Handling New Class in Online Label Shift

Yu-Yang Qian*, Yong Bai*, Zhen-Yu Zhang, Peng Zhao, Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
{qianyy, baiy, zhangzy, zhaop, zhouzh}@lamda.nju.edu.cn

***Abstract*—In many real-world applications, data are continuously accumulated within open environments. For instance, in disease diagnosis, the prevalence of diseases can vary across seasons, and new types of diseases can emerge. This paper investigates the problem of learning from unlabeled data where the label distribution evolves over time, and meanwhile, previously unseen new class appears in the data stream. To handle the new class in online label shift, we first design a novel risk estimator by unbiased risk rewriting and mixture proportion estimation. Subsequently, we employ the online ensemble paradigm for model updating to handle unknown distribution shifts. The proposed approach enjoys a theoretical guarantee of dynamic regret, ensuring its effectiveness in adapting to the changing label distribution and the presence of the new class in streams. Experiments conducted on diverse benchmark datasets and two real-world applications demonstrate the effectiveness of the proposed algorithm.**

***Index Terms*—data stream, distribution shift, new class, weakly supervised learning, online label shift**

I. Introduction

Machine learning algorithms have made significant successes across various applications. These approaches, such as deep learning [1], typically rely on the assumption that the training and testing data are generated from an identical distribution. However, in many real-world tasks, the testing data are constantly collected from open environments, resulting in a distribution mismatch between the training and testing data, and the data distribution of testing data can even change over time [2, 3, 4, 5, 6]. Further, owing to the streaming nature of data, new class data could appear, presenting instances that were not encountered previously. Therefore, it is essential to adaptively learn from unlabeled data streams with changing distributions, particularly with the presence of new classes.

In this paper, we investigate the problem of handling new class in online label shift. Specifically, the learner can have some offline labeled data for model training. However, during the online testing phase, *unlabeled data* continuously arrives with its *distribution changing* over time; simultaneously, *new class data* may appear in online unlabeled data stream, as shown in Figure 1. The learner is required to continuously adapt to the changing distribution and accommodate the arrival of new class. This problem is crucial because it encompasses various real-world tasks. For instance, considering disease diagnosis tasks, the prevalence of diseases may vary across seasons [7], which induces continuous label shifts. Moreover, the appearance of new diseases, such as COVID-19, that were not encountered in the historical labeled data, poses a significant challenge in handling these new classes.

Existing approaches primarily focused on either handling online label shift within a fixed label space, or dealing with new classes while employing a fixed classifier for known classes. As a typical kind of distribution change, online label shift, characterized by continuous changes in the label distribution of unlabeled data stream, has garnered substantial interest in the literature [8, 9]. This line of research firstly estimates the underlying loss of online unlabeled data in an unbiased manner, followed by formulating the problem as an online convex optimization problem. However, these studies do not consider the appearance of new classes in the open environments, which is a common occurrence in various real-world tasks. Furthermore, research on handling the new classes focuses on handling only the new classes within unlabeled data stream [10, 11, 12]. This line of research uses various anomaly detectors to detect new classes and updates models accordingly. However, these studies mainly concentrate on detecting new classes while disregarding distribution changes within the known class data, which may cause a degradation in the overall performance. It is noteworthy that in numerous real-world scenarios, the issue of label shifts and new class occurs simultaneously, posing potential challenges to the existing algorithms. Additionally, while these methods show remarkable performances, their theoretical properties remain unclear.

We initiate and investigate the problem of handling New class in Online Label Shift (N-OLS), which encompasses a wide range of real-world tasks. Although previous works have studied the new class and label shift problems separately, the *conjunction* of online label shift and the new class presents new challenges, especially under unlabeled data streams. On one hand, the presence of the new class can introduce bias to the estimator that is solely trained on known classes. On the other hand, label shifts in known classes data can worsen the identification of the new classes. Therefore, it is crucial to adaptively learn the model in the online label shift setting with the new class. To this end, we explore the unlabeled data and develop a novel risk estimator for this problem employing unbiased risk rewriting and mixture proportion estimation techniques, enabling updates of the model under unknown level of distribution shift. To adapt to the continuous label shift in data streams, we employ the *online ensemble* paradigm [13], which maintains a group of base learners and adaptively combines their outputs to track the changing distribution. The proposed algorithm, HAndling New class in Online Label shift (HANOL), enjoys a theoretical guarantee of dynamic regret, ensuring its effectiveness in adapting to the evolving distribution. Extensive experiments are conducted, including

*Equal contribution.

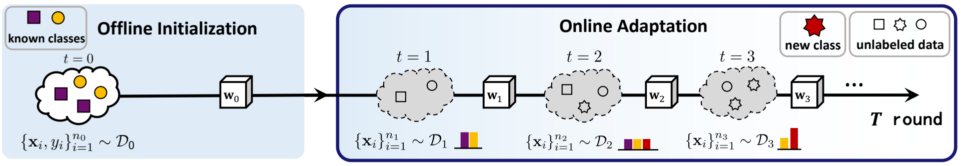

known classes
Offline Initialization
$t=0$
$\mathbf{w}_0$
$\{\mathbf{x}_i, y_i\}_{i=1}^{n_0} \sim \mathcal{D}_0$
Online Adaptation
new class
unlabeled data
$t=1$
$t=2$
$t=3$
$\mathbf{w}_1$
$\mathbf{w}_2$
$\mathbf{w}_3$
...
T round
$\{\mathbf{x}_i\}_{i=1}^{n_1} \sim \mathcal{D}_1$
$\{\mathbf{x}_i\}_{i=1}^{n_2} \sim \mathcal{D}_2$
$\{\mathbf{x}_i\}_{i=1}^{n_3} \sim \mathcal{D}_3$

Fig. 1: Illustration of the N-OLS problem. In the offline initialization stage, the learner observes labeled data from known classes; in the online adaptation stage, the learner receives a few unlabeled data containing the new class, and the distribution is changing over time.

benchmark and two real-world applications SHL [14] and fMoW [15]. Our approach enhances the average accuracy by 11% on SHL and 10% on fMoW datasets.

II. Problem Formulation

We consider a multi-class classification setting. The feature space is denoted by $\mathcal{X} \subseteq \mathbb{R}^d$, where d represents the feature dimension. Here, we denote $[K] \triangleq \{1, \ldots, K\}$ as the known classes in the initial offline labeled data, and nc as the new class, which does not appear in the offline data but is encountered in online unlabeled data streams. Therefore, the total label space is $\mathcal{Y} = [K] \cup \{\mathsf{nc}\}$, classes $[K] \triangleq \{1, \ldots, K\}$, which consists of $K+1$ classes in total.

In addition to the presence of new class, we consider the issue of online label shift. Specifically, throughout the entire time horizon of unlabeled data stream, conditional distribution remains unchanged, i.e., $\mathcal{D}_t(\mathbf{x} \mid y) = \mathcal{D}_0(\mathbf{x} \mid y)$ for all $\mathbf{x} \in \mathcal{X}, y \in [K]$ and $t \in [T]$; $\mathcal{D}_t(\mathbf{x} \mid y) = \mathcal{D}_{t-1}(\mathbf{x} \mid y)$ for all $\mathbf{x} \in \mathcal{X}, y = K+1$ and $t \geq 2$. However, the label distribution can change dynamically, i.e., $\mathcal{D}_t(y = j) \neq \mathcal{D}_{t-1}(y = j)$ for $j \in [K+1]$. Additionally, for every $j \in [K]$, $\mathcal{D}_0(y = j) > 0$.

In this paper, we formulate the New class in Online Label Shift (N-OLS) problem into the following two phases:

- *Offline supervised initialization.* Before adaptation, the learner has a certain number of labeled data $S_0 = \{\mathbf{x}_i, y_i\}_{i=1}^{n_0}$ from the initial distribution $\mathcal{D}_0(\mathbf{x}, y)$ defined over the known classes $\mathcal{X} \times [K]$, $\mathcal{D}_0(\mathbf{x}) = \sum_{j=1}^{K} [\boldsymbol{\mu}_{y_0}]_j \cdot \mathcal{D}_0^j(\mathbf{x})$, where $[\boldsymbol{\mu}_{y_0}]_j = \mathcal{D}_0(y = j)$ is the label prior for the j-th class, $\mathcal{D}_0^j(\mathbf{x}) = \mathcal{D}_0(\mathbf{x} \,|\, y = j)$ is the marginal distribution of the feature $\mathbf{x}$ over the known class $j \in [K]$. As an initialization, we suppose that we can have a labeled training dataset to obtain a well-performed model $f_0 : \mathcal{X} \mapsto \mathcal{Y}$, which serves as a reliable classifier for known class data.
- *Online unsupervised adaptation.* After obtaining the initial model f_0, the learner deploys it to a fully unsupervised changing environment. At round $t \in [T]$, the learner can receive a small number of *unlabeled data* $S_t = \{\mathbf{x}_i\}_{i=1}^{n_t}$ drawn from the current distribution $\mathcal{D}_t(\mathbf{x})$. Note that the label distribution in the online adaptation phase comprises not only the known classes $y \in [K]$, but also a new class $y \in \mathsf{nc}$, absent in the offline data, and is changing over time. The learner must sequentially explore the unlabeled data stream to adaptively update the model $\mathbf{w}_t$ and make accurate predictions for each S_t.

III. Proposed Approach

In this section, we present our approach for the N-OLS problem, with the overall protocol illustrated in Figure 2.

A. Risk Estimator for N-OLS Problem

In this part, we propose a new risk estimator designed for the N-OLS problem, employed to update the model by leveraging both the unlabeled and offline data. The estimator is designed by exploiting unlabeled data stream via the risk rewriting technique. We denote $R_t^k(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_t(\mathbf{x} \mid y=k)}[\ell(f(\mathbf{w}, \mathbf{x}), k)]$ as the risk of the model $\mathbf{w}$ over the k-th class at round t, where $t \in \{0\} \cup [T]$. Then we have $R_t^k(\mathbf{w}) = R_0^k(\mathbf{w})$ for the known classes $k \in [K]$ due to the online label shift assumption $\mathcal{D}_t(\mathbf{x} \mid y) = \mathcal{D}_0(\mathbf{x} \mid y)$. However, since new class can appear in the online unlabeled data stream, label distribution of the new class is unavailable, making risk $R_0^{\mathsf{nc}}(\mathbf{w}) \triangleq R_0^{K+1}(\mathbf{w})$ unknown. To tackle this issue, we propose a novel estimator for the expected online risk R_t. Notice that the marginal distribution $\mathcal{D}_t(\mathbf{x})$ can be decomposed as

$$\mathcal{D}_t(\mathbf{x}) = (1 - \theta_t)\mathcal{D}_t^{\mathsf{nc}}(\mathbf{x}) + \theta_t \left(\sum_{j=1}^{K} [\boldsymbol{\mu}_{y_t}]_j \mathcal{D}_0(\mathbf{x} \,|\, j) \right), \quad (1)$$

where $\mathcal{D}_t^{\mathsf{nc}}$ is the distribution of the new class in $\mathcal{D}_t$, $\boldsymbol{\mu}_{y_t} \in \Delta_K$ is the label distribution vector of known classes, and $(1 - \theta_t) \in [0, 1]$ is the proportion of new class at round t. By Eqn. (1), we rewrite the risk associated with new class data $R_t^{\mathsf{nc}}(\mathbf{w})$ as

$$\begin{aligned}
&(1 - \theta_t) R_t^{\mathsf{nc}}(\mathbf{w}) \triangleq (1 - \theta_t) \mathbb{E}_{\mathcal{D}_t^{\mathsf{nc}}(\mathbf{x})}[\ell(f(\mathbf{w}, \mathbf{x}), \mathsf{nc})] \\
&= \mathbb{E}_{\mathcal{D}_t(\mathbf{x})}[\ell(f(\mathbf{w}, \mathbf{x}), \mathsf{nc})] - \theta_t \mathbb{E}_{\mathcal{D}_t^{\mathsf{kc}}(\mathbf{x})}[\ell(f(\mathbf{w}, \mathbf{x}), \mathsf{nc})] \\
&= \mathbb{E}_{\mathcal{D}_t(\mathbf{x})}[\ell(f(\mathbf{w}, \mathbf{x}), \mathsf{nc})] - \theta_t \sum_{j=1}^{K} [\boldsymbol{\mu}_{y_t}]_j \mathbb{E}_{\mathcal{D}_0^j(\mathbf{x})}[\ell(f(\mathbf{w}, \mathbf{x}), \mathsf{nc})].
\end{aligned}$$

The expected risk over distribution $\mathcal{D}_t(\mathbf{x})$ can be approximated by the empirical risk over the unlabeled data S_t, given by $1/n_t \sum_{\mathbf{x} \in S_t} \ell(f(\mathbf{w}, \mathbf{x}), \mathsf{nc})$, while the risk over distribution $\mathcal{D}_0^j(\mathbf{x}) \triangleq \mathcal{D}_0(\mathbf{x} \,|\, y = j)$ can be approximated by the empirical risk over offline data S_0. Hence, we can build an estimator $\widehat{R}_t(\mathbf{w})$ for the online expected risk $R_t(\mathbf{w})$ as follows.

$$\begin{aligned}
\widehat{R}_t(\mathbf{w}) &= \widehat{\theta}_t \widehat{R}_t^{\mathsf{kc}}(\mathbf{w}) + (1 - \widehat{\theta}_t) \widehat{R}_t^{\mathsf{nc}}(\mathbf{w}) \\
&= \widehat{\theta}_t \sum_{j=1}^{K} [\widehat{\boldsymbol{\mu}}_{y_t}]_j R_0^j(\mathbf{w}) + \mathbb{E}_{S_t(\mathbf{x})}[\ell(f(\mathbf{w}, \mathbf{x}), \mathsf{nc})] \\
&\quad - \widehat{\theta}_t \sum_{j=1}^{K} [\widehat{\boldsymbol{\mu}}_{y_t}]_j \mathbb{E}_{S_0^j(\mathbf{x})}[\ell(f(\mathbf{w}, \mathbf{x}), \mathsf{nc})]. \quad (2)
\end{aligned}$$

Overall, we build an estimator $\widehat{R}_t(\mathbf{w})$ for the N-OLS problem by leveraging online unlabeled data and offline labeled data. This estimator will be unbiased, provided that we accurately determine the values of $\widehat{\theta}_t$ and $\widehat{\boldsymbol{\mu}}_{y_t}$. The remaining question is how to estimate the parameters $\widehat{\theta}_t$ and $\widehat{\boldsymbol{\mu}}_{y_t}$. In the following, we use black box shift estimator (BBSE) [16] to estimate the label distribution $\boldsymbol{\mu}_{y_t}$, and mixture proportion estimation (MPE) methods [17, 18] to estimate θ_t given that we can empirically observe $\mathcal{D}_0(\mathbf{x} \,|\, j)$ and $\mathcal{D}_t(\mathbf{x})$.

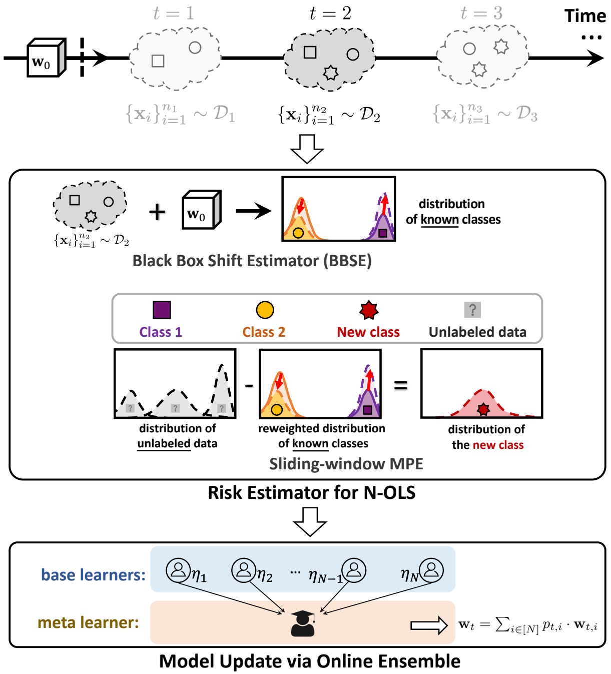

$t=1$
$t=2$
$t=3$
Time
...
$\mathbf{w}_0$
$\{\mathbf{x}_i\}_{i=1}^{n_1} \sim \mathcal{D}_1$
$\{\mathbf{x}_i\}_{i=1}^{n_2} \sim \mathcal{D}_2$
$\{\mathbf{x}_i\}_{i=1}^{n_3} \sim \mathcal{D}_3$
distribution of known classes
Black Box Shift Estimator (BBSE)
Class 1
Class 2
New class
Unlabeled data
distribution of unlabeled data
reweighted distribution of known classes
distribution of the new class
Sliding-window MPE
Risk Estimator for N-OLS
base learners:
η_1
η_2
η_{N-1}
η_N
meta learner:
$\mathbf{w}_t = \sum_{i \in [N]} p_{t,i} \cdot \mathbf{w}_{t,i}$
Model Update via Online Ensemble

Fig. 2: Overall protocol of our approach. We first develop a risk estimator by exploiting unlabeled data stream. Then, we employ the paradigm of online ensemble to adapt to the continuous label shift.

B. Label Distribution Estimation with Unlabeled Data

In this part, we introduce the details of estimating the changing label distribution $\boldsymbol{\mu}_{y_t}$ for known classes and the proportion of the new class θ_t in the N-OLS problem.

- *Estimate Proportion for Known Classes Data.* We use BBSE [16] to estimate the class prior of known classes $\boldsymbol{\mu}_{y_t}$ via solving $\widehat{\boldsymbol{\mu}}_{y_t} = C_0^{-1} \cdot \widehat{\boldsymbol{\mu}}_{\widehat{y}_t}$, where $C_0 \in \mathbb{R}^{K \times K}$ is the classifier's confusion matrix with $[C_0]_{i,j} \triangleq \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_0(\mathbf{x} \mid y=j)} \left[\mathbb{1}(f_0(\mathbf{x}) = i) \right]$ being the classification rate that the initial model f_0 predicts samples from class i as class j, and $\widehat{\boldsymbol{\mu}}_{\widehat{y}_t} \in \Delta_K$ with $[\widehat{\boldsymbol{\mu}}_{\widehat{y}_t}]_j = 1/n_t \cdot \sum_{\mathbf{x} \in S_t} [f_0(\mathbf{x})]_j$ is the estimated class prior of the prediction $f_0(\mathbf{x})$. Benefit from the benign properties of BBSE [16], we can guarantee that the estimation $\widehat{\boldsymbol{\mu}}_{y_t}$ satisfies $\mathbb{E}[\widehat{\boldsymbol{\mu}}_{y_t}] = \boldsymbol{\mu}_{y_t}$, where $\boldsymbol{\mu}_{y_t} \triangleq C_0^{-1} \boldsymbol{\mu}_{\widehat{y}_t} = \mathcal{D}_t^{\mathsf{kc}}(y)$ is the ground-truth label distribution.
- *Estimate Proportion for New Class Data.* Notice that the construction of the risk estimator $\widehat{R}_t$ requires to estimate the proportion θ_t, which is known as the problem of mixture proportion estimation (MPE) [17, 18], where one aims to estimate the proportion of a certain class in the whole distribution given their empirical observations. To tackle the challenge of limited data availability in the online stream, we propose a sliding window-based MPE algorithm. Specifically, we maintain a window queue of length L. At time t, following the first-in-first-out principle, the current round sample S_t is added to the queue, and a certain number of samples are removed from the queue's front. At each time step, inspired by the recently proposed Best Bin Estimation (BBE) technique [19], we utilize resampled offline labeled data and online unlabeled data in the sliding window to estimate the proportion of new class. We first train a well-performed classifier $h : \mathcal{X} \mapsto [0,1]$, where 0 means the data is sampled from the online unlabeled data, and 1 means data is sampled from labeled data, then we get
$$q_p(z) = \frac{\sum_{\mathbf{x}_i \in S_0} \mathbb{I}\left[h(\mathbf{x}_i) \geqslant z\right]}{|S_0|}, q_u(z) = \frac{\sum_{\mathbf{x}_i \in S_{\text{win}}} \mathbb{I}\left[h(\mathbf{x}_i) \geqslant z\right]}{|S_{\text{win}}|},$$
where S_0 is the offline dataset, and S_{win} is the online unlabeled data in the sliding window. By solving
$$\widehat{c} = \underset{c \in [0,1]}{\arg\min} \frac{q_u(c)}{q_p(c)} + \frac{1+\gamma}{q_p(c)} \left(\sqrt{\frac{\log(4/\delta)}{2 S_{\text{win}}}} + \sqrt{\frac{\log(4/\delta)}{2 S_0}} \right),$$
where $0 < \delta, \gamma < 1$ are the hyper-parameters, then, we can estimate new class proportion by $\widehat{\theta}_t = q_u(\widehat{c})/q_p(\widehat{c})$. The estimated proportion by sliding-window MPE enjoys a convergence rate of $\mathcal{O}(\min(|S_0|, |S_{\text{win}}|)^{-1/2})$ [19], thus can obtain the proportion of new class with a small variance.

Algorithm 1 HANOL: HAndling New class in Online Label shift

```
Require: step size pool $\mathcal{H}$; learning rate $\varepsilon$; step size $\eta_i \in \mathcal{H}$
initialization: get $\mathbf{w}_1^i \in \mathcal{W}$ by offline supervised initialization; $\forall i \in [N], p_1^i = 1/N$
for $t = 2$ to $T$ do
    for $i = 1$ to $N$ do
        construct risk estimator $\widehat{R}_t(\mathbf{w}_t^i)$ as (2)
        update the $i$-th base model $\mathbf{w}_t^i$ by (3)
        update the weight $p_t^i$ according to (4)
    end for
    output final model $\mathbf{w}_t = \sum_{i=1}^N p_t^i \cdot \mathbf{w}_t^i$
end for
```

Remark 1. Although the proposed algorithm primarily focuses on scenarios involving a *single* new class, it possesses practical potential in managing scenarios involving the *emerging* new classes [4], where more and more unseen new classes *successively* arise in the data stream as time evolves. Moreover, techniques such as core set and sketching can be employed to enhance the sample storage efficiency of S_0. Further details will be presented in the extended version.

C. Adaptation via Online Ensemble

Based on the risk estimator $\widehat{R}_t(\mathbf{w})$ constructed in Section III-A and the parameter estimating approach in Section III-B, we then design online adaptation algorithms to adapt model $\mathbf{w}_t$ to the changing distribution $\mathcal{D}_t$. A natural choice is to minimize the risk estimator $\widehat{R}_t(\mathbf{w})$ from scratch, which means $\mathbf{w}_t \in \arg\min_{\mathbf{w}} \widehat{R}_t(\mathbf{w})$. Whereas, $\widehat{R}_t(\mathbf{w})$ can suffer from high variance due to the small online sample size n_t, which may lead to poor generalization performance. To this end, we turn to reusing historical information via online gradient descent (OGD). However, OGD with a fixed step size may not be able to adapt to the changing distribution $\mathcal{D}_t$. To handle this issue, we propose an adaptive algorithm with a two-layer structure, which can adaptively track the suitable step size to the distribution $\mathcal{D}_t$, as shown in Algorithm 1.

In order to adapt to changing distributions, we employ the paradigm of online ensemble learning [20, 13]:

- *Construct base learners with multiple step sizes.* At round t, with risk estimator $\widehat{R}_t(\mathbf{w})$ in (2), we can obtain the estimated gradient $\nabla \widehat{R}_t(\mathbf{w})$ and update the model $\mathbf{w}_t$ by gradient descent, given by the following update schedule $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta \nabla \widehat{R}_t(\mathbf{w}_t)]$, where $\Pi_{\mathcal{W}}[\cdot]$ denotes the projection onto the domain $\mathcal{W}$ and $\eta > 0$ is the step size.

However, this OGD algorithm with a single step size cannot adapt to the changing distribution $\mathcal{D}_t$. To this end, following the idea of ensemble learning, we propose an online ensemble algorithm to adaptively track the changing distribution $\mathcal{D}_t$. Specifically, we maintain a set of base learners with different step sizes $\mathcal{H} = \{\eta_i\}_{i=1}^N$. At round t, we update the i-th base learner $\mathbf{w}_t^i$ by

$$\mathbf{w}_{t+1}^i = \Pi_{\mathcal{W}}\big[\mathbf{w}_t^i - \eta_i \nabla \widehat{R}_t(\mathbf{w}_t^i)\big]. \tag{3}$$

Thus, we obtain a group of base models $\{\mathbf{w}_t^i\}_{i=1}^N$, where different models excel in handling different shift intensities. By the following meta learner that adaptively assigns weights to different base models, our approach effectively tracks the varying environments in real time.

- *Combine the outputs by meta learner.* We maintain a meta learner that combines the outputs of multiple base learners through weighted averaging to obtain the final model, given by $\mathbf{w}_t = \sum_{i=1}^N p_t^i \cdot \mathbf{w}_t^i$. The weights $p_t^i \in [0,1]$ denotes the extent of utilization for the i-th base learner, and statistically satisfies $\sum_{i=1}^N p_t^i = 1$. Since the environment changes dynamically, the weights p_t^i should be adaptively updated according to the performance of the base learners,

$$p_t^i \propto \exp\big(-\varepsilon \sum\nolimits_{s=1}^{t-1} \widehat{R}_s\left(\mathbf{w}_s^i\right)\big), \tag{4}$$

where $\varepsilon > 0$ is a hyper-parameter that controls sensitivity of the meta learner to performances of base learners. Intuitively, meta learner assigns higher weights to base learners that exhibit better cumulative performance, thereby enabling adaptively tracking of the optimal base learner.

IV. Theoretical Analysis

In this section, we analyze the theoretical properties of our algorithm. We consider the convex feasible domain and loss functions. Our goal is to obtain a sequence of online model parameters $\{\mathbf{w}_t\}_{t=1}^T$ that can minimize the cumulative expected risk over the whole time horizon: $\sum_{t=1}^T R_t(\mathbf{w}_t)$. The excepted risk $R_t(\mathbf{w})$ at each round is defined as $\mathbb{E}_{\mathcal{D}_t(\mathbf{x},y)}[\ell(f(\mathbf{w},\mathbf{x}),y)]$, where $\ell : \mathbb{R}^{K+1} \times \mathcal{Y} \mapsto \mathbb{R}$ is any convex loss function and $f(\mathbf{w},\mathbf{x})$ is the prediction of the model $\mathbf{w} \in \mathcal{W}$ on the feature $\mathbf{x}$. We adopt the dynamic regret $\mathbf{Reg}_T^{\mathbf{d}}$ as the measure [9]. It is defined as the difference between the cumulative expected risk of the predictive model sequence $\{\mathbf{w}_t\}_{t=1}^T$ with $\{\mathbf{w}_t^*\}_{t=1}^T$:

$$\mathbf{Reg}_T^{\mathbf{d}} \triangleq \sum\nolimits_{t=1}^T R_t(\mathbf{w}_t) - \sum\nolimits_{t=1}^T R_t(\mathbf{w}_t^*),$$

where model parameter $\mathbf{w}_t^* \in \arg\min_{\mathbf{w}\in\mathcal{W}} R_t(\mathbf{w})$ is the best model at each round t. A small dynamic regret indicates that the algorithm can adapt to a changing environment and achieve a performance competitive with the best model sequence.

We denote the upper bound of gradient norm by $G \triangleq \sup_{\mathcal{X},\mathcal{Y},\mathcal{W}} \|\nabla \ell(f(\mathbf{w},\mathbf{x}),y)\|_2$ and denote the diameter of the convex parameter space $\mathcal{W}$ by $\Gamma \triangleq \sup_{\mathbf{w}_1,\mathbf{w}_2\in\mathcal{W}} \|\mathbf{w}_1 - \mathbf{w}_2\|_2$. We use $B \triangleq \sup_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y},\mathbf{w}\in\mathcal{W}} |\ell(f(\mathbf{w},\mathbf{x}),y)|$ as the upper bound of loss function value, and σ as the minimum singular value of the confusion matrix C_0. Under the assumption that the confusion matrix C_0 is invertible, i.e., $\sigma > 0$, our proposed algorithm enjoys the following dynamic regret guarantee.

Theorem 1 (Dynamic Regret)**.** *Suppose the confusion matrix C_0 is invertible. Set the step size pool as $\mathcal{H} = \{\eta_i = \frac{\sigma\Gamma}{2G\sqrt{(K+1)T}} \cdot 2^{i-1} \mid i \in [N]\}$, where $N = 1 + \lceil \frac{1}{2}\log_2(1+2T)\rceil$ is the number of base-learners. Our* Hanol *ensures that*

$$\mathbb{E}[\mathbf{Reg}_T^{\mathbf{d}}] \leq \mathcal{O}(\max\{V_T^{1/3}T^{2/3}, \sqrt{T}\}),$$

or simplified as $\mathcal{O}(V_T^{1/3}T^{2/3})$ for non-degenerated cases of $V_T \geq \Theta(T^{-\frac{1}{2}})$, where $V_T = \sum_{t=2}^T \|\mathcal{D}_t(y) - \mathcal{D}_{t-1}(y)\|_1$ measures the intensity of label distributions variation.

Proof Sketch. We convert the overall dynamic regret into two components: *meta regret* and *base regret*. The meta regret quantifies the gap between the ensemble model and individual base models, which is bounded by $\frac{2B}{\sigma}\sqrt{(\ln N + 2)(K+1)T}$. The base regret measures the gap between base model and the optimal model sequence, for which we introduce a piecewise stationary reference sequence with a total change count of Δ and decompose the base regret into two parts. The first part is the gap between the base model sequence and reference sequence, which can be bounded by $\frac{3G\Gamma}{\sigma}\sqrt{(K+1)T\left(1+\frac{2T}{\Delta}\right)}$. The second part is the gap between the reference sequence and optimal model sequence, which can be bounded by $2\Delta B V_T$. By combining upper bound of the meta regret with the base regret and tuning the change count Δ, we can establish the overall dynamic regret guarantee. Further, we theoretically demonstrate that our algorithm can track the optimal base learner adaptively by only maintaining about $\log T$ base learners. Detailed proofs will be provided in the extended version.

V. Experiments

In this section, we present the empirical evaluation of our approach, which aims to answer the following questions:

- **Q1:** Does Hanol outperform other contenders in the N-OLS problem when confronted with various types of shifts?
- **Q2:** Does Hanol show effectiveness in real-world tasks with the arrival of new classes and continuous label shift?
- **Q3:** Does Hanol correctly detect the shifts and estimate the proportion of the new class? Is it efficient?

A. Benchmark Datasets

This section seeks to answer **Q1**. We compare our proposed approach Hanol with seven competing methods using five benchmark datasets in the N-OLS scenario. The competing methods comprise a baseline approach (*FIX*), two for managing distribution shifts (*FTFWH* [8] and *ASL* [21]), two for handling the new classes in data streams (*SENC-F* [10] and *KNNENS* [22]), and two originally designed methods to tackle the offline N-OLS problem (*Self-N* and *PULSE* [23]). The details of the competitors are deferred to the extended version.

We generate a changing environment where the label distributions shift over time, and new class data appear in online stage, which is not contained in the offline training data.

TABLE I: Average error (%) of different algorithms on benchmark datasets with different types of environmental shifts, where HANOL represents our method. We report the mean and standard deviation over five runs. The best algorithms are emphasized in bold. "○" indicates the algorithm is significantly inferior to our algorithms by paired t-test at a 5% significance level. The online sample size is set as $n_t = 10$.

Gradual Shift

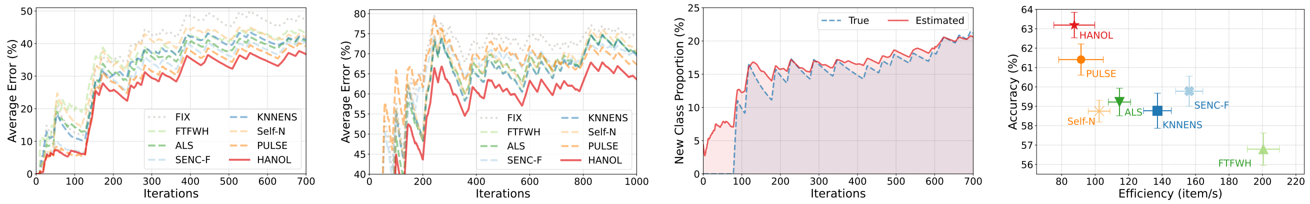
	FIX	FTFWH	ASL	SENC-F	KNNENS	Self-N	PULSE	HANOL
CIFAR10	22.89 ± 0.81 ○	19.01 ± 0.73 ○	19.23 ± 0.97 ○	18.92 ± 0.89 ○	19.23 ± 0.79 ○	19.11 ± 0.85 ○	18.71 ± 0.85	**18.52** ± **0.89**
CINIC10	35.57 ± 1.03 ○	30.08 ± 1.08 ○	31.45 ± 0.24 ○	30.25 ± 0.93 ○	30.12 ± 1.05 ○	31.95 ± 0.86 ○	29.89 ± 1.12 ○	**28.82** ± **0.96**
EuroSAT	16.23 ± 0.03 ○	10.82 ± 0.21 ○	11.24 ± 0.13 ○	10.55 ± 0.04 ○	10.91 ± 0.06 ○	11.13 ± 0.25 ○	**9.62** ± **0.16**	9.73 ± 0.06
Fashion	13.34 ± 0.13 ○	12.59 ± 0.16 ○	11.35 ± 0.23 ○	12.13 ± 0.51 ○	11.91 ± 0.32 ○	11.72 ± 0.05 ○	10.01 ± 0.09	**9.89** ± **0.02**
MNIST	4.98 ± 0.17 ○	3.12 ± 0.02 ○	2.56 ± 0.78	3.01 ± 0.09 ○	2.87 ± 0.17 ○	2.98 ± 0.05 ○	**2.43** ± **0.14**	2.56 ± 0.06

Periodical Shift

	FIX	FTFWH	ASL	SENC-F	KNNENS	Self-N	PULSE	HANOL
CIFAR10	24.28 ± 0.72 ○	20.19 ± 0.82 ○	20.98 ± 0.79 ○	20.20 ± 0.77 ○	20.43 ± 0.77 ○	20.56 ± 0.81 ○	20.11 ± 0.81	**19.94** ± **0.74**
CINIC10	36.82 ± 1.03 ○	31.24 ± 0.91 ○	33.31 ± 0.52 ○	31.46 ± 1.15 ○	31.52 ± 1.15 ○	32.29 ± 0.88 ○	31.31 ± 1.12	**30.88** ± **1.04**
EuroSAT	17.72 ± 0.29 ○	12.12 ± 0.16 ○	11.89 ± 0.49 ○	11.72 ± 0.21 ○	12.22 ± 0.11 ○	12.61 ± 0.11 ○	10.84 ± 0.05 ○	**9.93** ± **0.17**
Fashion	14.75 ± 0.19 ○	14.04 ± 0.36 ○	12.96 ± 0.32 ○	13.56 ± 0.45 ○	13.27 ± 0.48 ○	12.67 ± 0.25 ○	**10.82** ± **0.12**	11.02 ± 0.44
MNIST	6.41 ± 0.15 ○	4.36 ± 0.16 ○	4.23 ± 0.15 ○	4.44 ± 0.02 ○	3.93 ± 0.09	4.02 ± 0.19 ○	3.82 ± 0.06	**3.75** ± **0.04**

Sudden Shift

	FIX	FTFWH	ASL	SENC-F	KNNENS	Self-N	PULSE	HANOL
CIFAR10	23.58 ± 0.74 ○	19.24 ± 0.87 ○	19.56 ± 0.35 ○	19.23 ± 0.88	19.54 ± 0.82 ○	19.45 ± 0.76 ○	19.39 ± 0.88 ○	**18.88** ± **0.86**
CINIC10	36.21 ± 0.89 ○	33.33 ± 1.15 ○	32.41 ± 0.72 ○	30.77 ± 1.12	30.64 ± 0.94	32.45 ± 0.94 ○	**30.55** ± **1.12**	31.26 ± 0.82
EuroSAT	16.79 ± 0.16 ○	11.15 ± 0.16	11.23 ± 0.45 ○	11.12 ± 0.07	11.55 ± 0.08 ○	11.36 ± 0.23 ○	10.18 ± 0.01	**10.06** ± **0.21**
Fashion	13.64 ± 0.24 ○	12.96 ± 0.37 ○	12.12 ± 0.07 ○	12.62 ± 0.01 ○	12.21 ± 0.37 ○	12.09 ± 0.05 ○	11.61 ± 0.26 ○	**10.92** ± **0.23**
MNIST	5.52 ± 0.05 ○	3.48 ± 0.13 ○	3.23 ± 0.23	3.45 ± 0.16 ○	3.44 ± 0.16 ○	3.24 ± 0.21	3.11 ± 0.13	**3.08** ± **0.09**



(a) accuracy curve on SHL

(b) accuracy curve on fMoW

(c) new class estimation on SHL

(d) efficiency comparison

Fig. 3: (a) & (b) Comparison of overall performances on the real-world tasks. (c) Accuracy of the estimated new class proportion of our sliding-window MPE module. (d) Evaluation of efficiency and accuracy (defined as 100% - average error) of different algorithms. We report the mean and standard deviation over five runs. An algorithm closer to the top-right corner indicates superior efficiency and performance.

In the online adaptation stage, the learner can only observe unlabeled data streams. Specifically, we randomly choose two classes as the new class for each benchmark dataset. The label distribution at round t is a mixture of two different constant distributions $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \Delta_{K+1}$ with a time-varying coefficient α_t, i.e., $\mathcal{D}_t(y) = (1-\alpha_t)\boldsymbol{\mu} + \alpha_t \boldsymbol{\mu}'$, where $\boldsymbol{\mu}_{y_t}$ denotes the distribution at round t and α_t controls the intensity of distribution changes. We only observe the label distribution $\boldsymbol{\mu}_0 \in \Delta_K$ for known classes in the offline training data. We simulate three representative types of distribution shifts commonly encountered in real-world tasks; the details regarding the simulation of these shifts can be found in the extended version. We evaluate all contenders by the average error over $T = 10,000$ rounds, with five benchmarks. More details of the benchmark datasets are deferred to the extended version.

Implementation Details. For the aforementioned five benchmark datasets, we employ a fine-tuned ResNet34 network for feature extraction. Images used to train the ResNet do not overlap with either the offline or online datasets. We sample 30, 000 data for offline initialization. We repeat all experiments for five times and evaluate the average error and standard deviation. The learning rates of the algorithms are set according to theoretical guidelines. The hyper-parameter ε for the meta learner is set as $\sqrt{(\ln N)/T}$. δ and γ in MPE are set as default values following [19], i.e., 0.1 and 0.01, respectively, without modification. The window size in sliding-window MPE is $L = 20$ by default, without deliberate selection. Enhanced performance could be potentially achieved by selecting window size using techniques such as cross-validation. All experiments are executed on a computer equipped with 2 Intel Xeon 8358 CPUs, each having 32 cores.

Results on Benchmark Datasets. The results in Table I demonstrate that our proposed algorithm effectively handles the new classes in the online label shift problem, outperforming other approaches. The baseline *FIX* is inferior to the online algorithms, highlighting the necessity of sequentially updated algorithms with online unlabeled data. Our approach surpasses both *FTFWH* and *ASL*, indicating that handling the new class is crucial in the N-OLS setting. Besides, compared with *SENC-F* and *KNNENS*, which primarily focus on managing new classes, our approach achieves better performance. This indicates that label shifts can lead to the misclassification of the new class, and our black box shift estimator effectively tackles this issue. Our HANOL algorithm consistently outperforms both *PULSE* and *Self-N*, showing the effectiveness of our online updating scheme with sliding window-based MPE and online ensemble. These results show the success of our approach in tackling the N-OLS problem.

TABLE II: Average error (%) of different algorithms on the real-world applications of SHL [14] and fMoW [15] datasets. The performance metrics reported include both the mean accuracy and the standard deviation of different algorithms over five separate runs.

	FIX	FTFWH	ASL	SENC-F	KNNENS	Self-N	PULSE	HANOL
SHL	47.32 ±1.05	43.21 ±1.67	40.78 ±1.42	40.22 ±1.55	41.23 ±1.81	41.25 ±1.12	38.19 ±1.61	**36.81** ±1.32
fMoW	73.15 ±3.31	69.38 ±2.64	69.54 ±2.13	68.87 ±3.34	69.23 ±1.81	70.37 ±2.84	66.32 ±2.71	**63.16** ±3.01

B. Real-world Applications

In this part, we aim to answer **Q2** and **Q3**. We compare the proposed approach with other contenders on two real-world applications: (i) the SHL locomotion recognition dataset [14], and (ii) the Functional Map of the World (fMoW) dataset [15], a sequential satellite image recognition task. We report the average error of various algorithms on the SHL and fMoW datasets in Table II, along with their respective timely performance depicted in Figure 3(a) and Figure 3(b). As shown in these empirical studies, our proposed method exhibits superior performance compared to the *FTFWH* and *ASL* methods, highlighting the significance of addressing the arrival of new classes in real-world tasks. Moreover, our HANOL effectively adapts to label shift by the black box shift estimator and constructing a risk estimator for the N-OLS problem through the use of unlabeled data, thereby outperforming the *SENC-F* and *KNNENS* methods. Our approach also surpasses the *PULSE* and *Self-N* methods, thanks to the benefits of the online updating scheme and the proposed sliding-window MPE mechanism, which alleviate the lack of labeled data problem in online data streams.

Modular Analysis. As demonstrated in Figure 3(c), our proposed sliding-window MPE module is capable of accurately estimating the proportion of new classes, thereby managing the issue of the new class in the N-OLS problem effectively.

Efficiency Comparison. We also compare the efficiency of different algorithms. Specifically, we evaluate the efficiency (items processed per second) and accuracy (defined as 100% - average error) of various algorithms. An algorithm that plots closer to the top-right corner indicates superior efficiency and performance since it achieves a better performance with higher efficiency. As demonstrated in Figure 3(d), the moving average-based *FTFWH* is the most efficient, but it yields the poorest performance. Though ensemble-based methods, *ASL* and *KNNENS*, exhibit slower speed, they accomplish superior performance. Our approach, albeit with a slight compromise on efficiency, attains the best performance among all.

VI. Conclusion

In this paper, we investigate the problem of handling new class in online label shift. We proposed a novel method, called HANOL, to tackle both online label shift and the presence of the new class in unlabeled data stream. In HANOL, we first build a risk estimator for unlabeled data stream via risk rewriting and mixture proportion estimation to handle both the presence of new class and the distribution shift. Then, we employ the paradigm of online ensemble to adapt to the unknown continuous label shift. The proposed method enjoys a theoretical guarantee of dynamic regret, affirming its effectiveness in adapting to changing distributions. We conduct experiments on five benchmark datasets and two real-world applications to validate the effectiveness of our HANOL.

Acknowledgements

This research was supported by National Key R&D Program of China (2022ZD0114800) and NSFC (62206125). The authors thank Yu-Jie Zhang for insightful discussions.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[2] M. Sugiyama and M. Kawanabe, *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press, 2012.

[3] Y. Bengio, Y. Lecun, and G. Hinton, "Deep learning for AI," *Communication of ACM*, vol. 64, no. 7, p. 58–65, 2021.

[4] Z.-H. Zhou, "Open-environment machine learning," *National Science Review*, vol. 9, no. 8, p. nwac123, 2022.

[5] Z.-H. Zhou, "Stream efficient learning," *ArXiv preprint*, vol. arXiv:2305.02217, 2023.

[6] C. Hou, S. Gu, C. Xu, and Y. Qian, "Incremental learning for simultaneous augmentation of feature and class," *TPAMI*, vol. in press, 2023.

[7] J. Kang, A. Kang, A. Green, K. Gwee, and K. Ho, "Systematic review: worldwide variation in the frequency of coeliac disease and changes over time," *Alimentary Pharmacology and Therapeutics*, vol. 38, no. 3, pp. 226–245, 2013.

[8] R. Wu, C. Guo, Y. Su, and K. Q. Weinberger, "Online adaptation to label distribution shift," in *NeurIPS*, 2021, pp. 11 340–11 351.

[9] Y. Bai, Y.-J. Zhang, P. Zhao, M. Sugiyama, and Z.-H. Zhou, "Adapting to online label shift with provable guarantees," in *NeurIPS*, 2022, pp. 29 960–29 974.

[10] X. Mu, K.-M. Ting, and Z.-H. Zhou, "Classification under streaming emerging new classes: A solution using completely-random trees," *TKDE*, vol. 29, no. 8, pp. 1605–1618, 2017.

[11] X.-Q. Cai, P. Zhao, K.-M. Ting, X. Mu, and Y. Jiang, "Nearest neighbor ensembles: An effective method for difficult problems in streaming classification with emerging new classes," in *ICDM*, 2019, pp. 970–975.

[12] B. Lu, X. Gan, L. Yang, W. Zhang, L. Fu, and X. Wang, "Geometer: Graph few-shot class-incremental learning via prototype representation," in *KDD*, 2022, pp. 1152–1161.

[13] P. Zhao, Y.-J. Zhang, L. Zhang, and Z.-H. Zhou, "Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization," *ArXiv preprint*, vol. arXiv:2112.14368, 2021.

[14] H. Gjoreski, M. Ciliberto, L. Wang, F. J. O. Morales, S. Mekki, S. Valentin, and D. Roggen, "The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices," *IEEE Access*, vol. 6, pp. 42 592–42 604, 2018.

[15] G. A. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *CVPR*, 2018, pp. 6172–6180.

[16] Z. C. Lipton, Y.-X. Wang, and A. J. Smola, "Detecting and correcting for label shift with black box predictors," in *ICML*, 2018, pp. 3128–3136.

[17] H. Ramaswamy, C. Scott, and A. Tewari, "Mixture proportion estimation via kernel embeddings of distributions," in *ICML*, 2016, pp. 2052–2060.

[18] Y.-J. Zhang, P. Zhao, L. Ma, and Z.-H. Zhou, "An unbiased risk estimator for learning with augmented classes," in *NeurIPS*, 2020, pp. 10 247–10 258.

[19] S. Garg, Y. Wu, A. J. Smola, S. Balakrishnan, and Z. C. Lipton, "Mixture proportion estimation and PU learning: A modern approach," in *NeurIPS*, 2021, pp. 8532–8544.

[20] L. Zhang, S. Lu, and Z.-H. Zhou, "Adaptive online learning in dynamic environments," in *NeurIPS*, 2018, pp. 1330–1340.

[21] H. Yan, Y. Guo, and C. Yang, "Augmented self-labeling for source-free unsupervised domain adaptation," in *NeurIPS Workshop*, 2021.

[22] J. Zhang, T. Wang, W. W. Y. Ng, and W. Pedrycz, "Knnens: A k-nearest neighbor ensemble-based method for incremental learning under data stream with emerging new classes," *TNNLS*, pp. 1–8, 2022.

[23] S. Garg, S. Balakrishnan, and Z. C. Lipton, "Domain adaptation under open set label shift," in *NeurIPS*, 2022.