# SYNERGY BETWEEN SYNAPTIC CONSOLIDATION AND EXPERIENCE REPLAY FOR GENERAL CONTINUAL LEARNING

**Fahad Sarfraz,**∗ **Elahe Arani**∗**, Bahram Zonooz**
Advanced Research Lab, NavInfo Europe, The Netherlands
{fahad.sarfraz,elahe.arani}@navinfo.eu, bahram.zonooz@gmail.com

## ABSTRACT

Continual learning (CL) in the brain is facilitated by a complex set of mechanisms. This includes the interplay of multiple memory systems for consolidating information as posited by the complementary learning systems (CLS) theory and synaptic consolidation for protecting the acquired knowledge from erasure. Thus, we propose a general CL method that creates a synergy between SYNaptic consolidation and dual memory Experience Replay (SYNERgy). Our method maintains a semantic memory that accumulates and consolidates information across the tasks and interacts with the episodic memory for effective replay. It further employs synaptic consolidation by tracking the importance of parameters during the training trajectory and anchoring them to the consolidated parameters in the semantic memory. To the best of our knowledge, our study is the first to employ dual memory experience replay in conjunction with synaptic consolidation that is suitable for general CL whereby the network does not utilize task boundaries or task labels during training or inference. Our evaluation on various challenging CL scenarios and characteristics analyses demonstrate the efficacy of incorporating both synaptic consolidation and CLS theory in enabling effective CL in DNNs.[1]

## 1 INTRODUCTION

A major challenge towards achieving general intelligence and making DNNs suitable for deployment in an ever-changing environment is the ability to continuously acquire, retain and consolidate knowledge (Parisi et al., 2019). However, DNNs exhibit catastrophic forgetting whereby the network's performance on previously learned tasks drops drastically as they learn a new task (McCloskey & Cohen, 1989). This can be majorly attributed to the violation of the strong i.i.d assumption in the sequential learning of tasks involved in CL (Hadsell et al., 2020). Furthermore, catastrophic forgetting is also considered as an inevitable feature of the standard training procedure on connectionist models (Kirkpatrick et al., 2017) whereby all the weights of the model are updated to learn a new task, resulting in overwriting of important weights for the previous tasks. To avoid catastrophic forgetting without hampering the learning of new tasks, a CL method needs to be plastic enough to adapt to the new data while being stable enough to not forget the already acquired knowledge.

In the brain, a delicate balance between plasticity and stability is maintained by a rich set of neurophysiological mechanisms (Parisi et al., 2019; Zenke et al., 2017) and facilitated by multiple memory systems (Hassabis et al., 2017). In particular, synaptic consolidation (Clopath, 2012) retains information over longer periods by decreasing the rates of plasticity in a proportion of strengthened synapses that persist despite learning of other tasks (Cichon & Gan, 2015). Furthermore, the CLS theory posits that intelligent agents must possess two complementary learning systems: instance-based hippocampal system and parametric neocortical system. The Hippocampus rapidly encodes novel information which is gradually consolidated into the structured knowledge representation of the neocortex through the replay of neural activity patterns that accompanied the learning event (Tulving, 2002; Kumaran et al., 2016). The synergy between these mechanisms plays a crucial role in enabling efficient lifelong learning in the brain. Therefore, we hypothesize that an efficient method of employing a dual memory replay mechanism and incorporating synaptic consolidation for mediating the plasticity of weights can enable effective CL in DNNs.

Several works have taken inspiration from the studies in neuroscience, but they have mainly focused on a single aspect. The replay mechanism in Hippocampus has inspired a series of rehearsal-based approaches (Chaudhry et al., 2018b; Lopez-Paz & Ranzato, 2017) which maintains an episodic memory buffer and replays samples from previous tasks while training on the new task. On the other hand, regularization-based approaches (Kirkpatrick et al., 2017; Ritter et al., 2018; Zenke et al., 2017) derive inspiration from synaptic consolidation and penalize changes in the network

---

∗ Contributed equally.     [1] We open source our code at https://github.com/NeurAI-Lab/SYNERgy.

weights. Both of these approaches have their merits and demerits. Rehearsal-based approaches have proven to be more effective (Farquhar & Gal, 2018) but require the amount of memory being stored and replayed to be proportional to the number of tasks which can be a limiting factor in learning longer sequences. On the other hand, while regularization-based approaches alleviate the need to store samples from previous tasks, they fail to estimate the joint distribution of all the encountered tasks on challenging CL settings (Farquhar & Gal, 2018) and often require information about the task boundaries and/or task labels. Research in both of these directions has mostly been orthogonal and an optimal method for incorporating both of these mechanisms for enabling CL in DNNs is not well studied. We, therefore, explore the design space of CL methods that employ experience replay to consolidate knowledge and synaptic consolidation to retain acquired knowledge over longer periods.

We propose SYNERgy which employs a dual memory replay mechanism in conjunction with a synaptic consolidation for general incremental learning (General-IL) (Figure 1). As the structural knowledge of the learned tasks is encoded in the weights of the working model (Kirkpatrick et al., 2017), our method maintains an additional semantic memory which aggregates and consolidates knowledge across the tasks by taking an exponential moving average of the working model's weights in a stochastic manner. The semantic memory interacts with the episodic memory to provide consolidated replay activations for enforcing a similarity structure on the update of the working model. This facilities knowledge consolidation and aligns the working model's decision boundary with the semantic memory. To further retain knowledge of previous tasks, SYNERgy incorporates synaptic consolidation by tracking the importance of parameters during the training trajectory and slowing down learning on parameters that are considered important for the acquired knowledge. Our work focuses on the realistic General-IL setting which involves training the method on a long sequence of tasks where the boundaries between the tasks are not distinct, the tasks themselves are not disjoint, and the method does not have access to task boundaries during training or testing (Arani et al., 2022). Therefore, SYNERgy does not utilize the task boundaries or make any assumption about the distribution of the data stream.

We extensively evaluate our method on a wide range of CL scenarios which covers a broad spectrum of the challenges a learning agent has to face in the real-world. The empirical results demonstrate the effectiveness and versatility of SYNERgy in acquiring, consolidating, and retaining knowledge over longer sequences. In addition to reducing forgetting, our method improves the reliability of the model, reduces the bias towards recent tasks, and converges to wider minima. Our work presents a compelling case for further exploring the design space of brain-inspired CL methods which incorporates both synaptic consolidation and complementary learning systems.

## 2 BACKGROUND

**Regularization-based methods** draw inspiration from synaptic consolidation in the brain to propose algorithms which control the rate of learning on model weights based on their importance to previously seen tasks (Kirkpatrick et al., 2017). oEWC (Schwarz et al., 2018) estimates the importance of each parameter with the empirical Fisher information matrix (FIM), aggregates the FIMs evaluated at the task boundaries, and subsequently slows down learning on the subset of the model weights which are considered important for previous tasks. SI (Zenke et al., 2017) computes an importance measure online along the entire learning trajectory. While it constitute a promising biologically plausible direction, they fail to learn the joint distribution of tasks and avoid catastrophic forgetting in more challenging settings (Farquhar & Gal, 2018; Hadsell et al., 2020). Additionally, they require task boundary and/or task label information which limits their application in the General-IL setting where task boundaries are not discrete.

**Rehearsal-based approaches** are inspired by the critical role that the replay of past neural activation patterns in the brain play in memory formation, consolidation, and retrieval. Experience Replay (ER) (Riemer et al., 2018) typically involves storing a subset of samples from previous tasks and mixing them with data from the new task to update the model parameters with an approximate joint distribution and has proven to be effective in reducing catastrophic forgetting under challenging settings (Farquhar & Gal, 2018). Several techniques have since been employed on top of ER to better utilize the memory samples (Lopez-Paz & Ranzato, 2017; Aljundi et al., 2019). Dark Experience Replay (DER++) samples logits during the entire optimization trajectory and adds a consistency loss on top of ER. A key limitation of experience replay methods is that the memories being stored and replayed need to be proportional to the number of tasks as for a fixed buffer size, the representation of earlier tasks in memory diminishes as we learn over longer sequences. Furthermore, an optimal mechanism for replay is still an open question (Hayes et al., 2021).

**CLS theory** has further inspired dual memory learning systems (Pham et al., 2021; Wang et al., 2022). CLS-ER (Arani et al., 2022) mimics the fast and slow learning in CLS by maintaining short- and long-term semantic memories which stochastically aggregates the weights of the working model using exponentially moving average at different rates. The semantic memories interact with the episodic memory to enforce a consistency loss for replay. We use their proposed Mean-ER approach as it provides an effective and simpler approach for studying the effect of synaptic consolidation in a dual memory replay architecture.
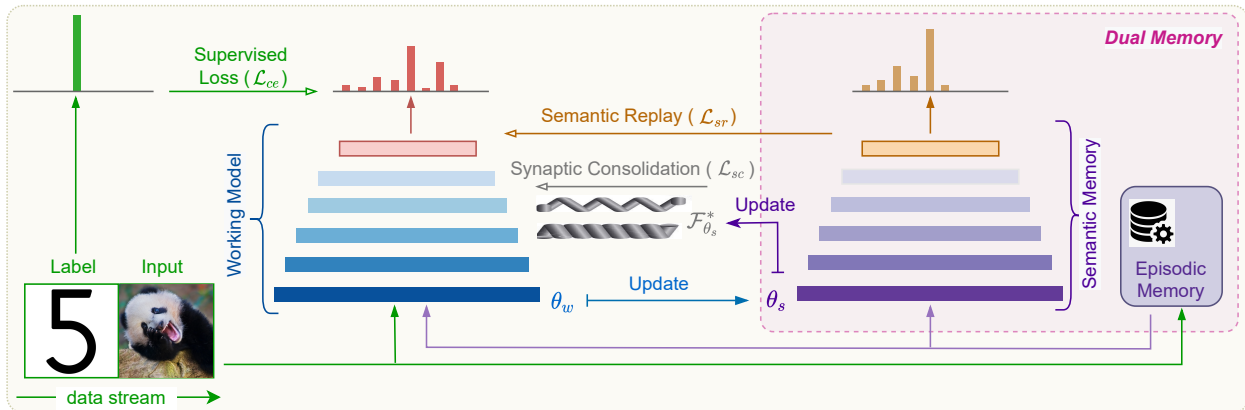
Figure 1: SYNERgy employs synaptic consolidation alongside an experience replay in a dual memory system. The semantic memory, $\theta_s$, is maintained by taking an exponential moving average of the working model, $\theta_w$, to aggregate knowledge across the tasks. The semantic memory then interacts with the episodic memory to extract consolidated replay logits for semantic replay in addition to episodic replay. This enables the working model to effectively consolidate knowledge across tasks by learning the joint distribution of tasks while aligning its decision boundary with the semantic memory. Furthermore, SYNERgy tracks the importance of the semantic memory parameters by aggregating Fisher information matrices, $\mathcal{F}_{\theta_S}^*$, and constrains the weight changes of the working model by anchoring them to the important parameters of the semantic memory through the synaptic consolidation loss $\mathcal{L}_{sc}$ (details in Algorithm 1).

While all these approaches have their merits and demerits and individually constitute a promising direction, research has been predominantly orthogonal. Our work draws inspiration from how these mechanisms together enable effective learning in the brain and explore the design space of CL methods that employ synaptic consolidation and experience replay in multiple memory systems for General-IL.

## 3 METHODOLOGY

The human brain has evolved to continually acquire, retain and consolidate knowledge. This ability to efficiently learn from an ever-changing environment is enabled by a rich set of neurophysiological mechanisms and multiple memory systems. Studies focusing on the structure and function of dendritic spines during learning show evidence for specialized mechanisms that protect knowledge about previous tasks from interference during learning on a new task (Cichon & Gan, 2015; Yang et al., 2009). Neurobiological models suggest that CL in the neocortex relies on a process of task-specific synaptic consolidation which involves rendering a proportion of synapses less plastic and therefore stable over long timescales (Benna & Fusi, 2016; Kirkpatrick et al., 2017). Another important characteristic of the brain is that it relies on multiple memory systems. The CLS theory (Kumaran et al., 2016) posits that intelligent agents must possess two learning systems: the Hippocampus quickly learns the specifics of individual experiences which are then gradually consolidated into the structured knowledge representations in the Neocortex. This process is accompanied by the replay of the structured patterns of neural activities that accompanied the learning event. We believe that exploring the design space of models that employ both synaptic consolidation and CLS theory mechanisms may hold the key to enabling more effective CL in DNNs. To this end, similar to the brain, our proposed method SYNERgy incorporates synaptic consolidation and experience replay in a dual memory manner for General-IL.

### 3.1 DUAL MEMORY SYSTEM

Inspired by the interplay of the rapid instance-based system (Hippocampus) and the slow parametric system (Neocortex) in the brain, our method employs an instance-based episodic memory and a semantic memory which maintains consolidated structured knowledge across the learned tasks.

**Episodic Memory:** To replay samples from the previous tasks, we utilize a small episodic memory buffer as an instance-based memory system which can be thought of as a very primitive hippocampus, permitting a form of complementary learning (Hassabis et al., 2017). As our focus is on General-IL, we employ *Reservoir sampling* (Vitter, 1985) to maintain the memory buffer which does not utilize the task boundaries to select the memory samples. It instead attempts to approximately match the distribution of the incoming stream by assigning equal probability to

each incoming sample for being added to the buffer and replacing a random sample in the buffer. The probability for an incoming sample to be added to the buffer is given by $\frac{\mathcal{B}}{N_s}$, where $\mathcal{B}$ is the buffer size and $N_s$ is the total number of seen samples. Note that we opt for Reservoir sampling because of its simplicity, effectiveness, and suitability for General-IL, however, any method for memory selection and retrieval can be employed in our method.

**Semantic Memory:** The acquired knowledge of the tasks in DNNs is represented in the learned weights (Krishnan et al., 2019). Therefore, we aim to progressively accumulate and consolidate knowledge across the tasks encoded in the subsequent learned weights of the working model ($\theta_w$) throughout the training trajectory. To this end, we follow the approach in Arani et al. (2022) to maintain our semantic memory (parametrized by $\theta_s$) by taking the exponential moving average of the working model as they provides an efficient approach for aggregating weights:

$$\theta_s \leftarrow \alpha_s \theta_s + (1 - \alpha_s)\, \theta_w, \quad if\ \ r_s > a \sim U(0,1) \tag{1}$$

where $\alpha_s$ is the decay parameter and $r_s$ is the update rate. As the semantic memory is designed to stochastically accumulate knowledge across the entire training trajectory by aggregating the weights of the working model, it can also be considered as an ensemble of several student models with varying degrees of specialization for different tasks. The semantic memory loosely mimics the Neocortex as it constitutes a parametric memory system that builds structural representations for generalization that interacts with the instance level episodic memory (Section 3.3) to efficiently accumulate and consolidate knowledge.

## 3.2 SYNAPTIC CONSOLIDATION

In the brain, synaptic consolidation supports CL by reducing the plasticity of synapses that are considered important to previously learned tasks. This is enabled by a complex molecular machinery in biological synapses that allows multiples scales of learning by mediating the plasticity of individual synapses (Redondo & Morris, 2011). In DNNs, however, weights are only described by a scalar value and employing synaptic consolidation requires an additional estimate of the importance of synapses for the previous tasks. Thus, we utilize the Fisher information matrix (FIM) to estimate the importance of each parameter (Kirkpatrick et al., 2017). We maintain a running approximation of FIM, similar to the online variants of EWC (Chaudhry et al., 2018a; Schwarz et al., 2018), but make several modifications to make the algorithm suitable for General-IL and enable effective consolidation of information across the tasks:

i. We do not utilize the task boundaries and instead update the running estimate of FIM throughout the training trajectory stochastically using an update probability.

ii. Instead of evaluating the FIM on the training data of the current task, we use the samples in the memory buffer as they approximate the joint distribution of all the seen tasks. This provides a better estimate of the importance of a parameter for all the tasks and avoids the update from biasing the running estimate towards the current task.

iii. As the semantic memory effectively consolidates information across the tasks, and provides a better estimate of the optimal parameters for the joint distribution of tasks seen so far (Figure S1 in Appendix), we use it to evaluate the FIM instead of the working model.

iv. Subsequently, the working model weights are anchored to the important parameters of the semantic memory. Since the synaptic regularization term acts as a proxy for the gradients of the previous tasks Hadsell et al. (2020), the consolidated weights of the semantic memory can provide a better estimate and more optimal constraints for the working model update. The use of semantic memory as anchor parameters has the additional benefits of not requiring to save the model's state which saves memory and not over constraining the model by anchoring it to parameters at the task boundary which may be unreliable and may hamper consolidation of information.

v. Finally, we note that one shortcoming of the regularization-based approaches is that they do not consider the structural functionality of filters in the CNNs. Each filter consists of a unit that extracts a specific feature from the input. Therefore, allowing large changes in some parameters while penalizing changes in others fails to prevent a drift in the functionality of the filter unit which might be important for the previous tasks. For instance, a horizontal edge detector requires each parameter of the filter to have certain values and changes in even a few parameters changes the functionality of the filter. To address this, we modify the FIM such that a uniform penalty is applied to all parameters of a filter. Specifically, we adjust the FIM for penalizing the weight changes such that the importance of each filter in the convolutional layer is given by the mean importance value of its parameters.

## 3.3 FORMULATION

SYNERgy employs a dual memory experience replay mechanism in conjunction with synaptic consolidation. It maintains a fixed episodic memory $\mathcal{M}$ with budget size $\mathcal{B}$ through *Reservoir sampling* and an additional semantic memory $\theta_s$ which aggregates the weights of the working model $\theta_w$ by taking an exponential moving average using Eq.

---

**Algorithm 1** SYNERgy algorithm for general continual learning

---

    **Input:** data stream $\mathcal{D}$; learning rate $\eta$; penalty terms $\lambda$, $\beta$; update probabilities $r_s$, $r_F$; decay parameters $\alpha_S$, $\alpha_F$
    **Initialize:** $\theta_s = \theta_w$
    $\mathcal{M} \leftarrow \{\}$
1: **while** Training **do**
2:      Sample data: $(x_b, y_b) \sim \mathcal{D}$ and $(x_m, y_m) \sim \mathcal{M}$
3:      Calculate adjusted Fisher information matrix: $\mathcal{F}_{adj}$
4:      Calculate loss: $\mathcal{L} = \mathcal{L}_{sl} + \lambda \mathcal{L}_{sr} + \beta \mathcal{L}_{sc}$    (Eqs. 2, 4 & 7)
5:      Update working model: $\theta_w \leftarrow \theta_w - \eta \nabla_{\theta_w} \mathcal{L}$
6:      Update semantic memory: $\theta_s \leftarrow \alpha_s \theta_s + (1 - \alpha_s) \theta_w$,   $if\ \ r_s > a \sim U(0,1)$    (Eq. 1)
7:      Update Fisher information matrix: $\mathcal{F}_{\theta_s}^* \leftarrow \alpha_F \mathcal{F}_{\theta_s}^* + (1 - \alpha_F) \mathcal{F}_{\theta_s}$,   $if\ \ r_F > a \sim U(0,1)$   (Eqs. 5 & 6)
8:      Update episodic memory: $\mathcal{M} \leftarrow \text{Reservoir}(\mathcal{M}, (x_b, y_b))$
    **return** $\theta_s$

---

1. While training, samples from the current task, $(x_b, y_b) \sim \mathcal{D}_t$, are interleaved with the memory buffer samples, $(x_m, y_m) \sim \mathcal{M}$. The overall loss function for updating the working model comprises of three components as follows:

**Supervised Loss:** consists of learning from the ground truth labels of the incoming stream and the memory buffer samples through the task loss, $\mathcal{L}_t$, and episodic replay loss, $\mathcal{L}_{er}$. This ensures that the model learns from the approximate joint distribution of all the tasks seen so far. We use the standard Cross-Entropy loss for object recognition for both the task and episodic replay losses.

$$\mathcal{L}_{sl} = \mathcal{L}_t(f(x_b; \theta_w, y_b)) + \mathcal{L}_{er}(f(x_m; \theta_w, y_m)) \tag{2}$$

$$\mathcal{L}_{ce} = \mathcal{L}_t(f(x_b; \theta_w, y_b)) + \mathcal{L}_{er}(f(x_m; \theta_w, y_m)) \tag{3}$$

**Semantic Replay:** To provide additional structural information about the previous tasks, we utilize the semantic memory to extract the consolidated logits for effective replay of the memory samples. As the semantic memory aggregates information across the tasks, it takes into account the required adaptations in the decision boundary and the feature space to provide consolidated logits for replay, which facilitates the consolidation of knowledge (Arani et al., 2022). Therefore, we apply semantic replay between the logits of the working model and the semantic memory on the memory samples, which encourages the model to maintain the similarity structure between the classes.

$$\mathcal{L}_{sr} = \mathbb{E}_{x_m \sim M}[(f(x_m; \theta_w) - f(x_m; \theta_s))^2]. \tag{4}$$

**Synaptic Consolidation:** To incorporate synaptic consolidation, we utilize the empirical FIM, $\mathcal{F}_{\theta_s}$, to estimate the importance of each parameter for the encountered tasks. At task t, we use the semantic memory to evaluate the FIM on all the samples in the memory buffer to approximate the expectation over the joint distribution of tasks seen so far, $\cup_{i=1}^{t} \mathcal{D}_i$. Consequently, the semantic memory is used as an anchor for constraining the network update. Our approach provides a better estimate of the parameter's importance as the semantic memory maintains consolidated information across the tasks and the Reservoir sampling in the episodic memory approximates the joint distribution of tasks,

$$\mathcal{F}_{\theta_s} = \mathbb{E}_{(x,y) \sim \mathcal{M}}\left[\left(\frac{\partial \log p(y|x; \theta_s)}{\partial \theta_s}\right)\left(\frac{\partial \log p(y|x; \theta_s)}{\partial \theta_s}\right)^T\right]. \tag{5}$$

As our focus is on General-IL, we avoid utilizing the task boundaries to calculate the FIM and save the model's state as anchor parameters. Instead, we take a stochastic approach to evaluate the FIM periodically throughout the training trajectory and aggregate them using an exponential moving average,

$$\mathcal{F}_{\theta_s}^* \leftarrow \alpha_F \mathcal{F}_{\theta_s}^* + (1 - \alpha_F) \mathcal{F}_{\theta_s}, \quad if\ \ r_F > a \sim U(0,1) \tag{6}$$

where $\alpha_F$ is the decay parameter to control the strength of update and $r_F$ is the update rate.

Furthermore, to maintain the structural functionality of filters in CNNs as noted in Section 3.2, we adjust the importance estimate such that importance is defined at a filter level instead of individual parameters. We use the mean of the filter parameters in the aggregated FIM, $\mathcal{F}_{\theta_s}^*$, to estimate the importance of a filter (the resulting adjusted FIM is denoted by $\mathcal{F}_{adj}$). The synaptic consolidation loss is then given by the quadratic penalty weighted by $\mathcal{F}_{adj}$,

$$\mathcal{L}_{sc} = \| \theta_w - \theta_s \|_{\mathcal{F}_{adj}}^2. \tag{7}$$

**Overall Loss:** The working model is then updated by the weighted sum of individual losses,

$$\mathcal{L} = \mathcal{L}_{sl} + \lambda \mathcal{L}_{sr} + \beta \mathcal{L}_{sc} \tag{8}$$

where $\lambda$ and $\beta$ control the strength of semantic replay and synaptic consolidation losses, respectively. Further details of the method are given in Algorithm 1. Note that for inference, we use the semantic memory as it gradually consolidates structural representations for generalization across the tasks (see Figure 4).

Table 1: Comparison with prior works on various CL settings. The baseline results for Class-IL and Domain-IL on 200 and 500 buffer sizes are from Buzzega et al. (2020) and MNIST-360 results are from Arani et al. (2022).

| Buffer | Method | Class-IL | | Domain-IL | General-IL | | |
|---|---|---|---|---|---|---|---|
| | | S-CIFAR10 | S-TinyImg | R-MNIST | MNIST360 | GCIL-U | GCIL-L |
| – | JOINT | $92.20_{\pm0.15}$ | $59.99_{\pm0.19}$ | $95.76_{\pm0.04}$ | $82.98_{\pm3.24}$ | $58.36_{\pm1.02}$ | $56.94_{\pm1.56}$ |
| | SGD | $19.62_{\pm0.05}$ | $7.92_{\pm0.26}$ | $67.66_{\pm8.53}$ | $19.09_{\pm0.69}$ | $12.67_{\pm0.24}$ | $22.88_{\pm0.53}$ |
| – | oEWC | $19.49_{\pm0.12}$ | $7.58_{\pm0.10}$ | $77.35_{\pm5.77}$ | - | - | - |
| | SI | $19.48_{\pm0.17}$ | $6.58_{\pm0.31}$ | $71.91_{\pm5.83}$ | - | - | - |
| | LwF | $19.61_{\pm0.05}$ | $8.46_{\pm0.22}$ | - | - | - | - |
| 200 | ER | $44.79_{\pm1.86}$ | $8.49_{\pm0.16}$ | $85.01_{\pm1.90}$ | $49.27_{\pm2.25}$ | $16.40_{\pm0.37}$ | $19.27_{\pm0.77}$ |
| | DER++ | $64.88_{\pm1.17}$ | $10.96_{\pm1.17}$ | $90.43_{\pm1.87}$ | $54.16_{\pm3.02}$ | $18.84_{\pm0.60}$ | $26.94_{\pm1.27}$ |
| | CLS-ER | $66.19_{\pm0.75}$ | $\mathbf{21.95}_{\pm0.26}$ | $\mathbf{92.26}_{\pm0.18}$ | $\mathbf{66.37}_{\pm0.83}$ | $25.06_{\pm0.81}$ | $28.54_{\pm0.87}$ |
| | SYNERgy | $\mathbf{66.27}_{\pm1.49}$ | $21.08_{\pm0.98}$ | $92.02_{\pm0.83}$ | $65.23_{\pm0.52}$ | $\mathbf{30.00}_{\pm2.63}$ | $\mathbf{32.24}_{\pm2.14}$ |
| 500 | ER | $57.74_{\pm0.27}$ | $9.99_{\pm0.29}$ | $88.91_{\pm1.44}$ | $65.04_{\pm1.53}$ | $28.21_{\pm0.69}$ | $20.30_{\pm0.63}$ |
| | DER++ | $72.70_{\pm1.36}$ | $19.38_{\pm1.41}$ | $92.77_{\pm1.05}$ | $69.62_{\pm1.59}$ | $32.92_{\pm0.74}$ | $25.82_{\pm0.83}$ |
| | CLS-ER | $\mathbf{75.22}_{\pm0.71}$ | $29.61_{\pm0.54}$ | $\mathbf{94.06}_{\pm0.07}$ | $75.70_{\pm0.41}$ | $36.34_{\pm0.59}$ | $28.63_{\pm0.68}$ |
| | SYNERgy | $73.85_{\pm0.73}$ | $\mathbf{31.26}_{\pm0.67}$ | $93.37_{\pm0.45}$ | $\mathbf{76.48}_{\pm0.83}$ | $\mathbf{38.45}_{\pm0.46}$ | $\mathbf{33.07}_{\pm2.28}$ |
| 1000 | ER | $70.64_{\pm1.13}$ | $13.23_{\pm0.39}$ | $89.44_{\pm1.34}$ | $75.18_{\pm1.50}$ | $31.98_{\pm0.72}$ | $34.13_{\pm0.83}$ |
| | DER++ | $77.42_{\pm1.06}$ | $24.76_{\pm1.21}$ | $93.72_{\pm0.56}$ | $76.03_{\pm1.61}$ | $38.95_{\pm0.56}$ | $33.64_{\pm0.88}$ |
| | CLS-ER | $76.44_{\pm0.44}$ | $34.09_{\pm0.37}$ | $\mathbf{93.73}_{\pm0.20}$ | $79.54_{\pm0.34}$ | $39.69_{\pm0.66}$ | $39.52_{\pm0.91}$ |
| | SYNERgy | $\mathbf{77.47}_{\pm1.15}$ | $\mathbf{36.49}_{\pm0.58}$ | $93.63_{\pm0.49}$ | $\mathbf{80.96}_{\pm0.33}$ | $\mathbf{41.08}_{\pm0.59}$ | $\mathbf{39.96}_{\pm1.59}$ |

## 4 EMPIRICAL EVALUATION

We compare SYNERgy with regularization- (oEWC, SI), rehearsal- (ER, DER++), knowledge distillation-(LwF; Li & Hoiem (2017)) and multiple memory-based (CLS-ER) methods under uniform experimental settings on varying CL scenarios (details of CL protocols and experimental setup are provided in Appendix A and E). The lower bound is provided by standard training without any measure to avoid forgetting (SGD) and the upper bound is given by the joint training of all tasks (JOINT).

Table 1 shows the performance of the models on different datasets with varying complexities in the Class-IL, Domain-IL and General-IL settings. SYNERgy performs on par with the state-of-the-art rehearsal based approaches or provides generalization gains in almost all the considered scenarios. Particularly in the challenging Class-IL setting with low buffer size and higher data complexity, SYNERgy improves the performance gains. The considerable performance gap between the regularization- and rehearsal-based approaches present a compelling case for the necessity of a small memory buffer to learn the joint distribution of the sequential tasks. This can be attributed to modeling the weight importance based on the local information at the task boundaries which can be unreliable and over-restrictive for constraining the model update. Furthermore, they fail to account for the required consolidation of information across the tasks and modifications in the representation space to discriminate across all the tasks and not only among each task. Synaptic consolidation in our method addresses these shortcomings.

On the other hand, the gap between the baseline rehearsal approach ER and our method in the more practical lower buffer regime (200 and 500) shows that merely intertwining the memory samples and current task samples is not sufficient to learn the joint distribution. The network can benefit both from extracting and replaying more information (e.g. neural activations) from the previous state as well as restricting the mode update to avoid overwriting important parameters. DER++ addresses the former by saving the soft-targets (logits) alongside the samples in memory and the performance improvement shows the effectiveness of providing additional information about the similarity structure of the classes. However, the local logits from the previous state of the model in their method fail to account for the adaptation of the decision boundaries and representation space required to distinguish classes in previous tasks from the new classes. Therefore, enforcing the model to match the sub-optimal logits and thereby a similarity structure that may not hold hampers the consolidation of information. This is more pronounced in the challenging S-TinyImageNet (referred to as S-TinyImg) setting where both the number of classes and inter-task similarity is high. In contrast, the aggregating and consolidating information across the tasks in the semantic memory provides a global context and an updated similarity structure for replay in the logits. The performance of CLS-ER and our method highlights the benefits of multiple memory systems in CL. While their method benefits from maintaining two semantic memories, it
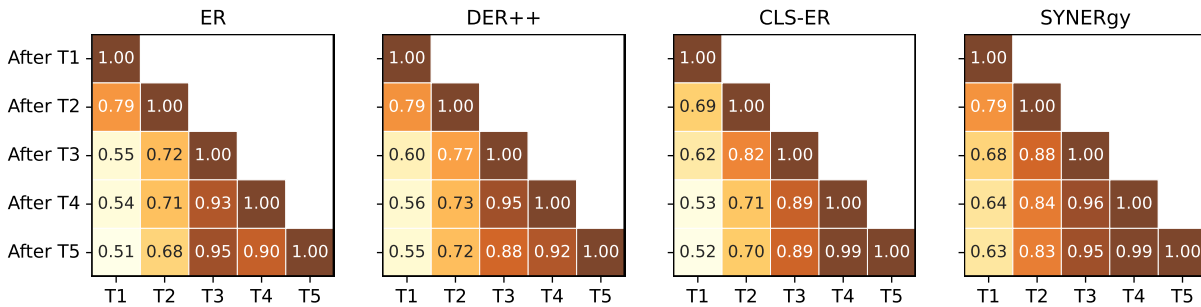
Figure 2: Drift in the learned parameters as training progresses. We report the average cosine similarity between the layers of task models (x-axis) at the end of each task (y-axis). Synaptic consolidation enables the model to consolidate knowledge while keeping the parameters closer to their optimal task parameters.
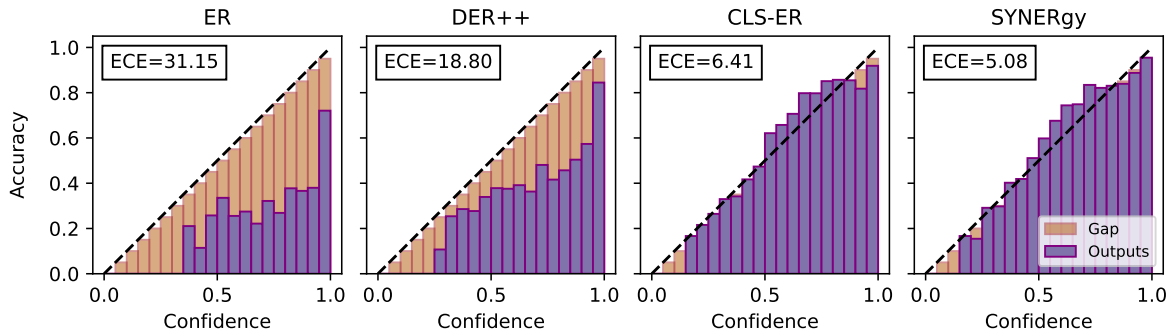


Figure 3: Reliability plots and expected calibration error: SYNERgy improves the reliabiliy of predictions even with lower buffer size. See Figures S2 and S3 for other dataset and buffer sizes.

lacks a mechanism for synaptic consolidation which our method addresses. The task performance analysis in Section 5 shows that SYNERgy provides a better balance between the plasticity and stability of the model.

We also compare the performance of the models suitable for General-IL on the challenging MNIST360 and GCIL datasets. SYNERgy provides significant performance improvement in both of these settings where the model needs to effectively accumulate and transfer knowledge of a class across repeated exposure and deal with multiple distribution shifts. The sub-optimal logits in DER++ fail to take advantage of the improved representations from the additional samples of the recurring classes in later tasks whereas the semantic memory consolidates information across different recurrences. The performance improvement over CLS-ER in this challenging setting, highlights the benefit of incorporating synaptic consolidation in the muliple memory systems for enabling general incremental learning. SYNERgy effectively employs a mechanism for synaptic consolidation that is suitable for the blurry task boundaries in General-IL which was missing in literature, and opens up a new direction of research that incorporates both these mechanisms which are critical components of the learning machinery of the brain into DNNs to enable effective CL.

## 5 MODEL CHARACTERISTICS

Here, we probe into the characteristics distilled in the model trained on S-CIFAR-100 with 500 buffer size with multiple memory-based replay and synaptic consolidation and compare it with ER, DER++, and CLS-ER.

**Drift in Parameters:** An important aspect of enabling effective CL is to avoid the drift in the optimal parameters of the model for previous tasks to mitigate forgetting. To analyze the impact of synaptic consolidation in SYNERgy, we evaluate the change in the learned parameters at the end of each task as training progresses. We evaluate the similarity between the parameters by averaging the cosine similarity between the layers of the model. To account for the scale differences and linear changes, we normalize the weights of each layer by dividing by the maximum weight. Figure 2 shows that SYNERgy effectively reduces the changes in the parameters. In particular, the final parameters of SYNERgy (T5) are considerably closer to the earliest tasks (T1 and T2) compared to the baselines. This suggests

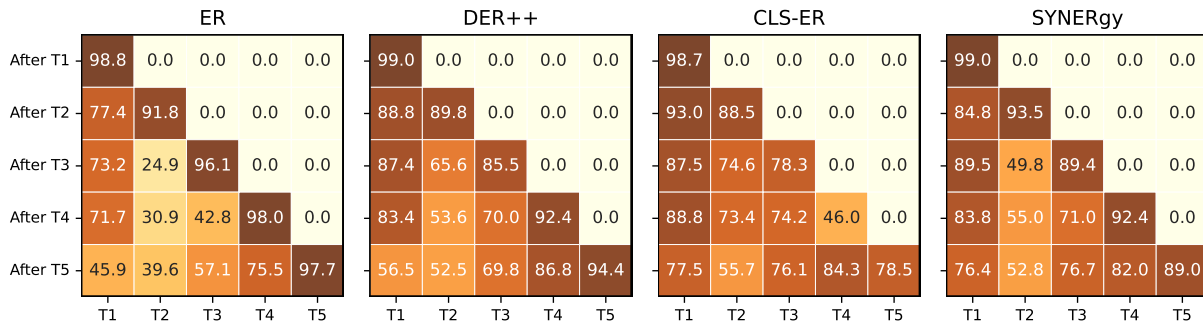|  | ER |  |  |  |  |  | DER++ |  |  |  |  |  | CLS-ER |  |  |  |  |  | SYNERgy |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| After T1 | 98.8 | 0.0 | 0.0 | 0.0 | 0.0 |  | 99.0 | 0.0 | 0.0 | 0.0 | 0.0 |  | 98.7 | 0.0 | 0.0 | 0.0 | 0.0 |  | 99.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| After T2 | 77.4 | 91.8 | 0.0 | 0.0 | 0.0 |  | 88.8 | 89.8 | 0.0 | 0.0 | 0.0 |  | 93.0 | 88.5 | 0.0 | 0.0 | 0.0 |  | 84.8 | 93.5 | 0.0 | 0.0 | 0.0 |
| After T3 | 73.2 | 24.9 | 96.1 | 0.0 | 0.0 |  | 87.4 | 65.6 | 85.5 | 0.0 | 0.0 |  | 87.5 | 74.6 | 78.3 | 0.0 | 0.0 |  | 89.5 | 49.8 | 89.4 | 0.0 | 0.0 |
| After T4 | 71.7 | 30.9 | 42.8 | 98.0 | 0.0 |  | 83.4 | 53.6 | 70.0 | 92.4 | 0.0 |  | 88.8 | 73.4 | 74.2 | 46.0 | 0.0 |  | 83.8 | 55.0 | 71.0 | 92.4 | 0.0 |
| After T5 | 45.9 | 39.6 | 57.1 | 75.5 | 97.7 |  | 56.5 | 52.5 | 69.8 | 86.8 | 94.4 |  | 77.5 | 55.7 | 76.1 | 84.3 | 78.5 |  | 76.4 | 52.8 | 76.7 | 82.0 | 89.0 |
|  | T1 | T2 | T3 | T4 | T5 |  | T1 | T2 | T3 | T4 | T5 |  | T1 | T2 | T3 | T4 | T5 |  | T1 | T2 | T3 | T4 | T5 |

Figure 4: Task-wise performance of methods: Heatmap shows the performance of tasks (x-axis) evaluated at the end of training each task in the sequence (y-axis). The consolidated structural representation in SYNERgy generalizes well across the tasks and significantly reduces forgetting.

that our method effectively employs synaptic consolidation to reduce forgetting by keeping the parameters closer to the optimal parameters of the previous tasks.

**Model Calibration** measures how indicative the probability of predictions are of their accuracies. DNNs tend to be poorly calibrated which reduces the reliability of their predictions. Combined with the bias towards recent tasks, this issue is even more pronounced in CL. Following Guo et al. (2017), we plot accuracy as a function of confidence (reliability plot) and calculate the weighted average over the absolute accuracy/confidence difference (Expected Calibration Error). Figure 3 shows that SYNERgy improves the calibration of model, hence the reliability of its predictions.

**Task Performances:** The average accuracy across the tasks alone fail to evaluate how well a CL method is able to maintain a balance between avoiding forgetting (stability) and allowing the model to learn a new task (plasticity). To this end, we evaluate the test set task-wise performance at the end of training of new tasks. The difference between the performance of the model when it was trained on the task and at the end of training shows the degree of forgetting whereas the performance of the model when it was trained shows how well the model is able to learn a new task. Figure 4 shows that SYNERgy provides a better balance between stability and plasticity compared to the baselines even with lower buffer size. While CLS-ER is able to reduce forgetting, the performance of the model on the current task suffers perhaps due to the lower rate of updating the stable model in the method. On Task 5, SYNERgy provides 89% accuracy compared to 78% in CLS-ER while retaining comparable performance on previous tasks. This suggests that synaptic consolidation in conjunction with dual memory experience replay provides a viable approach for tackling the stability and plasticity dilemma in CL.

**Task Probabilities:** The sequential learning of tasks in CL biases the model towards the current task (Wu et al., 2019) which affects performance on earlier tasks considerably. To test whether synaptic consolidation in SYNERgy reduces the bias towards recent tasks, we evaluate the probability of predicting each task at the end of training. We take the softmax output of each sample in the test set and average the probabilities of the classes in each task. Figure 5b shows that while ER and DER++ show considerable bias towards the last task when trained with low buffer size, the multiple memories in CLS-ER and SYNERgy provide more uniform prediction probabilities across the tasks. SYNERgy. CLS-ER reduces the probability of predicting the last task which affects the performance of the model on recent tasks. SYNERgy, on the other hand, leads to more uniform task probabilities without compromising on the recent task.

**Stability-Plasticity Trade-off:** To dissect how well SYNERgy tackles the stability-plasticity dilemma that lies at the core of CL, we take a closer look at the task-wise performance matrix, $\mathcal{T}$, where $\mathcal{T}_{i,j}$ corresponds to the test accuracy of task $j$ after learning task $i$. At a given task, $t$, in sequence, we consider *stability*, $\mathcal{S}$, as the average performance of all the previous $t-1$ tasks after learning task $t$, $mean(\mathcal{T}_{t,1:t-1})$. *Plasticity*, $\mathcal{P}$, is measured by the average performance of the tasks when they are first learned, $mean(Diag(\mathcal{T}))$. Hence, *stability* measures how well the model avoids forgetting and retains performance on the previous tasks whereas *plasticity* measures the capability of the model to learn the new task. Effective CL requires an optimal balance between stability and plasticity. However, there isn't a standardized way of measuring this trade-off. We, therefore, propose a trade-off measure that provides a formal way of comparing how well different methods maintain the balance. The *Trade-off* is measured as follow:

$$Trade\text{-}off = \frac{2 \times \mathcal{S} \times \mathcal{P}}{\mathcal{S} + \mathcal{P}} \tag{9}$$

Figure 5a compares the stability, plasticity and trade-off of the different models evaluated at the end of training all $T$ tasks in sequence. SYNERgy maintains a better trade-off compared to the baselines. CLS-ER provides higher stability

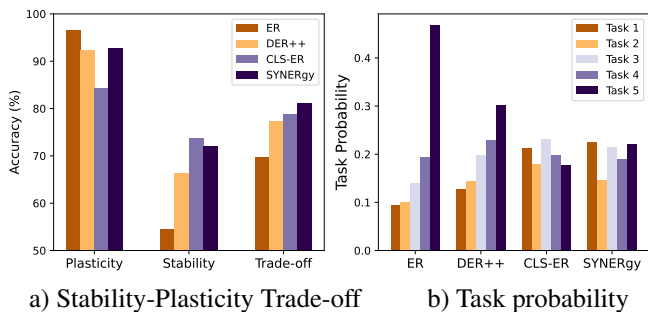a) Stability-Plasticity Trade-off     b) Task probability

Figure 5: a) SYNERgy provides a better trade-off between the stability and plasticity of the model. b) It also reduces the bias towards recent tasks and provides more uniform predictions across the tasks. See Figures S4 and S5 for other buffer sizes and datasets.

Table 2: **Ablation Study:** DM and TB denote the usage of dual memories and task boundary information for training, respectively. SC indicates which samples ($\mathcal{M}$ or $\mathcal{D}_t$) and model ($\theta_s$ or $\theta_w$) are utilized to evaluate FIM and whether it is adjusted $\mathcal{F}^{adj}$. All the components contribute to the performance gains.

| | DM | TB | SC | Accuracy |
|---|---|---|---|---|
| SYNERgy | ✓ | ✗ | $\mathcal{M}, \mathcal{F}^{adj}_{\theta_s}$ | $\mathbf{73.85}_{\pm 0.73}$ |
| $-\mathcal{F}_{adj}$ | ✓ | ✗ | $\mathcal{M}, \mathcal{F}^{*}_{\theta_s}$ | $72.88_{\pm 1.63}$ |
| $-\mathcal{M}, \theta_S$ | ✓ | ✗ | $\mathcal{D}_t, \mathcal{F}^{adj}_{\theta_w}$ | $72.26_{\pm 0.82}$ |
| Mean-ER | ✓ | ✗ | ✗ | $71.99_{\pm 1.26}$ |
| + oEWC | ✓ | ✓ | $\mathcal{D}_t, \mathcal{F}^{*}_{\theta_w}$ | $67.73_{\pm 1.37}$ |
| ER | ✗ | ✗ | ✗ | $57.74_{\pm 0.27}$ |

largely due to the lower frequency of updates for the stable model but it comes at the cost of plasticity. SYNERgy, on the other hand, not only provides comparable plasticity to DER++ but also considerably higher stability.

**Ablation Study:** To gain insights into the effect of the different components of SYNERgy, we conduct a series of ablation experiments. We systematically remove the components and show their effect on the performance of the models trained in Table 2. We can derive several insights from the results. Both synaptic consolidation and dual memory experience replay play a critical role in reducing forgetting. Mean-ER provides a significant performance improvement over typical ER. Our proposed Synaptic consolidation in SYNERgy provides further performance gain and improves the stability of the model. Furthermore, evaluating the Fisher information matrix using the semantic memory on the memory buffer instead of the working model on the current task training data provides a better estimate of the importance of parameters for all encountered tasks. It also shows that not accounting for the functional integrity of the convolution filter ($-\mathcal{F}_{adj}$) reduces performance. Our results suggest that more specialized algorithms for calculating the parameter importance at the filter level instead of post-processing may further improve regularization. Importantly, merely plugging in regularization in Mean-ER (+oEWC) not only adds the reliance on task boundaries but is also detrimental to learning. oEWC anchors the weights of the model to their state at the task boundary which over-constrains the model update and hampers knowledge consolidation. Furthermore, the importance estimates are biased towards the current task. Our proposed modifications, in addition to making the method suitable for general continual learning, enable knowledge consolidation by maintaining an updated estimate of parameter importance across the encountered tasks and anchoring the weights to the consolidated weights in the semantic memory. Also, our method does not require additional memory to save the reference weights. Our study, therefore, not only highlights the benefits of incorporating both experience replay and regularization in a dual-memory architecture but also emphasizes the importance of design choices to enable consolidation as the de-facto combination is detrimental to knowledge consolidation.

## 6 CONCLUSION

Inspired by the mechanisms employed by the brain to enable efficient consolidation of information and protection from the erasure of past memories, we proposed a novel dual memory replay mechanism in synergy with synaptic consolidation for general incremental learning. Our proposed method, SYNERgy, maintains an additional semantic memory that interacts with the episodic memory for effective semantic replay to consolidate information across the tasks. Moreover, synaptic consolidation is employed by tracking the importance of parameters throughout the learning trajectory and penalizing changes in parameters that are considered important for previous tasks. We proposed several modifications to make the regularization required for synaptic consolidation suitable for general incremental learning such that the method does not utilize the task boundaries at training and inference. We showed strong empirical results on various challenging CL settings and demonstrate the effectiveness of our method in providing a better balance between stability and plasticity of the model. Furthermore, SYNERgy improves the reliability of the model, reduces the drift in the parameters of the previous tasks and alleviates the bias towards recent tasks. Our results show the potential of employing synaptic consolidation in conjunction with experience replay in a dual memory system and motivates further work in this promising direction.

REFERENCES

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems*, pp. 11816–11825, 2019.

Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uxxFrDwrE7Y.

Marcus K Benna and Stefano Fusi. Computational principles of synaptic memory consolidation. *Nature neuroscience*, 19(12):1697–1706, 2016.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *arXiv preprint arXiv:2004.07211*, 2020.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 532–547, 2018a.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018b.

Joseph Cichon and Wen-Biao Gan. Branch-specific dendritic ca 2+ spikes cause persistent synaptic plasticity. *Nature*, 520(7546):180–185, 2015.

Claudia Clopath. Synaptic consolidation: an approach to long-term learning. *Cognitive neurodynamics*, 6(3):251–257, 2012.

Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Tyler L Hayes, Giri P Krishnan, Maxim Bazhenov, Hava T Siegelmann, Terrence J Sejnowski, and Christopher Kanan. Replay in deep learning: Current approaches and missing biological elements. *Neural Computation*, 33(11):2908–2950, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Giri P Krishnan, Timothy Tadros, Ramyaa Ramyaa, and Maxim Bazhenov. Biologically inspired sleep algorithm for artificial neural networks. *arXiv preprint arXiv:1908.02240*, 2019.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in neural information processing systems*, pp. 6467–6476, 2017.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Fei Mi, Lingjing Kong, Tao Lin, Kaicheng Yu, and Boi Faltings. Generalized class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 240–241, 2020.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Quang Pham, Chenghao Liu, and Steven Hoi. Dualnet: Continual learning, fast and slow. *Advances in Neural Information Processing Systems*, 34, 2021.

Hadi Pouransari and Saman Ghili. Tiny imagenet visual recognition challenge. *CS231N course, Stanford Univ., Stanford, CA, USA*, 2015.

Roger L Redondo and Richard GM Morris. Making memories last: the synaptic tagging and capture hypothesis. *Nature Reviews Neuroscience*, 12(1):17–30, 2011.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.

Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, pp. 3738–3748, 2018.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pp. 4528–4537. PMLR, 2018.

Endel Tulving. Episodic memory: From mind to brain. *Annual review of psychology*, 53(1):1–25, 2002.

Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57, 1985.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. *arXiv preprint arXiv:2204.04799*, 2022.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 374–382, 2019.

Guang Yang, Feng Pan, and Wen-Biao Gan. Stably maintained dendritic spines are associated with lifelong memories. *Nature*, 462(7275):920–924, 2009.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

## A    EVALUATION PROTOCOLS

To gauge the effectiveness and versatility of our method, we test it under varying CL scenarios, each of which presents its own set of challenges that the CL method needs to tackle. In this study, we consider the following CL scenarios.

**Class-IL** involves learning a new set of disjoint classes in subsequent tasks and the agent is required to distinguish among all the classes seen so far. Class-IL measures the ability of the agent to learn izable features, which requires consolidating acquired knowledge with new tasks to optimally adapt the representation space and decision boundaries. We consider CIFAR10 (Krizhevsky et al., 2009) and TinyImageNet (Pouransari & Ghili, 2015) split into 5 and 10 tasks, respectively, and hence cover settings with varying datasets and task length complexities.

**Domain-IL** evaluates the ability of the model to tackle changes in the input data distribution with the number of classes remaining fixed. It measures the ability of the model to learn features that are robust to the distribution shift. We consider R-MNIST (Lopez-Paz & Ranzato, 2017) which requires the agent to classify MNIST digits for 20 subsequent tasks with images rotated by a random angle between 0 and 180 degrees.

**General-IL** attempts to simulate the challenges in the real-world and involves training the method on longer sequences where the boundaries between the tasks are not distinct, the tasks themselves are not disjoint, and there are multiple sources of data distributions shifts (Arani et al., 2022). In addition to preventing catastrophic forgetting, the learning agent has to tackle several challenges: class imbalance, sample efficiency, learning an object over multiple recurrences and multiple distribution shifts. We consider two General-IL settings which present different sets of challenges. MNIST360 (Buzzega et al., 2020) exposes the model to both a sharp distribution shift in classes and a smooth rotational distribution shift by presenting batches of two consecutive MNIST images (e.g. $\{0, 1\}$, $\{1, 2\}$) and the samples are rotated at an increasing angle. GCIL (Mi et al., 2020) utilizes probabilistic modeling to sample the classes and data distributions from CIFAR100 dataset (Krizhevsky et al., 2009) in each task. Hence, the number of classes in each task is not fixed, the classes can overlap and the sample size for each class can vary. GCIL contains a sequence of 20 tasks where each task has 1000 samples with a maximum of 50 classes and the sample distribution is either uniform (U) or longtail (L).

## B    WORKING MODEL PERFORMANCE

Here, we provide the performance of the working model on each of the evaluation setting and compare with the semantic memory. Table S1 indicates that the semantic memory is able to effectively aggregate knowledge from the working model and build consolidated features for generalization across the tasks. Particularly, for the challenging data and memory-restrictive settings, the performance gap between the semantic memory and working model is considerable. Figure S1 further compares the task-wise performance and show that the semantic memory generalizes well across the tasks whereas the working model is more optimized for the recent task. These results highlight the effectiveness of maintaining semantic memory using the exponential moving average over the weights of the working model.

Table S1: Performance of the working model and semantic memory on the different evaluation settings. Semantic memory provides better generalization across all the considered setting especially in challenging dataset and memory restrictive scenarios.

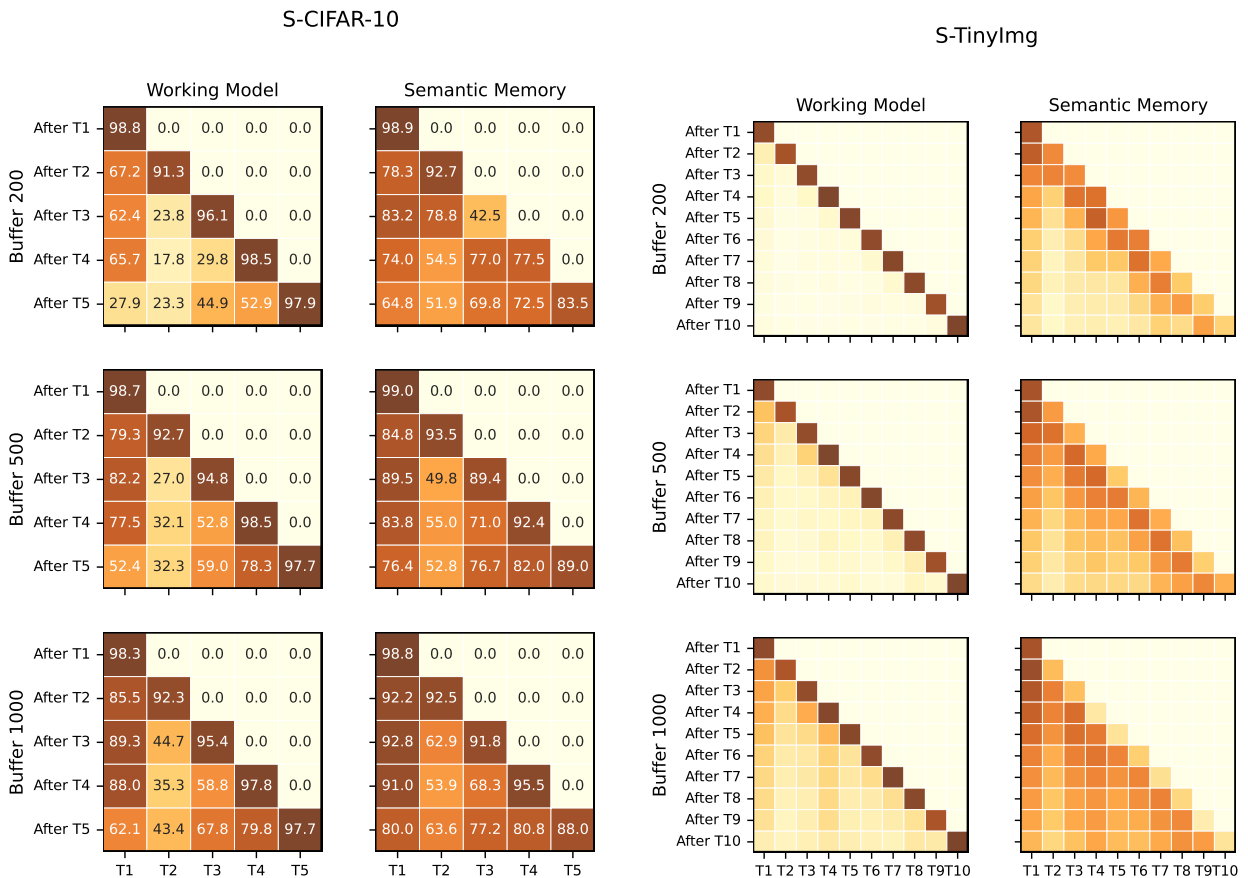| Buffer size | Dataset | Semantic memory | Working model | Dataset | Semantic memory | Working model |
|---|---|---|---|---|---|---|
| 200 | | $66.27_{\pm1.4}$ | $51.13_{\pm1.34}$ | | $65.23_{\pm0.52}$ | $59.84_{\pm0.87}$ |
| 500 | S-CIFAR-10 | $73.85_{\pm0.73}$ | $63.80_{\pm1.82}$ | MNIST-360 | $76.48_{\pm0.83}$ | $73.49_{\pm0.99}$ |
| 1000 | | $77.47_{\pm1.15}$ | $70.70_{\pm1.24}$ | | $80.96_{\pm0.33}$ | $78.50_{\pm0.89}$ |
| 200 | | $21.68_{\pm1.51}$ | $10.46_{\pm0.33}$ | | $30.00_{\pm2.63}$ | $25.27_{\pm2.82}$ |
| 500 | S-TinyImg | $31.26_{\pm0.67}$ | $14.98_{\pm0.37}$ | GCIL-U | $38.45_{\pm0.47}$ | $36.72_{\pm0.70}$ |
| 1000 | | $36.49_{\pm0.58}$ | $22.44_{\pm0.57}$ | | $41.08_{\pm059}$ | $40.27_{\pm0.69}$ |
| 200 | | $92.02_{\pm0.83}$ | $91.74_{\pm0.83}$ | | $32.24_{\pm2.14}$ | $27.84_{\pm1.17}$ |
| 500 | R-MNIST | $93.37_{\pm0.45}$ | $93.18_{\pm0.49}$ | GCIL-L | $33.07_{\pm2.28}$ | $29.60_{\pm1.14}$ |
| 1000 | | $93.63_{\pm0.49}$ | $93.51_{\pm0.48}$ | | $39.96_{\pm1.59}$ | $39.77_{\pm1.46}$ |

Figure S1: Task-wise performance of working model and semantic memory evaluated on the test set of S-CIFAR10 and TinyImageNet datasets. Heatmap shows the performance of tasks (x-axis) evaluated at the end of training each task in the sequence (y-axis). Semantic memory buils consolidated structural representation which generalizes well across the tasks and is therefore suitable for inference.

## C    COMPARISON WITH CLS-ER

CLS-ER provides us with an effective framework for studying the effect of synaptic consolidation in a multiple memory replay framework. We built our approach on top of their Mean-ER method instead of CLS-ER (which is a much stronger baseline) because of it's simplicity and parallels with the CLS theory in the brain. While the multiple semantic memories in CLS-ER provides generalization gain, it also introduces additional complexities and makes it more challenging to study the effect of synaptic consolidation.

The goal of our study was not to outperform the existing approaches but rather to investigate the complementary nature of synaptic consolidation and dual memory experience replay similar to the brain. Interestingly, we found that incorporating synaptic consolidation with even one semantic memory (as in Mean-ER) enables SYNERgy to perform comparable to CLS-ER (with two semantic memories) in simpler settings and provides considerable gains in the more realistic and challenging GCIL-CIFAR-100 longtail and uniform settings. Additionally, our stability plasticity analysis shows that SYNERgy maintains a better trade-off compared to CLS-ER. CLS-ER provides higher stability largely due to the lower frequency of updates for the stable model but it comes at a considerable cost of plasticity whereas SYNERgy does not compromise the plasticity of the model. Furthermore, SYNERgy requires less computations as it involves one additional forward pass for the memory samples instead of two in CLS-ER and for the complex settings the frequency of fisher information matrix update is quite low .

We believe that SC can further improve the performance of CLS-ER and provides an interesting research avenue for employing SC in a multiple semantic memories setup.

## D CHARACTERISTICS ANALYSIS

To further study the characteristics of the model, we compare the task probabilities and calibration of the models trained on different datasets with varying memory budgets.

### D.1 MODEL CALIBRATION

Figures S2 and S3 provides the reliability plots and ECE of the different models trained on S-CIFAR-10 and S-TinyImageNet respectively with varying memory buffer sizes. CLS-ER and SYNERgy have significantly lower calibration error compared to ER and DER, particularly with lower buffer sizes on S-TinyImageNet. On TinyImageNet with buffer size 200 and 500, CLS-ER provides lower calibration whereas for all the other settings, SYNERgy provides lower calibration error.
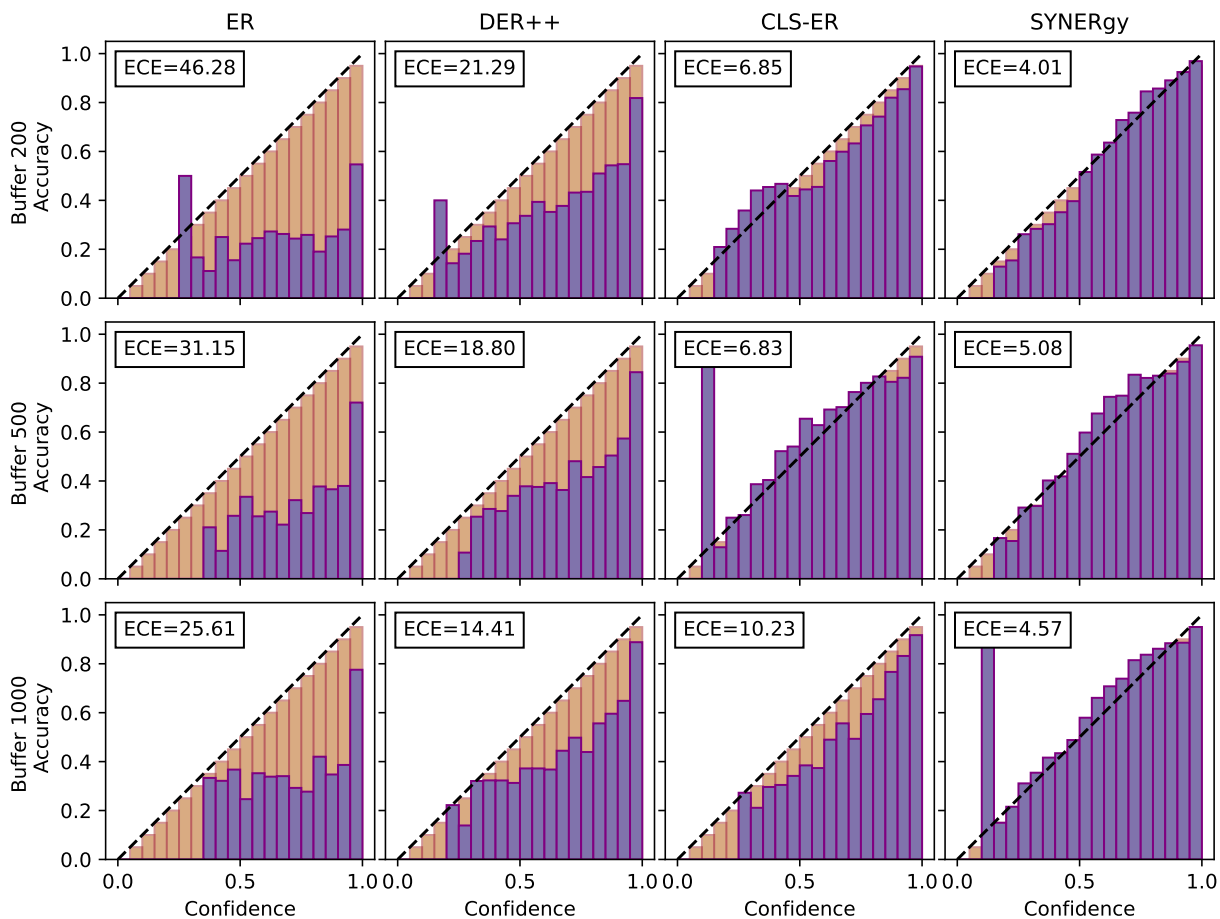


Figure S2: Reliability plots and ECE on S-CIFAR-10 with varying memory budget.

### D.2 TASK PROBABILITIES

Figure S4 and S5 provides the task probabilities of the different models trained on S-CIFAR-10 and S-TinyImageNet respectively with varying memory buffer sizes. SYNERgy reduces the bias towards the recent task and provides more uniform probability of prediction across the tasks. On S-TinyImageNet with 500 buffer size, CLS-ER has a much lower probability for the last task which is also not desirable and significantly reduces the performance of the model on the recent task.
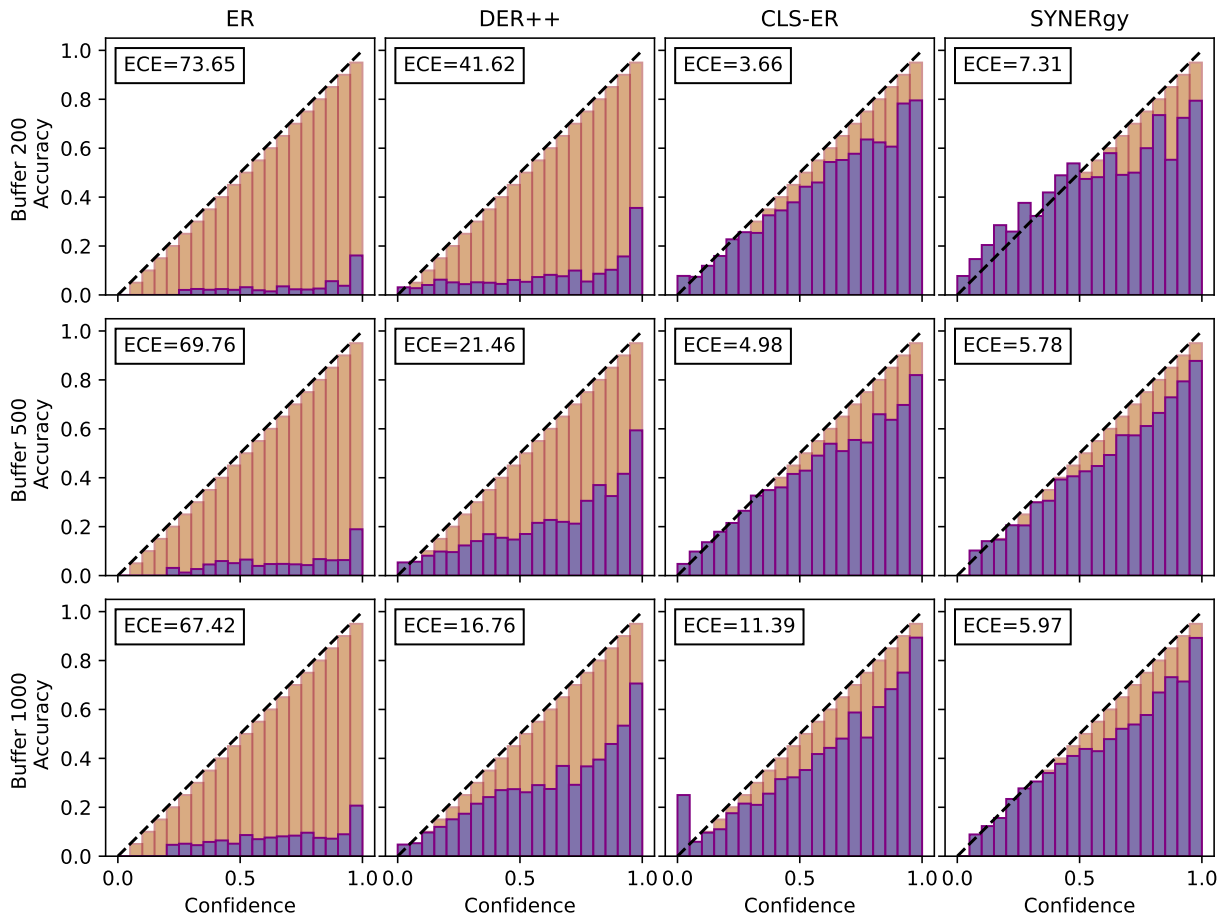
Figure S3: Reliability plots and ECE on S-TinyImageNet with varying memory budget.
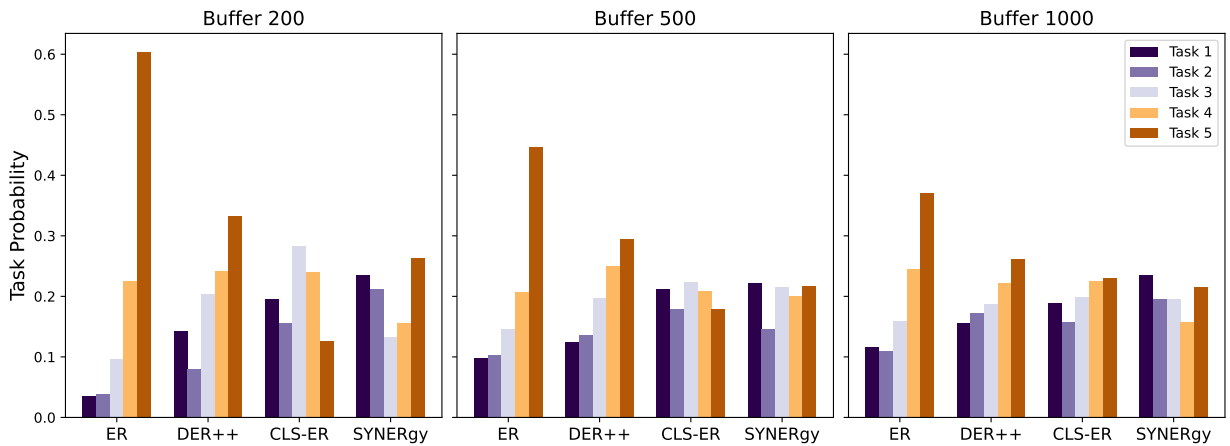


Figure S4: Task probabilities on S-CIFAR-10: SYNERgy reduces the bias towards recent task across the different memory budget.
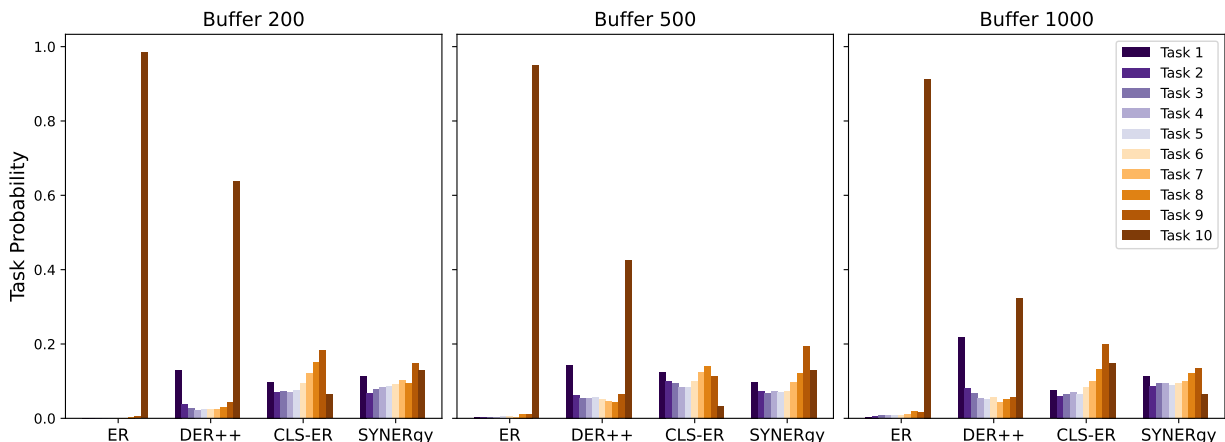
Figure S5: Task probabilities on S-TinyImageNet: SYNERgy reduces the bias towards recent task across the different memory budget.

## E    EXPERIMENTAL SETUP AND TRAINING DETAILS

To compare our method with different CL methods under common experimental settings and evaluation framework, we extended the Mammoth framework (Buzzega et al., 2020). For each of our evaluation settings, we use the same training scheme (SGD optimizer, input stream batch size, memory buffer batch size, and the number of training epochs) and find the optimal hyperparameters for our method using a grid search over $\lambda$, $\beta$, $\alpha_S$, $\alpha_F$, $r_S$, and $r_F$ on a small validation set. The selected hyperparameters for each setting are provided in Table S2. Note that for a given dataset, the selected hyperparameters do not vary much. We use the same backbone for each dataset as in Buzzega et al. (2020): a fully connected network with two hidden layers, each with 100 ReLU units on MNIST dataset variants and ResNet-18 (He et al., 2015) for the other datasets. We apply random crop and random horizontal flip on the input stream and buffer samples for S-CIFAR10, S-TinyImageNet, and GCIL-CIFAR100. While also being interested in the online CL setting, similar to Buzzega et al. (2020); Arani et al. (2022), we take into account the dataset complexity to disentangle the effect of underfitting from forgetting. We use a single epoch for MNIST-based settings whereas, for the other complex settings, we increase the number of epochs to 50 epochs. Note that CLS-ER and DER++ use 100 epochs on GCIL Arani et al. (2022). For all of our experiments, we report the mean and one standard deviation of the average task accuracy of 10 differently initialized runs.

### E.1    GCIL-CIFAR100

Following Arani et al. (2022), we consider the uniform and longtail GCIL-CIFAR-100 setting with a sequence of 20 tasks, each with 1000 samples and a maximum of 50 classes. Because of the probabilistic sampling, different dataset seeds results in vastly different tasks. For reproducibility and fair comparison, we set the GIL-seed to 1993 for all our experiments and run with 10 differently initialized models to measure the variance of the different models on a fixed setting. We train the baselines (ER, DER++ and CLS-ER) with the hyper-parameters provided in Arani et al. (2022).

### E.2    1000 BUFFER SIZE

We consider that the 5120 buffer size used in Arani et al. (2022); Buzzega et al. (2020) for S-Cifar-10, S-TinyImageNet and Rot-MNIST to be too large for these settings to distinguish the effectiveness of the CL method. Also, there is a huge jump form the modest 500 buffer size and we believe that the performance of the model on 1000 buffer can provide more insights for the research community. We use the same hyperparameters for 1000 buffer size as reported in Arani et al. (2022); Buzzega et al. (2020) for buffer size 5120.

## F    EFFECT OF HYPERPARAMETERS

SYNERgy involves involves the interplay between the semantic memory and the episodic memory to enforce a consistency loss while also anchoring the working model's weights to the important parameters of the semantic memory.

Table S2: The selected hyperparameters for SYNERgy in each of the experimental settings.

| Dataset | Buffer size | batch_size | #epoch | $\eta$ | $\lambda$ | $\beta$ | $\alpha_s$ | $r_s$ | $\alpha_F$ | $r_F$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 200 | 32 | 50 | 0.05 | 0.2 | 5 | 0.999 | 0.2 | 0.999 | 0.01 |
| S-CIFAR-10 | 500 | 32 | 50 | 0.05 | 0.2 | 5 | 0.999 | 0.4 | 0.999 | 0.007 |
| | 1000 | 32 | 50 | 0.05 | 0.2 | 5 | 0.999 | 0.9 | 0.999 | 0.001 |
| | 200 | 32 | 50 | 0.05 | 0.1 | 0.1 | 0.999 | 0.05 | 0.999 | 0.0005 |
| S-TinyImg | 500 | 32 | 50 | 0.05 | 0.1 | 0.1 | 0.999 | 0.05 | 0.999 | 0.0004 |
| | 1000 | 32 | 50 | 0.05 | 0.1 | 0.1 | 0.999 | 0.05 | 0.999 | 0.0004 |
| | 200 | 128 | 1 | 0.2 | 1 | 1 | 0.99 | 1 | 0.99 | 0.4 |
| R-MNIST | 500 | 128 | 1 | 0.2 | 1 | 1 | 0.99 | 1 | 0.99 | 0.8 |
| | 1000 | 128 | 1 | 0.2 | 1 | 1 | 0.99 | 1 | 0.99 | 0.4 |
| | 200 | 16 | 1 | 0.2 | 1 | 1 | 0.99 | 0.8 | 0.99 | 0.9 |
| MNIST-360 | 500 | 16 | 1 | 0.2 | 1 | 1 | 0.99 | 0.8 | 0.99 | 0.8 |
| | 1000 | 16 | 1 | 0.2 | 0.8 | 1 | 0.99 | 0.8 | 0.99 | 0.8 |
| | 200 | 32 | 50 | 0.05 | 0.2 | 1 | 0.999 | 0.3 | 0.999 | 0.005 |
| GCIL-U | 500 | 32 | 50 | 0.05 | 0.2 | 1 | 0.999 | 0.3 | 0.999 | 0.005 |
| | 1000 | 32 | 50 | 0.05 | 0.2 | 1 | 0.999 | 0.4 | 0.999 | 0.005 |
| | 200 | 32 | 50 | 0.05 | 0.2 | 1 | 0.999 | 0.2 | 0.999 | 0.005 |
| GCIL-L | 500 | 32 | 50 | 0.05 | 0.2 | 1 | 0.999 | 0.2 | 0.999 | 0.005 |
| | 1000 | 32 | 50 | 0.05 | 0.2 | 1 | 0.999 | 0.4 | 0.999 | 0.005 |

Given the nature of the interactions and the formulation, the different components of the method are complementary. Table S3 shows that our hyperparameters are complementary in nature, allowing us to keep some parameters fixed and only fine tune the other ones to facilitate hyperparamter tuning. The method is not highly sensitive to a specific set of hyperparameters for each setting as a number of settings provide similar performance.

Table S3: Effect of hyperparameters on SYNERgy performance (both semantic model and working model) with 500 buffer_size, $\eta = 0.05$, and $\lambda = 0.2$ on GCIL-L. The values are averaged over 3 runs.

| $\beta$ | $r_s$ | $r_F$ | Semantic memory | Working model | $\beta$ | $r_s$ | $r_F$ | Semantic memory | Working model |
|---|---|---|---|---|---|---|---|---|---|
| | | 0.001 | $37.25_{\pm0.50}$ | $32.94_{\pm0.55}$ | | | 0.001 | $36.50_{\pm0.40}$ | $32.88_{\pm0.24}$ |
| | 0.2 | 0.005 | $35.09_{\pm0.21}$ | $32.56_{\pm0.65}$ | | 0.2 | 0.005 | $33.62_{\pm2.53}$ | $30.08_{\pm1.62}$ |
| | | 0.010 | $36.12_{\pm0.17}$ | $32.06_{\pm0.82}$ | | | 0.010 | $35.98_{\pm0.84}$ | $31.71_{\pm0.90}$ |
| | | 0.001 | $32.55_{\pm4.83}$ | $32.44_{\pm0.70}$ | | | 0.001 | $35.47_{\pm1.06}$ | $33.05_{\pm0.40}$ |
| 0.5 | 0.4 | 0.005 | $34.11_{\pm2.59}$ | $32.04_{\pm1.02}$ | 1 | 0.4 | 0.005 | $31.91_{\pm1.82}$ | $29.72_{\pm1.74}$ |
| | | 0.010 | $35.02_{\pm0.67}$ | $32.11_{\pm0.33}$ | | | 0.010 | $35.32_{\pm1.29}$ | $32.83_{\pm1.07}$ |
| | | 0.001 | $33.92_{\pm0.92}$ | $31.95_{\pm0.66}$ | | | 0.001 | $33.37_{\pm1.03}$ | $31.21_{\pm1.01}$ |
| | 0.6 | 0.005 | $33.76_{\pm1.37}$ | $31.66_{\pm1.05}$ | | 0.6 | 0.005 | $33.67_{\pm0.76}$ | $31.46_{\pm0.88}$ |
| | | 0.010 | $33.02_{\pm1.33}$ | $31.61_{\pm0.40}$ | | | 0.010 | $33.66_{\pm0.37}$ | $31.97_{\pm0.39}$ |