SUPPLEMENTARY FILES

We will begin by restating some preliminaries, these are the exact copy of the initial text in Section 3 of the paper.

A PRELIMINARIES

Let x and y be random variables corresponding to input and output probability spaces with support $\mathcal{X}$ and $\mathcal{Y}$ and $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{Y})$ representing the corresponding Borel algebras. Define t as a random variable denoting the joint space of $x \times y$ with a model $f_{(w,h)} : \mathcal{X} \rightarrow \mathcal{Y}$ being specified using weights w and hyperparameters h. Given compact sets $\mathcal{W}$ over w and $\mathcal{H}$ over h, the goal is to learn the weights by searching over the hypothesis space $f = \{f_{(w,h)}, \forall h \in \mathcal{H}, w \in \mathcal{W}\}$ through a loss function $\ell_{w,h}(t)$. In this paper, we will assume that the hyperparameter/architecture is fixed and therefore, will drop the notation h and denote loss simply as $\ell_w(t)$. Throughout the paper, we will assume $x_k = x(k)$ and use them interchangeably, and $\mathbf{k} = [1, 2, \cdots, k]$. In this context, we characterize the effective model capacity as follows.

Effective Model Capacity: We will assume that $\ell_w(t)$ is continuous and twice differentiable over the support $\mathcal{X} \times \mathcal{Y}$ or $\mathcal{X}$, and the compact set $\mathcal{W}$. Under these assumptions, let $\ell_{min} = \mathcal{O}_{\mathcal{W}}(T) = min_{w \in \mathcal{W}} \quad E_{t \in T}[\ell_w(t)]$ be the optimization procedure with T being a dataset of samples t with $T \subset \mathcal{B}(\mathcal{T})$. Then, given the best hyperparameter/architecture configurations, the optimization procedure $\mathcal{O}_{\mathcal{W}}$ seeks to find the weights $w^* \in \mathcal{W}$ that minimizes the loss over a dataset. Given this setting, we define the effective model capacity (the upper/lower bounds derived in the appendix) as the smallest achievable loss value using $\mathcal{O}_{\mathcal{W}}$ that remains unchanged even when additional data or training is used.

Definition 1 [Effective Model Capacity (EMC)] Given $\mathcal{W}$ as the weight space and $T \in \mathcal{B}(\mathcal{T})$ with an optimization procedure $\mathcal{O}_{\mathcal{W}}(T)$, the EMC of the model f is given as

$$\epsilon = \underset{T \in \mathcal{B}(\mathcal{T})}{min} \quad \big[\mathcal{O}_{\mathcal{W}}(T)\big] = \underset{T \in \mathcal{B}(\mathcal{T})}{min} \quad \big[\underset{w \in \mathcal{W}}{min} \quad \underset{t \in T}{E}[\ell_w(t)]\big]$$

Given a feasible weight set $\mathcal{W}_k$, and loss function $\ell_{w_k}(t), t \in \mathcal{T}_k$, the model at k is denoted by f_{w_k}, the goal of CL is to maintain memory of all observed tasks, then, the CL forgetting cost for the interval $\mathbf{k} = [1, k]$ is given as

$$\underset{w_k \in \mathcal{W}_k}{min} J_{w_k}(\mathbf{T}_k) = \underset{w_k \in \mathcal{W}_k}{min} \sum_{i=1}^{k} \gamma_i \left[\underset{t \in T(i)}{E}\left[\ell_{w_k}(t)\right]\right] \;, \quad \forall T(i) \in \mathbf{T}_k,$$

where, γ ensures boundedness of $J_{w_k}(\mathbf{T}_k)$ (see [45], Lemma 1). For a fixed $h \in \mathcal{H}$, the complete CL problem is

$$V^{(*)}(u_k) = \underset{u_k}{min} \quad \sum\nolimits_{i=k}^{K} \left[J_{w_i}(\mathbf{T}_i)\right], u_k = \{w_i, i = k, k+1, \cdots K\}$$

CL Effective Model Capacity and Balance Point: For ease of exposition, we begin by stating

Definition 2 [Forgetting Effective Model Capacity (FEMC)] For task $k \in [1, K]$, dataset $\mathbf{T}_k$, weight space $\mathcal{W}_k$, optimization procedure $\mathcal{O}_{\mathcal{W}_k}(\mathbf{T}_k)$, EMC at k, $\epsilon_k = min_{\mathbf{T}_k, w_k} J_{w_k}(\mathbf{T}_k)$, we define FEMC at task k as:

$$\text{FEMC}(k) = \max_{\mathbf{k}} \epsilon_{\mathbf{k}} = \max\{\epsilon_1, \epsilon_2, \cdots, \epsilon_k\}$$

$FEMC(k)$ at each k is defined by the highest forgetting loss in the interval $[1, k]$. We now define CL effective model capacity as follows.

Definition 3 [Effective Model Capacity for CL (CLEMC)] For a task $k \in [1, K]$, we define CLEMC as the sum of FEMC across all possible tasks as

$$\epsilon_k^{(*)} = \sum_{i=k}^{K} \text{FEMC}(i) = \sum_{i=k}^{K} \max_{\mathbf{i}} \; \epsilon_{\mathbf{i}}$$

We will derive the notion of first difference in capacity as a function of the forgetting cost

B First Difference

Lemma 1. For $k \in [1, K]$, let $u_k = \{w_i, i = k, k+1, \cdots K\}$ be weight sequences from k with $\mathcal{U}(k) = \{\mathcal{W}_i, i = k, k+1, \cdots\}$– the compact sets. Next define (J_F), (CL) and (CLEMC) to write

$$\epsilon^{(*)}_{k+1} - \epsilon^{(*)}_{k} = min_{\mathbf{k}} \{ \max_{\mathbf{T}_i} \{ \langle \partial_{w_k} V^{(*)}(u_k), dw_k \rangle + \sum_{T \in \mathbf{T}_k} \langle \partial_T V^{(*)}(u_k), dT \rangle \} \}$$

Proof. We first derive the current forgetting cost as a function of infinitesimal change in $V^{(*)}(u_k)$ in the following technical lemma.

Lemma. *Consider $k \in [0, K]$ with the forgetting cost as in (J_F) and CL problem in* (CL)*. Then,*

$$-\min_{w_k} J_{w_k}(\mathbf{T}_k) = \Big\langle \partial_{w_k} V^{(*)}(u_k), dw_k \Big\rangle + \sum_{T \in \mathbf{T}_k} \Big\langle \partial_T V^{(*)}(u_k), dT \Big\rangle + \mathcal{O}(2) \tag{1}$$

where d is the first difference operator, ∂ refers to the first derivative and $\mathcal{O}(2)$ represent the higher order derivative terms.

Proof. Let $u_k = \{w_i, i = k, k+1, \cdots K\}$ be the sequence of weights starting from k with $\mathcal{U}(k) = \{\mathcal{W}_i, i = k, k+1, \cdots\}$ being the sequence of their respective compact sets. Under the assumption that the optimal cost $V^{(*)}(u_k)$ is given by the optimal trajectory of weights u_k corresponding to the tasks sets $\{T_i \in \mathcal{T}_i, i = k, k+1, \cdots K\}$, we can write the following system of recursive equations

$$V^{(*)}(u_k) = \min_{u_k \in \mathcal{U}_k} \quad \sum\nolimits_{i=k}^{K} [J_{w_i}(\mathbf{T}_i)] \tag{2a}$$

$$V^{(*)}(u_{k+1}) = \min_{u_{k+1} \in \mathcal{U}_{k+1}} \sum\nolimits_{i=k+1}^{K} [J_{w_i}(\mathbf{T}_i)] \tag{2b}$$

$$V^{(*)}(u_k) = \min_{w_k} J_{w_k}(\mathbf{T}_k) + V^{(*)}(u_{k+1}) \tag{2c}$$

where (2a) and (2b) follow directly from using (J_F) and (2c) is obtained by simply rewriting (2a) using (2b).

Now, given two trajectories u_k and u_{k+1}, the change introduced by u_{k+1} to $V^{(*)}(u_k)$ is given by Taylor series approximation of $V^{(*)}(u_k)$ around w_k and $\mathbf{T}_k$ as,

$$V^{(*)}(u_{k+1}) = V^{(*)}(u_k) + \Big\langle \partial_{w_k} V^{(*)}(u_k), dw_k \Big\rangle + \sum_{T \in \mathbf{T}_k} \Big\langle \partial_T V^{(*)}(u_k), dT \Big\rangle + \mathcal{O}(2) \tag{3}$$

where dT_k and dw_k are the infinitesimal perturbations to data and weights respectively and $\mathcal{O}(2)$ represent higher order derivative terms. Substituting (3) into (2c) to get

$$\cancel{V^{(*)}(u_k)} = \min_{w_k} J_{w_k}(\mathbf{T}_k) + \cancel{V^{(*)}(u_k)} + \Big\langle \partial_{w_k} V^{(*)}(u_k), dw_k \Big\rangle + \sum_{T \in \mathbf{T}_k} \Big\langle \partial_T V^{(*)}(u_k), dT \Big\rangle + \mathcal{O}(2)$$

which proves the result stated in the technical Lemma. □

Using the above result, we can now prove Lemma (1). Towards this end, we begin by writing,

$$\epsilon^{(*)}_{k} = \sum_{i=k}^{K} max_{\mathbf{i}}\, \epsilon_{\mathbf{i}} = max_{\mathbf{k}}\, \epsilon_{\mathbf{k}} + \epsilon^{(*)}_{k+1} \tag{4a}$$

$$\epsilon^{(*)}_{k+1} - \epsilon^{(*)}_{k} = min_{\mathbf{k}} \{ -\epsilon_{\mathbf{k}} \} = min_{\mathbf{k}} \{ -\{ \min_{\mathbf{T}_i} \min_{w_i} J_{w_i}(\mathbf{T}_i) \},\ i \in \mathbf{k} \} \tag{4b}$$

$$= min_{\mathbf{k}} \{ \max_{\mathbf{T}_i} (-\min_{w_i} J_{w_i}(\mathbf{T}_i)),\ i \in \mathbf{k} \} \tag{4c}$$

where (4a) is obtained by applying (CLEMC), and (4b) is obtained by rewriting $\epsilon_{\mathbf{k}}$ using (J_F). Substituting (1) into (4c), and ignoring the higher order derivative terms denoted by $\mathcal{O}(2)$ [3], we obtain the result as

$$\epsilon_{k+1}^{(*)} - \epsilon_k^{(*)} = min_{k\in\mathbf{k}} \{\underset{\mathbf{T}_i}{max} (-\underset{w_i}{min}\, J_{w_i}(\mathbf{T}_i)), i \in \mathbf{k}\} \tag{5a}$$

$$= min_{k\in\mathbf{k}} \{\underset{\mathbf{T}_i}{max} \{\Big\langle \partial_{w_k} V^{(*)}(u_k), dw_k \Big\rangle + \sum_{T\in\mathbf{T}_k} \Big\langle \partial_T V^{(*)}(u_k), dT \Big\rangle\} \tag{5b}$$

□

Next, we will derive the lower bound on the first difference in capacity which stems from a lower and upperbound on capacity.

C LOWER BOUND ON FIRST DIFFERENCE

Theorem 1. The first difference in CLEMC (FD) is lower bounded as

$$\epsilon_k^{(*)} - \epsilon_{k+1}^{(*)} \geq \underset{k\in\mathbf{k}}{max} \{\underset{\mathbf{T}_i}{min} \{\|\partial_{w_k} J_{w_k^*}(\mathbf{T}_i)\|\|dw_k^*\| + \sum_{T(k)\in\mathbf{T}_i} \sum_{i=k}^{K} \|\partial_{T(k)} E_{t\in T(i)} \ell_{w_i^*}(t)\|\|dT(k)\|\}\},$$

Proof. From Lemma (1) we get

$$\epsilon_{k+1}^{(*)} - \epsilon_k^{(*)} = min_{k\in\mathbf{k}} \{\underset{\mathbf{T}_k}{max} \{\Big\langle \partial_{w_k} V^{(*)}(u_k), dw_k \Big\rangle + \sum_{T\in\mathbf{T}_k} \Big\langle \partial_T V^{(*)}(u_k), dT \Big\rangle\}\}$$

$$\leq min_{k\in\mathbf{k}} \{\underset{\mathbf{T}_k}{max} \{\|\partial_{w_k} V^{(*)}(u_k)\|\|dw_k\| + \sum_{T\in\mathbf{T}_k} \|\partial_T V^{(*)}(u_k)\|\|dT\|\}\} \tag{6a}$$

where (6a) is obtained using Cauchy-Schwarz inequality, $\langle a, b\rangle \leq \|a\|\|b\|$. We then bound both the gradient norm terms in (6a) as follows.

For the first gradient norm term, we assume that the optimal cost, $V^{(*)}$, is given by the weight trajectory u_k with $u_k = \{w_i, i = k, k+1, \cdots K\}$. We can then bound it through the following inequalities.

$$\|\partial_{w_k} V^{(*)}(u_k)\| = \|\partial_{w_k} \min_{u_k} \sum_{i=k}^{K} J_{w_i}(\mathbf{T}_i)\| \tag{7a}$$

$$\leq \|\partial_{w_k} \sum_{i=k}^{K} min_{w_i} J_{w_i}(\mathbf{T}_i)\| \tag{7b}$$

$$\leq \|\sum_{i=k}^{K} \partial_{w_k} min_{w_i} J_{w_i}(\mathbf{T}_i)\| \tag{7c}$$

$$\leq \|\partial_{w_k} min_{w_k} J_{w_k}(\mathbf{T}_i)\| \tag{7d}$$

where (7b) is because the norm of the gradient, with respect to weights, at the optimal cost (due to an optimal trajectory) is always less than the norm of the gradient, with respect to the weights, at a forgetting cost corresponding to any arbitrary weight trajectory. (7c) follows from the sum rule of derivatives and (7d) is because all terms from w_{k+1} onwards vanish due to lack of dependence on w_k.

For the second norm of the gradient term in (6a), we again write the optimal cost $V^{(*)}(u_k) = \sum_{i=k}^{K} min_{w_i} J_{w_i}(\mathbf{T}_i)$ such that $\mathbf{T}_i = \{T(1), \cdots T(i)\}$. We further observe that if the optimal cost is differentiated with respect to $T(k)$ only the k^{th} term in the inner sum will remain. We can then bound it through the following inequalities.

$$\|\partial_{T(k)} V^{(*)}(u_k)\| \leq \|\partial_{T(k)} \sum_{i=k}^{K} min_{w_i} J_{w_i}(\mathbf{T}_i)\| \tag{8a}$$

$$\leq \|\sum_{i=k}^{K} \partial_{T(k)} min_{w_i} \sum_{p=1}^{i} E_{t\in T(p)} \ell_{w_i}(t)\| \tag{8b}$$

$$\leq \|\sum_{i=k}^{K} \partial_{T(k)} E_{t\in T(i)} \ell_{w^*(i)}(t)\| \tag{8c}$$

Then, upon substituting (7d) and (8c) into (6a) we get,

$$\epsilon_{k+1}^{(*)} - \epsilon_k^{(*)} \leq \min_{k\in\mathbf{k}} \{\max_{\mathbf{T}_i} \{\|\partial_{w_k} J_{w_k^*}(\mathbf{T}_i)\|\|dw_k^*\| + \sum_{T(k)\in\mathbf{T}_i} \sum_{i=k}^{K} \|\partial_{T(k)} E_{t\in T(i)} \ell_{w_i^*}(t)\|\|dT(k)\|\}\}, \tag{9}$$

where we have replaced the inner minimization problem with respect to weights by the corresponding w^*. Multiplication with -1 provides the lower bound as

$$\epsilon_k^{(*)} - \epsilon_{k+1}^{(*)} \geq \max_{k\in\mathbf{k}} \{\min_{\mathbf{T}_i} \{\|\partial_{w_k} J_{w_k^*}(\mathbf{T}_i)\|\|dw_k^*\| + \sum_{T(k)\in\mathbf{T}_i} \sum_{i=k}^{K} \|\partial_{T(k)} E_{t\in T(i)} \ell_{w_i^*}(t)\|\|dT(k)\|\}\}, \tag{10}$$

□

This lower bound then leads to the conclusion that capacity is non-stationary and diverges with increase in weight update or divergence between subsequent tasks. This non-stationarity extends to experience replay and experience replay with regularization

D DIVERGENCE WITH RESPECT TO WEIGHTS

Theorem 2. Fix $k \in \mathbb{N}$ and I, the number of weight updates required to obtain the optimal value. Assume that $\|\partial_{w_k} J_{w_k^*}(\mathbf{T}_i)\| \geq \Phi_w$, $\|\partial_{T(k)} E_{t\in T(i)} \ell_{w_i^*}(t)\| \geq \Phi_T$, and let the smallest value of $min_{T(k)}\|dT(k)\| \geq \Phi_{dT}$. Let $L, \mathcal{R}$ be the Lipschitz constants for the cost function and the regularization function respectively with α_{MIN} being the smallest learning rate. Then, $\sum_k^K d\epsilon_k^{(*)}$ diverges as a function of K, and I with and without the regularization factor.

Proof. We first prove the technical Lemma below.

Lemma. *Fix $k \in \mathbb{N}$ and let the weights at any task k be updated for a total of I steps. Assume $T(k)$ is provided through a series of batches such that $T(k) = \{t_k^{(i)}, i = 1, \cdots, I\}$ with $t_k^{(i)}$ be a tensor corresponding to batch of data at the i^{th} step for the k^{th} task, sampled uniformly from the underlying support. For the i^{th} update step of the k^{th} task, let the forgetting cost be denoted by $J_{w_k}(\mathbf{T}_k)$, gradient be denoted by $g_k^{(i)}$, and learning rate by $\alpha_k^{(i)}$. Then,*

$$dw_k^* = -\sum_{i=0}^{I-1} \alpha_k^{(i)} g_k^{(i)} \tag{11}$$

Proof. Note now that, we abuse notation to define $dw_k^* = w_k^* - w_k^{(0)} = w_k^{(I)} - w_k^{(0)}$ assuming that the optimal point is achieved after I updates (indicated by parenthesis). Then, at any particular update step, we obtain

$$w_k^{(i+1)} = w_k^{(i)} - \alpha_k^{(i)} g_k^{(i)} \tag{12}$$

where $g_k^{(i)}$ is the update gradient at the this step.

$$w_k^{(i+1)} = w_k^{(i)} - \alpha_k^{(i)} g_k^{(i)} \tag{13}$$

We may now write the sum over the I steps at a

$$w_k^{(1)} = w_k^{(0)} - \alpha_k^{(0)} g_k^{(0)} \tag{14a}$$
$$w_k^{(2)} = w_k^{(1)} - \alpha_k^{(1)} g_k^{(1)} \tag{14b}$$
$$\vdots \tag{14c}$$
$$w_k^{(I)} = w_k^{(I-1)} - \alpha_k^{(I-1)} g_k^{(I-1)} \tag{14d}$$

Adding all these terms to write

$$dw_k^* = -\sum_{i=0}^{I-1} \alpha_k^{(i)} g_k^{(i)} \tag{15}$$

□

Given the first difference in capacity from the technical Lemma above, and under the assumption that $\|\partial_{w_k} J_{w_k^*}(\mathbf{T}_i)\| \geq \Phi_w$ and $\|\partial_{T(k)} E_{t \in T(i)} \ell_{w_i^*}(t)\| \geq \Phi_T$

$$\epsilon_k^{(*)} - \epsilon_{k+1}^{(*)} \geq \underset{k \in \mathbf{k}}{max} \{ \underset{\mathbf{T}_i}{min} \{ \|\partial_{w_k} J_{w_k^*}(\mathbf{T}_i)\| \|dw_k^*\| + \sum_{T(k) \in \mathbf{T}_i} \sum_{i=k}^{K} \|\partial_{T(k)} E_{t \in T(i)} \ell_{w_i^*}(t)\| \|dT(k)\| \} \}$$

$$\geq \underset{k \in \mathbf{k}}{max} \{ \underset{\mathbf{T}_i}{min} \{ \Phi_w \|dw_k^*\| + \sum_{T(k) \in \mathbf{T}_i} \sum_{i=k}^{K} \Phi_T \|dT(k)\| \} \} \tag{16a}$$

$$\geq \underset{k \in \mathbf{k}}{max} \{ \Phi_w \|dw_k^*\| + \underset{\mathbf{T}_i}{min} \sum_{T(k) \in \mathbf{T}_i} \sum_{i=k}^{K} \Phi_T \|dT(k)\| \} \} \tag{16b}$$

$$\geq \underset{k \in \mathbf{k}}{max} \{ \Phi_w \|dw_k^*\| + \sum_{T(k) \in \mathbf{T}_i} \sum_{i=k}^{K} \Phi_T \underset{T(k)}{min} \|dT(k)\| \} \} \tag{16c}$$

Let the smallest value of $\underset{T(k)}{min} \|dT(k)\| \geq \Phi_{dT}$, then, we can write

$$\epsilon_k^{(*)} - \epsilon_{k+1}^{(*)} \geq \underset{k \in \mathbf{k}}{max} \{ \Phi_w \|dw_k^*\| + \sum_{T(k) \in \mathbf{T}_i} \sum_{i=k}^{K} \Phi_T \Phi_{dT} \} \tag{17}$$

$$\geq \underset{k \in \mathbf{k}}{max} \{ \Phi_w \|dw_k^*\| \} + \underset{k \in \mathbf{k}}{max} \{ \sum_{T(k) \in \mathbf{T}_i} \sum_{i=k}^{K} \Phi_T \Phi_{dT} \} \tag{18}$$

Taking sum from k to K provides with the fact that each $\mathbf{T}_k$ has a total of k sub datasets.

$$\epsilon_k^{(*)} - \epsilon_K^{(*)} \geq \sum_k^K \left[\underset{k \in \mathbf{k}}{max} \{ \Phi_w \|dw_k^*\| \} + \underset{k \in \mathbf{k}}{max} \{ \sum_{T(k) \in \mathbf{T}_k} \sum_{i=k}^{K} \Phi_T \Phi_{dT} \} \right] \tag{19}$$

$$\geq \sum_k^K \underset{k \in \mathbf{k}}{max} \{ \Phi_w \|dw_k^*\| \} + \sum_k^K \underset{k \in \mathbf{k}}{max} \{ \sum_{T(k) \in \mathbf{T}_k} (K-k) \Phi_T \Phi_{dT} \} \tag{20}$$

$$\geq \sum_k^K \underset{k \in \mathbf{k}}{max} \{ \Phi_w \|dw_k^*\| \} + k(K-k)^2 \underset{k \in \mathbf{k}}{max} \{ \Phi_T \Phi_{dT} \} \tag{21}$$

Since, $\underset{k \in \mathbf{k}}{max} \{ \Phi_T \Phi_{dT} \} = \Phi_T \Phi_{dT}$, $\underset{k \in \mathbf{k}}{max} \{ \Phi_w \Phi_{dw} \} = \Phi_w \Phi_{dw}$ and $\|dw_k^*\| \geq \Phi_{dw}$, we write

$$\epsilon_k^{(*)} - \epsilon_K^{(*)} \geq \sum_k^K \underset{k \in \mathbf{k}}{max} \{ \Phi_w \Phi_{dw} \} + k(K-k)^2 \, \Phi_T \Phi_{dT} \geq \sum_k^K \Phi_w \Phi_{dw} + k(K-k)^2 \, \Phi_T \Phi_{dT} \tag{22a}$$

We will now assume that the changes introduced by the task are bounded over all future and past tasks. Given that $K > 0, k > 0, \Phi_w > 0, \Phi_T > 0, \Phi_{dT} > c$, we obtain

$$\epsilon_k^{(*)} - \epsilon_K^{(*)} \geq (K-k)\Phi_w\Phi_{dw} + k(K-k)^2\ \Phi_T c \tag{23}$$

Now, by assumption that, for each task, the optimal value of weight is obtained after updating the weights for a total of I steps provides $\Phi_{dw} \geq -\sum_{i=0}^{I-1} \alpha_k^{(i)} g_k^{(i)} \geq -\sum_{i=0}^{I-1} \alpha_k^{(i)}(-L) \geq \sum_{i=0}^{I-1} \alpha_{\text{MIN}} L \geq I\alpha_{\text{MIN}} L$. Thus, we obtain

$$\epsilon_k^{(*)} - \epsilon_K^{(*)} \geq (K-k)\Phi_w I\alpha_{\text{MIN}} L + k(K-k)^2\ \Phi_T c \tag{24}$$

Then $\epsilon_k^{(*)} - \epsilon_K^{(*)}$ diverges as a function of K, k, I, c.

Similarly, for the case with regularization we may write $d\Phi_{dw} \geq -\sum_{i=0}^{I} \alpha_k^{(i)} g_k^{(i)} \geq -\sum_{i=0}^{I-1} \alpha_k^{(i)} - (L + \beta\mathcal{R}) \geq \sum_{i=0}^{I-1} \alpha_{\text{MIN}}(L + \beta\mathcal{R}) \geq I\alpha_{\text{MIN}}(L + \beta\mathcal{R})$, where $L, \mathcal{R}$ are the Lipschitz bounds on the gradients and regularizer function respectively and $\beta > 0$ is a coefficient. Thus, we obtain

$$\epsilon_k^{(*)} - \epsilon_K^{(*)} \geq (K-k)\Phi_w I\alpha_{\text{MIN}}(L + \beta\mathcal{R}) + k(K-k)^2\ \Phi_T c \tag{25}$$

and we observe divergence as a function of K, k. □

Finally we show our main result, that is, if a small change is introduced by every task, it accumulate to result in a divergent capacity.

E DIVERGENCE WITH RESPECT TO TASKS

Theorem 3. Under the condition of Theorem 2, let the maximum change in subsequent tasks and weights be given by $\underset{k\in\mathbf{k}}{max}\ \{\Phi_T\Phi_{dT}\} = c$. Then, the $\sum_k^K d\epsilon_k^{(*)}$ diverges as a function of K, and I without any assumptions on the weight updates.

Proof. Given the first difference in capacity, and under the assumption that $\|\partial_{w_k} J_{w_k^*}(\mathbf{T}_i)\| \geq \Phi_w$ and $\|\partial_{T(k)} E_{t\in T(i)} \ell_{w_i^*}(t)\| \geq \Phi_T$

$$\begin{aligned}
\epsilon_k^{(*)} - \epsilon_{k+1}^{(*)} &\geq \underset{k\in\mathbf{k}}{max}\ \{\underset{\mathbf{T}_i}{min}\ \{\|\partial_{w_k} J_{w_k^*}(\mathbf{T}_i)\|\|dw_k^*\| + \sum_{T(k)\in\mathbf{T}_i}\sum_{i=k}^{K} \|\partial_{T(k)} E_{t\in T(i)}\ell_{w_i^*}(t)\|\|dT(k)\|\}\} \\
&\geq \underset{k\in\mathbf{k}}{max}\ \{\underset{\mathbf{T}_i}{min}\ \{\Phi_w\|dw_k^*\| + \sum_{T(k)\in\mathbf{T}_i}\sum_{i=k}^{K} \Phi_T\|dT(k)\|\}\} && \text{(26a)} \\
&\geq \underset{k\in\mathbf{k}}{max}\ \{\Phi_w\|dw_k^*\| + \underset{\mathbf{T}_i}{min} \sum_{T(k)\in\mathbf{T}_i}\sum_{i=k}^{K} \Phi_T\|dT(k)\|\}\} && \text{(26b)} \\
&\geq \underset{k\in\mathbf{k}}{max}\ \{\Phi_w\|dw_k^*\| + \sum_{T(k)\in\mathbf{T}_i}\sum_{i=k}^{K} \Phi_T \underset{T(k)}{min}\|dT(k)\|\}\} && \text{(26c)}
\end{aligned}$$

Let the smallest value of $\underset{T(k)}{min}\|dT(k)\| \geq \Phi_{dT}$, then, we can write

$$\begin{aligned}
\epsilon_k^{(*)} - \epsilon_{k+1}^{(*)} &\geq \underset{k\in\mathbf{k}}{max}\ \{\Phi_w\|dw_k^*\| + \sum_{T(k)\in\mathbf{T}_i}\sum_{i=k}^{K} \Phi_T\Phi_{dT}\} \\
&\geq \underset{k\in\mathbf{k}}{max}\ \{\Phi_w\|dw_k^*\|\} + \underset{k\in\mathbf{k}}{max}\ \{\sum_{T(k)\in\mathbf{T}_i}\sum_{i=k}^{K} \Phi_T\Phi_{dT}\} && \text{(27)}
\end{aligned}$$

Taking sum from k to K provides with the fact that each $\mathbf{T}_k$ has a total of k sub datasets.

$$\epsilon_k^{(*)} - \epsilon_K^{(*)} \geq \sum_k^K \left[\underset{k \in \mathbf{k}}{max} \{\Phi_w \| dw_k^* \|\} + \underset{k \in \mathbf{k}}{max} \{ \sum_{T(k) \in \mathbf{T}_k} \sum_{i=k}^K \Phi_T \Phi_{dT} \} \right] \tag{28a}$$

$$\geq \sum_k^K \underset{k \in \mathbf{k}}{max} \{\Phi_w \| dw_k^* \|\} + \sum_k^K \underset{k \in \mathbf{k}}{max} \{ \sum_{T(k) \in \mathbf{T}_k} (K-k) \Phi_T \Phi_{dT} \} \tag{28b}$$

$$\geq \sum_k^K \underset{k \in \mathbf{k}}{max} \{\Phi_w \| dw_k^* \|\} + k(K-k)^2 \underset{k \in \mathbf{k}}{max} \{\Phi_T \Phi_{dT}\} \tag{28c}$$

Since, $\underset{k \in \mathbf{k}}{max} \{\Phi_T \Phi_{dT}\} = \Phi_T \Phi_{dT}$, $\underset{k \in \mathbf{k}}{max} \{\Phi_w \Phi_{dw}\} = \Phi_w \Phi_{dw}$ and $\| dw_k^* \| \geq \Phi_{dw}$, we write

$$\epsilon_k^{(*)} - \epsilon_K^{(*)} \geq \sum_k^K \underset{k \in \mathbf{k}}{max} \{\Phi_w \Phi_{dw}\} + k(K-k)^2 \ \Phi_T \Phi_{dT} \tag{29a}$$

Assuming that the changes introduced by the task are bounded over all future and past tasks, i.e., $\Phi_{dT} > c$, we get

$$\epsilon_k^{(*)} - \epsilon_K^{(*)} \geq \sum_k^K \Phi_w \Phi_{dw} + k(K-k)^2 \Phi_T \ c \tag{30a}$$

Even for a constant change in task, $\epsilon_k^{(*)} - \epsilon_K^{(*)}$ diverges as a function of K. □

F Details for Case Study 4

We used the following configuration of a transformer block to instantiate the 8M model.

Embedding layer: (32000, 128); Attention layer: (k, q, v, o): (128, 128); MLP layer: gate_projection (128, 256), up_projection (128, 256), down_projection (256, 128); Activation function: SiLU; Layernorm: RMSNorm ; Head layer: (128, 32000); Attention heads: 2 ; Layers: 2 Hidden size: 128.

We used the following configuration of a transformer block to instantiate the 134M model.

Embedding layer: (32000, 768); Attention layer: (k, q, v, o): (768, 768); MLP layer: gate_projection (768, 2048), up_projection (768, 2048), down_projection (2048, 768); Activation function: SiLU; Layernorm: RMSNorm ; Head layer: (768, 32000); Attention heads: 12 ; Layers: 12 Hidden size: 768.

Pre-training Data Mix:

- `wiki`: 0.28
- `git`: 0.28
- `arxiv`: 0.16
- `books`: 0.28

Experience Replay - Data Mix:

- `wiki`: 1.0
- `wiki`: 0.2, `git`: 0.8
- `wiki`: 0.1, `git`: 0.1, `arxiv`: 0.8
- `wiki`: 0.06, `git`: 0.07, `arxiv`: 0.07, `books`: 0.8