A Complete Proof of Theoretical Analysis

A.1 Proof of Proposition 1

Proposition 1. *Let Θ be a cover of a parameter space with VC dimension d. If $\mathbb{D}_1, \cdots, \mathbb{D}_t$ are the distributions of the continually learned $1:t$ tasks, then for any $\delta \in (0,1)$ with probability at least $1-\delta$, for every solution $\theta_{1:t}$ of the continually learned $1:t$ tasks in parameter space Θ, i.e., $\theta_{1:t} \in \Theta$:*

$$\mathcal{E}_{\mathbb{D}_t}(\theta_{1:t}) < \hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta_{1:t}) + \frac{1}{2(t-1)} \sum_{k=1}^{t-1} \mathrm{Div}(\mathbb{D}_k, \mathbb{D}_t) + \sqrt{\frac{d \ln(N_{1:t-1}/d) + \ln(1/\delta)}{N_{1:t-1}}}, \quad (1)$$

$$\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta_{1:t}) < \hat{\mathcal{E}}^b_{D_t}(\theta_{1:t}) + \frac{1}{2(t-1)} \sum_{k=1}^{t-1} \mathrm{Div}(\mathbb{D}_t, \mathbb{D}_k) + \sqrt{\frac{d \ln(N_t/d) + \ln(1/\delta)}{N_t}}, \quad (2)$$

where $\mathrm{Div}(\mathbb{D}_i, \mathbb{D}_j) := 2 \sup_{h \in \mathcal{H}} |\mathcal{P}_{\mathbb{D}_i}(I(h)) - \mathcal{P}_{\mathbb{D}_j}(I(h))|$ is the $\mathcal{H}$-divergence for the distribution $\mathbb{D}_i$ and $\mathbb{D}_j$ ($I(h)$ is the characteristic function). $N_{1:t-1} = \sum_{k=1}^{t-1} N_k$ is the total number of training samples over all old tasks.

We assume that a distribution $\mathbb{D}$ is with input space $\mathcal{X}$ and a global label function $h : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ denotes a label space, and $h(x)$ generates target label for all the input, i.e., $y = h(x)$. Consider a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, c]$ (where c is the upper bound), such that $\ell(y_1, y_2) = 0$ holds if and only if $y_1 = y_2$. Then, we define a population loss over the distribution $\mathbb{D}$ by $\mathcal{E}_{\mathbb{D}}(\theta) = \mathcal{E}_{\mathbb{D}}(f_\theta, h) := \mathbb{E}_{(x,y)\sim\mathbb{D}}[\ell(f_\theta(x), h(x))]$. Let D denote a training set following the distribution $\mathbb{D}$ with N data-label pairs. To minimize $\mathcal{E}_{\mathbb{D}}(\theta)$, we can minimize an empirical risk over the training set D in a parameter space, i.e., $\min_\theta \hat{\mathcal{E}}_D(\theta)$. Further, to find a flat solution, we define a robust empirical risk by the worst case of the neighborhood in parameter space as $\hat{\mathcal{E}}^b_D(\theta) := \max_{\|\Delta\| \le \mathrm{b}} \hat{\mathcal{E}}_{\mathrm{D}}(\theta + \Delta)$, where b is the radius around θ and $\|\cdot\|$ denotes the L2 norm.

Below are one important definition and three critical lemmas for the proof of Proposition 1.

Definition 1. *(Based on Definition 1 of [2]) Given two distributions, $\mathbb{T}$ and $\mathbb{S}$, let $\mathcal{H}$ be a hypothesis class on input space $\mathcal{X}$ and denote by $I(h)$ the set for which $h \in \mathcal{H}$ is the characteristic function: that is, $x \in I(h) \Leftrightarrow h(x) = 1$. The $\mathcal{H}$-divergence between $\mathbb{T}$ and $\mathbb{S}$ is*

$$\mathrm{Div}(\mathbb{T}, \mathbb{S}) = 2 \sup_{h \in \mathcal{H}} |\mathcal{P}_{\mathbb{T}}(I(h)) - \mathcal{P}_{\mathbb{S}}(I(h))|. \quad (3)$$

Lemma 1. *Let $\mathbb{S} = \{\mathbb{S}_i\}_{i=1}^s$ and $\mathbb{T}$ be s source distributions and the target distribution, respectively. The $\mathcal{H}$-divergence between $\{\mathbb{S}_i\}_{i=1}^s$ and $\mathbb{T}$ is bounded as follows:*

$$\mathrm{Div}(\mathbb{S}, \mathbb{T}) \le \frac{1}{\mathrm{s}} \sum_{i=1}^{s} \mathrm{Div}(\mathbb{S}_\mathrm{i}, \mathbb{T}). \quad (4)$$

Proof. By the definition of $\mathcal{H}$-divergence,

$$\begin{aligned}
\mathrm{Div}(\mathbb{S},\mathbb{T}) &= 2\sup_{h\in\mathcal{H}}|\mathcal{P}_{\mathbb{S}}(I(h))-\mathcal{P}_{\mathbb{T}}(I(h))| \\
&= 2\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{s}\frac{1}{s}(\mathcal{P}_{\mathbb{S}_i}(I(h))-\mathcal{P}_{\mathbb{T}}(I(h)))\right| \\
&\leq 2\sup_{h\in\mathcal{H}}\sum_{i=1}^{s}\frac{1}{s}|\mathcal{P}_{\mathbb{S}_i}(I(h))-\mathcal{P}_{\mathbb{T}}(I(h))| \\
&\leq 2\sum_{i=1}^{s}\frac{1}{s}\sup_{h\in\mathcal{H}}|\mathcal{P}_{\mathbb{S}_i}(I(h))-\mathcal{P}_{\mathbb{T}}(I(h))| \\
&= \frac{1}{s}\sum_{i=1}^{s}\mathrm{Div}(\mathbb{S}_{\mathrm{i}},\mathbb{T}),
\end{aligned} \tag{5}$$

where the first inequality is due to the triangle inequality and the second inequality is by the additivity of the sup function. This finishes the proof.

Lemma 2. *Given two distributions, $\mathbb{T}$ and $\mathbb{S}$, the difference between the population loss with $\mathbb{T}$ and $\mathbb{S}$ is bounded by the divergence between $\mathbb{T}$ and $\mathbb{S}$ as follows:*

$$|\mathcal{E}_{\mathbb{T}}(f_1,h_1)-\mathcal{E}_{\mathbb{S}}(f_1,h_1)|\leq\frac{1}{2}\mathrm{Div}(\mathbb{T},\mathbb{S}), \tag{6}$$

where $\mathrm{Div}(\mathbb{T},\mathbb{S}) := 2\sup_{h\in\mathcal{H}}|\mathcal{P}_{\mathbb{T}}(I(h))-\mathcal{P}_{\mathbb{S}}(I(h))|$ is the $\mathcal{H}$-divergence for the distribution $\mathbb{T}$ and $\mathbb{S}$ ($I(h)$ is the characteristic function).

Proof. By the definition of $\mathcal{H}$-divergence,

$$\begin{aligned}
\mathrm{Div}(\mathbb{T},\mathbb{S}) &= 2\sup_{h\in\mathcal{H}}|\mathcal{P}_{\mathbb{T}}(I(h))-\mathcal{P}_{\mathbb{S}}(I(h))| \\
&= 2\sup_{f_1,h_1\in\mathcal{H}}\left|\mathcal{P}_{(x,y)\sim\mathbb{T}}[f_1(x)\neq h_1(x)]-\mathcal{P}_{(x,y)\sim\mathbb{S}}[f_1(x)\neq h_1(x)]\right| \\
&= 2\sup_{f_1,h_1\in\mathcal{H}}\left|\mathbb{E}_{(x,y)\sim\mathbb{T}}[\ell(f_1(x),h_1(x))]-\mathbb{E}_{(x,y)\sim\mathbb{S}}[\ell(f_1(x),h_1(x))]\right| \\
&= 2\sup_{f_1,h_1\in\mathcal{H}}|\mathcal{E}_{\mathbb{T}}(f_1,h_1)-\mathcal{E}_{\mathbb{S}}(f_1,h_1)| \\
&\geq 2|\mathcal{E}_{\mathbb{T}}(f_1,h_1)-\mathcal{E}_{\mathbb{S}}(f_1,h_1)|.
\end{aligned} \tag{7}$$

It completes the proof.

Lemma 3. *Let Θ be a cover of a parameter space with VC dimension d. Then, for any $\delta\in(0,1)$ with probability at least $1-\delta$, for any $\theta\in\Theta$:*

$$|\mathcal{E}_{\mathbb{D}}(\theta)-\hat{\mathcal{E}}_D^b(\theta)|\leq\sqrt{\frac{d[\ln(N/d)]+\ln(1/\delta)}{2N}}, \tag{8}$$

where $\hat{\mathcal{E}}_D^b(\theta)$ is a robust empirical risk with N samples in its training set D, and b is the radius around θ.

Proof. For the distribution $\mathbb{D}$, we have

$$\mathcal{P}(|\mathcal{E}_{\mathbb{D}}(\theta) - \hat{\mathcal{E}}_D(\theta)| \geq \epsilon) \leq 2m_{\Theta}(N)\exp(-2N\epsilon^2), \tag{9}$$

where $m_{\Theta}(N)$ is the amount of all possible prediction results for N samples, which implies the model complexity in the parameter space Θ. We set $m_{\Theta}(N) = \frac{1}{2}\left(\frac{N}{d}\right)^d$ in our model, and assume a confidence bound $\epsilon = \sqrt{\frac{d[\ln(N/d)]+\ln(1/\delta)}{2N}}$. Then we get

$$\mathcal{P}(|\mathcal{E}_{\mathbb{D}}(\theta) - \hat{\mathcal{E}}_D(\theta)| \geq \epsilon) \leq \left(\frac{N}{d}\right)^d \exp(-2N\epsilon^2) = \delta. \tag{10}$$

Hence, the inequality $|\mathcal{E}_{\mathbb{D}}(\theta) - \hat{\mathcal{E}}_D(\theta)| \leq \epsilon$ holds with probability at least $1-\delta$. Further, based on the fact that $\hat{\mathcal{E}}^b_D(\theta) \geq \hat{\mathcal{E}}_D(\theta)$, we have

$$|\mathcal{E}_{\mathbb{D}}(\theta) - \hat{\mathcal{E}}^b_D(\theta)| \leq |\mathcal{E}_{\mathbb{D}}(\theta) - \hat{\mathcal{E}}_D(\theta)| \leq \epsilon. \tag{11}$$

It completes the proof.

Proof of Proposition 1 If we continually learn t tasks that follow the distribution $\mathbb{D}_1, \cdots, \mathbb{D}_t$, then a solution $\theta_{1:t}$ can be obtained. In addition, let θ_t denote a solution obtained over the distribution $\mathbb{D}_t$ only, and $\theta_{1:t-1}$ be a solution obtained over the set of distribution $\mathbb{D}_1, \cdots, \mathbb{D}_{t-1}$. Then, we have

$$\begin{aligned}
&\mathcal{E}_{\mathbb{D}_t}(\theta_{1:t-1}) \leq \mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta_{1:t-1}) + \frac{1}{2}\mathrm{Div}(\mathbb{D}_{1:t-1}, \mathbb{D}_t) \\
&\leq \hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta_{1:t-1}) + \frac{1}{2}\mathrm{Div}(\mathbb{D}_{1:t-1}, \mathbb{D}_t) + \sqrt{\frac{d[\ln(N_{1:t-1}/d)] + \ln(1/\delta)}{2N_{1:t-1}}} \\
&\leq \hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta_{1:t-1}) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k, \mathbb{D}_t) + \sqrt{\frac{d[\ln(N_{1:t-1}/d)] + \ln(1/\delta)}{2N_{1:t-1}}} \\
&\leq \hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta_{1:t-1}) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k, \mathbb{D}_t) + \sqrt{\frac{d[\ln(N_{1:t-1}/d)] + \ln(1/\delta)}{N_{1:t-1}}},
\end{aligned} \tag{12}$$

where the first three inequalities are from Lemma 2, Lemma 3 and Lemma 1, respectively. $\mathbb{D}_{1:t-1} := \{\mathbb{D}_k\}_{k=1}^{t-1}$ and we rewrite a mixture of all the $t-1$ distributions as $\mathbb{D}_{1:t-1} := \frac{1}{t-1}\sum_{k=1}^{t-1}\mathbb{D}_k$ using convex combination. $N_{1:t-1} = \sum_{k=1}^{t-1} N_k$ is the total number of training samples over all $t-1$ old tasks.

Further, we have

$$\begin{aligned}
&\mathcal{E}_{\mathbb{D}_t}(\theta_{1:t}) < \mathcal{E}_{\mathbb{D}_t}(\theta_{1:t-1}) \\
&\leq \hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta_{1:t-1}) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k, \mathbb{D}_t) + \sqrt{\frac{d[\ln(N_{1:t-1}/d)] + \ln(1/\delta)}{N_{1:t-1}}} \\
&\leq \hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta_{1:t}) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k, \mathbb{D}_t) + \sqrt{\frac{d[\ln(N_{1:t-1}/d)] + \ln(1/\delta)}{N_{1:t-1}}}.
\end{aligned} \tag{13}$$

Likewise, we get

$$\begin{aligned}
&\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta_t) \leq \mathcal{E}_{\mathbb{D}_t}(\theta_t) + \frac{1}{2}\mathrm{Div}(\mathbb{D}_t, \mathbb{D}_{1:t-1}) \\
&\leq \hat{\mathcal{E}}^b_{D_t}(\theta_t) + \frac{1}{2}\mathrm{Div}(\mathbb{D}_t, \mathbb{D}_{1:t-1}) + \sqrt{\frac{d[\ln(N_t/d)] + \ln(1/\delta)}{2N_t}} \\
&\leq \hat{\mathcal{E}}^b_{D_t}(\theta_t) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t, \mathbb{D}_k) + \sqrt{\frac{d[\ln(N_t/d)] + \ln(1/\delta)}{2N_t}} \\
&\leq \hat{\mathcal{E}}^b_{D_t}(\theta_t) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t, \mathbb{D}_k) + \sqrt{\frac{d[\ln(N_t/d)] + \ln(1/\delta)}{N_t}}.
\end{aligned} \tag{14}$$

Further, we have

$$\begin{aligned}
&\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta_{1:t}) < \mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta_t) \\
&\leq \hat{\mathcal{E}}^b_{D_t}(\theta_t) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t, \mathbb{D}_k) + \sqrt{\frac{d[\ln(N_t/d)] + \ln(1/\delta)}{N_t}} \\
&\leq \hat{\mathcal{E}}^b_{D_t}(\theta_{1:t}) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t, \mathbb{D}_k) + \sqrt{\frac{d[\ln(N_t/d)] + \ln(1/\delta)}{N_t}},
\end{aligned} \tag{15}$$

where N_t is the number of training samples over the distribution $\mathbb{D}_t$.

Combining all the inequalities above finishes the proof.

A.2 Proof of Proposition 2

Proposition 2. *Let $\hat{\theta}^b_{1:t}$ denote the optimal solution of the continually learned $1:t$ tasks by robust empirical risk minimization over the current task, i.e., $\hat{\theta}^b_{1:t} = \arg\min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{D_t}(\theta)$, where Θ denotes a cover of a parameter space with VC dimension d. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$:*

$$\mathcal{E}_{\mathbb{D}_t}(\hat{\theta}^b_{1:t}) - \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_t}(\theta) \leq \min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta) - \min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_{1:t-1}}(\theta) + \frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k, \mathbb{D}_t) + \lambda_1, \tag{16}$$

$$\mathcal{E}_{\mathbb{D}_{1:t-1}}(\hat{\theta}^b_{1:t}) - \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta) \leq \min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{D_t}(\theta) - \min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_t}(\theta) + \frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t, \mathbb{D}_k) + \lambda_2, \tag{17}$$

where $\lambda_1 = 2\sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}}$, $\lambda_2 = 2\sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}}$, and $\mathrm{Div}(\mathbb{D}_i, \mathbb{D}_j) := 2\sup_{h\in\mathcal{H}}|\mathcal{P}_{\mathbb{D}_i}(I(h)) - \mathcal{P}_{\mathbb{D}_j}(I(h))|$ is the $\mathcal{H}$-divergence for the distribution $\mathbb{D}_i$ and $\mathbb{D}_j$ ($I(h)$ is the characteristic function).

Proof of Proposition 2 Let $\hat{\theta}^b_{1:t}$ denote the optimal solution of the continually learned $1:t$ tasks by robust empirical risk minimization over the new task, i.e., $\hat{\theta}^b_{1:t} = \arg\min_{\theta\in\Theta} \hat{\mathcal{E}}^b_{D_t}(\theta)$, where Θ denotes a cover of a parameter space with VC dimension d. Likewise, let $\hat{\theta}^b_t$ be the optimal solution by robust empirical risk minimization over the distribution $\mathbb{D}_t$ only, and $\hat{\theta}^b_{1:t-1}$ over the set of distribution $\mathbb{D}_1, \cdots, \mathbb{D}_{t-1}$. That is, $\hat{\theta}^b_t = \arg\min_\theta \hat{\mathcal{E}}^b_{D_t}(\theta)$ and $\hat{\theta}^b_{1:t-1} = \arg\min_\theta \hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta)$.

Then, let θ_t be the optimal solution over the distribution $\mathbb{D}_t$ only, i.e., $\theta_t = \arg\min_\theta \mathcal{E}_{\mathbb{D}_t}(\theta)$. From Lemma 3, the following inequality holds with probability at least $1-\frac{\delta}{2}$,

$$\begin{aligned} |\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta_t) - \hat{\mathcal{E}}_{D_{1:t-1}}(\theta_t)| &\le \sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{2N_{1:t-1}}} \\ &\le \sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}}, \end{aligned} \tag{18}$$

where $N_{1:t-1} = \sum_{k=1}^{t-1} N_k$ is the total number of training samples over all $t-1$ old tasks. Then, we have

$$\begin{aligned} \min_{\theta\in\Theta} \hat{\mathcal{E}}_{D_{1:t-1}}(\theta) &\le \hat{\mathcal{E}}_{D_{1:t-1}}(\theta_t) \\ &\le \mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta_t) + \sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}} \\ &\le \mathcal{E}_{\mathbb{D}_t}(\theta_t) + \frac{1}{2}\mathrm{Div}(\mathbb{D}_{1:t-1}, \mathbb{D}_t) + \sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}} \\ &= \min_{\theta\in\Theta} \mathcal{E}_{\mathbb{D}_t}(\theta) + \frac{1}{2}\mathrm{Div}(\mathbb{D}_{1:t-1}, \mathbb{D}_t) + \sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}} \\ &\le \min_{\theta\in\Theta} \mathcal{E}_{\mathbb{D}_t}(\theta) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k, \mathbb{D}_t) + \sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}}, \end{aligned} \tag{19}$$

where the third inequality holds from Lemma 2, and the final inequality is from Lemma 1.

From Proposition 1, the following inequality holds with probability at least $1-\frac{\delta}{2}$,

$$\mathcal{E}_{\mathbb{D}_t}(\hat{\theta}^b_{1:t-1}) < \hat{\mathcal{E}}^b_{D_{1:t-1}}(\hat{\theta}^b_{1:t-1}) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k, \mathbb{D}_t) + \sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}}. \tag{20}$$

Combining Eqn. 19 and Eqn. 20, we get

$$\begin{aligned}
&\mathcal{E}_{\mathbb{D}_t}(\hat{\theta}^b_{1:t}) - \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_t}(\theta) \leq \mathcal{E}_{\mathbb{D}_t}(\hat{\theta}^b_{1:t-1}) - \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_t}(\theta) \\
&\leq \hat{\mathcal{E}}^b_{D_{1:t-1}}(\hat{\theta}^b_{1:t-1}) - \min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_{1:t-1}}(\theta) + \frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k,\mathbb{D}_t) + 2\sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}} \\
&= \min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta) - \min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_{1:t-1}}(\theta) + \frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k,\mathbb{D}_t) + 2\sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}}.
\end{aligned} \tag{21}$$

This completes the first part of Proposition 2.

Similarly, let $\theta_{1:t}$ be the optimal solution over the distribution $\mathbb{D}_{1:t-1}$ only, i.e., $\theta_{1:t-1} = \arg\min_\theta \mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta)$. From Lemma 3, the following inequality holds with probability at least $1-\frac{\delta}{2}$,

$$\begin{aligned}
|\mathcal{E}_{\mathbb{D}_t}(\theta_{1:t-1}) - \hat{\mathcal{E}}_{D_t}(\theta_{1:t-1})| &\leq \sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{2N_t}} \\
&\leq \sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}},
\end{aligned} \tag{22}$$

where N_t is the number of training samples in the distribution $\mathbb{D}_t$. Then, we have

$$\begin{aligned}
\min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_t}(\theta) &\leq \hat{\mathcal{E}}_{D_t}(\theta_{1:t-1}) \\
&\leq \mathcal{E}_{\mathbb{D}_t}(\theta_{1:t-1}) + \sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}} \\
&\leq \mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta_{1:t-1}) + \frac{1}{2}\mathrm{Div}(\mathbb{D}_t,\mathbb{D}_{1:t-1}) + \sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}} \\
&= \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta) + \frac{1}{2}\mathrm{Div}(\mathbb{D}_t,\mathbb{D}_{1:t-1}) + \sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}} \\
&\leq \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t,\mathbb{D}_k) + \sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}},
\end{aligned} \tag{23}$$

where the third inequality holds from Lemma 2, and the final inequality is from Lemma 1. From Proposition 1, the following inequality holds with probability at least $1-\frac{\delta}{2}$,

$$\mathcal{E}_{\mathbb{D}_{1:t-1}}(\hat{\theta}^b_t) < \hat{\mathcal{E}}^b_{D_t}(\hat{\theta}^b_t) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t,\mathbb{D}_k) + \sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}}. \tag{24}$$

Combining Eqn. 23 and Eqn. 24, we get

$$\begin{aligned}
&\mathcal{E}_{\mathbb{D}_{1:t-1}}(\hat{\theta}^b_{1:t}) - \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta) \leq \mathcal{E}_{\mathbb{D}_{1:t-1}}(\hat{\theta}^b_t) - \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta) \\
&\leq \hat{\mathcal{E}}^b_{D_t}(\hat{\theta}^b_t) - \min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_t}(\theta) + \frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t,\mathbb{D}_k) + 2\sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}} \\
&= \min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{D_t}(\theta) - \min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_t}(\theta) + \frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t,\mathbb{D}_k) + 2\sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}}.
\end{aligned} \tag{25}$$

This completes the second part of Proposition 2.

A.3 Proof of Proposition 3

Proposition 3. *Let $\{\Theta_i \in \mathbb{R}^r\}_{i=1}^K$ be a set of K parameter spaces ($K > 1$ in general), d_i be a VC dimension of Θ_i, and $\Theta = \cup_{i=1}^K \Theta_i$ with VC dimension d. Based on Proposition 2, for $\hat{\theta}^b_{1:t} = \arg\min_{\bar{\theta}\in\Theta}\hat{\mathcal{E}}^b_{D_t}(\bar{\theta})$, the upper bound of generalization gap is further tighter with*

$$\lambda_1 = \max_{i\in[1,K]}\sqrt{\frac{d_i\ln(N_{1:t-1}/d_i)+\ln(2K/\delta)}{N_{1:t-1}}} + \sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}}, \tag{26}$$

and

$$\lambda_2 = \max_{i\in[1,K]}\sqrt{\frac{d_i\ln(N_t/d_i)+\ln(2K/\delta)}{N_t}} + \sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}}. \tag{27}$$

Below is one critical lemma for the proof of Proposition 3.

Lemma 4. *Let $\{\Theta_i \in \mathbb{R}^r\}_{i=1}^K$ be a set of K parameter spaces ($K > 1$ in general), d_i be a VC dimension of Θ_i, and $\Theta = \cup_{i=1}^K \Theta_i$ with VC dimension d. Let $\theta_i = \arg\max_{\theta\in\Theta_i}\mathcal{E}_{\mathbb{D}}(\theta)$ be a local maximum in the i-th parameter space (i.e., i-th ball). Then, for any $\delta \in (0,1)$ with probability at least $1-\delta$, for any $\theta \in \Theta$:*

$$|\mathcal{E}_{\mathbb{D}}(\theta) - \hat{\mathcal{E}}^b_D(\theta)| \leq \max_{i\in[1,K]}\sqrt{\frac{d_i\ln(N/d_i)+\ln(K/\delta)}{2N}}, \tag{28}$$

where $\hat{\mathcal{E}}^b_D(\theta)$ is a robust empirical risk with N samples in its training set D, and b is the radius around θ.

Proof. For the distribution $\mathbb{D}$, we have

$$\begin{aligned}
\mathcal{P}\left(\max_{i\in[1,K]}|\mathcal{E}_{\mathbb{D}}(\theta_i) - \hat{\mathcal{E}}_D(\theta_i)| \geq \epsilon\right) &\leq \sum_{i=1}^K \mathcal{P}\left(|\mathcal{E}_{\mathbb{D}}(\theta_i) - \hat{\mathcal{E}}_D(\theta_i)| \geq \epsilon\right) \\
&\leq \sum_{i=1}^K 2m_{\Theta_i}(N)\exp(-2N\epsilon^2),
\end{aligned} \tag{29}$$

where $m_{\Theta_i}(N)$ is the amount of all possible prediction results for N samples, which implies the model complexity in the parameter space Θ_i. We set $m_{\Theta_i}(N) = \frac{1}{2}\left(\frac{N}{d_i}\right)^{d_i}$ in our model, and assume a confidence bound $\epsilon_i = \sqrt{\frac{d_i[\ln(N/d_i)]+\ln(K/\delta)}{2N}}$, and $\epsilon = \max_{i\in[1,K]} \epsilon_i$. Then we get

$$\begin{aligned}
\mathcal{P}\left(\max_{i\in[1,K]} |\mathcal{E}_{\mathbb{D}}(\theta_i) - \hat{\mathcal{E}}_D(\theta_i)| \geq \epsilon\right) &\leq \sum_{i=1}^{K} 2m_{\Theta_i}(N)\exp(-2N\epsilon^2) \\
&= \sum_{i=1}^{K}\left(\frac{N}{d_i}\right)^{d_i}\exp(-2N\epsilon^2) \\
&\leq \sum_{i=1}^{K}\left(\frac{N}{d_i}\right)^{d_i}\exp(-2N{\epsilon_i}^2) \\
&= \sum_{i=1}^{K}\frac{\delta}{K} = \delta.
\end{aligned} \tag{30}$$

Hence, the inequality $|\mathcal{E}_{\mathbb{D}}(\theta) - \hat{\mathcal{E}}_D(\theta)| \leq \epsilon$ holds with probability at least $1-\delta$. Further, based on the fact that $\hat{\mathcal{E}}^b_D(\theta) \geq \hat{\mathcal{E}}_D(\theta)$, we have

$$|\mathcal{E}_{\mathbb{D}}(\theta) - \hat{\mathcal{E}}^b_D(\theta)| \leq |\mathcal{E}_{\mathbb{D}}(\theta) - \hat{\mathcal{E}}_D(\theta)| \leq \epsilon. \tag{31}$$

It completes the proof.

Proof of Proposition 3 Let $\{\Theta_i \in \mathbb{R}^r\}_{i=1}^K$ be a set of K parameter spaces ($K > 1$ in general), d_i be a VC dimension of Θ_i, and $\Theta = \cup_{i=1}^K \Theta_i$ with VC dimension d. Let $\hat{\theta}^b_{1:t}$ denote the optimal solution of the continually learned $1:t$ tasks by robust empirical risk minimization over the new task, i.e., $\hat{\theta}^b_{1:t} = \arg\min_{\theta\in\Theta} \hat{\mathcal{E}}^b_{D_t}(\theta)$, where Θ denotes a cover of a parameter space with VC dimension d. Likewise, let $\hat{\theta}^b_t$ be the optimal solution by robust empirical risk minimization over the distribution $\mathbb{D}_t$ only, and $\hat{\theta}^b_{1:t-1}$ over the set of distribution $\mathbb{D}_1, \cdots, \mathbb{D}_{t-1}$. That is, $\hat{\theta}^b_t = \arg\min_\theta \hat{\mathcal{E}}^b_{D_t}(\theta)$ and $\hat{\theta}^b_{1:t-1} = \arg\min_\theta \hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta)$.

Then, let θ_t be the optimal solution over the distribution $\mathbb{D}_t$ only, i.e., $\theta_t = \arg\min_\theta \mathcal{E}_{\mathbb{D}_t}(\theta)$. From Lemma 3 and Proposition 2, the following inequality holds with probability at least $1 - \frac{\delta}{2}$,

$$|\mathcal{E}_{\mathbb{D}_t}(\theta_{1:t-1}) - \hat{\mathcal{E}}_{D_t}(\theta_{1:t-1})| \leq \sqrt{\frac{d\ln(N_t/d) + \ln(2/\delta)}{N_t}}, \tag{32}$$

where $N_{1:t-1} = \sum_{k=1}^{t-1} N_k$ is the total number of training samples over all $t-1$ old tasks. Then, we have

$$\min_{\theta\in\Theta} \hat{\mathcal{E}}_{D_{1:t-1}}(\theta) \leq \min_{\theta\in\Theta} \mathcal{E}_{\mathbb{D}_t}(\theta) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k, \mathbb{D}_t) + \sqrt{\frac{d\ln(N_t/d) + \ln(2/\delta)}{N_t}}. \tag{33}$$

From Proposition 1 and Lemma 4, the following inequality holds with probability at least $1-\frac{\delta}{2}$,

$$\begin{aligned}\mathcal{E}_{\mathbb{D}_t}(\hat{\theta}^b_{1:t-1}) < \hat{\mathcal{E}}^b_{D_{1:t-1}}(\hat{\theta}^b_{1:t-1}) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k,\mathbb{D}_t) \\ + \max_{i\in[1,K]}\sqrt{\frac{d_i\ln(N_{1:t-1}/d_i)+\ln(2K/\delta)}{2N_{1:t-1}}}.\end{aligned} \tag{34}$$

Combining Eqn. 33 and Eqn. 34, we get

$$\begin{aligned}&\mathcal{E}_{\mathbb{D}_t}(\hat{\theta}^b_{1:t}) - \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_t}(\theta) \le \mathcal{E}_{\mathbb{D}_t}(\hat{\theta}^b_{1:t-1}) - \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_t}(\theta) \\ &\le \hat{\mathcal{E}}^b_{D_{1:t-1}}(\hat{\theta}^b_{1:t-1}) - \min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_{1:t-1}}(\theta) + \frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k,\mathbb{D}_t) \\ &+ \max_{i\in[1,K]}\sqrt{\frac{d_i\ln(N_{1:t-1}/d_i)+\ln(2K/\delta)}{2N_{1:t-1}}} + \sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}} \\ &= \min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta) - \min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_{1:t-1}}(\theta) + \frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k,\mathbb{D}_t) \\ &+ \max_{i\in[1,K]}\sqrt{\frac{d_i\ln(N_{1:t-1}/d_i)+\ln(2K/\delta)}{2N_{1:t-1}}} + \sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}}.\end{aligned} \tag{35}$$

This completes the first part of Proposition 3.

Similarly, let $\theta_{1:t}$ be the optimal solution over the distribution $\mathbb{D}_{1:t-1}$ only, i.e., $\theta_{1:t-1} = \arg\min_\theta \mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta)$. From Lemma 3 and Proposition 2, the following inequality holds with probability at least $1-\frac{\delta}{2}$,

$$|\mathcal{E}_{\mathbb{D}_t}(\theta_{1:t-1}) - \hat{\mathcal{E}}_{D_t}(\theta_{1:t-1})| \le \sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}}, \tag{36}$$

where N_t is the number of training samples in the distribution $\mathbb{D}_t$. Then, we have

$$\min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_t}(\theta) \le \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t,\mathbb{D}_k) + \sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}}. \tag{37}$$

From Proposition 1 and Lemma 4, the following inequality holds with probability at least $1-\frac{\delta}{2}$,

$$\mathcal{E}_{\mathbb{D}_{1:t-1}}(\hat{\theta}^b_t) < \hat{\mathcal{E}}^b_{D_t}(\hat{\theta}^b_t) + \frac{1}{2(t-1)}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t,\mathbb{D}_k) + \max_{\mathrm{i}\in[1,\mathrm{K}]}\sqrt{\frac{d_i\ln(N_t/d_i)+\ln(2K/\delta)}{2N_t}}. \tag{38}$$

Combining Eqn. 37 and Eqn. 38, we get

$$\begin{aligned}
&\mathcal{E}_{\mathbb{D}_{1:t-1}}(\hat{\theta}^b_{1:t}) - \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta) \leq \mathcal{E}_{\mathbb{D}_{1:t-1}}(\hat{\theta}^b_t) - \min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta) \\
&\leq \hat{\mathcal{E}}^b_{D_t}(\hat{\theta}^b_t) - \min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_t}(\theta) + \frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t,\mathbb{D}_k) \\
&+ \max_{i\in[1,K]}\sqrt{\frac{d_i\ln(N_t/d_i)+\ln(2K/\delta)}{2N_t}} + \sqrt{\frac{d\ln(N_t/d)+\ln(1/\delta)}{N_t}} \qquad (39)\\
&= \min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{D_t}(\theta) - \min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_t}(\theta) + \frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t,\mathbb{D}_k) \\
&+ \max_{i\in[1,K]}\sqrt{\frac{d_i\ln(N_t/d_i)+\ln(2K/\delta)}{2N_t}} + \sqrt{\frac{d\ln(N_t/d)+\ln(1/\delta)}{N_t}}.
\end{aligned}$$

This completes the second part of Proposition 3.

Discrepancy between task distributions: Below are three important lemmas to prove how cooperating multiple continual learners can optimize the discrepancy between task distributions, which is measured by $\mathcal{H}$-divergence.

Lemma 5. *(Based on Theorem 3.4 of [7] and Lemma 1 of [2]) Let Θ be a cover of a parameter space with VC dimension d. If T and S are samples of size N from two distributions $\mathbb{T}$ and $\mathbb{S}$, respectively, and $\hat{\mathrm{Div}}(T,S)$ is the empirical $\mathcal{H}$-divergence between samples, then for any $\delta\in(0,1)$, with probability at least $1-\delta$,*

$$\mathrm{Div}(\mathbb{T},\mathbb{S}) \leq \hat{\mathrm{Div}}(T,S) + 4\sqrt{\frac{d\ln(2N)+\ln(2/\delta)}{N}}. \qquad (40)$$

Lemma 6. *(Based on Lemma 2 of [2]) Let T and S be samples of size N from two distributions $\mathbb{T}$ and $\mathbb{S}$, respectively. Then the empirical $\mathcal{H}$-divergence between samples, i.e., $\hat{\mathrm{Div}}(T,S)$ can be computed by finding a classifier which attempts to separate one distribution from the other. That is,*

$$\hat{\mathrm{Div}}(T,S) = 2\left(1 - \frac{1}{N}\min_{\theta\in\Theta}\left[\sum_{x:p_\theta(x)=0} I[x\in\mathbb{S}] + \sum_{x:p_\theta(x)=1} I[x\in\mathbb{T}]\right]\right), \qquad (41)$$

where $I[x\in\mathbb{S}]$ is the binary indicator variable which is 1 when the input $x\in\mathbb{S}$, and 0 when $x\in\mathbb{T}$. $p_\theta(\cdot)$ is the learned prediction function.

Of note, Lemma 6 implies we first find a solution in parameter space which has minimum error for the binary problem of distinguishing source from target distributions. By cooperating K parameter spaces, i.e., $\Theta = \cup_{i=1}^K\Theta_i$, we can improve classification errors so as to decrease $\mathcal{H}$-divergence.

Lemma 7. *Let $\{\Theta_i\in\mathbb{R}^r\}_{i=1}^K$ be a set of K parameter spaces ($K>1$ in general), d_i be a VC dimension of Θ_i, and $\Theta=\cup_{i=1}^K\Theta_i$ with VC dimension d. If T and S*

are samples of size N from two distributions $\mathbb{T}$ and $\mathbb{S}$, respectively, and $\hat{\mathrm{Div}}(T,S)$ is the empirical $\mathcal{H}$-divergence between samples, then in the parameter space Θ, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathrm{Div}(\mathbb{T},\mathbb{S}) \leq \hat{\mathrm{Div}}(T,S) + \max_{i\in[1,K]} 4\sqrt{\frac{d_i \ln(2N) + \ln(2K/\delta)}{2N}}. \tag{42}$$

Proof. For two distributions $\mathbb{T}$ and $\mathbb{S}$, we have

$$\begin{aligned} &\mathcal{P}\left(\max_{i\in[1,K]} |\mathrm{Div}_{\Theta_i}(\mathbb{T},\mathbb{S}) - \hat{\mathrm{Div}}_{\Theta_i}(T,S)| \geq \epsilon\right) \\ &\leq \sum_{i=1}^{K} \mathcal{P}\left(|\mathrm{Div}_{\Theta_i}(\mathbb{T},\mathbb{S}) - \hat{\mathrm{Div}}_{\Theta_i}(T,S)| \geq \epsilon\right) \leq \sum_{i=1}^{K} 2m_{\Theta_i}(N)\exp(-2N\epsilon^2), \end{aligned} \tag{43}$$

where $m_{\Theta_i}(N)$ is the amount of all possible predictions for N samples, which implies the model complexity in the parameter space Θ_i. We set $m_{\Theta_i}(N) = 16\,(2N)^{d_i}$ in our model, and assume a confidence bound $\epsilon_i = 4\sqrt{\frac{d_i \ln(2N)+\ln(2K/\delta)}{2N}}$, and $\epsilon = \max_{i\in[1,K]} \epsilon_i$. Then we get

$$\begin{aligned} \mathcal{P}\left(\max_{i\in[1,K]} |\mathrm{Div}_{\Theta_i}(\mathbb{T},\mathbb{S}) - \hat{\mathrm{Div}}_{\Theta_i}(T,S)| \geq \epsilon\right) &\leq \sum_{i=1}^{K} 2m_{\Theta_i}(N)\exp(-2N\epsilon^2) \\ &= \sum_{i=1}^{K} 32\,(2N)^{d_i}\exp(-2N\epsilon^2) \\ &\leq \sum_{i=1}^{K} 32\,(2N)^{d_i}\exp(-2N{\epsilon_i}^2) \\ &= \sum_{i=1}^{K} \frac{\delta}{K} = \delta. \end{aligned} \tag{44}$$

Hence, the inequality $|\mathrm{Div}_{\Theta_i}(\mathbb{T},\mathbb{S}) - \hat{\mathrm{Div}}_{\Theta_i}(T,S)| \leq \epsilon$ holds with probability at least $1-\delta$. It completes the proof.

Comparing Lemma 5 and Lemma 7, it can be found that by cooperating K parameter spaces, our proposal can mitigate the discrepancy between tasks, i.e., $\mathrm{Div}(\mathbb{T},\mathbb{S})$, by decreasing the empirical $\mathcal{H}$-divergence (i.e., $\hat{\mathrm{Div}}_{\Theta_i}(T,S)$) and another factor.

B Experiment Details

B.1 Implementation

We follow the implementation of [6, 3, 15] for supervised continual learning. For CIFAR-100-SC and CIFAR-100-RS, we use an Adam optimizer of initial learning

rate 0.001 and train all methods with batch size of 256 for 100 epochs. For CUB-200-2011 and Tiny-ImageNet, we use a SGD optimizer of initial learning rate 0.005 and momentum 0.9, and train all methods with batch size of 64 for 40 epochs.

We follow the implementation of [11] for unsupervised continual learning on CIFAR-100-RS (which is called Split CIFAR-100 in [11]). We use a SGD optimizer of initial learning rate 0.03, momentum 0.9 and weight decay 5e-4, and train all methods with batch size of 256 for 200 epochs.

B.2 Hyperparameter

For CIFAR-100-SC, CIFAR-100-RS and CUB-200-2011, we adopt the same hyperparameters for the baselines used in [15]. While for other experiments (e.g., Tiny-ImageNet) and baselines (e.g., CPR [3]), we make an extensive hyperparameter search to make the comparison as fair as possible. The hyerparameters for supervised continual learning are summarized in Table 1.

Table 1. Hyperparamters for supervised continual learning. $^*\lambda$ is the same as the corresponding baseline approach.

Methods	CIFAR-100-SC	CIFAR-100-RS	CUB-200-2011	Tiny-ImageNet
AGS-CL [6]	λ(3200), μ(10), ρ(0.3)	λ(1600), μ(10), ρ(0.3)	–	–
HAT [14]	c(500), smax(200)	c(500), smax(200)	–	–
EWC [8]	λ(40000)	λ(10000)	λ(1)	λ(80)
MAS [1]	λ(16)	λ(4)	λ(0.01)	λ(0.1)
SI [16]	λ(8)	λ(10)	λ(6)	λ(0.8)
RWALK [4]	λ(128)	λ(6)	λ(48)	λ(5)
P&C [13]	λ(40000)	λ(20000)	λ(1)	λ(80)
*AFEC [15]	λ_e(1)	λ_e(1)	λ_e(0.001)	λ_e(0.1)
*CPR [3]	β(1.5)	β(1.5)	β(1)	β(0.6)
*CoSCL (Ours)	γ(0.02), s(100)	γ(0.02), s(100)	γ(0.0001), s(100)	γ(0.001), s(100)

B.3 Architecture

The network architectures used for the main experiments are detailed in Table 2, 3 (the output head is not included).

B.4 Evaluation Metric

We use three metrics to evaluate the performance of continual learning, including averaged accuracy (AAC), forward transfer (FWT) and backward transfer (BWT) [10]:

$$\mathrm{AAC} = \frac{1}{\mathrm{T}} \sum_{\mathrm{i}=1}^{\mathrm{T}} \mathrm{A}_{\mathrm{T},\mathrm{i}}, \tag{45}$$

$$\mathrm{FWT} = \frac{1}{\mathrm{T}-1} \sum_{\mathrm{i}=2}^{\mathrm{T}} \mathrm{A}_{\mathrm{i}-1,\mathrm{i}} - \hat{\mathrm{A}}_{\mathrm{i}}, \tag{46}$$

Table 2. Network architecture for CIFAR-100-SC and CIFAR-100-RS. We set $nc = 32$ for a single continual learner (#Param=837K) while $nc = 8$ for 5 small continual learners in CoSCL (#Param=773K).

Layer	Channel	Kernel	Stride	Padding	Dropout
Input	3				
Conv 1	nc	3×3	1	1	
Conv 2	nc	3×3	1	1	
MaxPool			2	0	0.25
Conv 3	$2nc$	3×3	1	1	
Conv 4	$2nc$	3×3	1	1	
MaxPool			2	0	0.25
Conv 5	$4nc$	3×3	1	1	
Conv 6	$4nc$	3×3	1	1	
MaxPool			2	1	0.25
Dense 1	256				

Table 3. Network architecture for CUB-200-2011 and Tiny-ImageNet. We set $nc = 64$ for a single continual learner (#Param=57.8M) while $nc = 34$ for 5 small continual learners in CoSCL (#Param=57.2M).

Layer	Channel	Kernel	Stride	Padding	Dropout
Input	3				
Conv 1	nc	11×11	4	2	
MaxPool		3×3	2	0	0
Conv 2	$3nc$	5×5	1	2	
MaxPool		3×3	2	0	0
Conv 3	$6nc$	3×3	1	1	
Conv 4	$4nc$	3×3	1	1	
Conv 5	$4nc$	3×3	1	1	
MaxPool		3×3	2	0	0
Dense 1	$64nc$				0.5
Dense 2	$64nc$				0.5

$$\mathrm{BWT} = \frac{1}{\mathrm{T}-1}\sum_{\mathrm{i}=1}^{\mathrm{T}-1} \mathrm{A}_{\mathrm{T,i}} - \mathrm{A}_{\mathrm{i,i}}, \tag{47}$$

where $\mathrm{A}_{\mathrm{t,i}}$ is the test accuracy of task i after incrementally learning task t, and $\hat{\mathrm{A}}_{\mathrm{i}}$ is the test accuracy of each task i learned from random initialization. Averaged accuracy (ACC) is the averaged performance of all the tasks ever seen. Forward transfer (FWT) evaluates the averaged influence of remembering the old tasks to each new task. Backward transfer (BWT) evaluates the averaged influence of learning each new task to the old tasks.

C Additional Results

C.1 Diversity of Expertise across Tasks

To evaluate the diversity of expertise across tasks, we use the feature representations of each continual learner to make predictions with the shared output head, and calculate the relative accuracy. As shown in Fig. 1, the solution learned by each continual learner varies significantly across tasks and complement with each other.

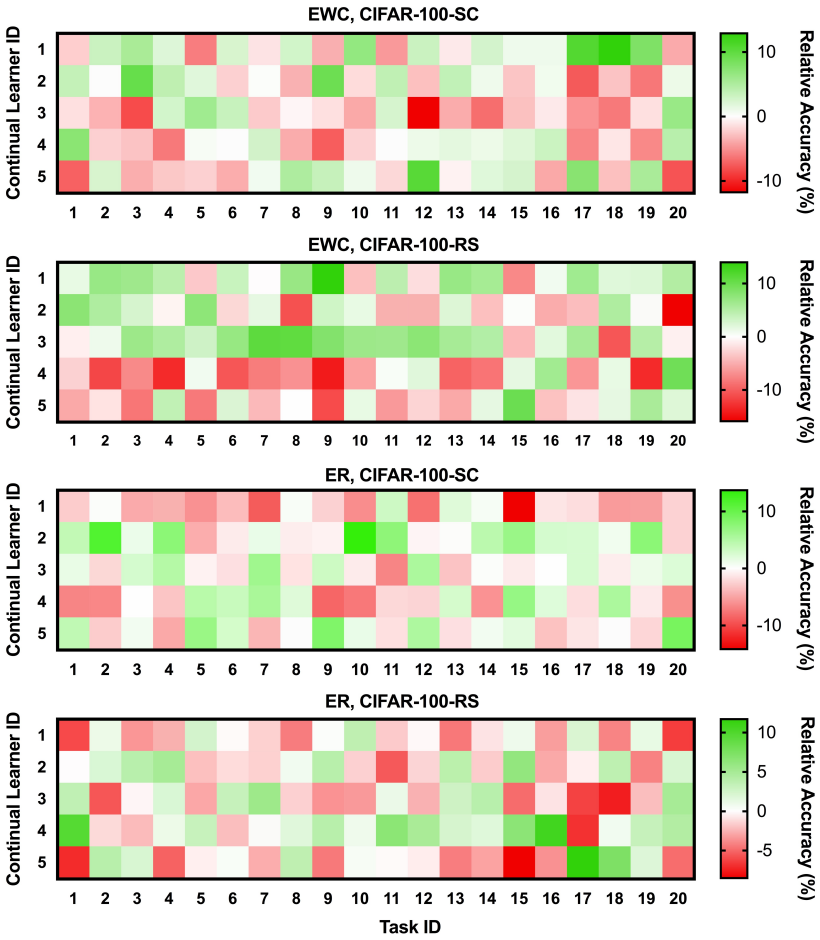

Fig. 1. Diversity of expertise across tasks. Here we use EWC [8] or Experience Replay (ER) [12] as the default continual learning method. The relative accuracy for each task is calculated by subtracting the performance of each learner from the averaged performance of all learners.

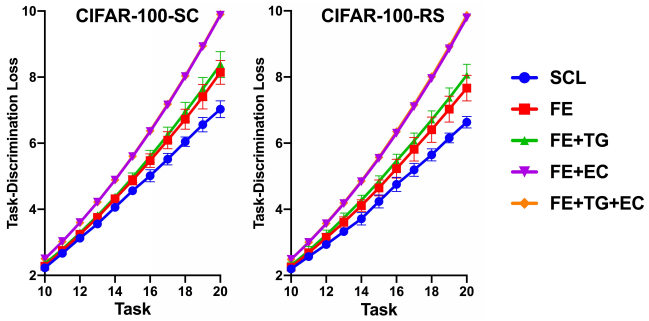

Fig. 2. Task-discrimination loss in feature space. We plot all baselines from the tenth task, where significant differences start to arise. Larger loss indicates a smaller $\mathcal{H}$-divergence. SCL: single continual learner; FE: feature ensemble; TG: task-adaptive gates; EC: ensemble cooperation loss.

C.2 Discrepancy between Task Distributions

To empirically approximate the $\mathcal{H}$-divergence between tasks in feature space, we train a discriminator with a fully-connected layer to distinguish whether the features of input images belong to a task or not [9]. Specifically, the discriminator is trained with the features of training data and the binary cross-entropy loss. We use Adam optimizer and initial learning rate 0.0001 with batch size of 256 for 10 epochs. Then we evaluate the $\mathcal{H}$-divergence between tasks with the features of test data, where a larger discrimination loss indicates a smaller $\mathcal{H}$-divergence. Since the discrimination becomes increasingly harder as more tasks are introduced, from the tenth task we start to observe significant differences between all the baselines. The proposed feature ensemble (FE) and ensemble cooperation (EC) can largely decrease the discrepancy between tasks, while the task-adaptive gates (TG) have a moderate effect.

C.3 Results of ResNet

In addition to regular CNN architectures, our method is also applicable to other architectures such as ResNet. We use a WideResNet-28-2 architecture to perform the task incremental learning experiments on CIFAR-100-RS, following a widely-used implementation code [5]. CoSCL (5 learners with accordingly-adjusted width) can improve the performance from 69.52% to 73.26% for EWC and from 62.23% to 68.69% for MAS.

References

1. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. In: Proceedings of the European Conference on Computer Vision. pp. 139–154 (2018)
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**(1), 151–175 (2010)
3. Cha, S., Hsu, H., Hwang, T., Calmon, F., Moon, T.: Cpr: Classifier-projection regularization for continual learning. In: Proceedings of the International Conference on Learning Representations. (2020)
4. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: Proceedings of the European Conference on Computer Vision. pp. 532–547 (2018)
5. Hsu, Y.C., Liu, Y.C., Ramasamy, A., Kira, Z.: Re-evaluating continual learning scenarios: A categorization and case for strong baselines. In: NeurIPS Continual learning Workshop (2018), https://arxiv.org/abs/1810.12488
6. Jung, S., Ahn, H., Cha, S., Moon, T.: Continual learning with node-importance based adaptive group sparse regularization. arXiv e-prints pp. arXiv–2003 (2020)
7. Kifer, D., Ben-David, S., Gehrke, J.: Detecting change in data streams. In: VLDB. vol. 4, pp. 180–191. Toronto, Canada (2004)
8. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017)
9. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: Proceedings of the International Conference on Machine Learning. pp. 97–105. PMLR (2015)
10. Lopez-Paz, D., et al.: Gradient episodic memory for continual learning. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 6467–6476 (2017)
11. Madaan, D., Yoon, J., Li, Y., Liu, Y., Hwang, S.J.: Rethinking the representational continuity: Towards unsupervised continual learning. arXiv preprint arXiv:2110.06976 (2021)
12. Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., Tesauro, G.: Learning to learn without forgetting by maximizing transfer and minimizing interference. arXiv preprint arXiv:1810.11910 (2018)
13. Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y.W., Pascanu, R., Hadsell, R.: Progress & compress: A scalable framework for continual learning. In: Proceedings of the International Conference on Machine Learning.. pp. 4528–4537. PMLR (2018)
14. Serra, J., Suris, D., Miron, M., Karatzoglou, A.: Overcoming catastrophic forgetting with hard attention to the task. In: Proceedings of the International Conference on Machine Learning.. pp. 4548–4557. PMLR (2018)
15. Wang, L., Zhang, M., Jia, Z., Li, Q., Bao, C., Ma, K., Zhu, J., Zhong, Y.: Afec: Active forgetting of negative transfer in continual learning. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 34 (2021)
16. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: Proceedings of the International Conference on Machine Learning.. pp. 3987–3995. PMLR (2017)