the experiments.

Index	Datasets	# class	# dim	$\|C\|$		
				min	max	avg
1	usps	10	256	708	1553	929
2	segment	7	19	330	330	330
3	satimage	6	36	626	1533	1073
4	optdigits	10	64	554	572	562
5	pendigits	10	16	1055	1144	1099
6	SenseVeh	3	100	12316	26423	20527
7	landset	6	73	626	1533	1073
8	mnist	10	780	6313	7877	7000
9	shuttle	7	9	10	45586	8286

Another two algorithms exploiting unlabeled data are,

- **LACU-SVM** (Da et al., 2014) is an SVM based algorithm that utilizes the geometry property of unlabeled data to tune the decision boundaries of classifiers.
- **PAC-iForest** (Liu et al., 2018) is an iForest(Liu et al., 2008) based method, which selects the rejection threshold by using unlabeled data to ensure a desired novelty detection ratio. We use PAC-iForest to detect new classes and SVM to classify known classes.

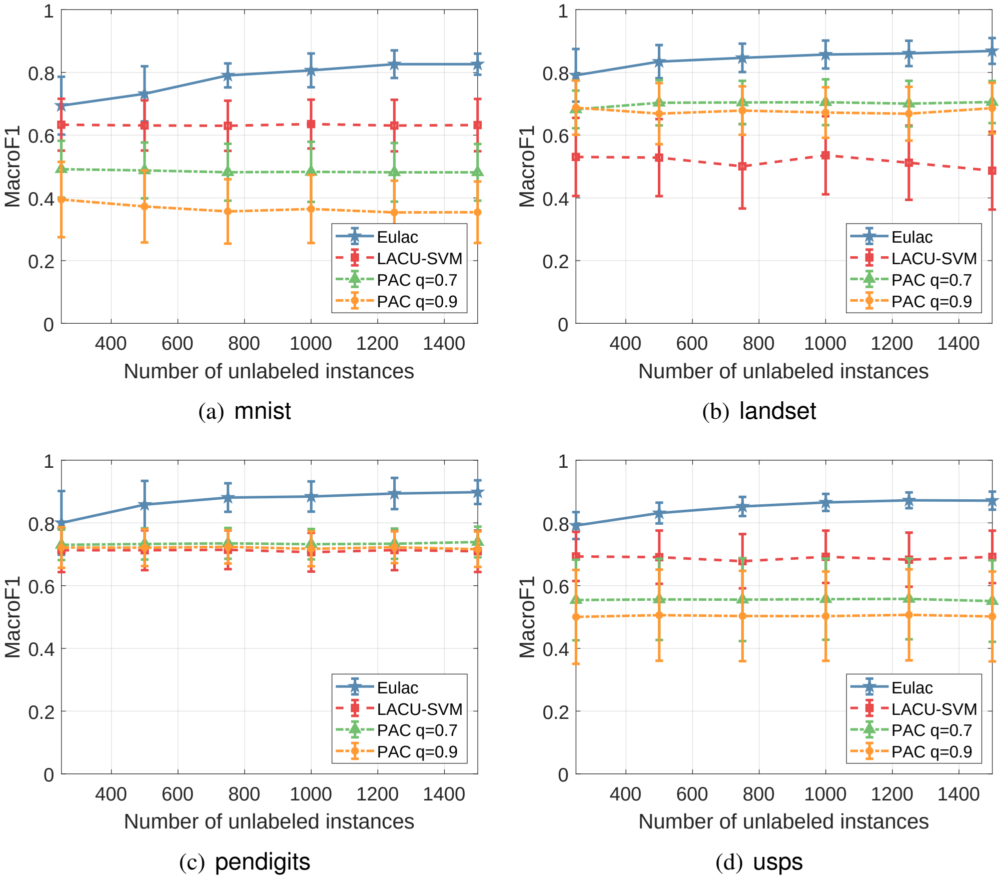
Parameters Setting. For all SVM-based algorithms (OVR-SVM, W-SVM, LACU-SVM, Eulac), we use the Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = \exp(\|\mathbf{x} - \mathbf{y}\|^2/2\sigma^2)$, where the bandwidth parameter σ for OVR-SVM, W-SVM and Eulac is selected from the candidate sets $\{10^{-2}, \ldots, 10\} \times$ median$_{i,j\in[n_l+n_u]}(\|\mathbf{x}_i - \mathbf{x}_j\|)$ and $\{10^{-2}, \ldots, 10\} \times$ median$_{i,j\in[n_l]}(\|\mathbf{x}_i - \mathbf{x}_j\|)$ by the 5-fold cross validation, respectively. For W-SVM and Eulac, the regularization parameter C is selected from the pool $\{10^{-3}, \ldots, 10^1\}$. While for OVR-SVM, this parameter is set as 1. Other parameters not specified, including parameters of LACU-SVM, are set according to corresponding papers. In the last, we use the square loss $\psi(z) = (1 - z)^2/4$ as the surrogate loss function for our algorithm.

For OSNN, we select the rejection threshold T by the cross validation method proposed by authors. When the number of known classes is 2, this parameter selection is not effective and we set $T = 0.75$. For EVM, we set $k = 4$ and the tail size τ is selected by cross validation. For PAC-iForest, it is unknown on how to set alien-detection rate q under the LAC setting, so we report results with $q = 0.7,\ 0.9$. Last, we adopt the approach of Ramaswamy et al. (2016) to estimate the mixture ratio θ, whose value is required by our algorithm and PAC-iForest as an input parameter.

Results. Table 2 reports performance in terms of Macro-F1 score. We can see that Eulac algorithm outperforms other contenders in most datasets, which validates the helpfulness of unlabeled data and effectiveness of our algorithm in exploiting them. Note that it is surprising that W-SVM and EVM achieve better performance than LACU-SVM and PAC-iForest, which are fed with unlabeled data. This indicates that the usage of unlabeled data does not necessarily improve performance in general. Another reason might be that these geometric-based methods require to set parameters empirically and the default one may not be proper in all datasets. By contrast, Eulac algorithm can perform an *unbiased* cross validation procedure to select proper parameters.

Table 2: Macro-F1 score comparisons on benchmark datasets. The best method is emphasized in bold. Besides, ● indicates that Eulac is significantly better than the compared methods (paired t-tests at 95% significance level) and – indicates numerical limits or errors.

Dataset	OVR-SVM	W-SVM	OSNN	EVM	LACU-SVM	PAC-iForest		Eulac
						$q = 0.7$	$q = 0.9$	
usps	75.42 ± 4.87 ●	79.77 ± 4.97 ●	63.14 ± 8.91 ●	61.14 ± 6.27 ●	69.20 ± 8.34 ●	55.69 ± 13.3 ●	50.27 ± 14.2 ●	**86.52** ± **2.72**
segment	71.78 ± 5.12 ●	80.82 ± 9.38 ●	85.10 ± 5.98	82.13 ± 5.88 ●	40.69 ± 12.5 ●	63.64 ± 13.1 ●	57.60 ± 17.7 ●	**86.17** ± **5.80**
satimage	54.67 ± 9.80 ●	76.29 ± 13.2 ●	62.48 ± 11.2 ●	72.10 ± 8.16 ●	51.56 ± 17.3 ●	60.76 ± 7.79 ●	56.94 ± 11.1 ●	**81.25** ± **6.18**
optdigits	80.11 ± 3.80 ●	87.82 ± 4.64 ●	86.97 ± 3.79 ●	72.00 ± 8.33 ●	80.92 ± 3.68 ●	71.65 ± 5.46 ●	69.54 ± 8.86 ●	**91.54** ± **2.95**
pendigits	72.78 ± 5.19 ●	87.79 ± 3.95	86.69 ± 3.39 ●	**89.94** ± **1.30**	70.66 ± 6.18 ●	73.21 ± 4.52 ●	71.74 ± 3.59 ●	88.41 ± 4.81
SenseVeh	48.07 ± 3.80 ●	45.96 ± 2.32 ●	49.91 ± 6.88 ●	51.24 ± 3.91 ●	51.61 ± 3.31 ●	54.12 ± 7.19 ●	33.63 ± 3.37 ●	**77.33** ± **2.17**
landset	60.43 ± 7.65 ●	68.91 ± 17.0 ●	73.25 ± 9.23 ●	76.00 ± 7.79 ●	53.59 ± 9.88 ●	70.50 ± 7.16 ●	67.20 ± 6.69 ●	**85.70** ± **4.46**
mnist	66.74 ± 2.76 ●	75.38 ± 4.62 ●	57.75 ± 10.9 ●	58.39 ± 5.94 ●	63.53 ± 7.58 ●	48.31 ± 9.62 ●	36.46 ± 10.5 ●	**80.66** ± **5.38**
shuttle	37.39 ± 14.1 ●	58.48 ± 34.5 ●	48.21 ± 16.4 ●	–	34.18 ± 13.4 ●	29.36 ± 8.70 ●	24.39 ± 13.5 ●	**66.49** ± **17.9**
Eulac w/ t/ l	9/ 0/ 0	8/ 1/ 0	8/ 1/ 0	8/ 1/ 0	9/ 0/ 0	9/ 0/ 0	9/ 0/ 0	rank first 8/ 9



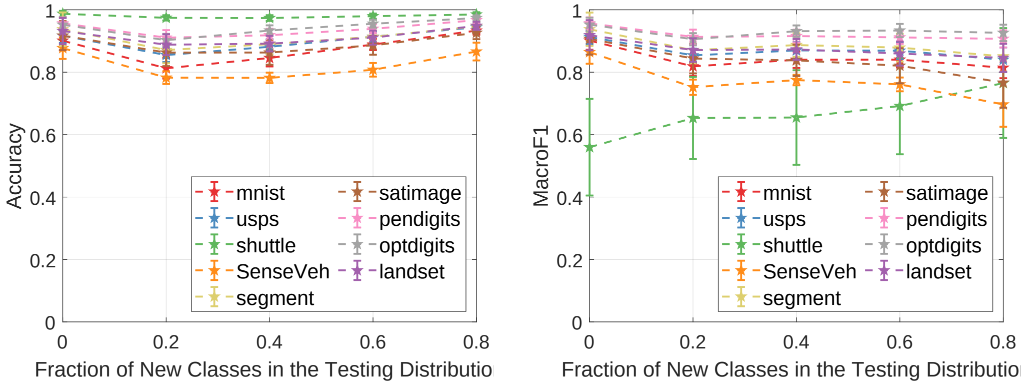
(a) mnist

(b) landset

(c) pendigits

(d) usps

Figure 2: Macro-F1 score comparisons of Eulac, LACU-SVM, and PAC-iForest when the number of unlabeled data increases.



(a) Accuracy

(b) Macro-F1

Figure 3: Performance of our approach in various environments.

6.2 Comparisons with Increasing Number of Unlabeled Data

We compare the Eulac algorithm with cnotenders to examine its effectiveness in exploiting unlabeled data. Concretely, we vary the size of unlabeled data from 250 to 1500 with an interval of 250 on 4 datasets: mnist, landset, pendigits, and usps. Figure 2 presents results of Macro-F1 score.

The results show that the score of LACU-SVM remains unchanged or even drops in four dataset, while performance of our algorithm improves when provided with more unlabeled data, which is in accordance with the theoretical analysis in Section 5. This again validates that our algorithm can exploit unlabeled data effectively. Notice that PAC-iForest also enjoys theoretical guarantees, nevertheless, the guarantees are only for the novelty detection ratio and thus the overall performance over the testing distribution is not promised to be improved, as validated in experiments.

6.3 Performance in Various Environments

We are curious about performance of our algorithm in various environments where the fraction of new classes increases. To this end, we conduct experiments on 9 datasets with the unknown class ratio ranging from 0, 0.2, 0.6, 0.8. The mixture ratio is supposed to be known in advance. Figures 3(a) and 3(b) show performance variation in terms of accuracy and Macro-F1.

We observe that our algorithm retains high performance in most cases with a changing mixture ratio θ, which verifies the adaptivity of our algorithm in various environments under different unknown class ratios.

7. Conclusion

In this paper, we discover that it is achievable to establish an unbiased risk estimator for the LAC problem by exploiting unlabeled data. The key observation is that the distribution of new classes can be effectively approximated by distributions of labeled and unlabeled data, which enables us to evaluate the risk of an classifier over testing distribution in the training stage. Subsequently, we develop the EULAC algorithm to perform empirical minimization of this unbiased risk, and the approach enjoys nice theoretical properties. Notably, we provide generalization bound over the *testing* distribution, which demonstrates that performance of our approach improves with an increasing number of labeled and *unlabeled* data, and thus theoretically justifies the effectiveness in exploiting unlabeled data. Extensive empirical studies also validate efficacy of the proposed approach.

References

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, pages 1823–1840, 2008.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning, Machine Learning Summer Schools 2003*, pages 169–207, 2003.

Xin-Qiang Cai, Peng Zhao, Kai Ming Ting, Xin Mu, and Yuan Jiang. Nearest neighbor ensembles: An effective method for difficult problems in streaming classification with emerging new classes. In *Proceedings of the 19th International Conference on Data Mining (ICDM)*, 2019.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, pages 41–46, 1970.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *Proceedings of International Conference on Algorithmic Learning Theory (ALT)*, pages 67–82, 2016a.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 1660–1668, 2016b.

Qing Da, Yang Yu, and Zhi-Hua Zhou. Learning with augmented class by exploiting unlabeled data. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*, pages 1760–1766, 2014.

Thomas G. Dietterich. Steps toward robust artificial intelligence. *AI Magazine*, pages 3–24, 2017.

Thomas G. Dietterich. Robust artificial intelligence and robust human organizations. *Frontiers Computer Science*, pages 1–3, 2019.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 703–711, 2014.

Michael Fink, Shai Shalev-Shwartz, Yoram Singer, and Shimon Ullman. Online multiclass learning by interclass hypothesis sharing. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 313–320, 2006.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2151–2159, 2019.

Radu Herbei and Marten H Wegkamp. Classification with reject option. *Canadian Journal of Statistics*, pages 709–721, 2006.

Hideko Kawakubo, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Computationally efficient class-prior estimation under class balance change using energy distance. *IEICE Transactions on Information and System*, pages 176–186, 2016.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, pages 413–422, 2008.

Si Liu, Risheek Garrepalli, Thomas G. Dietterich, Alan Fern, and Dan Hendrycks. Open category detection with PAC guarantees. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3175–3184, 2018.

Pedro Ribeiro Mendes-Junior, Roberto Medeiros de Souza, Rafael de Oliveira Werneck, Bernardo V. Stein, Daniel V. Pazinato, Waldir R. de Almeida, Otávio A. B. Penatti, Ricardo da Silva Torres, and Anderson Rocha. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, pages 359–386, 2017.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2012.

Xin Mu, Kai Ming Ting, and Zhi-Hua Zhou. Classification under streaming emerging new classes: A solution using completely-random trees. *IEEE Transactions on Knowledge and Data Engineering*, 29(8):1605–1618, 2017.

Michael D. Muhlbaier, Apostolos Topalis, and Robi Polikar. Learn^{++}.nc: Combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes. *IEEE Transactions on Neural Networks and Learning Systems*, pages 152–168, 2009.

Harish G. Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2052–2060, 2016.

Ryan M. Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, pages 101–141, 2004.

Ethan Rudd, Lalit P. Jain, Walter J. Scheirer, and Terrance Boult. The extreme value machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Walter J. Scheirer, Anderson Rocha, Archana Sapkota, and Terrance E. Boult. Towards open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

Walter J. Scheirer, Lalit P. Jain, and Terrance E. Boult. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2317–2324, 2014.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

Clayton Scott and Gilles Blanchard. Novelty detection: Unlabeled data definitely help. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics AISTATS*, pages 464–471, 2009.

Wenbo Wang and Xingye Qiao. Learning confidence sets using support vector machines. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 4934–4943, 2018.

Ming Yuan and Marten H. Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, pages 111–130, 2010.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, pages 1225–1251, 2004.

Zhi-Hua Zhou and Zhaoqian Chen. Hybrid decision tree. *Knowledge-Based Systems*, pages 515–528, 2002.

Appendix A. Proof of Proposition 1

Proof. For simplicity, we substitute f_{nc} by f_{K+1} in the proof. First, recall that the risk of OVR strategy is defined as,

$$R_\psi(f_1, \ldots, f_K, f_{K+1}) = \mathbb{E}_{(\mathbf{x},y)\sim P_{te}} \left[\psi(f_y(\mathbf{x})) + \sum_{k=1, k\neq y}^{K+1} \psi(-f_k(\mathbf{x})) \right].$$

According to the class shift condition, we have

$$\begin{aligned} R_\psi(f_1, \ldots, f_K, f_{nc}) &= \mathbb{E}_{(\mathbf{x},y)\sim P_{te}} \left[\psi(f_y(\mathbf{x})) + \sum_{k=1, k\neq y}^{K+1} \psi(-f_k(\mathbf{x})) \right] \\ &= \underbrace{\theta \cdot \mathbb{E}_{(\mathbf{x},y)\sim P_{tr}} \left[\psi(f_y(\mathbf{x})) + \sum_{k=1, k\neq y}^{K+1} \psi(-f_k(\mathbf{x})) \right]}_{\texttt{term (a)}} \\ &\quad + \underbrace{(1-\theta) \cdot \mathbb{E}_{(\mathbf{x},y)\sim P_{new}} \left[\psi(f_y(\mathbf{x})) + \sum_{k=1, k\neq y}^{K+1} \psi(-f_k(\mathbf{x})) \right]}_{\texttt{term (b)}}. \end{aligned} \tag{9}$$

Since $p_{XY}^{new}(\mathbf{x}, y) = 0$ holds for all $\mathbf{x} \in \mathcal{X}$ and $y \neq nc$, we can reform the $\texttt{term (b)}$ as,

$$\texttt{term (b)} = (1-\theta) \cdot \mathbb{E}_{\mathbf{x}\sim p_X^{new}(\mathbf{x})} \left[\psi(f_{K+1}(\mathbf{x})) + \sum_{k=1}^{K} \psi(-f_k(\mathbf{x})) \right].$$

Since the marginal distribution of new classes $p_X^{new}(\mathbf{x})$ is unknown, we reduce it to the difference of the marginal distributions of training and testing data. Under the class shift condition, we have,

$$p_{XY}^{te}(\mathbf{x}, y) = \theta \cdot p_{XY}^{tr}(\mathbf{x}, y) + (1-\theta) \cdot p_{XY}^{new}(\mathbf{x}, y).$$

By summing over the label space $\mathcal{Y}$, we obtain the marginal distribution $p_X^{new}(\mathbf{x})$,

$$(1-\theta)\cdot p_X^{new}(\mathbf{x}) = p_X^{te}(\mathbf{x}) - \theta \cdot p_X^{tr}(\mathbf{x}),$$

where the $\texttt{term (b)}$ can be further converted to the following form,

$$\begin{aligned}\texttt{term (b)} &= (1-\theta)\cdot \mathbb{E}_{\mathbf{x}\sim p_X^{new}(\mathbf{x})}\left[\psi(f_{K+1}(\mathbf{x})) + \sum_{k=1}^{K}\psi(-f_k(\mathbf{x}))\right] \\ &= \mathbb{E}_{\mathbf{x}\sim p_X^{te}(\mathbf{x})}\left[\psi(f_{K+1}(\mathbf{x})) + \sum_{k=1}^{K}\psi(-f_k(\mathbf{x}))\right] \\ &\quad - \theta\cdot \mathbb{E}_{\mathbf{x}\sim p_X^{tr}(\mathbf{x})}\left[\psi(f_{K+1}(\mathbf{x})) + \sum_{k=1}^{K}\psi(-f_k(\mathbf{x}))\right].\end{aligned} \tag{10}$$

We complete the proof by plugging (10) into (9). □

Appendix B. Proofs of Theorem 1 and Theorem 2

Before showing proofs of Theorem 1 and Theorem 2, for self-contentedness, we introduce results regarding infinite-sample consistency (ISC) of OVR strategy orignially provided by Zhang (2004).

Theorem 5 (Theorem 10 of Zhang (2004)). *Consider the OVR method, whose surrogate loss function is defined as $\Psi_y(\boldsymbol{f}) = \psi(f_y) + \sum_{k=1,k\neq y}^{K}\psi(-f_k)$. Assume ψ is convex, bounded below, differentiable, and $\psi(z) < \psi(-z)$ when $z > 0$. Then, OVR method is infinite-sample consistency (ISC) on $\Omega = \mathbb{R}^K$ with respect to 0-1 classification risk.*

Then, we present the relationship between the risk of an ISC method and the Bayes error,

Theorem 6 (Theorem 3 of Zhang (2004)). *Let $\mathcal{B}$ be the set of all vector Borel measurable functions, which take values in $\mathbb{R}^K$. For $\Omega \subset \mathbb{R}^K$, let $\mathcal{B}_\Omega = \{\boldsymbol{f} \in \mathcal{B} : \forall \mathbf{x}, \boldsymbol{f}(\mathbf{x}) \in \Omega\}$. If $[\Psi_y(\cdot)]$ is ISC on Ω with respect to 0-1 classification risk, then for any $\epsilon_1 > 0$, there exists $\epsilon_2 > 0$ such that for all underlying Borel probability measurable D, and $\boldsymbol{f}(\cdot) \in \mathcal{B}_\Omega$,*

$$\mathbb{E}_{(\mathbf{x},y)\sim D}[\Psi_y(\boldsymbol{f}(\mathbf{x}))] \leq \inf_{\boldsymbol{f}'\in\mathcal{B}_\Omega} \mathbb{E}_{(\mathbf{x},y)\sim D}[\Psi_y(\boldsymbol{f}'(\mathbf{x}))] + \epsilon_2$$

implies

$$R(T(\boldsymbol{f}(\cdot))) \leq R^* + \epsilon_1,$$

where $T(\cdot)$ is defined as $T(\boldsymbol{f}(\mathbf{x})) := \arg\max_{i=1,\ldots,K,nc} f_i(\mathbf{x})$, and R^* is the optimal Bayes error.

For the OVR strategy, we can further obtain a more quantitative bound.

Theorem 7 (Theorem 11 of Zhang (2004)). *Under the assumptions of Lemma 5. The function $V_\psi(q) = \inf_{z\in\mathbb{R}}[q\psi(z) + (1-q)\psi(-z)]$ is concave on $[0,1]$. Assume that there exists a constant $c_\psi > 0$ such that*

$$(q-q')^2 \leq c_\psi^2\left(2V_\psi\left(\frac{q+q'}{2}\right) - V_\psi(q) - V_\psi(q')\right),$$

then we know that for any $\boldsymbol{f}(\cdot)$,

$$R(T(\boldsymbol{f}(\cdot))) - R^* \leq c_\psi\left(\mathbb{E}_{(\mathbf{x},y)\sim D}[\Psi_y(\boldsymbol{f}(\mathbf{x}))] - \inf_{\boldsymbol{f}'\in\mathcal{B}_\Omega}\mathbb{E}_{(\mathbf{x},y)\sim D}[\Psi_y(\boldsymbol{f}'(\mathbf{x}))]\right)^{1/2}.$$

The proofs of Theorem 1 and Theorem 2 are consequences of Proposition 2 and the above theorems. We provide proofs in the following,

Proof of Theorem 1. According to Proposition 2, the LAC risk R_{LAC} equals to the risk of OVR strategy $R_\psi = \mathbb{E}_{(\mathbf{x},y)\sim P_{te}}[\Psi_y(\boldsymbol{f}(\mathbf{x}))]$. Thus, to prove the infinity sample consistency of LAC risk R_{LAC}, it is sufficient to demonstrate such a property of OVR strategy over distribution P_{te}, which is shown as Theorem 5 and Theorem 6. □

Proof of Theorem 2. To prove Theorem 2, we first show the consistency of OVR strategy with square loss. It is easy to verify that, when taking $\psi(z) = (1-z)^2/4$, we have $V_\psi(q) = \inf_{z\in\mathbb{R}}[q\psi(z) + (1-q)\psi(-z)] = q(1-q)$, which is concave on $[0,1]$. As a consequence, the inequality

$$(q-q') \leq c_\psi^2 \left(2V_\psi(\frac{q+q'}{2}) - V_\psi(q) - V_\psi(q')\right)$$

holds for all $q, q' \in \mathbb{R}$ with $c_\psi = \sqrt{2}$.

According to Theorem 7, the excess risk with respect to the 0-1 loss function over P_{te} is bounded by that of the OVR method,

$$\begin{aligned} R(T(\boldsymbol{f}(\cdot))) - R^* \leq & \sqrt{2\Big(\mathbb{E}_{(\mathbf{x},y)\sim P_{te}}[\Psi_y(\boldsymbol{f}(\mathbf{x}))] - \inf_{\boldsymbol{f}'} \mathbb{E}_{(\mathbf{x},y)\sim P_{te}}[\Psi_y(\boldsymbol{f}'(\mathbf{x}))]\Big)} \\ = & \sqrt{2\Big(R_\psi(\boldsymbol{f}) - R_\psi^*\Big)}, \end{aligned}$$

where $R_\psi^* = \inf_{\boldsymbol{f}'} R_\psi(\boldsymbol{f}')$. Then by applying the equality of the risk of OVR strategy R_ψ and that of our algorithm R_{LAC} from Proposition 2, we complete the proof. □

Appendix C. Proof of Theorem 3

Proof. Recall that,

$$R_{LAC} = \theta \underbrace{\mathbb{E}_{(\mathbf{x},y)\sim P_{tr}}\left[f_{nc}(\mathbf{x}) - f_y(\mathbf{x})\right]}_{R_{LAC}^A} + \underbrace{\mathbb{E}_{\mathbf{x}\sim p_X^{te}(\mathbf{x})}\left[\psi(f_{nc}(\mathbf{x})) + \sum_{k=1}^K \psi(-f_k(\mathbf{x}))\right]}_{R_{LAC}^B}, \tag{11}$$

and

$$\widehat{R}_{LAC} = \theta \underbrace{\frac{1}{n_l}\sum_{i=1}^{n_l}(f_{nc}(\mathbf{x}_i) - f_{y_i}(\mathbf{x}_i))}_{\widehat{R}_{LAC}^A} + \underbrace{\frac{1}{n_u}\sum_{i=1}^{n_u}\left(\psi(f_{nc}(\mathbf{x}_i)) + \sum_{k=1}^K \psi(-f_k(\mathbf{x}_i))\right)}_{\widehat{R}_{LAC}^B}. \tag{12}$$

To obtain the generalization bound of $\widehat{R}_{LAC}$, it is sufficient to establish the generalization bound of $\widehat{R}_{LAC}^A$ and $\widehat{R}_{LAC}^B$.

Firstly, we study the generalization bound of $\widehat{R}_{LAC}^A$. With the kernel-based hypothesis set $\mathcal{F} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \Phi(\mathbf{x})\rangle \mid \|\mathbf{w}\|_{\mathbb{F}} \leq \Lambda\}$ and $K(\mathbf{x},\mathbf{x}) < r^2$, according to McDiarmid's inequality and the standard analysis for generalization bound based on Rademacher complexity (Mohri et al., 2012, Theorem 3.1), we have that

$$R^A_{LAC}(f_1,\ldots,f_K,f_{nc}) \leq \widehat{R}^A_{LAC}(f_1,\ldots,f_K,f_{nc}) + 2\widehat{\mathfrak{R}}_{D_L}(\widetilde{\mathcal{F}}) + 6\Lambda r\sqrt{\frac{2\log(2/\delta')}{n_l}} \tag{13}$$

holds with probability at least $1-\delta'$, where

$$\widetilde{\mathcal{F}} = \{(\mathbf{x},y) \mapsto \langle \mathbf{w}^{nc}, \Phi(\mathbf{x})\rangle - \langle \mathbf{w}^{y}, \Phi(\mathbf{x})\rangle \mid \|\mathbf{w}^1\|_{\mathbb{F}},\ldots,\|\mathbf{w}^K\|_{\mathbb{F}}, \|\mathbf{w}^{nc}\|_{\mathbb{F}} \leq \Lambda\}.$$

The Rademacher complexity of hypothesis set $\widehat{\mathfrak{R}}_{D_L}(\widetilde{\mathcal{F}})$ can be further bounded by,

$$\begin{aligned}
\widehat{\mathfrak{R}}_{D_L}(\widetilde{\mathcal{F}}) &= \frac{1}{n_l}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f_1,\ldots,f_K,f_{nc}\in\mathcal{F}} \sum_{i=1}^{n_l} \sigma_i(f_{nc}(\mathbf{x}_i) - f_{y_i}(\mathbf{x}_i))\right] \\
&\leq \frac{1}{n_l}\mathbb{E}_{\sigma}\left[\sup_{f_{nc}\in\mathcal{F}} \sum_{i=1}^{n_l}\sigma_i f_{nc}(\mathbf{x}_i)\right] + \frac{1}{n_l}\mathbb{E}_{\sigma}\left[\sup_{f_1,\ldots,f_K\in\mathcal{F}}\sum_{i=1}^{n_l}\sigma_i f_{y_i}(\mathbf{x}_i)\right] \\
&= \widehat{\mathfrak{R}}_{D_L}(\mathcal{F}) + \frac{1}{n_l}\mathbb{E}_{\sigma}\left[\sup_{f_1,\ldots,f_K\in\mathcal{F}}\sum_{i=1}^{n_l}\sum_{j\in[K]}\sigma_i f_j(\mathbf{x}_i)\cdot \mathbb{1}(y_i=j)\right] \\
&\leq \widehat{\mathfrak{R}}_{D_L}(\mathcal{F}) + \frac{1}{n_l}\sum_{j\in[K]}\mathbb{E}_{\sigma}\left[\sup_{f_j\in\mathcal{F}}\sum_{i=1}^{n_l}\sigma_i f_j(\mathbf{x}_i)\cdot\mathbb{1}(y_i=j)\right] \\
&= \widehat{\mathfrak{R}}_{D_L}(\mathcal{F}) + \frac{1}{n_l}\sum_{j\in[K]}\mathbb{E}_{\sigma}\left[\sup_{f_j\in\mathcal{F}}\sum_{i=1}^{n_l}\sigma_i f_j(\mathbf{x}_i)\cdot\left(\frac{2\mathbb{1}(y_i=j)-1}{2}+\frac{1}{2}\right)\right] \\
&\leq \widehat{\mathfrak{R}}_{D_L}(\mathcal{F}) + \frac{1}{n_l}\sum_{j\in[K]}\mathbb{E}_{\sigma}\left[\sup_{f_j\in\mathcal{F}}\sum_{i=1}^{n_l}\sigma_i f_j(\mathbf{x}_i)\cdot\frac{2\mathbb{1}(y_i=j)-1}{2}\right] \\
&\qquad + \frac{1}{n_l}\sum_{j\in[K]}\mathbb{E}_{\sigma}\left[\sup_{f_j\in\mathcal{F}}\sum_{i=1}^{n_l}\frac{1}{2}\sigma_i f_j(\mathbf{x}_i)\right] \\
&= \widehat{\mathfrak{R}}_{D_L}(\mathcal{F}) + \frac{K}{2}\widehat{\mathfrak{R}}_{D_L}(\mathcal{F}) + \frac{K}{2}\widehat{\mathfrak{R}}_{D_L}(\mathcal{F}) \\
&= (K+1)\widehat{\mathfrak{R}}_{D_L}(\mathcal{F})
\end{aligned}$$

By Theorem 5.5 of Mohri et al. (2012), the Rademacher complexity of the kernel-based hypothesis set is bounded by $\frac{\Lambda r}{\sqrt{n_l}}$, i.e., $\widehat{\mathfrak{R}}_{D_L}(\mathcal{F}) \leq \frac{\Lambda r}{\sqrt{n_l}}$. As a consequence, we can get the generalization bound for R^A_{LAC},

$$R^A_{LAC}(f_1,\ldots,f_K,f_{nc}) \leq \widehat{R}^A_{LAC}(f_1,\ldots,f_K,f_{nc}) + \frac{2(K+1)\Lambda r}{\sqrt{n_l}} + 6\Lambda r\sqrt{\frac{2\log(2/\delta')}{n_l}} \tag{14}$$

holds with probability at least $1-\delta'$ for all $f_1,\ldots,f_K,f_{nc}\in\mathcal{F}$.

Next, we turn to bound the term R^B_{LAC}. An argument similar to the one used to obtain (13) shows that,

$$R^B_{LAC}(f_1,\dots,f_K,f_{nc}) \leq \widehat{R}^B_{LAC}(f_1,\dots,f_K,f_{nc}) + 2\widehat{\mathfrak{R}}_{D_L}(\widetilde{\mathcal{F}}_\Psi) + 3(K+1)B_\psi\sqrt{\frac{\log(2/\delta')}{n_u}} \quad (15)$$

holds with probability at least $1-\delta'$ for all $f_1,\dots,f_K,f_{nc}\in\mathcal{F}$, where $B_\psi = \sup_{a\in[-\Lambda r,\Lambda r]}\psi(a)$ and $\widetilde{\mathcal{F}}_\Psi = \{\mathbf{x}\mapsto\psi(f_{nc}(\mathbf{x})) - \sum_{k=1}^K\psi(-f_k(\mathbf{x})) \mid f_1,\dots,f_K,f_{nc}\in\mathcal{F}\}$. According to the Talagrand's comparison inequality (Koltchinskii, 2011) and the fact $\widehat{\mathfrak{R}}_{D_L}(\mathcal{F}_1+\mathcal{F}_2)\leq\widehat{\mathfrak{R}}_{D_L}(\mathcal{F}_1)+\widehat{\mathfrak{R}}_{D_L}(\mathcal{F}_2)$, we have,

$$\widehat{\mathfrak{R}}_{D_L}(\widetilde{\mathcal{F}}_\Psi) \leq (K+1)L\widehat{\mathfrak{R}}_{D_L}(\mathcal{F}) = (K+1)L\frac{\Lambda r}{\sqrt{n_u}}, \quad (16)$$

where L is the Lipschitz constant of surrogate loss function ψ. Notice that some surrogate loss functions, whose first order derivative is unbounded (like square loss), are not Lipschitz continuous on $\mathbb{R}$. However, since $f_1,\dots,f_K,f_{nc}$ are bounded in $[-\Lambda r,\Lambda r]$, the Talagrand's Lemma is still applicable, where the term $R^B_{LAC}(f_1,\dots,f_K,f_{nc})$ can be also bounded following the same argument.

Plugging (16) into (15), we can get the generalization bound of R^B_{LAC} that

$$R^B_{LAC}(f_1,\dots,f_K,f_{nc}) \leq \widehat{R}^B_{LAC}(f_1,\dots,f_K,f_{nc}) + \frac{2(K+1)L\Lambda r}{\sqrt{n_u}} + 3(K+1)B_\psi\sqrt{\frac{\log(2/\delta')}{n_u}}. \quad (17)$$

holds with probability at least $1-\delta'$. Let $\delta'=\frac{\delta}{2}$ and sum (14) and (17) up, we can get that

$$\begin{aligned}
&R_{LAC}(f_1,\dots,f_K,f_nc)\\
=&\,\theta R^A_{LAC} + R^B_{LAC}\\
\leq&\widehat{R}_{LAC}(f_1,\dots,f_K,f_{nc}) + \theta\cdot\left(\frac{2(K+1)\Lambda r}{\sqrt{n_l}} + 6\Lambda r\sqrt{\frac{2\log(4/\delta)}{n_l}}\right)\\
&\quad + \frac{2(K+1)L\Lambda r}{\sqrt{n_u}} + 3(K+1)B_\psi\sqrt{\frac{\log(4/\delta)}{n_u}}\\
\leq&\widehat{R}_{LAC}(f_1,\dots,f_K,f_{nc}) + \frac{2(K+1)\Lambda r}{\sqrt{n_l}} + 6\Lambda r\sqrt{\frac{2\log(4/\delta)}{n_l}}\\
&\quad + \frac{2(K+1)L\Lambda r}{\sqrt{n_u}} + 3(K+1)B_\psi\sqrt{\frac{\log(4/\delta)}{n_u}}
\end{aligned}$$

holds with probability at least $1-\delta$, which finishes the proof. □

Appendix D. Proof of Corollary 4

Proof. Recall that optimization problem (5) is formulated as,

$$\min_{f_1,\dots,f_k,f_{nc}\in\mathbb{F}} \widehat{R}_{LAC}(f_1,\dots,f_K,f_{nc}) + \lambda\left(\sum_{i=1}^K\|f_i\|^2_{\mathbb{F}} + \|f_{nc}\|^2_{\mathbb{F}}\right),$$

whose regularization path (the set of solutions to these problems with varying regularization parameter λ) is identical to the corresponding optimization problem parameterized in terms of the constraint on the RKHS norm. Therefore, the optimal solution $(\widehat{f}_1, \ldots, \widehat{f}_K, \widehat{f}_{nc})$ of (5) with certain $\lambda > 0$ is also the solution for

$$\begin{aligned} \min_{f_1,\ldots,f_K,f_{nc}\in\mathbb{F}} \quad & \widehat{R}_{LAC}(f_1, \ldots, f_K, f_{nc}) \\ \text{s.t.} \quad & \sum_{k=1}^{K} \|f_k\|_{\mathbb{F}}^2 + \|f_{nc}\|_{\mathbb{F}}^2 \le c_\lambda^2, \end{aligned} \tag{18}$$

with certain $c_\lambda > 0$.

Thus, to obtain the estimation error bound of (5), it is equal to consider the optimization problem (18), where $\widehat{R}_{LAC}$ is minimized on the hypothesis set $\mathscr{F} = \{(f_1, \ldots, f_K, f_{nc}) \mid f_1, \ldots, f_K, f_{nc} \in \mathbb{F}, \sum_{k=1}^{K} \|f_k\|_{\mathbb{F}}^2 + \|f_{nc}\|_{\mathbb{F}}^2 \le c_\lambda^2\}$. Since

$$(f_1^*, \ldots, f_K^*, f_{nc}^*) = \arg\min_{\boldsymbol{f}\in\mathscr{F}} R_{LAC}(f_1, \ldots, f_K, f_{nc}),$$

where $\boldsymbol{f} = (f_1, \ldots, f_K, f_{nc})$, we have

$$\begin{aligned} & R_{LAC}(\widehat{f}_1, \ldots, \widehat{f}_K, \widehat{f}_{nc}) - R_{LAC}(f_1^*, \ldots, f_K^*, f_{nc}^*) \\ =& R_{LAC}(\widehat{f}_1, \ldots, \widehat{f}_K, \widehat{f}_{nc}) - \widehat{R}_{LAC}(\widehat{f}_1, \ldots, \widehat{f}_K, \widehat{f}_{nc}) \\ & \qquad\qquad + \widehat{R}_{LAC}(\widehat{f}_1, \ldots, \widehat{f}_K, \widehat{f}_{nc}) - R_{LAC}(f_1^*, \ldots, f_K^*, f_{nc}^*) \\ \le& R_{LAC}(\widehat{f}_1, \ldots, \widehat{f}_K, \widehat{f}_{nc}) - \widehat{R}_{LAC}(\widehat{f}_1, \ldots, \widehat{f}_K, \widehat{f}_{nc}) \\ & \qquad\qquad + \widehat{R}_{LAC}(f_1^* \ldots, f_K^*, f_{nc}^*) - R_{LAC}(f_1^*, \ldots, f_K^*, f_{nc}^*) \\ \le& 2 \sup_{\boldsymbol{f}\in\mathscr{F}} |R_{LAC}(f_1, \ldots, f_K, f_{nc}) - \widehat{R}_{LAC}(f_1, \ldots, f_K, f_{nc})| \end{aligned}$$

Let $\mathcal{F}_\lambda = \{\mathbf{x} \mapsto \langle \mathbf{w}, \Phi(\mathbf{x})\rangle \mid \|\mathbf{w}\|_{\mathbb{F}} \le c_\lambda\}$. Since $\mathscr{F} \subset \mathcal{F}_\lambda^{K+1}$, we have

$$\begin{aligned} & R_{LAC}(\widehat{f}_1, \ldots, \widehat{f}_K, \widehat{f}_{nc}) - R_{LAC}(f_1^*, \ldots, f_K^*, f_{nc}^*) \\ \le& 2 \sup_{\boldsymbol{f}\in\mathcal{F}_\lambda^{K+1}} |R_{LAC}(f_1, \ldots, f_K, f_{nc}) - \widehat{R}_{LAC}(f_1, \ldots, f_K, f_{nc})|. \end{aligned}$$

According to Theorem 3 and under the same assumptions, the right hand side can be further bounded by

$$\begin{aligned} & 2 \sup_{\boldsymbol{f}\in\mathcal{F}_\lambda^{K+1}} |R_{LAC}(f_1, \ldots, f_K, f_{nc}) - \widehat{R}_{LAC}(f_1, \ldots, f_K, f_{nc})| \\ \le& \frac{4(K+1)c_\lambda r}{\sqrt{n_l}} + 6c_\lambda r\sqrt{\frac{2\log(4/\delta)}{n_l}} + \frac{4(K+1)Lc_\lambda r}{\sqrt{n_u}} + 12(K+1)B_\psi\sqrt{\frac{\log(4/\delta)}{n_u}}, \end{aligned}$$

which completes the proof. □