# Supporting Document for the paper "Training a Dynamic Growing Mixture Model for Lifelong Learning"

Fei Ye and Adrian G. Bors

## APPENDIX A
## PROOF OF THEOREM 1

We adopt similar derivations to Theorem 8 from Domain Adaptation theory [1]. We consider fixing the classifier $h \in \mathcal{H}$. According to the triangle inequality property of $\mathcal{L}(\cdot,\cdot)$ (Definition 4) and the definition of the discrepancy distance $\mathcal{L}_d(\cdot,\cdot)$ form, we have [1]:

$$\begin{aligned}\mathcal{L}_{S_i}(h, f_{S_i}) &\leq \mathcal{L}_{S_i}(h, h_{\tilde{S}_i^{(t-i)}}) + \mathcal{L}_{S_i}(h_{S_i}, h_{\tilde{S}_i^{(t-i)}}) \\ &\quad + \mathcal{L}_{S_i}(h_{S_i}, f_{S_i}) \\ &\leq \mathcal{L}_{\tilde{S}_i^{(t-i)}}\big(h, f_{\tilde{S}_i^{(t-i)}}\big) + \mathcal{L}_d\big(S_{i,X}, \tilde{S}_{i,X}^{(t-i)}\big) \\ &\quad + \mathcal{L}_{S_i}(f_{S_i}, h_{S_i}) + \mathcal{L}_{\tilde{S}_i^{(t-i)}}(h_{S_i}, h_{\tilde{S}_i^{(t-i)}})\,.\end{aligned} \tag{1}$$

Then we rewrite Eq. (1) as :

$$\begin{aligned}\mathcal{L}_{S_i}\big(h, f_{S_i}\big) &\leq \mathcal{L}_{\tilde{S}_i^{(t-i)}}\big(h, f_{\tilde{S}_i^{(t-i)}}\big) + \mathcal{L}_{\tilde{S}_i^{(t-i)}}\big(h, h_{\tilde{S}_i^{(t-i)}}\big) \\ &\quad - \mathcal{L}_{\tilde{S}_i^{(t-i)}}\big(h, f_{\tilde{S}_i^{(t-i)}}\big) \\ &\quad + \mathcal{L}_{S_i}(f_{S_i}, h_{S_i}) + \mathcal{L}_{\tilde{S}_i^{(t-i)}}(h_{S_i}, h_{\tilde{S}_i^{(t-i)}}) \\ &\quad + \mathcal{L}_d\big(S_{i,X}, \tilde{S}_{i,X}^{(t-i)}\big)\end{aligned} \tag{2}$$

□

This proves Theorem 1.

## APPENDIX B
## PROOF OF THEOREM 2

Firstly, we take $\tilde{S}_i^{(t-i)}$ and $\tilde{S}_i^{(t-i-1)}$ to be the source and the target distribution and we then have a bound, according to *Theorem 1*:

$$\begin{aligned}&\mathcal{L}_{\tilde{S}_i^{(t-i-1)}}\big(h, f_{\tilde{S}_i^{(t-i-1)}}\big) \leq \mathcal{L}_{\tilde{S}_i^{(t-i)}}\big(h, f_{\tilde{S}_i^{(t-i)}}\big) \\ &+ \mathcal{L}_d\big(\tilde{S}_{i,X}^{(t-i-1)}, \tilde{S}_{i,X}^{(t-i)}\big) + f'\big(S_i^{(t-i-1)}, \tilde{S}_i^{(t-i)}\big).\end{aligned} \tag{3}$$

Furthermore, we take $\tilde{S}_i^{(t-i-1)}$ and $\tilde{S}_i^{(t-i-2)}$ to be the source and the target distribution and we then have a bound:

$$\begin{aligned}&\mathcal{L}_{\tilde{S}_i^{(t-i-2)}}\big(h, f_{\tilde{S}_i^{(t-i-2)}}\big) \leq \mathcal{L}_{\tilde{S}_i^{(t-i-1)}}\big(h, f_{\tilde{S}_i^{(t-i-1)}}\big) \\ &+ \mathcal{L}_d\big(\tilde{S}_{i,X}^{(t-i-2)}, \tilde{S}_{i,X}^{(t-i-1)}\big) + f'\big(S_i^{(t-i-2)}, \tilde{S}_i^{(t-i-1)}\big).\end{aligned} \tag{4}$$

$\tilde{S}_i^{(t-i-2)}$ is statistically closer to the distribution of the training set of the $i$-th task when comparing with $\tilde{S}_i^{(t-i-1)}$ due to the GRM process (See details in Definition 4 of the paper). So the risk bound is reasonable for Eq.(4). Similarly, we consider $\tilde{S}_i^{t-i-2}$ and $\tilde{S}_i^{t-i-3}$ as the source and target distribution and by the mathematical induction, we have the following risk bounds :

$$\begin{aligned}&\mathcal{L}_{\tilde{S}_i^{(t-i-3)}}\big(h, f_{\tilde{S}_i^{(t-i-3)}}\big) \leq \mathcal{L}_{\tilde{S}_i^{(t-i-2)}}\big(h, f_{\tilde{S}_i^{(t-i-2)}}\big) \\ &+ \mathcal{L}_d\big(\tilde{S}_{i,X}^{(t-i-3)}, \tilde{S}_{i,X}^{(t-i-2)}\big) + f'\big(S_i^{(t-i-3)}, \tilde{S}_i^{(t-i-2)}\big) \\ &\mathcal{L}_{\tilde{S}_i^{(t-i-4)}}\big(h, f_{\tilde{S}_i^{(t-i-4)}}\big) \leq \mathcal{L}_{\tilde{S}_i^{(t-i-3)}}\big(h, f_{\tilde{S}_i^{(t-i-3)}}\big) \\ &+ \mathcal{L}_d\big(\tilde{S}_{i,X}^{(t-i-4)}, \tilde{S}_{i,X}^{(t-i-3)}\big) + f'\big(S_i^{(t-i-4)}, \tilde{S}_i^{(t-i-3)}\big) \\ &\cdots \\ &\cdots \\ &\mathcal{L}_{\tilde{S}_i^{(0)}}\big(h, f_{\tilde{S}_i^{(0)}}\big) \leq \mathcal{L}_{\tilde{S}_i^{(1)}}\big(h, f_{\tilde{S}_i^{(1)}}\big) \\ &+ \mathcal{L}_d\big(\tilde{S}_{i,X}^{(0)}, \tilde{S}_{i,X}^{(1)}\big) + f'\big(S_i^{(0)}, \tilde{S}_i^{(1)}\big) \\ &\mathcal{L}_{\tilde{S}_i^{(-1)}}\big(h, f_{\tilde{S}_i^{(-1)}}\big) \leq \mathcal{L}_{\tilde{S}_i^{(0)}}\big(h, f_{\tilde{S}_i^{(0)}}\big) \\ &+ \mathcal{L}_d\big(\tilde{S}_{i,X}^{(-1)}, \tilde{S}_{i,X}^{(0)}\big) + f'\big(S_i^{(-1)}, \tilde{S}_i^{(0)}\big),\end{aligned} \tag{5}$$

where $\tilde{S}_i^{(-1)} = S_i$ and $\tilde{S}_i^{(0)}$ represent the distribution of the testing set and the training set of the $(i)$-th task, respectively. $h_{\tilde{S}_i^{(-1)}}$ is defined by $h_{\tilde{S}_i^{(-1)}} = \arg\min_{h_{\tilde{S}_i^{(-1)}} \in \mathcal{H}} \mathcal{L}_{\tilde{S}_i^{(-1)}}(h, \tilde{S}_i^{(-1)})$. Then we sum up the right-hand and left-hand sides of all equations (Eq. (3), Eq. (4) and Eq. (5)), resulting in :

$$\begin{aligned}&\mathcal{L}_{\tilde{S}_i^{(-1)}}\big(h, f_{\tilde{S}_i^{(-1)}}\big) \leq \mathcal{L}_{\tilde{S}_i^{(t-i)}}\big(h, f_{\tilde{S}_i^{(t-i)}}\big) \\ &+ \sum_{k=-1}^{t-i-1}\Big(\mathcal{L}_d(\tilde{S}_{i,X}^{(k)}, \tilde{S}_{i,X}^{(k+1)}) + f'(\tilde{S}_i^{(k)}, \tilde{S}_i^{(k+1)})\Big).\end{aligned} \tag{6}$$

Since we have $\tilde{S}_i^{(-1)} = S_i$, we can rewrite Eq. (6) as :

$$\begin{aligned}&\mathcal{L}_{S_i}\big(h, f_{S_i}\big) \leq \mathcal{L}_{S_i^{(t-i)}}\big(h, f_{\tilde{S}_i^{(t-i)}}\big) \\ &+ \sum_{k=-1}^{t-i-1}\Big(\mathcal{L}_d(\tilde{S}_{i,X}^{(k)}, \tilde{S}_{i,X}^{(k+1)}) + f'(\tilde{S}_i^{(k)}, \tilde{S}_i^{(k+1)})\Big)\end{aligned} \tag{7}$$

□

## APPENDIX C
## THE PROOF OF LEMMA 3

Firstly, we consider the risk bound for the tasks that are trained only once, we have :

$$\begin{aligned}&\sum_{i=1}^{\mathrm{card}(B)} \mathcal{L}_{S_{b_i}}\Big(h, h_{\zeta_{f_t(b_i)}})\Big) \leq \\ &\sum_{i=1}^{\mathrm{card}(B)} \Big\{\mathcal{L}_{\tilde{S}^{(0)}_{b_i}}\big(h_{\zeta_{f_t(b_i)}}, f_{\tilde{S}^{(0)}_{b_i}}\big) + f'\big(S_{b_i}, \tilde{S}^{(0)}_{b_i}\big) \\ &+ \mathcal{L}_d\big(S_{b_i,X}, \tilde{S}^{(0)}_{b_i,X}\big)\Big\}.\end{aligned} \tag{8}$$

Then we derive the risk bound for the tasks that are trained more than once and therefore have accumulated error terms, we have :

$$\begin{aligned}&\sum_{i=1}^{\mathrm{card}(B')} \mathcal{L}_{S_{b'_i}}\Big(h_{\zeta_{f_t(b'_i)}}, f_{S_{b'_i}}\Big) \leq \\ &\sum_{i=1}^{\mathrm{card}(B')} \left(\begin{array}{l} \mathcal{L}_{\tilde{S}^{(t-\hat{b}_i)}_{b'_i}}(h_{\zeta_{f_t(b'_i)}}, f_{\tilde{S}^{(t-\hat{b}_i)}_{b'_i}}) + \\ \sum_{k=-1}^{\hat{b}_i-1}\Big(\mathcal{L}_d(\tilde{S}^{k}_{b'_i,X}, \tilde{S}^{(k+1)}_{b'_i,X}) + f'(\tilde{S}^{k}_{b'_i}, \tilde{S}^{(k+1)}_{b'_i})\Big). \end{array}\right)\end{aligned} \tag{9}$$

Then we sum up Eq. (8) and Eq. (9), resulting in :

$$\begin{aligned}&\sum_{i=1}^{t}\Big\{\mathcal{L}_{S_i}\left(h_{\zeta_i}, S_i\right)\Big\} \leq \sum_{i=1}^{\mathrm{card}(B)} \Big\{\mathcal{L}_{\tilde{S}^{(0)}_{b_i}}\big(h_{\zeta_{f_t(b_i)}}, f_{\tilde{S}^{(0)}_{b_i}}\big) \\ &+ f'\big(S_{b_i}, \tilde{S}^{(0)}_{b_i}\big) + \mathcal{L}_d\big(S_{b_i,X}, \tilde{S}^{(0)}_{b_i,X}\big)\Big\} \\ &+ \sum_{i=1}^{\mathrm{card}(B')} \Big\{\mathcal{L}_{\tilde{S}^{(t-\hat{b}_i)}_{b'_i}}(h_{\zeta_{f_t(b'_i)}}, f_{\tilde{S}^{(t-\hat{b}_i)}_{b'_i}}) \\ &+ \sum_{k=-1}^{\hat{b}_i-1}\Big(\mathcal{L}_d(\tilde{S}^{k}_{b'_i,X}, \tilde{S}^{(k+1)}_{b'_i,X}) + f'(\tilde{S}^{k}_{b'_i}, \tilde{S}^{(k+1)}_{b'_i})\Big)\Big\},\end{aligned} \tag{10}$$

where we employ $\sum_{i=1}^{t}\Big\{\mathcal{L}_{S_i}\left(h_{\zeta_i}, S_i\right)\Big\}$ to represent the summation of the left hand sides of Eq. (8) and Eq. (9).

In the following, we prove $\mathcal{R}_{\text{single}} \geq \mathcal{R}_{\text{mixture}}$. We consider that the mixture model $\mathbf{M}$ only has a single component and we can derive the risk bound for all $t$ tasks as :

$$\begin{aligned}\mathcal{R}_{\text{single}} &= \mathcal{L}_{\tilde{S}^{(0)}_t}\big(h_{\zeta_1}, f_{\tilde{S}^{(0)}_t}\big) + f'\big(S_t, \tilde{S}^{(0)}_t\big) + \mathcal{L}_d\big(S_{t,X}, \tilde{S}^{(0)}_{t,X}\big) \\ &+ \sum_{i=1}^{t-i}\Big\{\mathcal{L}_{\tilde{S}^{(t-i)}_i}(h_{\zeta_1}, f_{\tilde{S}^{(t-i)}_i}) \\ &+ \sum_{k=-1}^{t-i-1}\Big(\mathcal{L}_d(\tilde{S}^{k}_{i,X}, \tilde{S}^{(k+1)}_{i,X}) + f'(\tilde{S}^{k}_{i}, \tilde{S}^{(k+1)}_{i})\Big)\Big\}.\end{aligned} \tag{11}$$

Then we consider an extreme case for the mixture model in which the number of components is only two. The first component learns the first task and is fixed in the following task learning. We then derive $\mathcal{R}_{\text{mixture}}$ as :

$$\begin{aligned}\mathcal{R}_{\text{mixture}} &= \mathcal{L}_{\tilde{S}^{(0)}_1}\big(h_{\zeta_1}, f_{\tilde{S}^{(0)}_1}\big) + f'\big(S_t, \tilde{S}^{(0)}_1\big) + \mathcal{L}_d\big(S_{1,X}, \tilde{S}^{(0)}_{1,X}\big) \\ &+ \mathcal{L}_{\tilde{S}^{(0)}_t}\big(h_{\zeta_2}, f_{\tilde{S}^{(0)}_t}\big) + f'\big(S_t, \tilde{S}^{(0)}_t\big) + \mathcal{L}_d\big(S_{t,X}, \tilde{S}^{(0)}_{t,X}\big) \\ &+ \sum_{i=2}^{t-i}\Big\{\mathcal{L}_{\tilde{S}^{(t-i)}_i}(h_{\zeta_2}, f_{\tilde{S}^{(t-i)}_i}) \\ &+ \sum_{k=-1}^{t-i-1}\Big(\mathcal{L}_d(\tilde{S}^{k}_{i,X}, \tilde{S}^{(k+1)}_{i,X}) + f'(\tilde{S}^{k}_{i}, \tilde{S}^{(k+1)}_{i})\Big)\Big\}.\end{aligned} \tag{12}$$

where the third and fourth rows in Eq. (12) are the risk bounds used for the following tasks. It clearly sees that Eq. (11) $\geq$ Eq. (12) since the first task in Eq. (11) has more accumulated errors while the first task in Eq. (12) does not suffer from forgetting.

TABLE I
THE NUMBER OF PARAMETERS FOR VARIOUS MODELS AFTER THE UNSUPERVISED LEARNING OF MSFIR AND CCCOS.

| Datasets | LGM [2] | CURL [3] | BE [4] | GMM | Stud |
|---|---|---|---|---|---|
| MSFIR | $3.3 \times 10^8$ | $2.3 \times 10^8$ | $3.6 \times 10^8$ | $2.1 \times 10^8$ | $1.4 \times 10^8$ |
| CCCOS | $1.9 \times 10^9$ | $2.0 \times 10^9$ | $2.0 \times 10^9$ | $7.2 \times 10^8$ | $1.7 \times 10^8$ |

TABLE II
THE NUMBER OF PARAMETERS FOR VARIOUS MODELS UNDER THE LIFELONG SUPERVISED LEARNING OF TASK SEQUENCE MSFIRC.

| Datasets | LGM [2] | CURL [3] | BE [4] | GMM | MRGANs [5] |
|---|---|---|---|---|---|
| MSFIRC | $5.9 \times 10^8$ | $3.3 \times 10^8$ | $3.9 \times 10^8$ | $3.4 \times 10^8$ | $3.3 \times 10^8$ |

## APPENDIX D
## ANALYZING MODEL'S COMPLEXITY

In the following we evaluate the model size for various methods. The number of parameters for unsupervised learning of MSFIR and CCCOS sequences of tasks is reported in Table VII, where 'Stud' denotes the number of parameters for the Student module of GMM. Meanwhile in Table II we provide the number of parameters for the supervised learning of MSFIRC set of tasks.

## APPENDIX E
## KNOWLEDGE ASSIMILATION BY THE STUDENT

Recent studies have extended the Knowledge Distillation (KD) as a method used in the continual learning [6]. This is the first study to investigate how a Student could forget information during its lifelong learning from a good Teacher. The proposed knowledge distillation approach, described in Section-E from the paper cannot absorb entirely the knowledge of all previously learned tasks because the information accuracy depends on the generating capacity of each individual component of the Teacher's mixture module. Using the theoretical framework from Section-C from the paper, we explain Student's learning limitations. First, we observe that the Student's loss function, defined by Eq.(27) from the paper, learns the prior information from all components of the GMM Teacher module, while its effectiveness on the target

set of prior tasks is inextricably linked to the quality of the approximation distribution which can be generated by each component. Thus, we evaluate a risk bound to assess the forgetting by the Student.

*Lemma 1:* Suppose that we have trained an optimal GMM in which the count of experts is the same with the number of tasks. The Student is implemented using a classifier $h_s \in \mathcal{H}$, trained by means of knowledge distillation. We derive a risk bound for the Student module during the $t$th task learning :

$$\sum_{i=1}^{t}\Big\{\mathcal{L}_{S_i}\left(h_s, f_{S_i}\right)\Big\} \leq \sum_{i=1}^{1}\{\mathcal{L}_{\tilde{S}_i^{(1)}}(h_s, f_{\tilde{S}_i^{(1)}}) + f'\big(S_i, \tilde{S}_i^{(1)}\big) + \mathcal{L}_d\big(S_{i,X}, \tilde{S}_{i,\mathcal{X}}^{(1)}\big)\} + \mathcal{L}_{\tilde{S}_t^{(0)}}\big(h_s, f_{\tilde{S}_t^{(0)}}\big) + f'\big(S_t, \tilde{S}_t^{(0)}\big) + \mathcal{L}_d\big(S_{t,X}, \tilde{S}_{t,\mathcal{X}}^{(0)}\big)\,. \tag{13}$$

Using Lemma 2 from Section III-C of the paper, we find that the optimal GMM can reach a tight risk bound, which is not true for the Student module. The reason for this can be explained by Eq. (13). The Student module is trained with the knowledge learned from each component in the GMM module. However, this knowledge represents the degenerate distributions $\{\tilde{S}^1_{1,\mathcal{X}}, \ldots, \tilde{S}^1_{(t-1),\mathcal{X}}\}$, while the Student does not have access to the real training samples from all the previously learned tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_{t-1}\}$. Moreover, the Student module is a static network architecture trained on multiple tasks involving different data domains. Such a static network architecture may also lead to a degraded performance in the target distribution due to the negative backward transfer [7]. A possible approach to reduce the degraded performance of the Student module consists in applying regularisation [7], which would regulate the network optimisation to reduce the negative transfer.

In practice, it would be hard for an GMM model to achieve an optimal network architecture following the lifelong learning procedure explained above. In the following, we analyze the Student's forgetting when considering that the Teacher continually modifies its network architecture, by deriving a new risk bound.

*Lemma 2:* Let $h_s \in \mathcal{H}$ be a Student trained on the knowledge learnt by the Teacher module (GMM) that has an arbitrary number of components during training (usually by training fewer components than the number of tasks $t$). The risk bound for $h_s$ when learning the $(t)$th task is :

$$\sum_{i=1}^{t}\Big\{\mathcal{L}_{S_i}\left(h_s, S_i\right)\Big\} \leq \sum_{i=1}^{\mathrm{card}(B)}\Big(\mathcal{L}_{\tilde{S}^{(1)}_{b_i}}\big(h_s, f_{\tilde{S}^{(1)}_{b_i}}\big) + f'\big(S_{b_i}, \tilde{S}^{(1)}_{b_i}\big) + \mathcal{L}_d\big(S_{b_i,\mathcal{X}}, \tilde{S}^{(1)}_{b_i,\mathcal{X}}\big)\Big) + \sum_{i=1}^{\mathrm{card}(B')}\Big\{\mathcal{L}_{\tilde{S}^{(t-\hat{b}_i+1)}_{b'_i}}(h_s, f_{\tilde{S}^{(t-\hat{b}_i+1)}_{b'_i}}) + \sum_{k=-1}^{\hat{b}_i}\Big(\mathcal{L}_d(\tilde{S}^k_{b'_i,\mathcal{X}}, \tilde{S}^{(k+1)}_{b'_i,\mathcal{X}}) + f'(\tilde{S}^k_{b'_i}, \tilde{S}^{(k+1)}_{b'_i})\Big)\Big\} + \mathcal{L}_{\tilde{S}_t^{(0)}}\big(h_s, f_{\tilde{S}_t^{(0)}}\big) + f'\big(S_t, \tilde{S}_t^{(0)}\big) + \mathcal{L}_d\big(S_{t,\mathcal{X}}, \tilde{S}_{t,\mathcal{X}}^{(0)}\big)\,, \tag{14}$$

where $B'$ does not involve the $(t)$-th task because the $(t)$-th task is only trained once. Let $\mathrm{R}_s^{\mathrm{mixture}}$ represent the right hand side of Eq. (14). According to Lemma 3 from the paper, we have $\mathrm{R}_s^{\mathrm{mixture}} \geq \mathrm{R}^{\mathrm{mixture}}$. We can easily prove Lemma 2 since the Student module $h_s$ is trained on samples drawn from the $\tilde{S}^2_{b_i}$ even if $b_i$th task is trained just once by one of the GMM' components.

We only consider implementing the Student, within the Teacher-Student architecture, for unsupervised learning. We observe that the Student module under unsupervised learning performs worse than the Teacher module in GMM. This is justified by Eq. (13) and (14), although these are mainly considered for supervised learning.

## REFERENCES


[1] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," in *Proc. Conf. on Learning Theory (COLT), arXiv preprint arXiv:2002.06715*, 2009.

[2] J. Ramapuram, M. Gregorova, and A. Kalousis, "Lifelong generative modeling," *Neurocomputing*, vol. 404, pp. 381–400, 2020.

[3] D. Rao, F. Visin, A. A. Rusu, Y. W. Teh, R. Pascanu, and R. Hadsell, "Continual unsupervised representation learning," in *Advances in Neural Information Proc. Systems (NeurIPS)*, 2019, pp. 7645–7655.

[4] Y. Wen, D. Tran, and J. Ba, "BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning," in *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:2002.06715*, 2020.

[5] C. Wu, L. Herranz, X. Liu, J. van de Weijer, and B. Raducanu, "Memory replay GANs: Learning to generate new categories without forgetting," in *Advances In Neural Inf. Proc. Systems (NIPS)*, 2018, pp. 5962–5972.

[6] M. Zhai, L. Chen, F. Tung, J. He, M. Nawhal, and G. Mori, "Lifelong GAN: Continual learning for conditional image generation," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 2759–2768.

[7] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6467–6476.