On the Convergence of Continual Learning with Adaptive Methods (Supplementary Material)

Seungyub Han[1] **Yeongmo Kim**[1] **Taehyun Cho**[1] **Jungwoo Lee**[1]

[1]Electrical and Computer Engineering Dept., Seoul National University, Seoul, Republic of Korea

A ADDITIONAL BACKGROUNDS AND EXTENDED DISCUSSION

A.1 SUMMARY OF NOTATIONS

Notations	Definitions	Notations	Definitions
x	model parameter	H_t	the union of I_t and J_t
$\mathcal{P}$	previous task	n_f	the number of data points in P
$\mathcal{C}$	current task	n_g	the number of data points in C
P	dataset of $\mathcal{P}$	$\langle \cdot, \cdot \rangle$	inner product
C	dataset of $\mathcal{C}$	L	L-smoothness constant
$h(x)$	mean loss of x on entire datasets	α_{H_t}	adaptive step size for f with H_t
$f(x)$	mean loss of x on P	β_{H_t}	adaptive step size for g with H_t
$g(x)$	mean loss of x on C	M_t	memory at time t
$f_i(x)$	loss of x on a data point $i \in P$	e_t	error of estimate f at time t
$g_j(x)$	loss of x on a data point $j \in C$	e_{M_t}	error of estimate f with M_t
$f_{I_t}(x)$	mini-batch loss of x on a batch I_t	f_{M_t}	mean loss of x with M_t
$g_{J_t}(x)$	mini-batch loss of x on a batch J_t	$M_{[t1:t2]}$	the history of memory from $t1$ to $t2$
I_t	minibatch sampled from P	B_t	memory bias term at t
J_t	minibatch sampled from C	Γ_t	forgetting term at t
$\mathbb{E}_t$	total expectation from 0 to time t	Λ_{H_t}	inner product between ∇f_{I_t} and ∇g_{J_t}

A.2 REVIEW OF TERMINOLOGY

(Restriction of f) If $f : A \to B$ and if A_0 is a subset of A, then the **restriction of f to A_0** is the function

$$f|_{A_0} : A_0 \to B$$

given by $f|_{A_0}(x) = f(x)$ for $x \in A_0$.

A.3 ADDITIONAL RELATED WORK

Regularization based methods. EWC has an additional penalization loss that prevent the update of parameters from losing the information of previous tasks. When we update a model with EWC, we have two gradient components from the current task and the penalization loss.

task-specific model components. SupSup learns a separate subnetwork for each task to predict a given data by superimposing all supermasks. It is a novel method to solve catastrophic forgetting with taking advantage of neural networks.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

SGD methods without expereince replay. stable SGD [Mirzadeh et al., 2020] and MC-SGD [Jin et al., 2021] show overall higher performance in terms of average accuracy than the proposed algorithm. For average forgetting, our method has the lowest value, which means that NCCL prevents catastrophic forgetting successfully with achieving the reasonable performance on the current task. We think that our method is focused on reducing catastrophic forgetting as we defined in the reformulated continual learning problem (12), so our method shows the better performance on average forgetting. Otherwise, MC-SGD finds a low-loss paths with mode-connectivity by updating with the proposed regularization loss. This procedure implies that a continual learning model might find a better local minimum point for the new (current) task than NCCL.

For non-memory based methods, the theoretical measure to observe forgetting and convergence during training does not exist. Our theoretical results are the first attempt to analyze the convergence of previous tasks during continual learning procedure. In future work, we can approximate the value of with fisher information for EWC and introduce Bayesian deep learning to analyze the convergence of each subnetworks for each task in the case of SupSup [Wortsman et al., 2020].

B ADDITIONAL EXPERIMENTAL RESULTS AND IMPLEMENTATION DETAILS

We implement the baselines and the proposed method on Tensorflow 1. For evaluation, we use an NVIDIA 2080ti GPU along with 3.60 GHz Intel i9-9900K CPU and 64 GB RAM.

B.1 ARCHITECTURE AND TRAINING DETAIL

For fair comparison, we follow the commonly used model architecture and hyperparameters of [Lee et al., 2020, Chaudhry et al., 2020]. For Permuted-MNIST and Split-MNIST, we use fully-connected neural networks with two hidden layers of $[400, 400]$ or $[256, 256]$ and ReLU activation. ResNet-18 with the number of filters $n_f = 64, 20$ [He et al., 2016] is applied for Split CIFAR-10 and 100. All experiments conduct a single-pass over the data stream. It is also called 1 epoch or 0.2 epoch (in the case of split tasks). We deal both cases with and without the task identifiers in the results of split-tasks to compare fairly with baselines. Batch sizes of data stream and memory are both 10. All reported values are the average values of 5 runs with diffrent seeds, and we also provide standard deviation. Other miscellaneous settings are the same as in [Chaudhry et al., 2020].

B.2 HYPERPARAMETER GRIDS

We report the hyper-paramters grid we used in our experiments below. Except for the proposed algorithm, we adopted the hyper-paramters that are reported in the original papers. We used grid search to find the optimal parameters for each model.

- finetune - learning rate [0.003, 0.01, 0.03 (CIFAR), 0.1 (MNIST), 0.3, 1.0]
- EWC - learning rate: [0.003, 0.01, 0.03 (CIFAR), 0.1 (MNIST), 0.3, 1.0] - regularization: [0.1, 1, 10 (MNIST,CIFAR), 100, 1000]
- A-GEM - learning rate: [0.003, 0.01, 0.03 (CIFAR), 0.1 (MNIST), 0.3, 1.0]
- ER-Ring - learning rate: [0.003, 0.01, 0.03 (CIFAR), 0.1 (MNIST), 0.3, 1.0]
- ORTHOG-SUBSPACE - learning rate: [0.003, 0.01, 0.03, 0.1 (MNIST), 0.2, 0.4 (CIFAR), 1.0]
- MER - learning rate: [0.003, 0.01, 0.03 (MNIST, CIFAR), 0.1, 0.3, 1.0] - within batch meta-learning rate: [0.01, 0.03, 0.1 (MNIST, CIFAR), 0.3, 1.0] - current batch learning rate multiplier: [1, 2, 5 (CIFAR), 10 (MNIST)]
- iid-offline and iid-online - learning rate [0.003, 0.01, 0.03 (CIFAR), 0.1 (MNIST), 0.3, 1.0]
- ER-Reservoir - learning rate: [0.003, 0.01, 0.03, 0.1 (MNIST, CIFAR), 0.3, 1.0]
- NCCL-Ring (default) - learning rate α: [0.003, 0.001(CIFAR), 0.01, 0.03, 0.1, 0.3, 1.0]
- NCCL-Reservoir - learning rate α: [0.003(CIFAR), 0.001, 0.01, 0.03, 0.1, 0.3, 1.0]

B.3 HYPERPARAMETER SEARCH ON β_{max} AND TRAINING TIME

Table 1: Permuted-MNIST (23 tasks 10000 examples per task), FC-[256,256] and Multi-headed split-CIFAR100, full size Resnet-18. Accuracies with different clipping rate on NCCL + Ring.

β_{max}	**Permuted-MNIST**	**Split-CIFAR100**
0.001	72.52(0.59)	49.43(0.65)
0.01	72.93(1.38)	56.95(1.02)
0.05	72.18(0.77)	56.35(1.42)
0.1	72.29(1.34)	58.20(0.155)
0.2	74.38(0.89)	57.60(0.36)
0.5	72.95(0.50)	59.06(1.02)
1	72.92(1.07)	57.43(1.33)
5	72.31(1.79)	57.75(0.24)

Table 2: Permuted-MNIST (23 tasks 10000 examples per task), FC-[256,256] and Multi-headed split-CIFAR100, full size Resnet-18. Training time.

Methods	**Training time [s]**	
	Permuted-MNIST	**Split-CIFAR100**
fine-tune	91	92
EWC	95	159
A-GEM	180	760
ER-Ring	109	129
ER-Reservoir	95	113
ORTHOG-SUBSPACE	90	581
NCCL+Ring	167	248
NCCL+Reservoir	168	242

B.4 ADDITIONAL EXPERIMENT RESULTS

Table 3: Permuted-MNIST (23 tasks 60000 examples per task), FC-[256,256].

Method	**memory size**	**1**		**5**	
	memory	accuracy	forgetting	accuracy	forgetting
multi-task	✗	83	-	83	-
Fine-tune	✗	53.5 (1.46)	0.29 (0.01)	47.9	0.29 (0.01)
EWC	✗	63.1 (1.40)	0.18 (0.01)	63.1 (1.40)	0.18 (0.01)
stable SGD	✗	80.1 (0.51)	0.09 (0.01)	80.1 (0.51)	0.09 (0.01)
MC-SGD	✗	85.3 (0.61)	0.06 (0.01)	85.3 (0.61)	0.06 (0.01)
MER	✓	69.9 (0.40)	0.14 (0.01)	78.3 (0.19)	0.06 (0.01)
A-GEM	✓	62.1 (1.39)	0.21 (0.01)	64.1 (0.74)	0.19 (0.01)
ER-Ring	✓	70.2 (0.56)	0.12 (0.01)	75.8 (0.24)	0.07 (0.01)
ER-Reservoir	✓	68.9 (0.89)	0.15 (0.01)	76.2 (0.38)	0.07 (0.01)
ORHOG-subspace	✓	84.32 (1.10)	0.12 (0.01)	84.32 (1.1)	0.11 (0.01)
NCCL + Ring	✓	74.22 (0.75)	0.13 (0.007)	84.41 (0.32)	0.053 (0.002)
NCCL+Reservoir	✓	79.36 (0.73)	**0.12 (0.007)**	**88.22 (0.26)**	**0.028 (0.003)**

Table 4: Multi-headed split-CIFAR100, reduced size Resnet-18 $n_f = 20$.

Method	**memory size**	**1**		**5**	
	memory	accuracy	forgetting	accuracy	forgetting
EWC	✗	42.7 (1.89)	0.28 (0.03)	42.7 (1.89)	0.28 (0.03)
Fintune	✗	40.4 (2.83)	0.31 (0.02)	40.4 (2.83)	0.31 (0.02)
Stable SGD	✗	59.9 (1.81)	0.08 (0.01)	59.9 (1.81)	0.08 (0.01)
MC-SGD	✗	63.3 (2.21)	0.06 (0.03)	63.3 (2.21)	0.06 (0.03)
A-GEM	✓	50.7 (2.32)	0.19 (0.04)	59.9 (2.64)	0.10 (0.02)
ER-Ring	✓	56.2 (1.93)	0.13 (0.01)	62.6 (1.77)	0.08 (0.02)
ER-Reservoir	✓	46.9 (0.76)	0.21 (0.03)	65.5 (1.99)	0.09 (0.02)
ORTHOG-subspace	✓	58.81 (1.88)	0.12 (0.02)	64.38 (0.95)	0.055 (0.007)
NCCL + Ring	✓	54.63 (0.65)	**0.059 (0.01)**	61.09 (1.47)	**0.02 (0.01)**
NCCL + Reservoir	✓	52.18 (0.48)	0.118 (0.01)	63.68 (0.18)	0.028 (0.009)

Table 5: Multi-headed split-MiniImagenet, full size Resnet-18 $n_f = 64$. Accuracy and forgetting results.

Method	memory size	1	
	memory	accuracy	forgetting
Fintune	✗	36.1(1.31)	0.24(0.03)
EWC	✗	34.8(2.34)	0.24(0.04)
A-GEM	✓	42.3(1.42)	0.17(0.01)
MER	✓	45.5(1.49)	0.15(0.01)
ER-Ring	✓	49.8(2.92)	0.12(0.01)
ER-Reservoir	✓	44.4(3.22)	0.17(0.02)
ORTHOG-subspace	✓	51.4(1.44)	0.10(0.01)
NCCL + Ring	✓	45.5(0.245)	**0.041(0.01)**
NCCL + Reservoir	✓	41.0(1.02)	**0.09(0.01)**

Table 6: Multi-headed split-CIFAR100, full size Resnet-18 $n_f = 64$. Accuracy and forgetting results.

Method	**memory size**	**1**		**5**	
	memory	accuracy	forgetting	accuracy	forgetting
Fintune	✗	42.6 (2.72)	0.27 (0.02)	42.6 (2.72)	0.27 (0.02)
EWC	✗	43.2 (2.77)	0.26 (0.02)	43.2 (2.77)	0.26 (0.02)
ICRAL	✓	46.4 (1.21)	0.16 (0.01)	-	-
A-GEM	✓	51.3 (3.49)	0.18 (0.03)	60.9 (2.5)	0.11 (0.01)
MER	✓	49.7 (2.97)	0.19 (0.03)	-	-
ER-Ring	✓	59.6 (1.19)	0.14 (0.01)	67.2 (1.72)	0.06 (0.01)
ER-Reservoir	✓	51.5 (2.15)	0.14 (0.09)	62.68 (0.91)	0.06 (0.01)
ORTHOG-subspace	✓	64.3 (0.59)	0.07 (0.01)	67.3 (0.98)	0.05 (0.01)
NCCL + Ring	✓	59.06 (1.02)	0.03 (0.02)	66.58 (0.12)	0.004 (0.003)
NCCL + Reservoir	✓	54.7 (0.91)	0.083 (0.01)	66.37 (0.19)	0.004 (0.001)

Table 7: permuted-MNIST (23 tasks 10000 examples per task), FC-[256,256]. Accuracy and forgetting results.

Method	**memory size**	**1**		**5**	
	memory	accuracy	forgetting	accuracy	forgetting
multi-task	✗	91.3	-	83	-
Fine-tune	✗	50.6 (2.57)	0.29 (0.01)	47.9	0.29 (0.01)
EWC	✗	68.4 (0.76)	0.18 (0.01)	63.1 (1.40)	0.18 (0.01)
MER	✓	78.6 (0.84)	0.15 (0.01)	88.34 (0.26)	0.049 (0.003)
A-GEM	✓	78.3 (0.42)	0.21 (0.01)	64.1 (0.74)	0.19 (0.01)
ER-Ring	✓	79.5 (0.31)	0.12 (0.01)	75.8 (0.24)	0.07 (0.01)
ER-Reservoir	✓	68.9 (0.89)	0.15 (0.01)	76.2 (0.38)	0.07 (0.01)
ORHOG-subspace	✓	86.6 (0.91)	0.04 (0.01)	87.04 (0.43)	0.04 (0.003)
NCCL + Ring	✓	74.38 (0.89)	0.05 (0.009)	83.76 (0.21)	0.014 (0.001)
NCCL+Reservoir	✓	76.48 (0.29)	0.1 (0.002)	86.02 (0.06)	0.013 (0.002)

Table 8: Single-headed split-MNIST, FC-[256,256]. Accuracy and forgetting results.

Method	**memory size**	**1**		**5**		**50**	
	memory	accuracy	forgetting	accuracy	forgetting	accuracy	forgetting
multi-task	✗	95.2	-	-	-	-	-
Fine-tune	✗	52.52 (5.24)	0.41 (0.06)	-	-	-	-
EWC	✗	56.48 (6.46)	0.31 (0.05)	-	-	-	-
A-GEM	✓	34.04 (7.10)	0.23 (0.11)	33.57 (6.32)	0.18 (0.03)	33.35 (4.52)	0.12 (0.04)
ER-Reservoir	✓	34.63 (6.03)	0.79 (0.07)	63.60 (3.11)	0.42 (0.05)	86.17 (0.99)	0.13 (0.016)
NCCL + Ring	✓	34.64 (3.27)	0.55 (0.03)	61.02 (6.21)	0.207 (0.07)	81.35 (8.24)	-0.03 (0.1)
NCCL+Reservoir	✓	37.02 (0.34)	0.509 (0.009)	65.4 (0.7)	0.16 (0.006)	88.9 (0.28)	-0.125 (0.004)

Table 9: Single-headed split-MNIST, FC-[400,400] and mem. size=500(50 / cls.). Accuracy and forgetting results.

Method	**accuracy**
multi-task	96.18
Fine-tune	50.9 (5.53)
EWC	55.40 (6.29)
A-GEM	26.49 (5.62)
ER-Reservoir	85.1 (1.02)
CN-DPM	93.23
Gdumb	91.9 (0.5)
NCCL + Reservoir	95.15 (0.91)

Table 10: Single-headed split-CIFAR10, full size Resnet-18 and mem. size=500(50 / cls.). Accuracy and forgetting results.

Method	**accuracy**
iid-offline	93.17
iid-online	36.65
Fine-tune	12.68
EWC	53.49 (0.72)
A-GEM	54.28 (3.48)
GSS	33.56
Reservoir Sampling	37.09
CN-DPM	41.78
NCCL + Ring	54.63 (0.76)
NCCL + Reservoir	55.43 (0.32)

Table 11: Single-headed split-CIFAR100, Resnet18 with $n_f = 20$. Memory size = 10,000. We conduct the experiment with the same setting of GMED [Jin et al., 2021].

Methods	**accuracy**
Finetune	3.06(0.2)
iid online	18.13(0.8)
iid offline	42.00(0.9)
A-GEM	2.40(0.2)
GSS-Greedy	19.53(1.3)
BGD	3.11(0.2)
ER-Reservoir	20.11(1.2)
ER-Reservoir + GMED	20.93(1.6)
MIR	20.02(1.7)
MIR + GMED	21.22(1.0)
NCCL-Reservoir	**21.95(0.3)**

C THEORETICAL ANALYSIS

In this section, we provide the proofs of the results for nonconvex continual learning. We first start with the derivation of Equation 5 in Assumption 3.1.

C.1 ASSUMPTION AND ADDITIONAL LEMMA

Derivation of Equation 5. Recall that

$$|f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle| \leq \frac{L}{2}\|x - y\|^2. \tag{1}$$

Note that f_i is differentiable and nonconvex. We define a function $g(t) = f_i(y + t(x - y))$ for $t \in [0, 1]$ and an objective function f_i. By the fundamental theorem of calculus,

$$\int_0^1 g'(t)dt = f(x) - f(y). \tag{2}$$

By the property, we have

$$\begin{aligned}
&|f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle| \\
&= \left| \int_0^1 \langle \nabla f_i(y + t(x - y)), x - y \rangle dt - \langle \nabla f_i(y), x - y \rangle \right| \\
&= \left| \int_0^1 \langle \nabla f_i(y + t(x - y)) - \nabla f_i(y), x - y \rangle dt \right|.
\end{aligned}$$

Using the Cauchy-Schwartz inequality,

$$\begin{aligned}
&\left| \int_0^1 \langle \nabla f_i(y + t(x - y)) - \nabla f_i(y), x - y \rangle dt \right| \\
&\leq \left| \int_0^1 \|\nabla f_i(y + t(x - y)) - \nabla f_i(y)\| \cdot \|x - y\| dt \right|.
\end{aligned}$$

Since f_i satisfies Equation 4, then we have

$$\begin{aligned}
&|f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle| \\
&\leq \left| \int_0^1 L\|y + t(x - y) - y\| \cdot \|x - y\| dt \right| \\
&= L\|x - y\|^2 \left| \int_0^1 t dt \right| \\
&= \frac{L}{2}\|x - y\|^2.
\end{aligned}$$

□

Lemma C.1. *Let $p = [p_1, \cdots p_D],\ q = [q_1, \cdots, q_D]$ be two statistically independent random vectors with dimension D. Then the expectation of the inner product of two random vectors $\mathbb{E}[\langle p, q \rangle]$ is $\sum_{d=1}^{D} \mathbb{E}[p_d]\mathbb{E}[q_d]$.*

Proof. By the property of expectation,

$$\begin{aligned}
\mathbb{E}[\langle p, q \rangle] &= \mathbb{E}[\sum_{d=1}^{D} p_d q_d] \\
&= \sum_{d=1}^{D} \mathbb{E}[p_d q_d] \\
&= \sum_{d=1}^{D} \mathbb{E}[p_d]\mathbb{E}[q_d].
\end{aligned}$$

□

C.2 PROOF OF MAIN RESULTS

We now show the main results of our work.

***Proof of Lemma 4.1**.* To clarify the issue of $\mathbb{E}_{M_t}\left[\mathbb{E}_{I_t}\left[e_t|M_t\right]\right]=0$, let us explain the details of constructing replay-memory as follows. We have considered episodic memory and reservoir sampling in the paper. We will first show the case of episodic memory by describing the sampling method for replay memory. We can also derive the case of reservoir sampling by simply applying the result of episodic memory.

Episodic memory (ring buffer). We divide the entire dataset of continual learning into the previous task P and the current task C on the time step $t=0$. For the previous task P, the data stream of P is i.i.d., and its sequence is random on every trial (episode). The trial (episode) implies that a continual learning agent learns from an online data stream with two consecutive data sequences of P and C. Episodic memory takes the last data points of the given memory size m by the First In First Out (FIFO) rule, and holds the entire data points until learning on C is finished. Then, we note that $M_t=M_0$ for all $t\geq 0$ and M_0 is uniformly sampled from the i.i.d. sequence of P. By the law of total expectation, we derive $\mathbb{E}_{M_0\subset P}\left[\mathbb{E}_{I_t}\left[\nabla f_{I_t}(x^t)|M_0\right]\right]$ for any $x^t,\ \forall t\geq 0$.

$$\mathbb{E}_{M_0\subset P}\left[\mathbb{E}_{I_t}\left[\nabla f_{I_t}(x^t)|M_0\right]\right]=\mathbb{E}_{M_0\subset P}\left[\nabla f_{M_0}(x^t)\right].$$

It is known that M_0 was uniformly sampled from P on each trial before training on the current task C. Then, we take expectation with respect to every trial that implies the expected value over the memory distribution M_0. We have

$$\mathbb{E}_{M_0\subset P}\left[\nabla f_{M_0}(x^t)\right]=\nabla f(x^t)$$

for any $x^t,\ \forall t$. We can consider $\nabla f_{M_t}(x^t)$ as a sample mean of P on every trial for any $x^t,\ \forall t\geq 0$. Although x^t is constructed iteratively, the expected value of the sample mean for any x^t, $\mathbb{E}_{M_0\subset P}\left[\nabla f_{M_0}(x^t)\right]$ is also derived as $\nabla f(x^t)$.

Reservoir sampling. To clarify the notation for reservoir sampling first, we denote the expectation with respect to the history of replay memory $M_{[0:t]}=(M_0,\cdots,M_t)$ as $\mathbb{E}_{M_{[0:t]}}$. This is the revised version of $\mathbb{E}_{M_t}$. Reservoir sampling is a trickier case than episodic memory, but $\mathbb{E}_{M_{[0:t]}}\left[\mathbb{E}_{I_t}\left[e_t|M_t\right]\right]=0$ still holds. Suppose that M_0 is full of the data points from P as the episodic memory is sampled and the mini-batch size from C is 1 for simplicity. The reservoir sampling algorithm drops a data point in M_{t-1} and replaces the dropped data point with a data point in the current mini-batch from C with probability $p=m/n$, where m is the memory size and n is the number of visited data points so far. The exact pseudo-code for reservoir sampling is described in [1]. The replacement procedure uniformly chooses the data point which will be dropped. We can also consider the replacement procedure as follows. The memory M_t for P is reduced in size 1 from M_{t-1}, and the replaced data point d_C from C contributes in terms of $\nabla g_{d_C}(x^t)$ if d_C is sampled from the replay memory. Let $M_{t-1}=[d_1,\cdots,d_{|M_{t-1}|}]$ where $|\cdot|$ denotes the cardinality of the memory. The sample mean of M_{t-1} is given as

$$\nabla f_{M_{t-1}}(x^{t-1})=\frac{1}{|M_{t-1}|}\sum_{d_i}\nabla f_{d_i}(x^{t-1}). \tag{3}$$

By the rule of reservoir sampling, we assume that the replacement procedure reduces the memory from M_{t-1} to M_t with size $|M_{t-1}|-1$ and the set of remained upcoming data points $C_t\in C$ from the current data stream for online continual learning is reformulated into $C_{t-1}\cup[d_C]$. Then, d_C can be resampled from $C_{t-1}\cup[d_C]$ to be composed of the minibatch of reservoir sampling with the dfferent probability. However, we ignore the probability issue now to focus on the effect of replay-memory on ∇f. Now, we sample M_t from M_{t-1}, then we get the random vector $\nabla f_{M_t}(x^t)$ as

$$\nabla f_{M_t}(x^t)=\frac{1}{|M_t|}\sum_{j=1}^{|M_{t-1}|}W_{ij}\nabla f_{d_j}(x^t), \tag{4}$$

where the index i is uniformly sampled from $i\sim[1,\cdots,|M_{t-1}|]$, and W_{ij} is the indicator function that W_{ij} is 0 if $i=j$ else 1.

The above description implies the dropping rule, and M_t can be considered as an uniformly sampled set with size $|M_t|$ from M_{t-1}. There could also be $M_t=M_{t-1}$ with probability $1-p=1-m/n$. Then the expectation of $\nabla f_{M_t}(x^t)$ given M_{t-1}

is derived as

$$\begin{aligned}\mathbb{E}_{M_t}[\nabla f_{M_t}(x^t)|M_{t-1}] &= p\left(\frac{1}{|M_{t-1}|}\sum_{i}^{|M_{t-1}|}\frac{1}{|M_t|}\sum_{j=1}^{|M_{t-1}|} W_{ij}\nabla f_{d_j}(x^t)\right) + (1-p)\left(\nabla f_{M_{t-1}}(x^t)\right)\\ &= \nabla f_{M_{t-1}}(x^t).\end{aligned}$$

When we consider the mini-batch sampling, we can formally reformulate the above equation as

$$\mathbb{E}_{M_t\sim p(M_t|M_{t-1})}\left[\mathbb{E}_{I_t\subset M_t}\left[\nabla f_{I_t}(x^t)|M_t\right]|M_{t-1}\right] = \nabla f_{M_{t-1}}(x^t). \tag{5}$$

Now, we apply the above equation recursively. Then,

$$\mathbb{E}_{M_1\sim p(M_1|M_0)}\left[\cdots\mathbb{E}_{M_t\sim p(M_t|M_{t-1})}\left[\mathbb{E}_{I_t\subset M_t}\left[\nabla f_{I_t}(x^t)|M_t\right]|M_{t-1}\right]\cdots|M_0\right] = \nabla f_{M_0}(x^t). \tag{6}$$

Similar to episodic memory, M_0 is uniformly sampled from P. Therefore, we conclude that

$$\mathbb{E}_{M_0,\cdots,M_t}[\nabla f_{M_t}(x^t)] = \nabla f(x^t) \tag{7}$$

by taking expectation over the history $M_{[0:t]} = (M_1, M_2, \cdots, M_t)$.

Note that taking expectation iteratively with respect to the history $M_{[t]}$ is needed to compute the expected value of gradients for M_t. However, the result $\mathbb{E}_{M_0,\cdots,M_t}[\mathbb{E}_{I_t}[e_t|M_t]] = 0$ still holds in terms of expectation.

Furthermore, we also discuss that the effect of reservoir sampling on the convergence of C. Unlike we simply update $g(x)$ by the stochastic gradient descent on C, the datapoints $d \in M \cap C$ have a little larger sampling probability than other datapoints $d_{C-M} \in C - M$. The expectation of gradient norm on the averaged loss $\mathbb{E}\|\nabla g(x^t)\|^2$ is based on the uniform and equiprobable sampling over C, but the nature of reservoir sampling distort this measure slightly. In this paper, we focus on the convergence of the previous task C while training on the current task C with several existing memory-based methods. Therefore, analyzing the convergence of reservoir sampling method will be a future work.

□

Proof of Lemma 4.2. We analyze the convergence of nonconvex continual learning with replay memory here. Recall that the gradient update is the following

$$x^{t+1} = x^t - \alpha_{H_t}\nabla f_{I_t}(x^t) - \beta_{H_t}\nabla g_{J_t}(x^t)$$

for all $t \in \{1, 2, \cdots, T\}$. Let $e_t = \nabla f_{I_t}(x^t) - \nabla f(x^t)$. Since we assume that $f,\ g$ is L-smooth, we have the following inequality by applying Equation 5:

$$\begin{aligned}
f(x^{t+1}) &\leq f(x^t) + \langle\nabla f(x^t), x^{t+1} - x^t\rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2\\
&= f(x^t) - \langle\nabla f(x^t), \alpha_{H_t}\nabla f_{I_t}(x^t) + \beta_{H_t}\nabla g_{J_t}(x^t)\rangle + \frac{L}{2}\|\alpha_{H_t}\nabla f_{I_t}(x^t) + \beta_{H_t}\nabla g_{J_t}(x^t)\|^2\\
&= f(x^t) - \alpha_{H_t}\langle\nabla f(x^t), \nabla f_{I_t}(x^t)\rangle - \beta_{H_t}\langle\nabla f(x^t), \nabla g_{J_t}(x^t)\rangle\\
&\quad + \frac{L}{2}\alpha_{H_t}^2\|\nabla f_{I_t}(x^t)\|^2 + \frac{L}{2}\beta_{H_t}^2\|\nabla g_{J_t}(x^t)\|^2 + L\alpha_{H_t}\beta_{H_t}\langle\nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle\\
&= f(x^t) - \alpha_{H_t}\langle\nabla f(x^t), \nabla f(x^t)\rangle - \alpha_{H_t}\langle\nabla f(x^t), e_t\rangle - \beta_{H_t}\langle\nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle + \beta_{H_t}\langle\nabla g_{J_t}(x^t), e_t\rangle\\
&\quad + \frac{L\alpha_{H_t}^2}{2}\|\nabla f(x^t)\|^2 + L\alpha_{H_t}^2\langle\nabla f(x^t), e_t\rangle + \frac{L\alpha_{H_t}^2}{2}\|e_t\|^2 + \frac{L\beta_{H_t}^2}{2}\|\nabla g_{J_t}(x^t)\|^2 + L\alpha_{H_t}\beta_{H_t}\langle\nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle\\
&= f(x^t) - \left(\alpha_{H_t} - \frac{L}{2}\alpha_{H_t}^2\right)\|\nabla f(x^t)\|^2 + \frac{L}{2}\beta_{H_t}^2\|\nabla g_{J_t}(x^t)\|^2 - \beta_{H_t}(1 - \alpha_{H_t}L)\langle\nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle\\
&\quad + \left(L\alpha_{H_t}^2 - \alpha_{H_t}\right)\langle\nabla f(x^t), e_t\rangle + \beta_{H_t}\langle\nabla g_{J_t}(x^t), e_t\rangle + \frac{L}{2}\alpha_{H_t}^2\|e_t\|^2.
\end{aligned} \tag{8}$$

To show the proposed theoretical convergence analysis of nonconvex continual learning, we define the catastrophic forgetting term Γ_t and the overfitting term B_t as follows:

$$\begin{aligned}
B_t &= (L\alpha_{H_t}^2 - \alpha_{H_t})\langle \nabla f(x^t), e_t\rangle + \beta_{H_t}\langle \nabla g_{J_t}(x^t), e_t\rangle,\\
\Gamma_t &= \frac{\beta_{H_t}^2 L}{2}\|\nabla g_{J_t}(x^t)\|^2 - \beta_{H_t}(1-\alpha_{H_t}L)\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle.
\end{aligned}$$

Then, we can rewrite Equation 8 as

$$f(x^{t+1}) \le f(x^t) - \left(\alpha_{H_t} - \frac{L}{2}\alpha_{H_t}^2\right)\|\nabla f(x^t)\|^2 + \Gamma_t + B_t + \frac{L}{2}\alpha_{H_t}^2\|e_t\|^2. \tag{9}$$

We first note that B_t is dependent of the error term e_t with the batch I_t. In the continual learning step, an training agent cannot access $\nabla f(x^t)$, then we cannot get the exact value of e_t. Furthermore, Γ_t is dependent of the gradients $\nabla f_{I_t}(x^t)$, $\nabla g_{I_t}(x^t)$ and the learning rates $\alpha_{H_t}, \beta_{H_t}$.

Taking expectations with respect to I_t on both sides given J_t, we have

$$\begin{aligned}
\mathbb{E}_{I_t}\left[f(x^{t+1})\right] &\le \mathbb{E}_{I_t}\left[f(x^t) - \left(\alpha_{H_t} - \frac{L}{2}\alpha_{H_t}^2\right)\|\nabla f(x^t)\|^2 + \Gamma_t + B_t + \frac{L}{2}\alpha_{H_t}^2\|e_t\|^2 \Big| J_t\right]\\
&\le \mathbb{E}_{I_t}\left[f(x^t) - \left(\alpha_{H_t} - \frac{L}{2}\alpha_{H_t}^2\right)\|\nabla f(x^t)\|^2 + \frac{L}{2}\alpha_{H_t}^2\|e_t\|^2\right] + \mathbb{E}_{I_t}\left[\Gamma_t + B_t \Big| J_t\right].
\end{aligned}$$

Now, taking expectations over the whole stochasticity we obtain

$$\mathbb{E}\left[f(x^{t+1})\right] \le \mathbb{E}\left[f(x^t) - \left(\alpha_{H_t} - \frac{L}{2}\alpha_{H_t}^2\right)\|\nabla f(x^t)\|^2 + \Gamma_t + B_t + \frac{L}{2}\alpha_{H_t}^2\|e_t\|^2\right].$$

Rearranging the terms and assume that $\frac{1}{1-L\alpha_{H_t}/2} > 0$, we have

$$\left(\alpha_{H_t} - \frac{L}{2}\alpha_{H_t}^2\right)\mathbb{E}\|\nabla f(x^t)\|^2 \le \mathbb{E}\left[f(x^t) - f(x^{t+1}) + \Gamma_t + B_t + \frac{L}{2}\alpha_{H_t}^2\|e_t\|^2\right]$$

and

$$\begin{aligned}
\mathbb{E}\|\nabla f(x^t)\|^2 &\le \mathbb{E}\left[\frac{1}{\alpha_{H_t}(1-\frac{L}{2}\alpha_{H_t})}\big(f(x^t) - f(x^{t+1}) + \Gamma_t + B_t\big) + \frac{\alpha_{H_t}L}{2(1-\frac{L}{2}\alpha_{H_t})}\|e_t\|^2\right]\\
&\le \mathbb{E}\left[\frac{1}{\alpha_{H_t}(1-\frac{L}{2}\alpha_{H_t})}\big(f(x^t) - f(x^{t+1}) + \Gamma_t + B_t\big) + \frac{\alpha_{H_t}L}{2(1-\frac{L}{2}\alpha_{H_t})}\sigma_f^2\right].
\end{aligned}$$

□

***Proof of Theorem 4.3**.* Suppose that the learning rate α_{H_t} is a constant $\alpha = c/\sqrt{T}$, for $c > 0$, $1 - \frac{L}{2}\alpha = \frac{1}{A} > 0$. Then, by summing Equation 7 from $t = 0$ to $T-1$, we have

$$
\begin{aligned}
\min_t \mathbb{E}\|\nabla f(x^t)\|^2 &\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x^t)\|^2 \\
&\leq \frac{1}{1-\frac{L}{2}\alpha}\left(\frac{1}{\alpha T}\left(f(x^0)-f(x^T)+\sum_{t=0}^{T-1}\left(\mathbb{E}\left[B_t+\Gamma_t\right]\right)\right)+\frac{L}{2}\alpha\sigma_f^2\right) \\
&= \frac{1}{1-\frac{L}{2}\alpha}\left(\frac{1}{c\sqrt{T}}\left(\Delta_f+\sum_{t=0}^{T-1}\left(\mathbb{E}\left[B_t+\Gamma_t\right]\right)\right)+\frac{Lc}{2\sqrt{T}}\sigma_f^2\right) \\
&= \frac{A}{\sqrt{T}}\left(\frac{1}{c}\left(\Delta_f+\sum_{t=0}^{T-1}\mathbb{E}\left[B_t+\Gamma_t\right]\right)+\frac{Lc}{2}\sigma_f^2\right). \qquad (10)
\end{aligned}
$$

We note that a batch I_t is sampled from a memory $M_t \subset M$ which is a random vector whose element is a datapoint $d \in P \cup C$. Then, taking expectation over $I_t \subset M_t \subset P \cup C$ implies that $\mathbb{E}[B_t] = 0$. Therefore, we get the minimum of expected square of the norm of gradients

$$
\min_t \mathbb{E}\|\nabla f(x^t)\|^2 \leq \frac{A}{\sqrt{T}}\left(\frac{1}{c}\left(\Delta_f+\sum_{t=0}^{T-1}\mathbb{E}[\Gamma_t]\right)+\frac{Lc}{2}\sigma_f^2\right).
$$

□

Proof of Lemma 4.4. To simplify the proof, we assume that learning rates $\alpha_{H_t}, \beta_{H_t}$ are a same fixed value $\beta = c'/\sqrt{T}$. The assumption is reasonable, because it is observed that the RHS of Equation 7 is not perturbed drastically by small learning rates in $0 < \alpha_{H_t}, \beta_{H_t} \leq 2/L \ll 1$. Let us denote the union of M_t over time $0 \leq t \leq T-1$ as $M = \bigcup_t M_t$. By the assumption, it is equivalent to update on $M \cup C$. Then, the non-convex finite sum optimization is given as

$$
\min_{x\in\mathbb{R}^d} h|_{M\cup C}(x) = \frac{1}{n_g+|M|}\sum_{i\in M\cup C} h_i(x), \qquad (11)
$$

where $|M|$ is the number of elements in M. This problem can be solved by a simple SGD algorithm [Reddi et al., 2016]. Thus, we have

$$
\min_t \mathbb{E}\|\nabla h|_{M\cup C}(x^t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\nabla h|_{M\cup C}(x^t)\|^2 \leq \sqrt{\frac{2\Delta_{h|_{M\cup C}}L}{T}}\sigma_{h|_{M\cup C}}. \qquad (12)
$$

□

Lemma C.2. *For any $C \subset D \subset M \cup C$, define $\omega^2_{h|_D}$ as*

$$
\omega^2_{h|_D} = \sup_x \mathbb{E}_{j\in D}\|\nabla h_j(x^t) - \nabla h|_{M\cup C}(x^t)\|^2].
$$

Then, we have

$$
\mathbb{E}\|\nabla g_{J_t}(x^t)\|^2 \leq \mathbb{E}\|\nabla h|_{M\cup C}(x^t)\|^2 + \sup_{C\subset D\subset M\cup C}\omega^2_{h|_D}. \qquad (13)
$$

Proof of Lemma C.2. We arrive at the following result by Jensen's inequality

$$
\begin{aligned}
\sup_x \mathbb{E}_{J_t\subset C}\|\nabla g_{J_t}(x^t) - \nabla h|_{M\cup C}(x^t)\|^2 &= \sup_x \mathbb{E}_{J_t\subset C}\left[\|\mathbb{E}_{j\in J_t}[\nabla h_j(x^t)] - \nabla h|_{M\cup C}(x^t)\|^2\right] \qquad (14) \\
&\leq \sup_{C\subset D\subset M\cup C}\ \sup_x \mathbb{E}_{J_t\subset D}\left[\|\mathbb{E}_{j\in J_t}[\nabla h_j(x^t)] - \nabla h|_{M\cup C}(x^t)\|^2\right] \qquad (15) \\
&\leq \sup_{C\subset D\subset M\cup C}\left[\sup_x \mathbb{E}_{j\in D}[\ \|\nabla h_j(x^t) - \nabla h|_{M\cup C}(x^t)\|^2]\right] \qquad (16) \\
&= \sup_{C\subset D\subset M\cup C}\omega^2_{h|_D}. \qquad (17)
\end{aligned}
$$

By the triangular inequality, we get

$$\begin{aligned}\mathbb{E}\|\nabla g_{J_t}(x^t)\|^2 &\leq \mathbb{E}\|\nabla g_{J_t}(x^t) - \nabla h|_{M\cup C}(x^t)\|^2 + \mathbb{E}\|\nabla h|_{M\cup C}(x^t)\|^2 &(18)\\ &\leq \mathbb{E}\|\nabla h|_{M\cup C}(x^t)\|^2 + \sup_{C\subset D\subset M\cup C} \omega^2_{h|_D}. &(19)\end{aligned}$$

□

For continual learning, the model x^0 reaches to an ϵ-stationary point of $f(x)$ when we have finished to learn P and start to learn C. Now, we discuss the frequency of transfer and interference during continual learning before showing Lemma 4.5. It is well known that the frequencies between interference and transfer have similar values (the frequency of constraint violation is approximately 0.5 for AGEM) as shown in Appendix D of [Chaudhry et al., 2019]. Even if memory-based continual learning has a small memory buffer which contains a subset of P, random sampling from the buffer allows to have similar frequencies between interference and transfer.

In this paper, we consider two cases for the upper bound of $\mathbb{E}[\Gamma_t]$, the moderate case and the worst case. For **the moderate case**, which covers most continual learning scenarios, we assume that the inner product term $\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle$ has the same probabilities of being positive (transfer) and negative (interference). Then, we can approximate $\mathbb{E}[\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle] \approx 0$ over all randomness. For **the worst case**, we assume that all $\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle$ has negative values.

Proof of Lemma 4.5. For the moderate case, we derive the rough upper bound of $\mathbb{E}[\Gamma_t]$:

$$\begin{aligned}\mathbb{E}\left[\Gamma_t\right] &= \mathbb{E}\left[\frac{\beta^2_{H_t}L}{2}\|\nabla g_{J_t}(x^t)\|^2 - \beta_{H_t}(1-\alpha_{H_t}L)\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle\right] &(20)\\ &\approx \mathbb{E}\left[\frac{\beta^2_{H_t}L}{2}\|\nabla g_{J_t}(x^t)\|^2\right] &(21)\\ &= O\left(\mathbb{E}\left[\frac{\beta^2 L}{2}\|\nabla g_{J_t}(x^t)\|^2\right]\right) &(22)\end{aligned}$$

By plugging Lemma C.2 into $\mathbb{E}[\Gamma_t]$, we obtain that

$$\begin{aligned}\mathbb{E}[\Gamma_t] &\leq O\left(\mathbb{E}\left[\frac{\beta^2 L}{2}\|\nabla g_{J_t}(x^t)\|^2\right]\right) &(23)\\ &= O\left(\mathbb{E}\left[\frac{\beta^2 L}{2}\|\nabla h|_{M\cup C}(x^t)\|^2 + \frac{\beta^2 L}{2}\sup_{C\subset D\subset M\cup C}\omega^2_{h|_D}\right]\right). &(24)\end{aligned}$$

We use the technique for summing up in the proof of Theorem 1, then the cumulative sum of catastrophic forgetting term is derived as

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}[\Gamma_t] &\leq \sum_{t=0}^{T-1}\frac{\beta^2 L}{2}O\left(\mathbb{E}\left[\|h|_{M\cup C}(x^t)\|^2\right] + \sup_{C\subset D\subset M\cup C}\omega^2_{h|_D}\right) &(25)\\ &\leq \frac{\beta^2 L}{2}\sum_{t=0}^{T-1}O\left(\frac{1}{\beta}\left[h|_{M\cup C}(x^t) - h|_{M\cup C}(x^{t+1})\right] + \frac{L\beta}{2}\sigma^2_{h|_{M\cup C}} + \sup_{C\subset D\subset M\cup C}\omega^2_{h|_D}\right) &(26)\\ &\leq \frac{\beta^2 L}{2}O\left(\frac{1}{\beta}\Delta_{h|_{M\cup C}} + \frac{TL\beta}{2}\sigma^2_{h|_{M\cup C}} + T\sup_{C\subset D\subset M\cup C}\omega^2_{h|_D}\right) &(27)\\ &= O\left(\beta\Delta_{h|_{M\cup C}} + \frac{TL\beta^3}{2}\sigma^2_{h|_{M\cup C}} + T\beta^2\sup_{C\subset D\subset M\cup C}\omega^2_{h|_D}\right). &(28)\end{aligned}$$

Now, we consider the randomness of memory choice. Let D^* be as follows:

$$D^* = \operatorname*{arg\,max}_{C\subset D\subset P\cup C}\beta\Delta_{h|_D} + \frac{TL\beta^3}{2}\sigma^2_{h|_D}. \tag{29}$$

Then, we obtain the following inequality,

$$\sum_{t=0}^{T-1} \mathbb{E}[\Gamma_t] \leq O\left(\beta\Delta_{h|_{D^*}} + \frac{TL\beta^3}{2}\sigma^2_{h|_{D^*}} + T\beta^2 \sup_{C\subset D\subset M\cup C} \omega^2_{h|_D}\right) \tag{30}$$

$$\leq O\left(\beta\Delta_{h|_{D^*}} + \frac{TL\beta^3}{2}\sigma^2_{h|_{D^*}} + T\beta^2 \sup_{C\subset D\subset P\cup C} \omega^2_{h|_D}\right). \tag{31}$$

Rearranging the above equation, we get

$$\sum_{t=0}^{T-1} \mathbb{E}[\Gamma_t] \leq O\left(T\left(\frac{L\beta^3}{2}\sigma^2_{h|_{D^*}} + \beta^2 \sup_{C\subset D\subset P\cup C} \omega^2_{h|_D}\right) + \beta\Delta_{h|_{D^*}}\right). \tag{32}$$

For the moderate case, we provide the derivations of the convergence rate for two cases of β as follows.

When $\beta < \alpha = c/\sqrt{T}$, the upper bound always satisfies

$$\sum_{t=0}^{T-1} \frac{\mathbb{E}[\Gamma_t]}{\sqrt{T}} \leq \frac{1}{\sqrt{T}} O\left(\frac{1}{T}\left(\frac{L\beta}{2}\sigma^2_{h|_{D^*}} + \frac{1}{\sqrt{T}} \sup_{C\subset D\subset P\cup C} \omega^2_{h|_D}\right) + \frac{1}{\sqrt{T}}\Delta_{h|_{D^*}}\right) < O\left(\frac{1}{T^{3/2}} + \frac{1}{T}\right).$$

For $\beta \geq \alpha = c/\sqrt{T}$, we cannot derive a tighter bound, so we still have

$$\sum_{t=0}^{T-1} \frac{\mathbb{E}[\Gamma_t]}{\sqrt{T}} \leq \frac{1}{\sqrt{T}} O\left(T\left(\frac{L\beta^3}{2}\sigma^2_{h|_{D^*}} + \beta^2 \sup_{C\subset D\subset P\cup C} \omega^2_{h|_D}\right) + \beta\Delta_{h|_{D^*}}\right) = O\left(\sqrt{T} + \frac{1}{\sqrt{T}}\right).$$

For the worst case, we assume that there exists a constant $c_{f,g}$ which satisfies $c_{f,g}\|\nabla g_{J_t}(x^t)\| \geq \|\nabla f_{I_t}(x^t)\|$.

$$\mathbb{E}\left[\Gamma_t\right] = \mathbb{E}\left[\frac{\beta^2_{H_t} L}{2}\|\nabla g_{J_t}(x^t)\|^2 - \beta_{H_t}(1-\alpha_{H_t}L)\langle\nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle\right] \tag{33}$$

$$\leq \mathbb{E}\left[\frac{\beta^2_{H_t} L}{2}\|\nabla g_{J_t}(x^t)\|^2 + \beta_{H_t}(1-\alpha_{H_t}L)\|\nabla f_{I_t}(x^t)\|\|\nabla g_{J_t}(x^t)\|\right] \tag{34}$$

$$\leq \mathbb{E}\left[\frac{\beta^2 L}{2}\|\nabla g_{J_t}(x^t)\|^2 + \beta c_{f,g}\|\nabla g_{J_t}(x^t)\|^2\right] \tag{35}$$

$$= O\left(\mathbb{E}\left[\left(\beta^2+\beta\right)\|\nabla g_{J_t}(x^t)\|^2\right]\right). \tag{36}$$

By plugging Lemma C.2 into $\mathbb{E}[\Gamma_t]$, we obtain that

$$\mathbb{E}[\Gamma_t] \leq O\left(\mathbb{E}\left[\left(\beta^2+\beta\right)\|\nabla g_{J_t}(x^t)\|^2\right]\right) \tag{37}$$

$$= O\left(\left(\beta^2+\beta\right)\mathbb{E}\left[\|\nabla h|_{M\cup C}(x^t)\|^2 + \sup_{C\subset D\subset M\cup C} \omega^2_{h|_D}\right]\right). \tag{38}$$

We use the technique for summing up in the proof of Theorem 1, then the cumulative sum of catastrophic forgetting term is derived as

$$\sum_{t=0}^{T-1} \mathbb{E}[\Gamma_t] \leq \sum_{t=0}^{T-1} \left(\beta^2+\beta\right) O\left(\mathbb{E}\left[\|h|_{M\cup C}(x^t)\|^2\right] + \sup_{C\subset D\subset M\cup C} \omega^2_{h|_D}\right) \tag{39}$$

$$\leq \left(\beta^2+\beta\right) \sum_{t=0}^{T-1} O\left(\frac{1}{\beta}\left[h|_{M\cup C}(x^t) - h|_{M\cup C}(x^{t+1})\right] + \frac{L\beta}{2}\sigma^2_{h|_{M\cup C}} + \sup_{C\subset D\subset M\cup C} \omega^2_{h|_D}\right) \tag{40}$$

$$\leq \left(\beta^2+\beta\right) O\left(\frac{1}{\beta}\Delta_{h|_{M\cup C}} + \frac{TL\beta}{2}\sigma^2_{h|_{M\cup C}} + T \sup_{C\subset D\subset M\cup C} \omega^2_{h|_D}\right) \tag{41}$$

$$= O\left((\beta+1)\Delta_{h|_{M\cup C}} + \frac{TL\beta^2(\beta+1)}{2}\sigma^2_{h|_{M\cup C}} + T\beta(\beta+1) \sup_{C\subset D\subset M\cup C} \omega^2_{h|_D}\right). \tag{42}$$

For the worst case, we provide the derivations of the convergence rate for two cases of β as follows.

When $\beta < \alpha = c/\sqrt{T}$, the upper bound always satisfies

$$\sum_{t=0}^{T-1} \frac{\mathbb{E}[\Gamma_t]}{\sqrt{T}} \leq \frac{1}{\sqrt{T}} O\left(\frac{Lc + \sqrt{T}}{\sqrt{T}} \sigma^2_{h|_{D^*}} + (\sqrt{T} + c) \sup_{C \subset D \subset P \cup C} \omega^2_{h|_D} + \frac{\sqrt{T} + c}{\sqrt{T}} \Delta_{h|_{D^*}} \right) < O\left(\frac{1}{T} + \frac{1}{\sqrt{T}} + 1 \right).$$

For $\beta \geq \alpha = c/\sqrt{T}$, we cannot derive a tighter bound, so we still have

$$\sum_{t=0}^{T-1} \frac{\mathbb{E}[\Gamma_t]}{\sqrt{T}} \leq \frac{1}{\sqrt{T}} O\left(T \left(\frac{L\beta^2(\beta+1)}{2} \sigma^2_{h|_{D^*}} + \beta(\beta+1) \sup_{C \subset D \subset P \cup C} \omega^2_{h|_D} \right) + (\beta+1)\Delta_{h|_{D^*}} \right) = O\left(\sqrt{T} + \frac{1}{\sqrt{T}} \right).$$

□

Even if we consider the worst case, we still have $O(1)$ for the cumulative forgetting $\mathbb{E}[\Gamma_t]$ when $\beta < \alpha$. This implies that we have the theoretical condition for control the forgetting on $f(x)$ while evolving on C. In the main text, we only discuss the moderate case to emphasize $f(x)$ can be converged by the effect of transfer during continual learning, but we have also considered the worst case can be well treated by our theoretical condition by keeping the convergence of $f(x)$ over time as follows.

Proof of Corollary 4.6. By Lemma 4.5, we have

$$\sum_{t=0}^{T-1} \frac{\mathbb{E}[\Gamma_t]}{\sqrt{T}} < O\left(\frac{1}{T^{3/2}} + \frac{1}{T} \right)$$

for $\beta < \alpha$ for **the moderate case**. Then, we can apply the result into RHS of the inequality in Theorem 4.3 as follows.

$$\begin{aligned} \min_t \mathbb{E}\|\nabla f(x^t)\|^2 &\leq \frac{A}{\sqrt{T}} \left(\frac{1}{c} \left(\Delta_f + \sum_{t=0}^{T-1} \mathbb{E}\left[\Gamma_t\right] \right) + \frac{Lc}{2} \sigma_f^2 \right) \\ &= \frac{A/c}{\sqrt{T}} \left(\Delta_f + \frac{Lc^2}{2} \sigma_f^2 \right) + \frac{A/c}{\sqrt{T}} \sum_{t=0}^{T-1} \mathbb{E}[\Gamma_t] \\ &= O\left(\frac{1}{T^{3/2}} + \frac{1}{T} + \frac{1}{T^{1/2}} \right) = O\left(\frac{1}{\sqrt{T}} \right). \end{aligned}$$

In addition, we have the convergence rate of $f(x)$ for **the worst case** as follows:

$$\min_t \mathbb{E}\|\nabla f(x^t)\|^2 = O(1), \tag{43}$$

which implies that $f(x)$ can keep the convergence while evolving on C.

□

Proof of Corollary 4.7. To formulate the IFO calls, Recall that $T(\epsilon)$

$$T(\epsilon) = \min\left\{T : \ \min \ \mathbb{E}\|\nabla f(x^t)\|^2 \leq \epsilon\right\}.$$

A single IFO call is invested in calculating each step, and we now compute IFO calls to reach an ϵ-accurate solution.

$$\frac{A}{\sqrt{T}} \left(\frac{1}{c} \left(\Delta_f + \sum_{t=0}^{T-1} \mathbb{E}\left[\Gamma_t\right] \right) + \frac{Lc}{2} \sigma_f^2 \right) \to \epsilon.$$

When $\beta < \alpha$, we get

$$\text{IFO calls} = O\left(\frac{1}{\epsilon^2} \right).$$

Otherwise, when $\beta \geq \alpha$, we cannot guarantee the upper bound of stationary decreases over time. Then, we cannot compute IFO calls for this case.

□

D DERIVATION OF EQUATIONS IN ADAPTIVE METHODS IN CONTINUAL LEARNING

Derivation for A-GEM Let the surrogate $\nabla \tilde{g}_{J_t}(x^t)$ as

$$\nabla \tilde{g}_{J_t}(x^t) = \nabla g_{J_t}(x^t) - \left\langle \frac{\nabla f_{I_t}(x^t)}{\|\nabla f_{I_t}(x^t)\|}, \nabla g_{J_t}(x^t) \right\rangle \frac{\nabla f_{I_t}(x^t)}{\|\nabla f_{I_t}(x^t)\|}, \tag{44}$$

where $\alpha_{H_t} = \alpha(1 - \frac{\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle}{\|\nabla f_{I_t}(x^t)\|^2})$ and $\beta_{H_t} = \alpha$ for Equation 3.

Then, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_t] &= \mathbb{E}\left[\frac{\beta_{H_t}^2 L}{2} \|\nabla \tilde{g}_{J_t}(x^t)\|^2 - \beta_{H_t} \langle \nabla f_{I_t}(x^t), \nabla \tilde{g}_{J_t}(x^t) \rangle\right] \\
&= \mathbb{E}\left[\frac{\beta_{H_t}^2 L}{2} \left(\|\nabla g_{J_t}(x^t)\|^2 - 2\frac{\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle^2}{\|\nabla f_{I_t}(x^t)\|^2} + \frac{\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle^2}{\|\nabla f_{I_t}(x^t)\|^2}\right) - \beta_{H_t} \langle \nabla f_{I_t}(x^t), \nabla \tilde{g}_{J_t}(x^t) \rangle\right] \\
&= \mathbb{E}\left[\frac{\beta_{H_t}^2 L}{2} \left(\|\nabla g_{J_t}(x^t)\|^2 - \frac{\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle^2}{\|\nabla f_{I_t}(x^t)\|^2}\right) - \beta_{H_t} \left(\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle - \langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle\right)\right] \\
&= \mathbb{E}\left[\frac{\beta_{H_t}^2 L}{2} \left(\|\nabla g_{J_t}(x^t)\|^2 - \frac{\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle^2}{\|\nabla f_{I_t}(x^t)\|^2}\right)\right].
\end{aligned} \tag{45}$$

Now, we compare the catastrophic forgetting term between the original value with $\nabla g_{J_t}(x^t)$ and the above surrogate.

$$\mathbb{E}\left[\frac{\beta_{H_t}^2 L}{2} \left(\|\nabla g_{J_t}(x^t)\|^2 - \frac{\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle^2}{\|\nabla f_{I_t}(x^t)\|^2}\right)\right] < \mathbb{E}\left[\frac{\beta_{H_t}^2 L}{2} \|\nabla g_{J_t}(x^t)\|^2 - \beta_{H_t} \langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle\right].$$

Then, we can conclude that $\mathbb{E}[\Gamma_t]$ with the surrogate of A-GEM is smaller than the original $\mathbb{E}[\Gamma_t]$.

Derivation of optimal Γ_t^* and $\beta_{H_t}^*$ For a fixed learning rate α, we have

$$\begin{aligned}
0 = \frac{\partial \mathbb{E}[\Gamma_t]}{\partial \beta_{H_t}} &= \mathbb{E}\left[\frac{\partial \Gamma_t}{\partial \beta_{H_t}}\right] \\
&= \mathbb{E}\left[\beta_{H_t} L \|\nabla g_{J_t}(x^t)\| - (1 - \alpha L) \langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle\right].
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\beta_{H_t}^* &= \frac{(1 - \alpha_{H_t} L) \langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle}{L \|\nabla g_{J_t}(x^t)\|^2} = \frac{(1 - \alpha_{H_t} L) \Lambda_{H_t}}{L \|\nabla g_{J_t}(x^t)\|^2}, \\
\Gamma_t^* &= -\frac{(1 - \alpha_{H_t} L) \langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle}{2L \|\nabla g_{J_t}(x^t)\|^2} = -\frac{(1 - \alpha_{H_t} L) \Lambda_{H_t}}{2L \|\nabla g_{J_t}(x^t)\|^2}.
\end{aligned}$$

E OVERFITTING TO REPLAY MEMORY

In Lemma 4.2, we show the expectation of stepwise change of upper bound. Now, we discuss the distribution of the upper bound by analyzing the random variable B_t. As B_t is computed by getting

$$B_t = (L\alpha_{H_t}^2 - \alpha_{H_t}) \langle \nabla f(x^t), e_t \rangle + \beta_{H_t} \langle \nabla g_{J_t}(x^t), e_t \rangle.$$

The purpose of our convergence analysis is to compute the upper bound of Equation 7, then we compute the upper bound of B_t.

$$B_t \leq (L\alpha_{H_t}^2 - \alpha_{H_t}) \|\nabla f(x^t)\| \|e_t\| + \beta_{H_t} \|\nabla g_{J_t}(x^t)\| \|e_t\|.$$

It is noted that the upper bound is related to the distribution of the norm of e_t. We have already know that $\mathbb{E}[e_t] = 0$, so we consider its variance, Var($\|e_t\|$) in this section. Let us denote the number of data points of P in a memory M_0 as m_P. We assume that M_0 is uniformly sampled from P. Then the sample variance, Var($\|e_t\|$) is computed as

$$\text{Var}(\|e_t\|) = \frac{n_f - m_P}{(n_f - 1) m_P} \sigma_f^2$$

by the similar derivation with Equation 14. The above result directly can be applied to the variance of B_t. This implies m_t is a key feature which has an effect on the convergence rate. It is noted that the larger m_P has the smaller variance by applying schemes, such as larger memory. In addition, the distributions of e_t and $\nabla f_{I_t}(x^t)$ are different with various memory schemes. Therefore, we can observe that memory schemes differ the performance even if we apply same step sizes.

References

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Arslan Chaudhry, Naeemullah Khan, Puneet Dokania, and Philip Torr. Continual learning in low-rank orthogonal subspaces. *Advances in Neural Information Processing Systems*, 33, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Xisen Jin, Arka Sadhu, Junyi Du, and Xiang Ren. Gradient-based editing of memory examples for online task-free continual learning. *Advances in Neural Information Processing Systems*, 34:29193–29205, 2021.

Soochan Lee, Junsoo Ha, Dongsu Zhang, and Gunhee Kim. A neural dirichlet process mixture model for task-free continual learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.

Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 314–323, 2016. URL `http://proceedings.mlr.press/v48/reddi16.html`.

Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, and Ali Farhadi. Supermasks in superposition. *Advances in Neural Information Processing Systems*, 33:15173–15184, 2020.