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 :

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

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 :

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

We substract the KL divergence resulting in :

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

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

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

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 :

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

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 :

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

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

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

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

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

Eq. (8) proves Theorem 1 $\square$

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 :

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

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

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

then we have :

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

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

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

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 :

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

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

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

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

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

Then we replace $\mathrm{W}^\star_{\mathcal{L}}(\mathrm{P}_{\mathbf{x}},\mathrm{P}_{\mathrm{G}_\mathrm{i}})$ 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 ($\tilde{b}^q_i$-th task). We firstly consider to treat $\mathrm{P}_{\hat{\mathbf{x}}^{\tilde{b}^q_i}}$ and $\mathrm{P}_{\mathbf{x}^{\tilde{b}^q_i}}$ as the target and source domain respectively. Then we derive the bound as :

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

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 $\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}^q_i,1)}}$ for one time of the generative replay processes (see Definition 6). We treat $\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}^q_i,0)}} = \mathrm{P}_{\hat{\mathbf{x}}^{\tilde{b}^q_i}}$ and $\mathrm{P}_{\hat{\mathbf{x}}^{(\tilde{b}^q_i,1)}}$ 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 :

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

Through mathematical induction, we have the bounds :

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

where $\underset{\sim}{c^q_i}$ denotes the number of generative replay processes for the $\tilde{b}^q_i$-th task, achieved by the b_i-th component.

We then sum up all above inequalities, resulting in :

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

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

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

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

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

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

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

$\square$

Appendix D
Additional results on ablation study

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. 1 indicate that OCM with $\alpha = 10$ achieves the best results.

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

Using the cosine distance for sample selection. We consider 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 :

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

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.

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 $\mathbf{x}^e_{i,j}$ is far away from $\mathbf{x}^l_{i,u}$, we replace Eq. (23) by considering :

$$\mathrm{R}^{C}(\mathbf{x}^e_{i,j},\mathbf{x}^l_{i,u}) < \lambda \Rightarrow \mathcal{M}^l_i = \mathcal{M}^l_i \cup \mathbf{x}^e_{i,j}\,, \tag{29}$$

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.

References

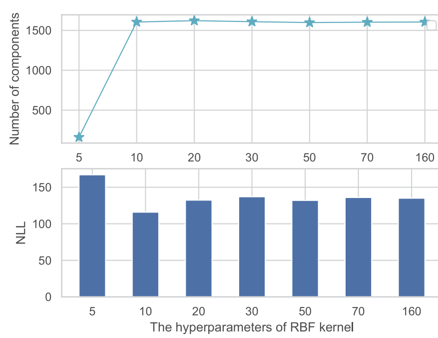

Fig. 1. The performance of the model when Changing α in Eq. (20) of the paper.