Supplemental Materials of
"Learning with Feature and Distribution Evolvable Streams"

Zhen-Yu Zhang [1] **Peng Zhao** [1] **Yuan Jiang** [1] **Zhi-Hua Zhou** [1]

A. Prerequisite Knowledge and Notations

In this section, we first introduce prerequisite knowledge and technical lemmas for proving the main results. Then we provide the detailed notations for the Feature space and Distribution Evolvable Stream Learning (FDESL) problem.

A.1. Rademacher Complexity

In this work, we use the Rademacher complexity (Bartlett & Mendelson, 2002) in proving generalization error bounds.

To simplify the presentation, we first introduce some notations. Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ be a sample of m points drawn independently and identically distributed according to the distribution $\mathcal{D}$, and then the *risk* and *empirical risk* of hypothesis g are defined as

$$R_{\mathcal{D}}(g) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(g(\mathbf{x}), y)], \quad \widehat{R}_S(g) = \frac{1}{m}\sum_{i=1}^{m} \ell(g(\mathbf{x}_i), y_i).$$

In order to prove the generalization error bounds proposed in the main paper, we use the notion of Rademacher complexity (Bartlett & Mendelson, 2002) to measure the complexity of the hypothesis set (family of decision functions) and use it to upper bound the generalization error.

Definition 1 (Rademacher Complexity)**.** Let $\mathcal{G}$ be a family of functions and a fixed sample of size m as $S = (\mathbf{z}_1, \cdots, \mathbf{z}_m)$. Then, the *empirical Rademacher complexity* of $\mathcal{G}$ with respect to the sample S is defined as

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(\mathbf{z}_i)\right].$$

Besides, the *Rademacher complexity* of $\mathcal{G}$ is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to $\mathcal{D}$,

$$\mathfrak{R}_m(\mathcal{G}) = \mathbb{E}_{S\sim\mathcal{D}^m}[\widehat{\mathfrak{R}}_S(\mathcal{G})].$$

Now we can establish a general margin-based generalization error bound for binary classification based on Rademacher complexity (Mohri et al., 2018).

Lemma 1 (Generalization Error Bound)**.** *Let $\mathcal{L}$ be the family of loss function associated to $\mathcal{G}$, i.e., $\mathcal{L} = \{\mathbf{x} \mapsto \ell(g(\mathbf{x}, y), g \in \mathcal{G}\}$. Suppose the loss function is L-Lipschitz, then, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $g \in \mathcal{G}$:*

$$R_{\mathcal{D}}(g) \leq \widehat{R}_S(g) + 2L\mathfrak{R}_m(\mathcal{L}) + \sqrt{\frac{\log(1/\delta)}{2m}},$$

where $\mathfrak{R}_m(\mathcal{L})$ is Rademacher complexity of loss function class $\mathcal{L}$ associated to $\mathcal{G}$, which can be bounded by using the celebrated Talagrand's lemma (Ledoux & Talagrand, 2013).

[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. Correspondence to: Yuan Jiang <jiangy@lamda.nju.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

A.2. Admissible Loss Function

In this part, we review the basic properties of the σ-admissible loss function family. Throughout this work, our analysis will assume that the loss function is convex and that it further verifies the following Lipschitz-like smoothness condition (Bousquet & Elisseeff, 2002; Mohri et al., 2018).

Definition 2 (σ-admissibility)**.** A loss function ℓ is σ-admissible with respect to the hypothesis class $\mathcal{G}$ if there exists $\sigma \in \mathbb{R}_+$ such that for any two hypotheses $g, g' \in \mathcal{G}$ and for all $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$,

$$|\ell(g(\mathbf{x}), y) - \ell(g'(\mathbf{x}), y)| \leq \sigma \left| g(\mathbf{x}) - g'(\mathbf{x}) \right| .$$

The admissibility property holds for most of common loss functions, including the quadratic loss and most other loss functions where the hypothesis set and the set of output labels are bounded by some $M \in \mathbb{R}_+ : \forall g \in \mathcal{G}, \forall \mathbf{x} \in \mathcal{X}, |g(\mathbf{x})| \leq M$ and $\forall y \in \mathcal{Y}, |y| \leq M$.

Example 1. We provide some examples of σ-admissible loss functions below (Mohri et al., 2018).

- Hinge loss $L_{hinge}(y', y) = \max(0, 1 - y'y)$ is σ-admissible with $\sigma = 1$.
- Least squares loss function $L_2(y', y) = (y' - y)^2$ that is bounded by M is σ-admissible with $\sigma = 2\sqrt{M}$.
- ϵ-insensitive loss $L_\epsilon(y', y) = |y' - y| - \epsilon$ and otherwise 0 if $|y' - y| \leq \epsilon$ is σ-admissible with $\sigma = 1$.

Proof. We show that the bounded least squares loss function is admissible, and other loss functions can be verified by an analogous argument. For and $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$, and $g, g' \in \mathcal{G}$,

$$\begin{aligned}
& |L_2(g(\mathbf{x}), y) - L_2(g'(\mathbf{x}), y) \\
= & |(g(\mathbf{x}) - y)^2 - (g'(\mathbf{x}) - y)^2| \\
= & |[(g(\mathbf{x}) - y) + (g'(\mathbf{x}) - y)][g(\mathbf{x}) - g'(\mathbf{x})]| \\
\leq & (|g(\mathbf{x}) - y| + |g'(\mathbf{x}) - y|)|g(\mathbf{x}) - g'(\mathbf{x})| \\
\leq & 2\sqrt{M}|g(\mathbf{x}) - g'(\mathbf{x})|.
\end{aligned}$$

In the second inequality, we use the M-boundedness property of the loss function. Therefore, we prove that the bounded least squares loss function is σ-admissible with $\sigma = 2\sqrt{M}$. □

By exploiting the admissibility of loss functions, we could align the hypothesis classes that lie in different feature spaces, when the data essentially share the same labels.

A.3. Discrepancy Minimization

In this part, we review the technical lemmas for the discrepancy minimization approaches. The discrepancy minimization approaches define the discrepancy of two distributions in a *fixed feature space*. A series of pioneering works propose different discrepancy measures to estimate the discrepancy between two distributions, including the $\mathcal{A}$-discrepancy (Kifer et al., 2004), $\mathcal{H}\Delta\mathcal{H}$-discrepancy (Ben-David et al., 2010), and $\mathcal{Y}$-discrepancy (Mohri & Medina, 2012), etc.

We propose the evolving discrepancy to measure the discrepancy of two consecutive batches with *different feature spaces* and data distributions. Our definition of the evolving discrepancy is a generalization to the feature and distribution evolving stream context of the $\mathcal{Y}$-discrepancy (Mohri & Medina, 2012),

Definition 3 ($\mathcal{Y}$-discrepancy)**.** Let $\mathcal{D}_P, \mathcal{D}_Q$ be two distributions over $\mathcal{X}$ and denote by f_P, f_Q the labeling functions over $\mathcal{D}_P$ and $\mathcal{D}_Q$, respectively. Given a hypothesis class $\mathcal{G}$ and the corresponding loss function ℓ, the $\mathcal{Y}$-discrepancy between $(\mathcal{D}_P, f_P)$ and $(\mathcal{D}_Q, f_Q)$ is defined as

$$\mathrm{disc}_{\mathcal{Y}}(\mathcal{D}_P, \mathcal{D}_Q) = \sup_{g \in \mathcal{G}} \left| R_{\mathcal{D}_P}(g, f_P) - R_{\mathcal{D}_Q}(g, f_Q) \right| .$$

As we only have the empirical data on hand, by introducing weights $\boldsymbol{\alpha}$ over the empirical data S_P sampled from $\mathcal{D}_P$, the weighted empirical $\mathcal{Y}$-discrepancy is denoted by

$$\mathrm{disc}_{\mathcal{Y}}(S_{P_{\boldsymbol{\alpha}}}, S_Q) = \sup_{g \in \mathcal{G}} \left| \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, f_P) - \widehat{R}_{S_Q}(g, f_Q) \right| .$$

With the definition of weighted empirical $\mathcal{Y}$-discrepancy, the generalization error on $(\mathcal{D}_Q, f_Q)$ can be bounded in terms of the risk over $(\mathcal{D}_{P_{\boldsymbol{\alpha}}}, f_P)$ and their $\mathcal{Y}$-discrepancy.

Lemma 2 (Proposition 5 of Cortes et al. (2019))**.** *Let $\mathcal{G}$ be a family of functions. Suppose the loss function associated to $\mathcal{G}$ is L-Lipschitz, then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample S_Q of size n, the following inequality holds for all $g \in \mathcal{G}$ and any weighted empirical distribution $P_{\boldsymbol{\alpha}}$ over the sample S_P,*

$$R_{\mathcal{D}_Q}(g, f_Q) \leq \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, f_P) + \mathrm{disc}_{\mathcal{Y}}(S_{P_{\boldsymbol{\alpha}}}, S_Q) + 2L\mathfrak{R}_n(\mathcal{G}) + M_Q\sqrt{\frac{\log(1/\delta)}{2n}},$$

where $M_Q = \sup_{\mathbf{x}\in\mathcal{X}, g\in\mathcal{G}} \ell(g(\mathbf{x}), f_Q)$.

Proof. Starting with a standard Rademacher complexity bound for $\mathcal{G}$, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $g \in \mathcal{G}$,

$$\begin{aligned} & R_{\mathcal{D}_Q}(g, f_Q) \\ \leq\ & \widehat{R}_{S_Q}(g, f_Q) + 2L\mathfrak{R}_n(\mathcal{G}) + M_Q\sqrt{\frac{\log(1/\delta)}{2n}} && (1) \\ =\ & \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, f_P) + \widehat{R}_{S_Q}(g, f_Q) - \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, f_P) + 2L\mathfrak{R}_n(\mathcal{G}) + M_Q\sqrt{\frac{\log(1/\delta)}{2n}} && (2) \\ \leq\ & \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, f_P) + \mathrm{disc}_{\mathcal{Y}}(S_{P_{\boldsymbol{\alpha}}}, S_Q) + 2L\mathfrak{R}_n(\mathcal{G}) + M_Q\sqrt{\frac{\log(1/\delta)}{2n}}, && (3) \end{aligned}$$

where the inequality (1) holds due to Lemma 1, a standard generalization error analysis based on the Rademacher complexity argument. Then we introduce the weighted empirical risk $\widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, f_P)$ in (2). The inequality (3) follows from the fact that $\widehat{R}_{S_Q}(g, f_Q) - \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, f_P) \leq \mathrm{disc}_{\mathcal{Y}}(S_{P_{\boldsymbol{\alpha}}}, S_Q)$ holds for any $S_{P_{\boldsymbol{\alpha}}}$ by Definition 3. □

A.4. Notations

Before proving our main results, we review the notations for the Feature space and Distribution Evolvable Stream Learning (FDESL) problems.

In streaming data learning, at each time, a batch of data is received where only their features are available. We require to predict their labels before receiving the true labels. In our scenario, both feature space and data distribution of the consecutive data batches might be changing. We state the specific setting in the following.

Consider the two consecutive batches in the data stream. Let $\mathcal{X}_P \subseteq \mathbb{R}^{d_1}$ be the feature space of the previous batch of size m and $\mathcal{X}_C \subseteq \mathbb{R}^{d_2}$ be the feature space of the current batch of size n, where $d_1 \neq d_2$. We denote by $\mathcal{Y}_P = \{-1, +1\}$ the label space for the previous batch. Following the pioneering work (Hou et al., 2017), we assume the existence of *evolving data* across two consecutive batches. By exploiting these evolving data, we can bridge the gap of two batches with different feature spaces. We split the two batches into three stages: the previous stage, the evolving stage, and the current stage.

- Previous stage: in the previous stage, we have labeled data $S_P = \{(\mathbf{x}_{P_1}, y_{P_1}), \ldots, (\mathbf{x}_{P_{m-k}}, y_{P_{m-k}})\}$, where $(\mathbf{x}_{P_i}, y_{P_i}) \in \mathcal{X}_P \times \mathcal{Y}_P$.
- Evolving stage: in the evolving stage, the data samples across two consecutive data batches have both feature representations in $\mathcal{X}_P$ and $\mathcal{X}_C$. We denote by $S_{\widetilde{P}} = \{\mathbf{x}_{\widetilde{P}_{m-k+1}}, \ldots, \mathbf{x}_{\widetilde{P}_m}\}$ the evolving data on previous data batch and $S_{\widetilde{C}} = \{\mathbf{x}_{\widetilde{C}_1}, \ldots, \mathbf{x}_{\widetilde{C}_k}\}$ on the current data batch, where $\mathbf{x}_{\widetilde{P}_i} \in \mathcal{X}_P$ and $\mathbf{x}_{\widetilde{C}_j} \in \mathcal{X}_C$.
- Current stage: in the current stage, we have unlabeled data $S_C = \{\mathbf{x}_{C_{k+1}}, \ldots, \mathbf{x}_{C_n}\}$, where $\mathbf{x}_{C_j} \in \mathcal{X}_C$.

Notice that the evolving stage does not last for a long time, namely, we have $k \ll m$ and $k \ll n$. In our formulation, we do not use the labels of $S_{\widetilde{P}}$ in the evolving stage, to avoid the potential problems caused by the non-stationary environment. That is, we exploit the aligned unlabeled data as a bridge to link the two batches with different feature spaces.

Moreover, besides the evolvable feature space, the data distribution could also change in the data stream, particularly when the data are collected from open and dynamic environments. The data distribution within each stage is supposed stationary, while distribution can change across the stages. Specifically, the distribution of $S_{\widetilde{P}}$ differs from that of S_P, and the distribution of $S_{\widetilde{C}}$ differs from that of S_C. Table 1 summarizes the main notations and corresponding definitions.

Table 1. Main Notations and Corresponding Definitions

Notation	Definition
$\boldsymbol{\alpha}$	Weights of data in the previous stage
$\boldsymbol{\beta}$	Weights of data in the evolving stage
$S_{P_{\boldsymbol{\alpha}}} = \{(\mathbf{x}_{P_{\boldsymbol{\alpha}}}, y_P)\} \in \mathbb{R}^{(m-k)\times d_1}$	Previous data in previous batch with weights $\boldsymbol{\alpha}$
$S_{\widetilde{P}_{\boldsymbol{\beta}}} = \{\mathbf{x}_{\widetilde{P}_{\boldsymbol{\beta}}}\} \in \mathbb{R}^{k\times d_1}$	Evolving data in previous batch with weights $\boldsymbol{\beta}$
$S_{\widetilde{C}_{\boldsymbol{\beta}}} = \{\mathbf{x}_{\widetilde{C}_{\boldsymbol{\beta}}}\} \in \mathbb{R}^{k\times d_2}$	Evolving data in current batch with weights $\boldsymbol{\beta}$
$S_C = \{\mathbf{x}_C\} \in \mathbb{R}^{(n-k)\times d_2}$	Current data in current batch
$g \in \mathcal{G}$	Classifier in previous feature space
$h \in \mathcal{H}$	Classifier in current feature space

B. Proofs of Main Results

In this section, we provide detailed proofs of main theoretical results in the paper, including Theorem 1 and Proposition 1.

B.1. Proof of Theorem 1

Theorem 1 (Restatement of Theorem 1)**.** *Let $\mathcal{G}$ and $\mathcal{H}$ be two families of classifiers, which might be associated with different feature spaces. Suppose that loss function ℓ is L-Lipschitz and σ-admissible. Then, for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$R_{\mathcal{D}_C}(h, f_C) \le \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) + \mathrm{disc}_E(S_P, S_C) + 2L\mathfrak{R}_n(\mathcal{H}) + M_C\sqrt{\frac{\log(1/\delta)}{2n}},$$

where $M_C = \sup_{\mathbf{x}_C \in \mathcal{X}, h\in\mathcal{H}} \ell(h(\mathbf{x}_C), y_C)$ and $\mathfrak{R}_n(\mathcal{H})$ is the Rademacher complexity of the function family $\mathcal{H}$.

Proof. We begin with standard generalization error bound on Rademacher complexity. We introduce the empirical weights $\boldsymbol{\alpha}$ over previous data S_P, $\boldsymbol{\beta}$ over evolving data $S_{\widetilde{P}}$ and $S_{\widetilde{C}}$. For any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds for all $g \in \mathcal{G}$ and $h \in \mathcal{H}$,

$$R_{\mathcal{D}_C}(h, f_C)$$

$$\le \widehat{R}_{S_C}(h, y_C) + 2L\mathfrak{R}_n(\mathcal{H}) + M_C\sqrt{\frac{\log(1/\delta)}{2n}} \tag{4}$$

$$= \widehat{R}_{S_C}(h, y_C) + \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}}) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}}) + \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_{\widetilde{C}}) - \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_{\widetilde{C}}) + 2L\mathfrak{R}_n(\mathcal{H}) + M_C\sqrt{\frac{\log(1/\delta)}{2n}} \tag{5}$$

$$\le \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}}) + \left|\widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}}) - \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_{\widetilde{C}})\right| + \sup_{h\in\mathcal{H}}\left|\widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_{\widetilde{C}}) - \widehat{R}_{S_C}(h, y_C)\right| + 2L\mathfrak{R}_n(\mathcal{H}) + M_C\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{6}$$

In inequality (4), we use Lemma 1, that is, a standard Rademacher complexity based generalization error bound. Then, we introduce the weighted empirical risk in the evolving stage in (5). The last inequality holds due to taking the absolute value and supremum over the hypothesis set.

We first upper bound the first term in (6). As we only have the labeled data in the previous stage, we rewrite the weighted empirical risk in the evolving stage in terms of the weighted empirical risk in the previous stage and their discrepancy,

$$\widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}}) \le \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) + \sup_{g\in\mathcal{G}}\left|\widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}})\right|. \tag{7}$$

Then, we upper bound the second term in (6). In the evolving stage, data essentially share the same label, i.e., $y_{\widetilde{P}_i} = y_{\widetilde{C}_i}, i = 1, \ldots, k$. By exploiting the σ-admissibility property of the loss functions defined in Definition 2, we have the following inequality,

$$\begin{aligned}
&\left|\widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}}) - \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_{\widetilde{C}})\right| \\
=& \left|\sum_{i=1}^{k} \beta_i \ell(g(\mathbf{x}_{\widetilde{P}_i}), y_{\widetilde{P}_i}) - \sum_{i=1}^{k} \beta_i \ell(h(\mathbf{x}_{\widetilde{C}_i}), y_{\widetilde{C}_i})\right| \\
\leq& \, \sigma \sum_{i=1}^{k} \beta_i \left|g(\mathbf{x}_{\widetilde{P}_i}) - h(\mathbf{x}_{\widetilde{C}_i})\right|.
\end{aligned} \tag{8}$$

Equation (8) shows that by exploiting the σ-admissible loss, we can align the classifiers in different feature spaces. Therefore, we can use historical data to learn a classifier in the current stage.

By combining (6), (7) and (8), then, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $g \in \mathcal{G}$ and $h \in \mathcal{H}$,

$$R_{\mathcal{D}_C}(h, f_C) \leq \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) + \mathrm{disc}_E(S_P, S_C) + 2L\mathfrak{R}_n(\mathcal{H}) + M_C\sqrt{\frac{\log(1/\delta)}{2n}},$$

where $\mathrm{disc}_E(S_P, S_C)$ is the evolving discrepancy of the two consecutive batches, defined as

$$\mathrm{disc}_E(S_P, S_C) = \sup_{g\in\mathcal{G}, h\in\mathcal{H}} \left\{\left|\widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}})\right| + \sigma \sum_{i=1}^{k} \beta_i \left|g(\mathbf{x}_{\widetilde{P}_i}) - h(\mathbf{x}_{\widetilde{C}_i})\right| + \left|\widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_{\widetilde{C}}) - \widehat{R}_{S_C}(h, y_C)\right|\right\}.$$

Now, we complete the proof of Theorem 1, which shows that the expected risk in the current stage can be bounded by the weighted empirical risk in the previous stage and their evolving discrepancy. □

B.2. Proof of Proposition 1

Proposition 1 (Restatement of Proposition 1)**.** *For any hypotheses sets $\mathcal{G}$ and $\mathcal{H}$, the evolving discrepancy* $\mathrm{disc}_E(S_P, S_C)$ *is upper bounded by*

$$\mathrm{disc}'_E(S_P, S_C) + \sigma(d_1(g, f_{\widetilde{P}}, \boldsymbol{\beta}) + d_1(\mathcal{G}, f_{\widetilde{P}}, \boldsymbol{\beta}) + d_1(\mathcal{H}, f_C)),$$

where $d_1(\mathcal{G}, f_{\widetilde{P}}, \boldsymbol{\beta}) = \min_{g\in\mathcal{G}} \mathbb{E}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}[|g(\mathbf{x}) - f_{\widetilde{P}}(\mathbf{x})|]$ with $f_{\widetilde{P}}$ being the concept function on the evolving stage, and $d_1(\mathcal{H}, f_C)$ follows a similar definition.

Proof. As the evolving discrepancy involves the unknown labeling function in the evolving stage and the current state, we upper bound the empirical evolving discrepancy by the introduced hypothesis classes. We first review the definition of the evolving discrepancy. For any hypotheses sets $\mathcal{G}$ and $\mathcal{H}$, the evolving discrepancy $\mathrm{disc}_E(S_P, S_C)$ is defined as

$$\begin{aligned}
&\mathrm{disc}_E(S_P, S_C) \\
=& \sup_{g\in\mathcal{G}, h\in\mathcal{H}} \left|\widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}})\right| + \sigma \sum_{i=1}^{k} \beta_i \left|g(\mathbf{x}_{\widetilde{P}_i}) - h(\mathbf{x}_{\widetilde{C}_i})\right| + \left|\widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_{\widetilde{C}}) - \widehat{R}_{S_C}(h, y_C)\right|.
\end{aligned} \tag{9}$$

In order to upper bound the evolving discrepancy, we introduce the notation of d_1 discrepancy. We denote by $d_1(\mathcal{G}, f_{\widetilde{P}}, \boldsymbol{\beta})$ the L_1 distance of hypothesis class $\mathcal{G}$ and $f_{\widetilde{P}}$ with weights $\boldsymbol{\beta}$ over the evolving data $S_{\widetilde{P}}$, and $d_1(\mathcal{H}, f_C)$ follows a similar definition of the L_1 distance of $\mathcal{H}$ and f_C:

$$d_1(\mathcal{G}, f_{\widetilde{P}}, \boldsymbol{\beta}) = \min_{g\in\mathcal{G}} \mathbb{E}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}} \left|g(\mathbf{x}) - f_{\widetilde{P}}(\mathbf{x})\right|, \quad d_1(\mathcal{H}, f_C) = \min_{h\in\mathcal{H}} \mathbb{E}_{S_C} \left|h(\mathbf{x}) - f_C(\mathbf{x})\right|.$$

We also denote by $d_1(g, h, \boldsymbol{\beta})$ the L_1 distance of $g \in \mathcal{G}$ and $h \in \mathcal{H}$ with weights $\boldsymbol{\beta}$ over the evolving data $S_{\widetilde{P}}$ and $S_{\widetilde{C}}$:

$$d_1(g, h, \boldsymbol{\beta}) = \sum_{i=1}^{k} \beta_i \left|g(\mathbf{x}_{\widetilde{P}_i}) - h(\mathbf{x}_{\widetilde{C}_i})\right|$$

For the first term in (9), by the triangle inequality and the σ-admissible loss function, we can write

$$
\begin{aligned}
&\sup_{g \in \mathcal{G}} \left| \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}}) \right| \\
\leq\ & \sup_{g \in \mathcal{G}} \left| \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{g_0}) \right| + \left| \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{g_0}) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}}) \right| && (10) \\
\leq\ & \sup_{g, g' \in \mathcal{G}} \left| \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{g'}) \right| + \left| \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{g_0}) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{\widetilde{P}}) \right| && (11) \\
\leq\ & \sup_{g, g' \in \mathcal{G}} \left| \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{g'}) \right| + \sigma \mathbb{E}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}} \left| y_{g_0} - y_{\widetilde{P}} \right| && (12) \\
\leq\ & \sup_{g, g' \in \mathcal{G}} \left| \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{g'}) \right| + \sigma d_1(\mathcal{G}, f_{\widetilde{P}}, \boldsymbol{\beta}). && (13)
\end{aligned}
$$

In inequality (10), we introduce the auxiliary hypothesis $g_0 \in \mathcal{G}$. By taking the supremum and using the admissibility of the loss function, we obtain (11) and (12). The hypothesis g_0 are later chosen to minimize the distance of $f_{\widetilde{P}}$ to $\mathcal{G}$ with weights $\boldsymbol{\beta}$ over the evolving empirical data, and we thus obtain (13).

A similar result can be obtained for the third term in (9). By the triangle inequality and σ-admissibility, we have

$$
\begin{aligned}
&\sup_{h \in \mathcal{H}} \left| \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_{\widetilde{C}}) - \widehat{R}_{S_C}(h, y_C) \right| \\
\leq & \sup_{h \in \mathcal{H}} \left| \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_{\widetilde{C}}) - \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_g) \right| + \left| \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_g) - \widehat{R}_{S_C}(h, y_{h_0}) \right| + \left| \widehat{R}_{S_C}(h, y_{h_0}) - \widehat{R}_{S_C}(h, y_C) \right| \\
\leq & \sup_{h, h' \in \mathcal{H}} \left| \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_g) - \widehat{R}_{S_C}(h, y_{h'}) \right| + \sigma \mathbb{E}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}} \left| y_g - y_{\widetilde{C}} \right| + \sigma \mathbb{E}_{S_C} \left| y_{h_0} - y_C \right| \\
\leq & \sup_{h, h' \in \mathcal{H}} \left| \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_g) - \widehat{R}_{S_C}(h, y_{h'}) \right| + \sigma d_1(g, f_{\widetilde{P}}, \boldsymbol{\beta}) + \sigma d_1(\mathcal{H}, f_C), && (14)
\end{aligned}
$$

where y_g is the pseudo labels of the evolving data provided by the hypothesis g. Here the hypothesis h_0 is later chosen to minimize the distance of f_C to $\mathcal{H}$. Term $d_1(g, f_{\widetilde{P}}, \boldsymbol{\beta})$ is the L_1 distance between g to the labeling function $f_{\widetilde{P}}$ with the weights $\boldsymbol{\beta}$ over the evolving empirical data. By minimizing the empirical risk in the previous stage and the evolving discrepancy, we can obtain a well-generalized classifier g in the evolving stage, and thus $d_1(g, f_{\widetilde{C}}, \boldsymbol{\beta})$ is relatively small.

Combining (9), (13) and (14), then, we can upper bound the evolving discrepancy as follows

$$
\begin{aligned}
&\mathrm{disc}_E(S_P, S_C) \\
\leq\ & \sup_{g, g' \in \mathcal{G}} \left| \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{g'}) \right| + \sigma d_1(g, h, \boldsymbol{\beta}) + \sup_{h, h' \in \mathcal{H}} \left| \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_g) - \widehat{R}_{S_C}(h, y_{h'}) \right| \\
&\qquad + \sigma (d_1(g, f_{\widetilde{P}}, \boldsymbol{\beta}) + d_1(\mathcal{G}, f_{\widetilde{P}}, \boldsymbol{\beta}) + d_1(\mathcal{H}, f_C)) \\
=\ & \mathrm{disc}'_E(S_P, S_C) + \sigma (d_1(g, f_{\widetilde{P}}, \boldsymbol{\beta}) + d_1(\mathcal{G}, f_{\widetilde{P}}, \boldsymbol{\beta}) + d_1(\mathcal{H}, f_C)),
\end{aligned}
$$

where $\mathrm{disc}'_E(S_P, S_C)$ is defined as

$$
\sup_{g, g' \in \mathcal{G}, h, h' \in \mathcal{H}} \left| \widehat{R}_{S_{P_{\boldsymbol{\alpha}}}}(g, y_P) - \widehat{R}_{S_{\widetilde{P}_{\boldsymbol{\beta}}}}(g, y_{g'}) \right| + \sigma d_1(g, h, \boldsymbol{\beta}) + \left| \widehat{R}_{S_{\widetilde{C}_{\boldsymbol{\beta}}}}(h, y_g) - \widehat{R}_{S_C}(h, y_{h'}) \right|.
$$

We now complete the proof of Proposition 1. Proposition 1 shows that we can optimize the upper bound of the evolving discrepancy by the empirical data on hand with hypothesis classes. □

C. Experiment Details

In this section, we present the implementation details of the experiments. We first review the detailed information of the four datasets in our empirical study. Then, we provide the details for the implementation of the proposed EDM algorithm.

For the FDESL tasks in the real-world applications, we perform the empirical studies on the following two datasets,

- RFID Dataset (Hou et al., 2017). The RFID dataset[1] contains the real-time data stream collected by the RFID technique. The RFID aerials record the signal as features, and we aim to predict the ground-truth location of the moving goods. We split all the position index into four categories and thus generate their labels. Before the aerials expired, we will arrange new aerials beside the old ones, which change the feature space and formulate the evolving data. We also record the time stamp when collecting the corresponding feature of each coming data. We chronologically generate the feature space and distribution evolvable streams with the batch data size of 800 and the evolving data size of 200.

- Amazon Dataset (McAuley et al., 2015). The Amazon dataset[2] contains the product's quality (label) from 2006 to 2008 according to the ratings of its users (feature). For each subset in our empirical study, e.g., Books, Movies, CDs, each product's label is its quality, which is calculated by the weighted combination of each user's rating. The weight of each rating is calculated by the quality of its user, and the quality of each user is calculated by the "helpfulness" (one of the attributes of the dataset) of the user's reviews. As time goes on, some users sign out while the new users are signing up. Thus, the feature space evolves, by the old features disappear and new features emerge. We find some periods in which both old and new features exist to formulate the evolving data. We split the user ratings into two categories and thus generate the binary streaming classification task. We generate the FDESL tasks with the batch data size of 1000 and the evolving data size of 200.

We further examine the performance of the EDM algorithm on more extensive scenarios, where the evolving stream is characterized by the textual information simulated by the real-world datasets. We conduct the empirical studies on the following two datasets,

- Reuters Multilingual Dataset (Amini et al., 2009). The Reuters multilingual dataset[3] contains about 11K articles from 6 classes in 5 languages so that we could simulate the evolving stream by different languages. As each document is translated into other languages, thus we simulate the evolving stage. For the evolving data across two batches, we simulate them by two different languages so that they share the different feature space. As the data distribution does not change naturally, we also simulate the distribution change issue by biased sampling with different class-prior probability. Namely, the class-prior in previous/current data and the one in the evolving data are different. We suppose that the data come in a batch-style streaming way and set the batch size as 3000, in which the size of the evolving data is 600. All documents are represented by using the TF-IDF feature.
- Cross-Language Dataset (Ng et al., 2012). The Cross-Language dataset[4] is a binary classification dataset contains documents from Google with English, Chinese, and French pages, so that we can simulate the evolving batches from any two of these three languages as they share the different feature spaces. We additionally supplement the two consecutive batches by crawled data from Wikipedia to simulate the evolving data, as each article in Wikipedia has multiple language versions. Thus, the distribution of evolving data also differs from the previous and current data, which simulates the issue of distribution change. We suppose that the data come in a batch-style streaming way and set the batch size as 3000, in which the size of the evolving data is 600.

In the following, we provide the details for the implementation of the proposed EDM algorithm. We use a full connection layer as the feature extractor with width 1024. We also set the main classifiers (min-player) and auxiliary classifiers (max-player) in the adversarial network as two full connection layers neural networks. For the optimization, we use the mini-batch SGD with initial learning rate 0.004 and the Nesterov momentum 0.9.

[1] http://www.lamda.nju.edu.cn/data_RFID.ashx

[2] http://jmcauley.ucsd.edu/data/amazon/links.html

[3] https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual,+Multiview+Text+Categorization+Test+collection

[4] https://hkumath.hku.hk/~mng/mng_files/cross-language.rar

References

Amini, M., Usunier, N., and Goutte, C. Learning from multiple partially observed views-an application to multilingual text categorization. In *Advances in Neural Information Processing Systems*, pp. 28–36, 2009.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

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

Cortes, C., Mohri, M., and Medina, A. M. Adaptation based on generalized discrepancy. *Journal of Machine Learning Research*, 20(1):1–30, 2019.

Hou, B.-J., Zhang, L., and Zhou, Z.-H. Learning with feature evolvable streams. In *Advances in Neural Information Processing Systems 30*, pp. 1417–1427, 2017.

Kifer, D., Ben-David, S., and Gehrke, J. Detecting change in data streams. In *Proceedings of the 30th International Conference on Very Large Data Bases*, pp. 180–191, 2004.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, 2013.

McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52, 2015.

Mohri, M. and Medina, A. M. New analysis and algorithm for learning with drifting distributions. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, pp. 124–138, 2012.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. The MIT Press, second edition, 2018.

Ng, M. K., Wu, Q., and Ye, Y. Co-transfer learning via joint transition probability graph based method. In *Proceedings of the 1st International Workshop on Cross Domain Knowledge Discovery in Web and Social Network Mining*, pp. 1–9. ACM, 2012.