A. Proof of Theorems

This section presents the omit details of regret theory analysis. We first give two useful lemmas, then we prove Theorem 1, Theorem 2, Theorem 3, respectively.

A.1. Useful lemmas

Lemma 1. *(Lemma 7 of [26]) Let $\mathcal{W}$ be a convex set in a Banach space $\mathcal{B}$. Then, any update of the form $\mathbf{w}^* = \Pi_{\mathcal{W}}[\mathbf{c} - \nabla]$ satisfies the following inequality*

$$\langle \mathbf{w}^* - \mathbf{u}, \nabla \rangle \leq \frac{1}{2}\|\mathbf{c} - \mathbf{u}\|^2 - \frac{1}{2}\|\mathbf{w}^* - \mathbf{u}\|^2 - \frac{1}{2}\|\mathbf{w}^* - \mathbf{c}\|^2 \tag{19}$$

for any $\mathbf{u} \in \mathcal{W}$.

Lemma 1 is shown to be useful in analyzing of Theorem 1.

Lemma 2. *(Theorem 19 of [32])The meta-regret of the Optimistic Hedge is upper bounded by*

$$\sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i} \leq \frac{2 + \ln N}{\varepsilon} + \varepsilon D_\infty, \tag{20}$$

which holds for any base learner $i \in [N]$. Besides, $D_\infty = \sum_{t=1}^{T} \|\boldsymbol{\ell}_t - \boldsymbol{m}_t\|_\infty^2$ measures the adaptivity.

Lemma 2 is the regret guarantee of Optimistic Hedge.

A.2. Proof of Theorem 1

Theorem 1. *Let $\boldsymbol{M}_t = \nabla R_t(\widehat{\mathbf{w}}_{t+1})$ and use $\widehat{R}_t(\mathbf{w})$ to replace $R_t(\mathbf{w})$. Under Assumption 1, Opt-OMD Eq. (12) satisfies universal dynamic regret*

$$\mathbf{Regret}_T^U \leq GD + \eta G^2 (4T + V_T) + \frac{1}{2\eta}(D^2 + 2DP_T), \tag{15}$$

for any comparator sequence $\mathbf{u}_1, ..., \mathbf{u}_T \in \mathcal{W}$, where $P_T = \sum_{t=2}^{T} \|\mathbf{u}_{t-1} - \mathbf{u}_t\|$ is the path-length of comparators $\mathbf{u}_1, ..., \mathbf{u}_T$, reflecting the non-stationarity of the environments, $V_T = \sum_{t=2}^{T} \|\widehat{\boldsymbol{P}}_{y_t} - \widehat{\boldsymbol{P}}_{y_{t-1}}\|_2^2$ measures the label shift intensity.

Proof. By the convexity of the risk estimator $\widehat{R}_t(\cdot)$, we have

$$\begin{aligned} &\widehat{R}_t(\mathbf{w}_{t,i}) - \widehat{R}_t(\mathbf{u}_t) \leq \langle \nabla \widehat{R}_t(\mathbf{w}_{t,i}), \mathbf{w}_{t,i} - \mathbf{u}_t \rangle \\ &= \langle \nabla \widehat{R}_t(\mathbf{w}_{t,i}) - \nabla \widehat{R}_{t-1}(\widehat{\mathbf{w}}_t), \mathbf{w}_{t,i} - \widehat{\mathbf{w}}_{t+1,i} \rangle + \langle \nabla \widehat{R}_{t-1}(\widehat{\mathbf{w}}_{t,i}), \mathbf{w}_{t,i} - \widehat{\mathbf{w}}_{t+1,i} \rangle + \langle \nabla \widehat{R}_t(\mathbf{w}_t), \widehat{\mathbf{w}}_{t+1,i} - \mathbf{u}_t \rangle. \end{aligned} \tag{21}$$

By using Lemma 1 to equation Eq. (12), we the dynamic regret in one round

$$\langle \mathbf{w}_{t+1,i} - \mathbf{u}_t, \eta \nabla \widehat{R}_t(\mathbf{w}_{t,i}) \rangle \leq \frac{1}{2}\|\mathbf{u}_t - \widehat{\mathbf{w}}_{t,i}\|_2^2 - \frac{1}{2}\|\mathbf{u}_t - \widehat{\mathbf{w}}_{t+1,i}\|_2^2 - \frac{1}{2}\|\widehat{\mathbf{w}}_{t+1,i} - \widehat{\mathbf{w}}_{t,i}\|_2^2, \tag{22}$$

$$\langle \mathbf{w}_{t,i} - \widehat{\mathbf{w}}_{t+1,i}, \eta \nabla \widehat{R}_{t-1}(\widehat{\mathbf{w}}_{t,i}) \rangle \leq \frac{1}{2}\|\widehat{\mathbf{w}}_{t,i} - \widehat{\mathbf{w}}_{t+1,i}\|_2^2 - \frac{1}{2}\|\mathbf{w}_{t,i} - \widehat{\mathbf{w}}_{t+1,i}\|_2^2 - \frac{1}{2}\|\mathbf{w}_{t,i} - \widehat{\mathbf{w}}_{t,i}\|_2^2. \tag{23}$$

Due to Hölder inequality, and the fact that $ab \leq \frac{a^2}{2\eta} + \frac{b^2\eta}{2}$ for any $a, b \geq 0$ and $\eta > 0$

$$\begin{aligned} \langle \nabla \widehat{R}_t(\mathbf{w}_{t,i}) - \nabla \widehat{R}_{t-1}(\widehat{\mathbf{w}}_{t,i}), \mathbf{w}_t - \widehat{\mathbf{w}}_{t+1} \rangle &\leq \|\nabla \widehat{R}_t(\mathbf{w}_{t,i}) - \nabla \widehat{R}_{t-1}(\widehat{\mathbf{w}}_{t,i})\|_2 \|\mathbf{w}_{t,i} - \widehat{\mathbf{w}}_{t+1,i}\|_2 \\ &\leq \frac{\|\mathbf{w}_{t,i} - \widehat{\mathbf{w}}_{t+1,i}\|_2^2}{2\eta} + \frac{\eta \|\nabla \widehat{R}_t(\mathbf{w}_{t,i}) - \nabla \widehat{R}_{t-1}(\widehat{\mathbf{w}}_{t,i})\|_2^2}{2}. \end{aligned} \tag{24}$$

Bringing Eq. (24), Eq. (22) and Eq. (23) into Eq. (21), we obtain

$$\begin{aligned} &\widehat{R}_t(\mathbf{w}_t) - \widehat{R}_t(\mathbf{u}_t) \leq \frac{\eta}{2}\|\nabla \widehat{R}_t(\mathbf{w}_t) - \nabla \widehat{R}_{t-1}(\widehat{\mathbf{w}}_{t,i})\|_2^2 + \frac{1}{2\eta}\|\mathbf{w}_{t,i} - \widehat{\mathbf{w}}_{t+1,i}\|_2^2 \\ &+ \frac{1}{2\eta}\left(\|\mathbf{u}_t - \widehat{\mathbf{w}}_{t,i}\|_2^2 - \|\mathbf{u}_t - \widehat{\mathbf{w}}_{t+1,i}\|_2^2 - \|\mathbf{w}_{t,i} - \widehat{\mathbf{w}}_{t+1,i}\|_2^2 - \|\mathbf{w}_{t,i} - \widehat{\mathbf{w}}_{t,i}\|_2^2\right) \\ &\leq \frac{\eta}{2}\|\nabla \widehat{R}_t(\mathbf{w}_{t,i}) - \nabla \widehat{R}_{t-1}(\widehat{\mathbf{w}}_{t,i})\|_2^2 + \frac{1}{2\eta}\left(\|\mathbf{u}_t - \widehat{\mathbf{w}}_{t,i}\|_2^2 - \|\mathbf{u}_t - \widehat{\mathbf{w}}_{t+1,i}\|_2^2\right). \end{aligned} \tag{25}$$

Considering all the rounds and summing, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\widehat{R}_t(\mathbf{w}_{t,i})-\sum_{t=1}^{T}\widehat{R}_t(\mathbf{u}_t)\leq \widehat{R}_1(\mathbf{w}_{1,i})-\widehat{R}_1(\mathbf{u}_1)+\sum_{t=2}^{T}\frac{\eta}{2}\|\nabla\widehat{R}_t(\mathbf{w}_{t,i})-\nabla\widehat{R}_{t-1}(\widehat{\mathbf{w}}_{t,i})\|_2^2\\
&+\sum_{t=2}^{T}\frac{1}{2\eta}\left(\|\mathbf{u}_t-\widehat{\mathbf{w}}_{t,i}\|_2^2-\|\mathbf{u}_t-\widehat{\mathbf{w}}_{t+1,i}\|_2^2\right)\\
&\leq GD+\underbrace{\frac{\eta}{2}\sum_{t=2}^{T}\|\nabla\widehat{R}_t(\mathbf{w}_{t,i})-\nabla\widehat{R}_{t-1}(\widehat{\mathbf{w}}_{t,i})\|_2^2}_{\text{term (a)}}+\underbrace{\frac{1}{2\eta}\sum_{t=2}^{T}\left(\|\mathbf{u}_t-\widehat{\mathbf{w}}_{t,i}\|_2^2-\|\mathbf{u}_t-\widehat{\mathbf{w}}_{t+1,i}\|_2^2\right)}_{\text{term (b)}},
\end{aligned}\tag{26}$$

where the second inequality due to the fact that $\widehat{R}_1(\mathbf{w}_{1,i})-\widehat{R}_1(\mathbf{u}_1)\leq\langle\nabla\widehat{R}_1(\mathbf{w}_{1,i}),\mathbf{w}_{1,i}-\mathbf{u}_1\rangle\leq GD$.

$$\begin{aligned}
\text{term (a)}&=\frac{\eta}{2}\sum_{t=2}^{T}\|\nabla\widehat{R}_t(\mathbf{w}_{t,i})-\nabla\widehat{R}_{t-1}(\widehat{\mathbf{w}}_{t,i})\|_2^2\\
&\leq\frac{\eta}{2}\sum_{t=2}^{T}2\left(\|\nabla\widehat{R}_t(\mathbf{w}_{t,i})-\nabla\widehat{R}_t(\widehat{\mathbf{w}}_{t,i})\|_2^2+\|\nabla\widehat{R}_t(\widehat{\mathbf{w}}_{t,i})-\nabla\widehat{R}_{t-1}(\widehat{\mathbf{w}}_{t,i})\|_2^2\right)\\
&\leq\eta\sum_{t=2}^{T}\left(4G^2+\sup_{\mathbf{w}\in\mathcal{W}}\|\nabla\widehat{R}_t(\mathbf{w})-\nabla\widehat{R}_{t-1}(\mathbf{w})\|_2^2\right)\\
&\leq\eta 4G^2T+\eta\sum_{t=2}^{T}\sup_{\mathbf{w}\in\mathcal{W}}\|\widehat{\boldsymbol{P}}_{y_t}-\widehat{\boldsymbol{P}}_{y_{t-1}}\|_2^2\|\nabla\widehat{\boldsymbol{R}}_0(\mathbf{w})\|_2^2\\
&=\eta 4G^2T+\eta G^2V_T,
\end{aligned}\tag{27}$$

where $V_T=\sum_{t=2}^{T}\|\widehat{\boldsymbol{P}}_{y_t}-\widehat{\boldsymbol{P}}_{y_{t-1}}\|_2^2$ disc the estimate of label shift intensity. The second inequality due to the Assumption 1. Term (b) can be bounded

$$\begin{aligned}
\text{term (b)}&=\frac{1}{2\eta}\sum_{t=2}^{T}\left(\|\mathbf{u}_t-\widehat{\mathbf{w}}_{t,i}\|_2^2-\|\mathbf{u}_t-\widehat{\mathbf{w}}_{t+1,i}\|_2^2\right)\\
&=\frac{1}{2\eta}\|\mathbf{u}_1-\widehat{\mathbf{w}}_{2,i}\|_2^2+\frac{1}{2\eta}\sum_{t=2}^{T}\left(\|\mathbf{u}_t-\widehat{\mathbf{w}}_{t,i}\|_2^2-\|\mathbf{u}_{t-1}-\widehat{\mathbf{w}}_{t,i}\|_2^2\right)\\
&\leq\frac{D^2}{2\eta}+\frac{1}{2\eta}\sum_{t=2}^{T}\|\mathbf{u}_t-\mathbf{u}_{t-1}\|_2\|\mathbf{u}_t-\widehat{\mathbf{w}}_{t,i}+\mathbf{u}_{t-1}-\widehat{\mathbf{w}}_{t,i}\|_2\\
&\leq\frac{D^2}{2\eta}+\frac{D}{\eta}\sum_{t=2}^{T}\|\mathbf{u}_t-\mathbf{u}_{t-1}\|_2.
\end{aligned}\tag{28}$$

Putting the above inequalities of term (a) and term (b) yields,

$$\begin{aligned}
\sum_{t=1}^{T}\widehat{R}_t(\mathbf{w}_{t,i})-\sum_{t=1}^{T}\widehat{R}_t(\mathbf{u}_t)&\leq GD+\eta 4G^2T+\eta G^2V_T+\frac{D^2}{2\eta}+\frac{D}{\eta}\sum_{t=2}^{T}\|\mathbf{u}_t-\mathbf{u}_{t-1}\|_2\\
&\leq GD+\eta G^2\left(4T+V_T\right)+\frac{1}{2\eta}(D^2+2DP_T).
\end{aligned}\tag{29}$$

This completes the proof. □

A.3. Proof of Theorem 2

Theorem 2. *Under Assumption 1, there exists a learning rate* $\varepsilon = \sqrt{\frac{2+\ln N}{B^2 T}}$*, such that the meta-regret is at most*

$$\sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_{t,i}) \leq \mathcal{O}(\sqrt{(2+\ln N)T}). \tag{16}$$

Proof. By the Jensen's inequality, the universal dynamic regret can be bounded by

$$\sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_t) - \widehat{R}_t(\mathbf{w}_{t,i}) \leq \sum_{t=1}^{T} \sum_{i=1}^{N} p_{t,i} \widehat{R}_t(\mathbf{w}_{t,i}) - \sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_{t,i}) = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \sum_{t=1}^{T} \ell_{t,i}, \tag{30}$$

where $\boldsymbol{p}_t$ is the weight vector of base learners, $\boldsymbol{\ell}_t = (\ell_{t,1}, \ell_{t,2}, ..., \ell_{t,N})$. Since the optimism is $\widehat{R}_t(\mathbf{w}_{t+1,i})$, Lemma 2 implies

$$\begin{aligned} \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \sum_{t=1}^{T} \ell_{t,i} &\leq \varepsilon \sum_{t=1}^{T} \|\boldsymbol{\ell}_t - \boldsymbol{m}_t\|_\infty^2 + \frac{2+\ln N}{\varepsilon} = \varepsilon \sum_{t=1}^{T} \left(\max_{i \in [N]} \{ \widehat{R}_t(\mathbf{w}_{t,i}) - \widehat{R}_{t-1}(\mathbf{w}_{t,i}) \} \right)^2 + \frac{2+\ln N}{\varepsilon} \\ &\leq \varepsilon B^2 T + \frac{2+\ln N}{\varepsilon}, \end{aligned} \tag{31}$$

where the second inequality holds due to Assumption 1. Since $\varepsilon = \sqrt{\dfrac{2+\ln N}{B^2 T}}$, we have

$$\sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_{t,i}) \leq 2GD\sqrt{2(2+\ln N)V_T^R} + 4\sqrt{2}D^2 L(2+\ln N) = 2B\sqrt{(2+\ln N)T} \tag{32}$$

This completes the proof. □

A.4. Proof of Theorem 3

Theorem 3. *Under Assumption 1, there exists a pool of candidate step sizes as*

$$\mathcal{H} = \left\{ \frac{D}{G} \sqrt{\frac{1}{2(4T+V_T)}} \cdot 2^{i-1}, \quad i \in [N] \right\}, \tag{17}$$

and the learning rate of the meta algorithm is $\varepsilon = \sqrt{\frac{2+\ln N}{B^2 T}}$*, such that, for any comparator sequence* $\mathbf{u}_1, ..., \mathbf{u}_T \in \mathcal{W}$*, the universal dynamic regret of LASU is*

$$\mathbf{Regret}_T^U \leq \mathcal{O}\left(\sqrt{(T+V_T)(1+P_T)} \right), \tag{18}$$

where $N = \lceil 2^{-1} \log_2(1 + \frac{2T}{D}) \rceil + 1$ *is the number of candidate step sizes.*

Proof. Notice that the universal dynamic regret can be decomposed intimate the following two parts

$$\sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widehat{R}_t(\mathbf{u}_t) = \underbrace{\sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_{t,i})}_{\text{meta-regret}} + \underbrace{\sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_{t,i}) - \sum_{t=1}^{T} \widehat{R}_t(\mathbf{u}_t)}_{\text{base-regret}}, \tag{33}$$

which holds for any base learner index $i \in [N]$. In above, $\{\mathbf{w}_t\}_{t=1,...,T}$ denotes the final output sequence, and $\{\mathbf{w}_{t,i}\}_{t=1,...,T}$ is the sequence of base the i-th base learner. The first part is meta-regret means the difference between the cumulative loss of final output and that of the base learner. The second part is base-regret means the universal dynamic regret of the base learner. In the following, we upper bound these two terms respectively.

Upper bound of meta-regret. The final decision $\mathbf{w}_t$ at round t is a weighted combination of predictions returned from the base learners, and the weight is updated by the online ensemble algorithm. Thus, we can apply Theorem 2 and obtain the upper bound of the meta-regret.

$$\text{meta-regret} = \sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_{t,i}) \leq 2B\sqrt{(2+\ln N)T}. \tag{34}$$

Upper bound of base-regret. We find the base $k \in [N]$ with the smallest base-regret to make the bound Eq. (33) tight. In other words, we need to identify the nearly optimal step size. Recall that the optimal step size is $\eta^* = \sqrt{\frac{D^2+2DP_T}{2G^2(4T+V_T)}}$, which is included by the step size pool

$$\mathcal{H} = \left\{ \frac{D}{G}\sqrt{\frac{1}{2(4T+V_T)}} \cdot 2^{i-1}, \quad i \in [N] \right\}, \tag{35}$$

where $N = \lceil 2^{-1}\log_2(1+\frac{2T}{D}) \rceil + 1$. By the construction of the candidate step size pool $\mathcal{H}$, we know that the step size therein is monotonically increasing with respect to the index. Therefore, we confirm that there exists an integer $k \in [N]$ such that $\eta_k \leq \eta^* \leq \eta_{k+1} = 2\eta_k$, where the optimal step size is $\eta^* = \sqrt{\frac{D^2+2DP_T}{2G^2(4T+V_T)}}$. The gap between the cumulative loss of final decisions and that of base learner k can be upper bounded as follows

$$\begin{aligned}
\text{base-regret} &= \sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_{t,k}) - \sum_{t=1}^{T} \widehat{R}_t(\mathbf{u}_t) \leq \frac{D^2+2DP_T}{2\eta_k} + \eta_k G^2(4T+V_T) + GD \\
&\leq \frac{D^2+2DP_T}{\eta^*} + \eta^* G^2(4T+V_T) + GD \\
&= \frac{3\sqrt{2}G}{2}\sqrt{(D^2+2DP_T)(4T+V_T)} + GD
\end{aligned} \tag{36}$$

Upper bound of universal dynamic regret. Combining Eq. (34) and Eq. (36), we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widehat{R}_t(\mathbf{u}_t) = \underbrace{\sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_{t,k})}_{\text{meta-regret}} + \underbrace{\sum_{t=1}^{T} \widehat{R}_t(\mathbf{w}_{t,k}) - \sum_{t=1}^{T} \widehat{R}_t(\mathbf{u}_t)}_{\text{base-regret}} \\
&\leq 2B\sqrt{(2+\ln N)T} + \frac{3\sqrt{2}G}{2}\sqrt{(D^2+2DP_T)(4T+V_T)} + GD \\
&\leq \sqrt{8B^2(2+\ln N)T + 9G^2(D^2+2DP_T)(4T+V_T)} + GD \\
&= \mathcal{O}\sqrt{(1+P_T)(V_T+T)}
\end{aligned} \tag{37}$$

The derivations uses the inequality of $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}, \forall a, b > 0$. Meanwhile, we treat the double logarithmic factor in T as a constant. This completes the proof. □

B. Details on Experiments

In this section, we present the details of experiments. We start with methods introduction and settings, followed by dataset and model details. Then we list different of label shift types.

B.1. Methods introduction and settings

To explain the superiority of our algorithm, we compare eight different online label shift algorithms with our algorithm, including:

- Fixed: The initial model trained with initial labeled data and used to predict online data.
- Oracle: It proposed by [3], accepting the true label at each online round and re-weighting the base model's predictions.
- FTH: Following the history algorithm proposed by [36] and the method make predictions with current estimates which base the average of all previous estimates.

- FTHWH: A variant of FTH proposed by [36] and the method make predictions with current estimates which are calculated by the average of recent slides window previous estimates. The length of the sliding window is 100.
- FLH-FTL: An ensemble method proposed by [3] which reformulates the adaptation problem as an online regression task and reweights the initial classifier.
- ROGD: A variant of OGD algorithm proposed by [4] and constructs the risk estimator with 0-1 loss to update its re-weighting vector.
- UOGD: A variant of OGD algorithm proposed by [4] and constructs unbiased risk estimator with surrogate loss to update its re-weighting vector.
- ATLAS: A meta-based construct method to run several UOGDs simultaneously as its base learners and then combine weights [4].

The learning rate in our OSCM-L algorithm is set as 0.01. The learning rates of ROGD, UOGD, ATLAS, and the step size pool of ATLAS are set as original works. In all experiments, all methods enjoy the same fixed diameter D estimated according to the parameter norm of the initial model f_0 for each dataset in decision domain Ω, the bound of loss function B, and lipschitz constant G can be estimated during training the initial model.

B.2. Dataset and model details

We conduct our experiments in 6 datasets and use different models to extract features. For all the datasets below, the initial labeled data are split by $4:1$ into training and validation data, where the former is used to train an initial model and the latter to estimate label shift and construct a convex risk.

- CIFAR-10 [19]: A dataset for classification consists of 60,000 colored images, categorized into ten different classes: airplane, automobile, ship, truck, bird, cat, deer, dog, frog, and horse. 50,000 samples are used as source data and 10,000 as target data to be sampled from during online learning.
- CINIC-10 [7]: A hybrid image classification consists of 60,000 colored images dataset that combines elements from both the CIFAR-10 and ImageNet datasets. 50,000 samples are used as source data and 10,000 as target data to be sampled from during online learning.
- EuroSat [18]: An image dataset of 10 types of land uses. Including 27,000 satellite images sourced from more than 30 European nations. These images are categorized into ten different types: industrial, residential, annual crop, permanent crop, river, sea and lake, herbaceous vegetation, highway, pasture, and forest. 60,000 samples are used as source data and 10,000 as target data to be sampled from during online learning.

For the above 3 benchmark datasets, we utilize a finetuned ResNet18 [17] to extract image features.

- MNIST [20]: A popular dataset of handwritten digits is comprised of 70,000 grayscale images, encompassing ten different categories of digits. 60,000 samples are used as source data and 10,000 as target data to be sampled from during online learning.
- Fashion [37]: A dataset comprises images of ten diverse fashion items: T-shirts, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots. 60,000 samples are used as source data and 10,000 as target data to be sampled from during online learning.
- SHL [14, 33]: A dataset designed to identify human locomotion in real-world settings. It includes multi-modal sensor data from a body-worn camera and four smartphones, which are carried at common body positions. These sensors capture data such as acceleration, gyroscope, magnetometer, orientation, gravity, pressure, altitude, and temperature. From an 11-day period, 30,000 initial and 77,000 online data samples are collected, covering six activities: still, walking, running, biking, car, and bus. Online samples are received in the order they occur, based on timestamps.

For above 3 benchmark datasets, we utilize a fine-tuned MLP to extract image features.

B.3. Label shift types

To simulate online label shift, we introduce four types of label shift [4] to capture varying distribution patterns, which include Linear Shift, Bernoulli Shift, Square Shift, and Sine Shift. The first two types represent shift intensities, while the rest denote the periodic nature of label shift. The label distribution in each shift is combined with two different margins $\boldsymbol{P}_1, \boldsymbol{P}_2 \in \Delta_{K-1}$ as $\boldsymbol{P}_{y_t} = (1-\alpha_t)\boldsymbol{P}_1 + \alpha_t \boldsymbol{P}_2$, where $\boldsymbol{P}_{y_t}$ denotes the label distribution at round t and the α_t controls the shift non-stationarity and patterns. The four different label shift types are as follows:

- Linear Shift: The parameter $\alpha_t = \frac{t}{T}$.
- Bernoulli Shift: At each round, we keep the parameter $\alpha_t = \alpha_{t-1}$ with probability $p \in [0, 1]$ and otherwise $\alpha_t = 1 - \alpha_{t-1}$. In the experiments, the parameter is set as $p = \frac{1}{\sqrt{T}}$.
- Square Shift: At every L rounds we set $\alpha_t = \alpha_{t-1}$. In the experiments, we set $L = \sqrt{T}$.

- Sinusoidal Shift: The parameter $\alpha_t = \sin \frac{i\pi}{L}$, where$i = t \mod L$.

C. More Discussion

This section first discuss the nonconvexity challenge in the previous OnLS models, and compare them with our solution. Then we clarify the implications of comparator sequence $\{\mathbf{u}_t\}_{t=1}^T$.

C.1. Nonconvexity challenge and our solution

Nonconvexity implies the risk estimator $\widehat{R}_t(\mathbf{w})$ is non-convex with respect to the parameter $\mathbf{w}$. Wu *et al*. [36] construct a risk estimator using non-convex operators, such as non-convex $0/1$ loss and argmax operation, making convexity verification difficult. Bai *et al*. [4] construct an estimator based on the inversion of confusion matrix $\widehat{\boldsymbol{C}}_f$, leading to nonconvexity. To obtain a convex estimator, we first propose the OSCM-L method, which estimates the label distribution using maximum likelihood. Then, a convex estimator is constructed as in Eq. (11) via the risk rewriting technique.

The approach proposed by Tachet des Combes *et al*. [9] can avoid the singular matrix by solving the QP. According to [1, 31], maximum likelihood-based methods generally outperform confusion matrix-based methods in handling label shift. Therefore, our LASU framework incorporates OSCM-L and an online ensemble method to adapt to OnLS. We compare our method with this approach, achieving an average error rate of 27.67% vs. 32.84% on the SHL dataset.

C.2. Implications of comparator sequence

Universal dynamic regret supports comparison with any feasible comparator sequence $\{\mathbf{u}_t\}_{t=1}^T$ [43, 46], making it ideal for analyzing non-stationary label shift. In contrast, static regret [48] lacks adaptivity, as it assumes a single fixed model performs well across all rounds in OnLS . Dynamic regret [45] is less robust and may overfit in non-stationary scenarios, particularly under random sampling. Our universal dynamic regret framework incorporates the term $P_T = \sum_{t=2}^T \|\mathbf{u}_{t-1} - \mathbf{u}_t\|$, which captures the non-stationarity of OnLS. However, since P_T is unknown due to environmental uncertainty, we design an online ensemble algorithm that adapts to OnLS, ensuring robustness and adaptivity in non-stationary environments.