Online Cooperative Memorization for Variational Autoencoders Supporting Document

Fei Ye and Adrian G. Bors

Appendix A
Proof of Theorem 1

When $p_\theta(\mathbf{x}\,|\,\mathbf{z})$ is the Gaussian decoder, the computation of $\log p_\theta(\mathbf{x}\,|\,\mathbf{z})$ involves the noise value σ :

$$\log p_\theta\left(\mathbf{x}\,|\,\mathbf{z}\right) = -\frac{1}{2\sigma^2}\|\mathbf{x}-\mu_\theta\left(\mathbf{z}\right)\|^2 - \frac{1}{2}\log 2\pi\sigma^2\,, \quad (1)$$

where $\mu_\theta(\mathbf{z})$ is the mean of distribution $p_\theta(\mathbf{x}\,|\,\mathbf{z})$. In order to simplify Eq. (1), the noise σ is set to $1/\sqrt{2}$, resulting in :

$$\log p_\theta\left(\mathbf{x}\,|\,\mathbf{z}\right) = -\|\mathbf{x}-\mu_\theta\left(\mathbf{z}\right)\|^2 - \frac{1}{2}\log \pi\,. \quad (2)$$

We substract the KL divergence resulting in :

$$\begin{aligned}&\log p_\theta\left(\mathbf{x}\,|\,\mathbf{z}\right) - D_{KL}(q_\omega(\mathbf{x}\,|\,\mathbf{z})\,|\,p(\mathbf{z})) = \\ &-\|\mathbf{x}-\mu_\theta\left(\mathbf{z}\right)\|_2^2 - D_{KL}(q_\omega(\mathbf{x}\,|\,\mathbf{z})\,|\,p(\mathbf{z})) - \frac{1}{2}\log \pi\,.\end{aligned} \quad (3)$$

Then we consider the expectation in both sides, resulting in :

$$\begin{aligned}&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}\big[\log p_\theta\left(\mathbf{x}\,|\,\mathbf{z}\right) - D_{KL}(q_\omega(\mathbf{x}\,|\,\mathbf{z})\,|\,p(\mathbf{z}))\big] \\ &= \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}\big[-\|\mathbf{x}-\mu_\theta\left(\mathbf{z}\right)\|_2^2 \\ &- D_{KL}(q_\omega(\mathbf{x}\,|\,\mathbf{z})\,|\,p(\mathbf{z})) - \frac{1}{2}\log \pi\big]\,.\end{aligned} \quad (4)$$

where the first term in the right-hand side of Eq. (4) can be rewritten as $\mathcal{L}(\mathbf{x}, \mathrm{G}_i(\mathbf{z}))$, and this relationship becomes :

$$\begin{aligned}&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}\big[\log p_\theta\left(\mathbf{x}\,|\,\mathbf{z}\right) - D_{KL}(q_\omega(\mathbf{x}\,|\,\mathbf{z})\,|\,p(\mathbf{z}))\big] \\ &= \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}\big[-\mathcal{L}(\mathbf{x}, \mathrm{G}_i(\mathbf{z})) \\ &- D_{KL}(q_\omega(\mathbf{x}\,|\,\mathbf{z})\,|\,p(\mathbf{z})) - \frac{1}{2}\log \pi\big]\,.\end{aligned} \quad (5)$$

where the first term in the left-hand side (LHS) of Eq. (5) is the ELBO, defined in Eq. (1) of the paper. Since the KL divergence $D_{KL}(\cdot)$ is equal or larger than 0, we have the following inequality :

$$\begin{aligned}&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] = \\ &\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x}, \mathrm{G}_i(\mathbf{z})) \\ &- D_{KL}(q_\omega(\mathbf{z}\,|\,\mathbf{x})\,||\,p(\mathbf{z})) - \frac{1}{2}\log \pi] \\ &\leq \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x}, \mathrm{G}_i(\mathbf{z}))] - \frac{1}{2}\log \pi\,,\end{aligned} \quad (6)$$

From the inequality from Eq. (8) from the paper after multiplying with -1 :

$$-\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x}, \mathrm{P}_{\mathbf{G}_i}) \geq \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x}, \mathrm{G}_i(\mathbf{z}))]\,, \quad (7)$$

and then rewrite Eq. (6) by considering Eq. (7), resulting in:

$$\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] \leq -\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x}, \mathrm{P}_{\mathrm{G}_i}) - \frac{1}{2}\log \pi\,. \quad (8)$$

Eq. (8) proves Theorem 1 □

Appendix B
Proof of Theorem 2

We consider Eq. (8) and add $-\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i}, \mathrm{P}_{\mathrm{G}_i})$ to both sides of resulting in :

$$\begin{aligned}&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i}, \mathrm{P}_{\mathrm{G}_i}) \leq \\ &- \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i}, \mathrm{P}_{\mathrm{G}_i}) - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x}, \mathrm{P}_{\mathrm{G}_i}) - \frac{1}{2}\log \pi\end{aligned} \quad (9)$$

The first term in the right-hand side (RHS) is bounded, similarly to Eq. (7), but on the memory buffer m_i :

$$\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x}, \mathrm{G}_i(\mathbf{z}))] \leq -\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i}, \mathrm{P}_{\mathrm{G}_i}), \quad (10)$$

then we have :

$$\begin{aligned}&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x}, \mathrm{G}_i(\mathbf{z}))] + \\ &\Big|\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x}, \mathrm{G}_i(\mathbf{z}))] - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i}, \mathrm{P}_{\mathrm{G}_i})\Big| \\ &\geq -\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i}, \mathrm{P}_{\mathrm{G}_i}).\end{aligned} \quad (11)$$

Then, by using Eq. (9), we derive :

$$\begin{aligned}&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i}, \mathrm{P}_{\mathrm{G}_i}) \\ &\leq \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x}, \mathrm{G}_i(\mathbf{z}))] - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x}, \mathrm{P}_{\mathbf{G}_i}) \\ &+ \Big|\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}\mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x}, \mathrm{G}_i(\mathbf{z}))] - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i}, \mathrm{P}_{\mathbf{G}_i})\Big| \\ &- \frac{1}{2}\log \pi\,.\end{aligned} \quad (12)$$

We then add the negative KL divergence term in both sides of Eq. (12) :

$$\begin{aligned}
&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i},\mathrm{P}_{\mathrm{G}_i}) \\
&- \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}[D_{KL}(q_\omega(\mathbf{z}\,|\,\mathbf{x})\,||\,p(\mathbf{z}))] \leq \\
&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}} \mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x},\mathrm{G}_i(\mathbf{z}))] \\
&- \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}[D_{KL}(q_\omega(\mathbf{z}\,|\,\mathbf{x})\,||\,p(\mathbf{z}))] - \frac{1}{2}\log\pi \\
&- \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x},\mathrm{P}_{\mathbf{G}_i}) + \Big| \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}} \mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[\\
&- \mathcal{L}(\mathbf{x},\mathrm{G}_i(\mathbf{z}))] - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i},\mathrm{P}_{\mathbf{G}_i}) \Big| ,
\end{aligned} \tag{13}$$

According to the definition of ELBO, this can be rewritten as :

$$\begin{aligned}
&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i},\mathrm{P}_{\mathrm{G}_i}) \\
&- \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}[D_{KL}(q_\omega(\mathbf{z}\,|\,\mathbf{x})\,||\,p(\mathbf{z}))] \leq \\
&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x},\mathrm{P}_{\mathbf{G}_i}) \\
&+ \Big| \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}} \mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x},\mathrm{G}_i(\mathbf{z}))] - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i},\mathrm{P}_{\mathrm{G}_i}) \Big| ,
\end{aligned} \tag{14}$$

Then we rewrite Eq. (14), resulting in :

$$\begin{aligned}
&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] \leq \\
&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] + \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i},\mathrm{P}_{\mathrm{G}_i}) \\
&- \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x},\mathrm{P}_{\mathrm{G_i}}) + \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}[D_{KL}(q_\omega(\mathbf{z}\,|\,\mathbf{x})\,||\,p(\mathbf{z}))] \\
&+ \Big| \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}} \mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x},\mathrm{G}_i(\mathbf{z}))] - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i},\mathrm{P}_{\mathrm{G}_i}) \Big| .
\end{aligned} \tag{15}$$

We consider that $\mathcal{L}(\cdot)$ satisfies the triangle inequality :

$$\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i},\mathrm{P}_{\mathrm{G_i}}) + \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x},\mathrm{P}_{\mathrm{G_i}}) \geq \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x},\mathrm{P}_{m_i}) \tag{16}$$

We move the second term from the LHS of Eq. (16) in the RHS :

$$\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x},\mathrm{P}_{\mathrm{G_i}}) \geq \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x},\mathrm{P}_{m_i}) - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i},\mathrm{P}_{\mathrm{G}_i}) \tag{17}$$

Then we replace $\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x},\mathrm{P}_{\mathrm{G_i}})$ from Eq. (15) by the expression of Eq. (17), resulting in :

$$\begin{aligned}
&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_\mathbf{x}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] \leq \\
&\inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] + 2\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i},\mathrm{P}_{\mathrm{G}_i}) \\
&- \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_\mathbf{x},\mathrm{P}_{m_i}) + \tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}_i},\mathrm{P}_{m_i}) ,
\end{aligned} \tag{18}$$

where $\tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}_i},\mathrm{P}_{m_i})$ is expressed as :

$$\begin{aligned}
\tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}_i},\mathrm{P}_{m_i}) &= \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}}[D_{KL}(q_\omega(\mathbf{z}\,|\,\mathbf{x})\,||\,p(\mathbf{z}))] \\
&+ \Big| \inf_{q_\omega(\mathbf{z})=p(\mathbf{z})} \mathbb{E}_{\mathrm{P}_{m_i}} \mathbb{E}_{q_\omega(\mathbf{z}\,|\,\mathbf{x})}[-\mathcal{L}(\mathbf{x},\mathrm{G}_i(\mathbf{z}))] \\
&- \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{m_i},\mathrm{P}_{\mathbf{G}_i}) \Big|
\end{aligned} \tag{19}$$

□

APPENDIX C
PROOF OF THEOREM 3

Let us firstly consider a certain component (a_i-th component) that has been trained only once. From Theorem 2 we derive the bound as follows :

$$\begin{aligned}
\mathbb{E}_{\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{a}_i}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] &\leq \mathbb{E}_{\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] \\
&+ 2\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}},\mathrm{P}_{\mathrm{G}^{a_i}}) - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{a}_i}},\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}}) \\
&+ \tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}^{a_i}},\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}}) ,
\end{aligned} \tag{20}$$

Eq. (20) holds because we treat $\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{a}_i}}$ and $\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}}$ as the target and source domain respectively. In the following, we consider a component (b_i-th component) that has been trained more than once. Since the b_i-th component would learn more than one task, we particularly focus on a certain task ($\tilde{b}_i^q$-th task). We firstly consider to treat $\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{b}_i^q}}$ and $\mathrm{P}_{\mathbf{x}^{\tilde{b}_i^q}}$ as the target and source domain respectively. Then we derive the bound as :

$$\begin{aligned}
\mathbb{E}_{\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{b}_i^q}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] &\leq \mathbb{E}_{\mathrm{P}_{\mathbf{x}^{\tilde{b}_i^q}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] \\
&+ 2\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\mathbf{x}^{\tilde{b}_i^q}},\mathrm{P}_{\mathrm{G}^{b_i}}) - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{b}_i^q}},\mathrm{P}_{\mathbf{x}^{\tilde{b}_i^q}}) \\
&+ \tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}^{b_i}},\mathrm{P}_{\mathbf{x}^{\tilde{b}_i^q}}) ,
\end{aligned} \tag{21}$$

We do not specify the state (the number of retraining processes) of each generator distribution $\mathrm{P}_{\mathrm{G}i}$ in order to simplify the notation. We have the empirical distribution $\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_i^q,1)}}$ for one time of the generative replay processes (see Definition 6). We treat $\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_i^q,0)}} = \mathrm{P}_{\hat{\mathbf{x}}^{\tilde{b}_i^q}}$ and $\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_i^q,1)}}$ as the target and source domain, respectively. We then derive the bound between $\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_{(i,q)},0)}}$ and $\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_{(i,q)},1)}}$ as follows :

$$\begin{aligned}
\mathbb{E}_{\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_i^q,0)}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] &\leq \mathbb{E}_{\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_i^q,1)}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] \\
&+ 2\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_{(i,q)},1)}},\mathrm{P}_{\mathrm{G}^{b_i}}) - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,1)}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,0)}}) \\
&+ \tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}^{b_i}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,1)}}) ,
\end{aligned} \tag{22}$$

Through mathematical induction, we have the bounds :

$$\begin{aligned}
\mathbb{E}_{\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_i^q,1)}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] &\leq \mathbb{E}_{\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_i^q,2)}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] \\
&+ 2\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,2)}},\mathrm{P}_{\mathrm{G}^{b_i}}) \\
&- \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,2)}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,1)}}) \\
&+ \tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}^{b_i}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,2)}}) \\
&\dots \\
\mathbb{E}_{\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_i^q,c_i^q-1)}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] &\leq \mathbb{E}_{\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_i^q,c_i^q)}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] \\
&+ 2\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,c_i^q)}},\mathrm{P}_{\mathrm{G}^{b_i}}) \\
&- \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,c_i^q)}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,c_i^q-1)}}) \\
&+ \tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}^{b_i}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,c_i^q)}},)
\end{aligned} \tag{23}$$

where c_i^q denotes the number of generative replay processes for the $\tilde{b}_i^q$-th task, achieved by the b_i-th component.

We then sum up all above inequalities, resulting in :

$$\begin{aligned}
\mathbb{E}_{\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{b}_i^q}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] &\leq \mathbb{E}_{\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,c_i^q)}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)] \\
&+ \sum_{s=0}^{\tilde{c}_{(i,q)}} \Big\{ 2\mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,s)}},\mathrm{P}_{\mathrm{G}^{b_i}}) - \mathrm{W}^\star_\mathcal{L}(\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}iq,s-1)}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,s)}}) \\
&+ \tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}^{b_i}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,s)}}) \Big\} .
\end{aligned} \tag{24}$$

Eq. (24) describes the bound for a single task. We then extend this bound to the components learning more than one task:

$$\begin{aligned}
&\sum_{i=1}^{|\mathcal{B}|}\Big\{\sum_{q=1}^{|\tilde{b}_i|}\{\mathbb{E}_{\mathrm{P}_{\tilde{\mathbf{x}}^{\tilde{b}_i^q}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)]\}\Big\}\leq\\
&\sum_{i=1}^{|\mathcal{B}|}\Big\{\sum_{q=1}^{|\tilde{b}_i|}\Big\{\mathbb{E}_{\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,c_i^q)}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)]\\
&+\sum_{s=0}^{c_i^q}\{2\mathrm{W}_{\mathcal{L}}^{\star}(\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,s)}},\mathrm{P}_{\mathrm{G}^{b_i}})\\
&-\mathrm{W}_{\mathcal{L}}^{\star}(\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_i^q,s-1)}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,s)}})+\tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}^{b_i}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,s)}})\}\Big\}\Big\},
\end{aligned}\tag{25}$$

We also extend the bound from Eq. (20) to components that would only learn one task each :

$$\begin{aligned}
&\sum_{i=1}^{|\mathcal{A}|}\Big\{\mathbb{E}_{\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{a}_i}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)]\Big\}\leq\\
&\sum_{i=1}^{|\mathcal{A}|}\Big\{\mathbb{E}_{\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)]+2\mathrm{W}_{\mathcal{L}}^{\star}(\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}},\mathrm{P}_{\mathrm{G}^{a_i}})\\
&-\mathrm{W}_{\mathcal{L}}^{\star}(\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{a}_i}},\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}})+\tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}^{a_i}},\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}})\Big\},
\end{aligned}\tag{26}$$

Eventually, the bound for all components is defined by considering both Eq. (25) and (26), resulting in :

$$\begin{aligned}
&\sum_{i=1}^{|\mathcal{A}|}\Big\{\mathbb{E}_{\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{a}_i}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)]\Big\}\\
&+\sum_{i=1}^{|\mathcal{B}|}\Big\{\sum_{q=1}^{|\tilde{b}_i|}\{\mathbb{E}_{\mathrm{P}_{\tilde{\mathbf{x}}^{\tilde{b}_i^q}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)]\}\Big\}\leq\\
&\sum_{i=1}^{|\mathcal{A}|}\Big\{\mathbb{E}_{\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)]+2\mathrm{W}_{\mathcal{L}}^{\star}(\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}},\mathrm{P}_{\mathrm{G}^{a_i}})\\
&-\mathrm{W}_{\mathcal{L}}^{\star}(\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{a}_i}},\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}})+\tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}^{a_i}},\mathrm{P}_{\mathbf{x}^{\tilde{a}_i}})\Big\}\\
&+\sum_{i=1}^{|\mathcal{B}|}\Big\{\sum_{q=1}^{|\tilde{b}_i|}\Big\{\mathbb{E}_{\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,c_i^q)}}}[\mathcal{L}_{ELBO}(\mathbf{x};\theta,\omega)]\\
&+\sum_{s=0}^{c_i^q}\{2\mathrm{W}_{\mathcal{L}}^{\star}(\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,s)}},\mathrm{P}_{\mathrm{G}^{b_i}})\\
&-\mathrm{W}_{\mathcal{L}}^{\star}(\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}_i^q,s-1)}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,s)}})+\tilde{\mathrm{F}}(\mathrm{P}_{\mathrm{G}^{b_i}},\mathrm{P}_{\tilde{\mathbf{x}}^{(\tilde{b}_i^q,s)}})\}\Big\}\Big\}
\end{aligned}\tag{27}$$

□

APPENDIX D
ALGORITHM 2 - TRAINING A DYNAMIC EXPANSION MODEL WITH OCM

Algorithm 2 Training a dynamic expansion model with OCM

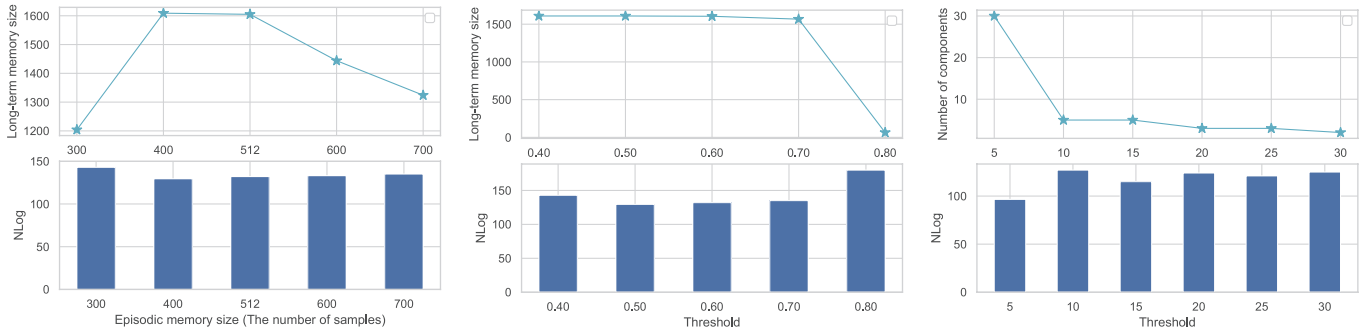
Input:$\mathcal{D}^S$ (Training dataset);

for $\mathcal{T}_i < \mathcal{T}_N$ **do**
  Step 1 (Learning:)
  $\mathbf{X}_b^i \sim \mathcal{D}^S$;
  $\mathbf{X}_b^i \in \mathcal{M}_i^e$;
  Train the model using samples from $\mathcal{M}_i^e$ and $\mathcal{M}_i^l$;
  if $\mathrm{Count}(\mathcal{M}_i^e) \geq n_{Max}^c$ **then**
    Step 2 (Evaluation:)
    $\mathbf{S}_i = \mathrm{F}_{\exp}\Big(-(\mathbf{Z}_i^e(-1\mathbf{Z}_i^l)^{\mathrm{T}}) \odot (\mathbf{Z}_i^e(-1\mathbf{Z}_i^l)^{\mathrm{T}})/2\alpha^2\Big)$. Calculate the graph relationship matrix;
    Step 3 (Selection:)
    for $j < N_i^e$ **do**
      $\mathrm{R}^S(\mathbf{x}_{i,j}^e) = \frac{1}{N_i^l}\sum_{k=1}^{N_i^l}\mathbf{S}_i(j,k)$; Calculate the average similarity score from $\mathbf{x}_{i,j}^e$ to LTM based on $\mathbf{S}_i$;
      if $\mathrm{R}^S(\mathbf{x}_{i,j}^e) > \lambda_1$ **then**
        $\mathcal{M}_i^l = \mathcal{M}_i^l \cup \mathbf{x}_{i,j}^e$; Add $\mathbf{x}_{i,j}^e$ into LTM memory;
      end if
    end for
    Check the expansion:
    R_i Calculated by Eq.(22) from the paper;
    if $|\mathrm{R}_i - \mathrm{R}_{last}| > \lambda_2$ **then**
      Build a new component;
      $\mathcal{M}_i^l = \emptyset$; Clear the LTM memory;
      $\mathrm{R}_{last} = \mathrm{R}_i$;
    end if
    $\mathcal{M}_i^e = \emptyset$; Clear the STM memory;
  end if
end for

APPENDIX E
ABLATION STUDY

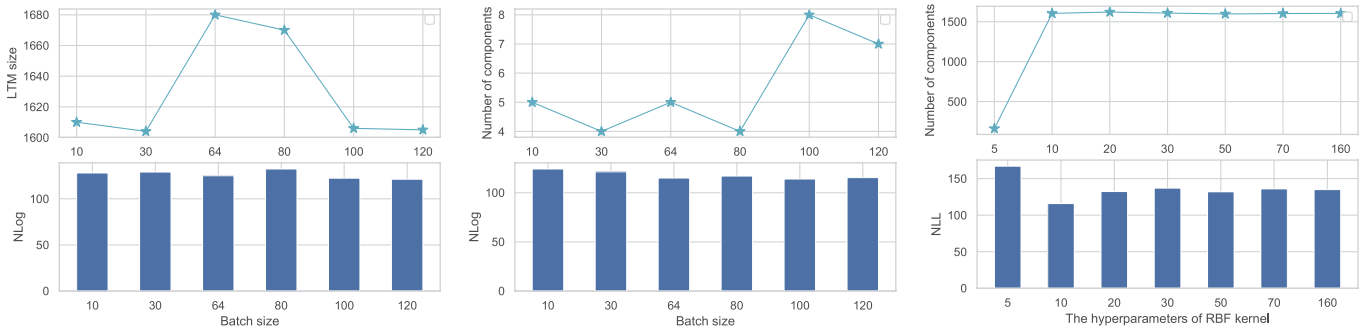
Ablation study for the hyperparameters. We firstly investigate the performance when varying the size of the STM, updated according to Eq. (19), from the paper, for a single VAE model for the lifelong learning of Split MNIST dataset. From the results provided in Fig. 1a, it can be observed that the model faces degenerated performance when the size of STM is very small (300) and such results are improved when considering additional samples in STM. The results when changing the threshold λ_1 from Eq. (23) from the paper, for transferring the data from the STM to the LTM are reported in Fig. 1b. These results indicate that a large λ_1 leads to a smaller memory size and consequently to a drop in performance.

Ablation study for the dynamic expansion. We investigate the performance of Dynamic-ELBO-OCM, by expanding the VAE's network when changing the threshold $\lambda_2 = \{5, 10, 15, 20, 25, 30\}$ in Eq. (25) from the paper, while $\lambda_1 = 0.6$ and the maximum memory size for STM is considered as 500. From the results reported in Fig. 1c we can observe



(a) Changing STM size. (b) Changing λ_1 in Eq. (23) from the paper. (c) Changing λ_2 in Eq. (25) from the paper.

Fig. 1. Assessment of the STM size and when changing the thresholds λ_1 and λ_2 for a single VAE model.



(a) Changing LTM size for a single VAE trained with OCM. (b) Batch size for the OCM expanding model. (c) Changing α in Eq. (**??**).

Fig. 2. Varying the hyperparameters when training OCM under Split MNIST.

that by increasing λ_2 we can reduce the number of components. If λ_2 is very small, such as $\lambda_2 = 5$, the model has more components and consequently its performance improves significantly.

Changing the batch size. We investigate the performance and the change in the memory buffers when varying the batch size. We consider batch sizes of 10, 30, 64, 80, 100, 120 for training a single VAE model with OCM under Split MNIST, and the results are reported in Fig. 2-a. From this plot we can observe that the change of batch size has only a minor influence in the LTM size as well as in the negative log-likelihood (NLog) result. In the following, we evaluate the performance of a dynamic expansion model with OCM under Split MNIST with different batch sizes and the results are reported in Fig. 2b. Changing the batch size does not have a significant influence, either on the performance or on the number of components. In Fig. 3 we present some memorized samples from the MNIST database which are stored in the LTM. The result shows that the proposed approach encourages LTM to store diverse data samples, using Eq. (19) during OCM learning.

RBF kernel scale. We investigate the performance of the proposed OCM framework when changing the hyperparameters of the RBF kernel in Eq. (20) from the paper. We vary the RBF scale $\alpha = \{5, 10, 20, 30, 50, 70, 100\}$ for the lifelong training a single VAE model trained with OCM under Split MNIST. The results presented in Fig. 2c indicate that OCM with $\alpha = 10$ achieves the best results.

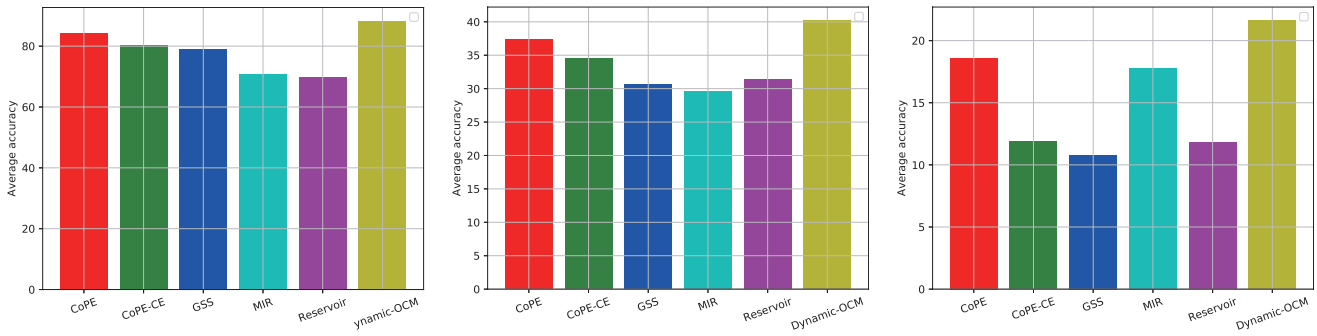
Using the cosine distance for sample selection. We consider

Fig. 3. Memorized samples drawn from LTM.

the cosine distance for evaluating the similarity in the proposed sample selection approach for LTM, instead of the graph based distance from Eq. (22) from the paper, defined as :

$$\mathrm{R}^C(\mathbf{x}^e_{i,j}, \mathbf{x}^l_{i,u}) := \frac{\mathbf{z}^e_{i,j} \cdot \mathbf{z}^l_{i,u}}{\left\|\mathbf{z}^e_{i,j}\right\| \left\|\mathbf{z}^l_{i,u}\right\|} = \frac{\sum_{i=1}^{d_z} \mathbf{z}^e_{i,j}(i)\mathbf{z}^l_{i,u}(i)}{\sqrt{\sum_{i=1}^{d_z} \left(\mathbf{z}^e_{i,j}(i)\right)^2} \sqrt{\sum_{i=1}^{d_z} \left(\mathbf{z}^l_{i,u}(i)\right)^2}}, \tag{28}$$

where the evaluation of similarity is based on the latent features $\mathbf{z}^l_{i,u}$ and $\mathbf{z}^e_{i,j}$, corresponding to the data $\mathbf{x}^l_{i,j}$, $\mathbf{x}^e_{i,u}$, from LTM and STM, respectively.



(a) Split MNIST. (b) Split CIFAR10. (c) Split CIFAR100.

Fig. 4. The results for the imbalanced benchmark where the results of baselines are cited from [1].

TABLE I
LOG-LIKELIHOOD ESTIMATION ON ALL TESTING SAMPLES BY USING THE IWVAE BOUND WITH $m = 1000$ IMPORTANCE SAMPLES IN EQ. (2) FROM THE PAPER.

Methods	Log	Memory	N
VAE-ELBO-OCM-COS	-137.92	1.6K	1
VAE-ELBO-OCM	-132.07	1.6K	1
VAE-IWVAE50-OCM	-127.11	1.6K	1
Dynamic-ELBO-OCM	**-115.89**	1.1K	5

We use "VAE-ELBO-OCM-COS" to represent a single VAE model trained with OCM, where the cosine distance is used as the criterion for the sample selection. Since a small measure in Eq. (26) means that $\mathbf{x}_{i,j}^e$ is far away from $\mathbf{x}_{i,u}^l$, we replace Eq. (23) by considering :

$$\mathrm{R}^C(\mathbf{x}_{i,j}^e, \mathbf{x}_{i,u}^l) < \lambda \Rightarrow \mathcal{M}_i^l = \mathcal{M}_i^l \cup \mathbf{x}_{i,j}^e \,, \tag{29}$$

where we set $\lambda = 0$. The results of various models trained under Split MNIST are provided in Table I, showing that the proposed kernel from Eq. (22) for sample selection outperforms the cosine distance.

Imbalanced Benchmark Results. We follow the imbalanced data stream setting from [1], where several selected datasets have more data samples then others. The network architecture for the imbalanced benchmark is the same as for the balanced setting except for the Split MNIST where we use an MLP network with two hidden layers with 100 units each, with a memory size of 3K. The imbalanced benchmark results are provided in Fig. 4, where the number of components used for the lifelong learning of Split MNIST, Split CIFAR10 and Split CIFAR100 is of 7,6 and 10, respectively. These results show that the proposed OCM with the dynamic expansion mechanism outperforms the state of the art models on the imbalanced data lifelong learning stream setting.

All the results indicate the advantages of using the dynamic expansion mechanism in the OCM and demonstrate the advantages in the improved performance when increasing the model's capacity, as indicated by Lemma 2 from the paper.

REFERENCES

[1] M. De Lange and T. Tuytelaars, "Continual prototype evolution: Learning online from non-stationary data streams," in *Proc. of the IEEE/CVF Int. Conference on Computer Vision (ICCV)*, 2021, pp. 8250–8259.