Rebuttal Response

Anonymous Authors[1]

1. Experiments of Class-incremental Learning

Table 1. Results of class-incremental experiments on Multiple Dataset (p denotes the noise rate).

Methods	$p = 0.0$		$p = 0.2$	
	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)
EWC	13.08±5.36	0.695±0.004	10.36±3.21	0.590±0.009
MAS	10.06±1.13	0.613±0.019	6.04±0.62	0.487±0.016
AGEM	13.67±1.24	0.654±0.018	12.10±1.24	0.548±0.007
OGD	13.00±0.74	0.707±0.005	9.75±0.44	0.586±0.005
DER	45.67±2.19	0.260±0.039	40.88±2.15	0.248±0.026
GDumb	40.69±3.32	**0.128±0.023**	36.25±2.23	**0.083±0.019**
MEGA	38.94±2.79	0.389±0.026	34.03±2.23	0.134±0.023
STREAM	**47.92±0.45**	0.153±0.018	**43.32±0.018**	0.367±0.013

Table 2. Results of class-incremental experiments on Split CIFAR-10 (p denotes the noise rate).

Methods	$p = 0.0$		$p = 0.2$	
	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)
EWC	18.36±0.23	0.714±0.018	18.01±0.23	0.717±0.007
MAS	14.36±1.13	**0.275±0.021**	11.66±1.59	0.546±0.020
AGEM	20.50±1.92	0.693±0.035	19.44±1.14	0.525±0.071
OGD	19.02±0.12	0.734±0.003	18.99±0.08	0.736±0.004
DER	29.87±2.52	0.586±0.084	22.62±3.27	0.685±0.029
GDumb	28.40±2.13	0.365±0.054	21.69±0.59	**0.227±0.010**
MEGA	33.16±3.12	0.471±0.106	24.84±2.26	0.475±0.091
STREAM	**35.35±3.52**	0.504±0.098	**26.17±1.32**	0.303±0.056

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Running Time Comparison

Table 3. Running time on Multiple Dataset and Split CIFAR-100.

Methods	Multiple Dataset (hours)	Split CIFAR-100 (hours)
EWC	0.16	1.31
MAS	0.17	1.31
AGEM	0.16	1.30
OGD	0.47	3.01
DER	0.18	1.27
GDumb	0.13	0.92
MEGA	0.15	1.05
STREAM	**0.11**	**0.79**

3. Failure Example

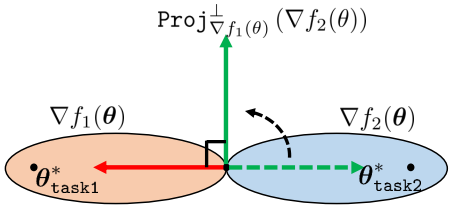
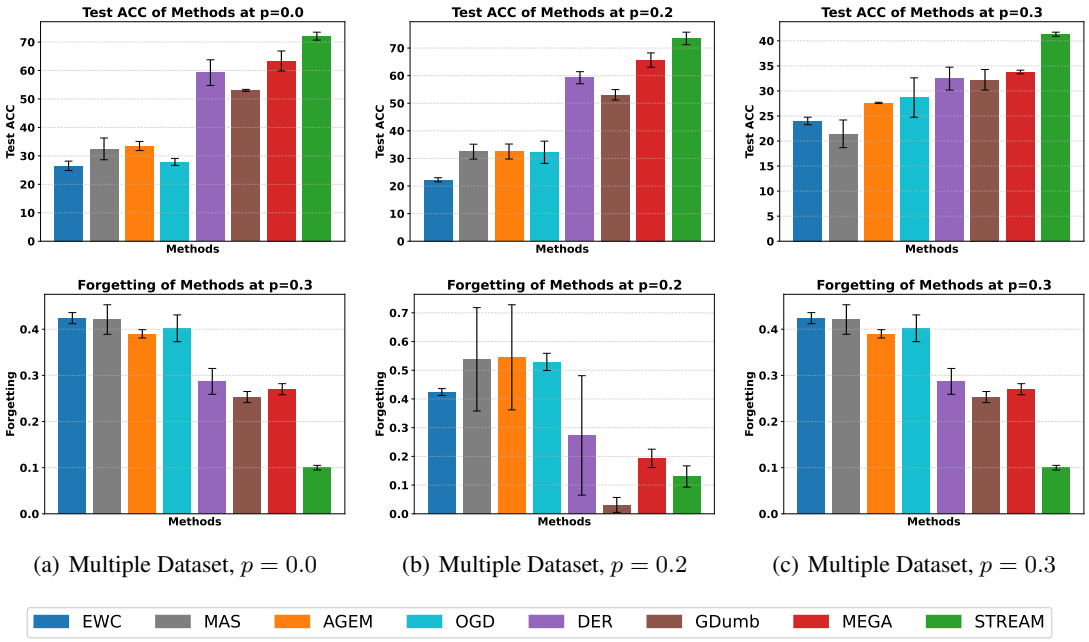


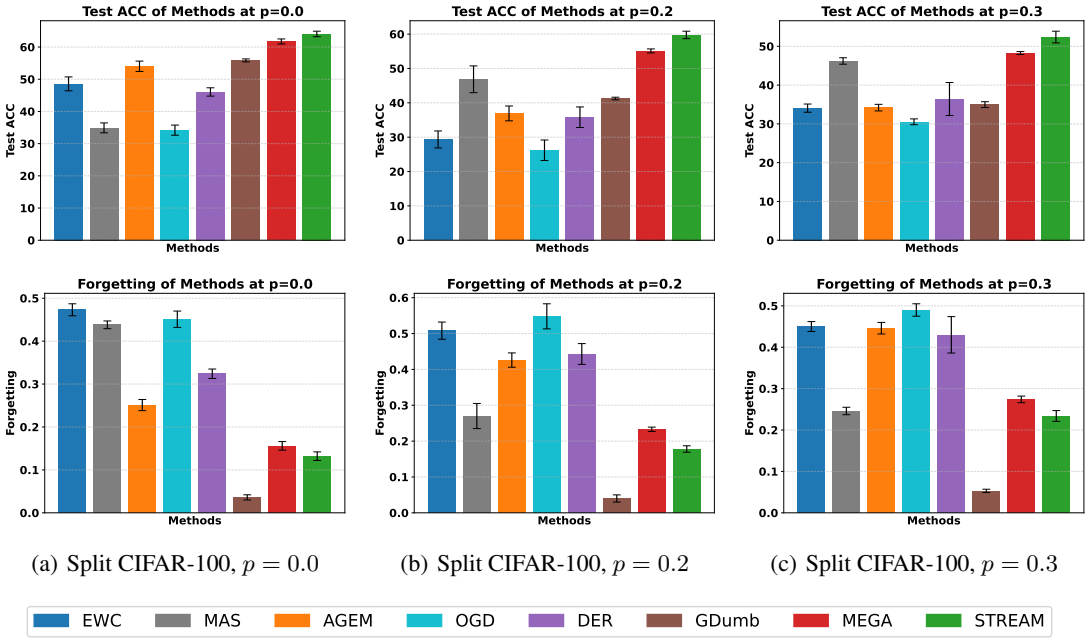
Figure 1. The Failure example for GEM, A-GEM and OGD, where the current gradient $\nabla f_2(\theta)$ is in the opposite direction to the memory gradient $\nabla f_1(\theta)$. The update direction is based on a rotation of the current gradient, which is orthogonal to the memory gradient: $\text{Proj}^{\perp}_{\nabla f_1(\theta)}(\nabla f_2(\theta))$. Therefore it cannot learn task 2.

4. Results with Multiple Runs



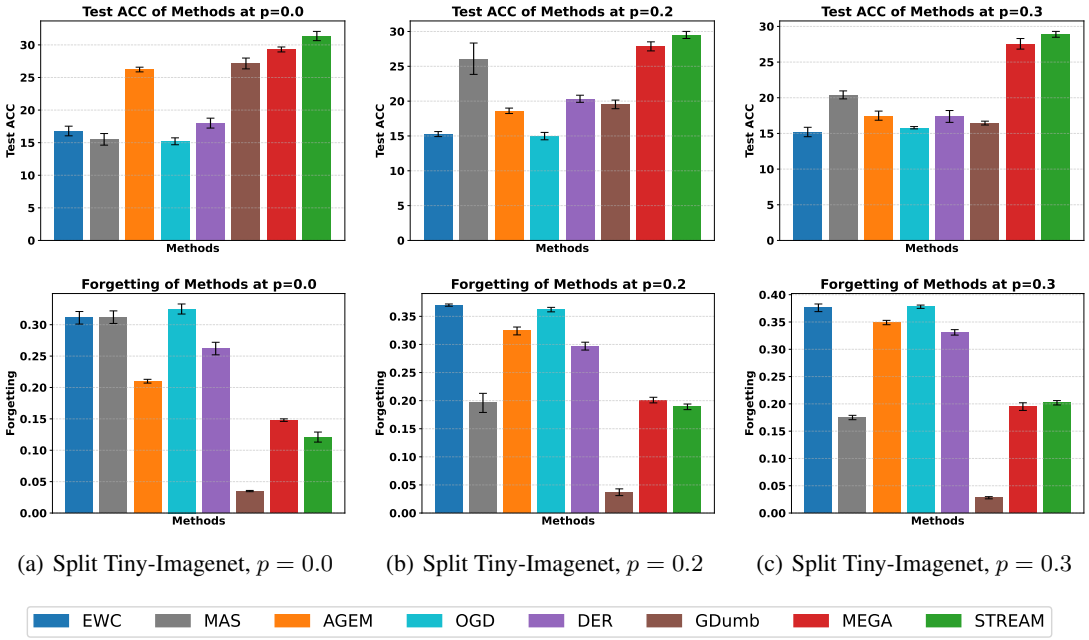
(a) Multiple Dataset, $p = 0.0$ (b) Multiple Dataset, $p = 0.2$ (c) Multiple Dataset, $p = 0.3$

Figure 2. Performance (test accuracy and forgetting) of continual learning methods on Multiple Dataset over 5 runs.



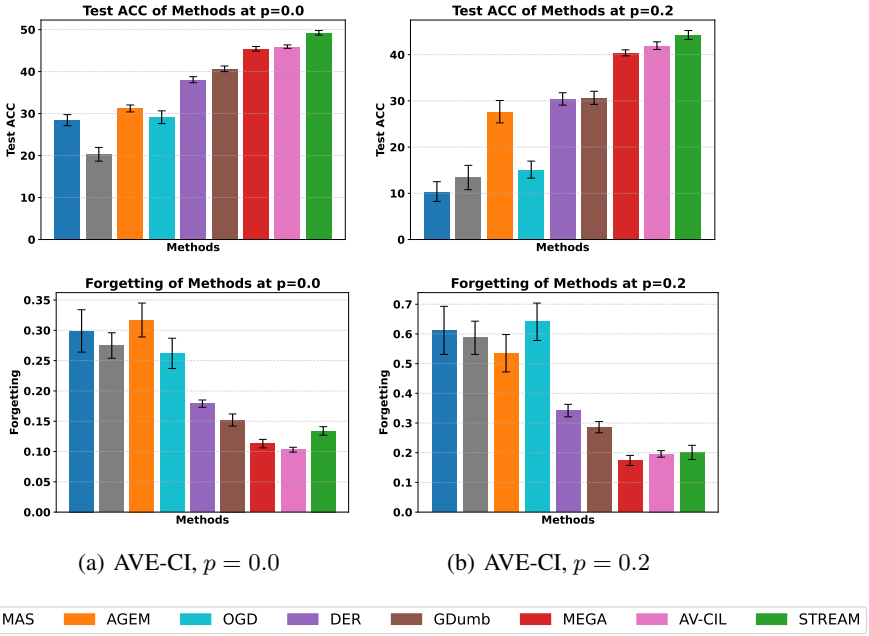
(a) Split CIFAR-100, $p = 0.0$ (b) Split CIFAR-100, $p = 0.2$ (c) Split CIFAR-100, $p = 0.3$

Figure 3. Performance (test accuracy and forgetting) of continual learning methods on Split CIFAR100 over 5 runs.



(a) Split Tiny-Imagenet, $p = 0.0$ (b) Split Tiny-Imagenet, $p = 0.2$ (c) Split Tiny-Imagenet, $p = 0.3$

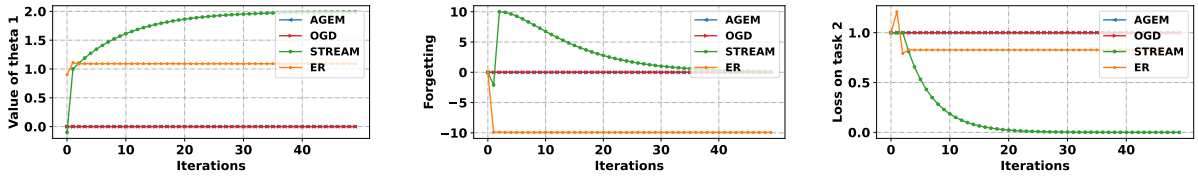
Figure 4. Performance (test accuracy and forgetting) of continual learning methods on Split Tiny-Imagenet over 5 runs.



(a) AVE-CI, $p = 0.0$ (b) AVE-CI, $p = 0.2$

Figure 5. Performance (test accuracy and forgetting) of continual learning methods on AVE-CI over 5 runs

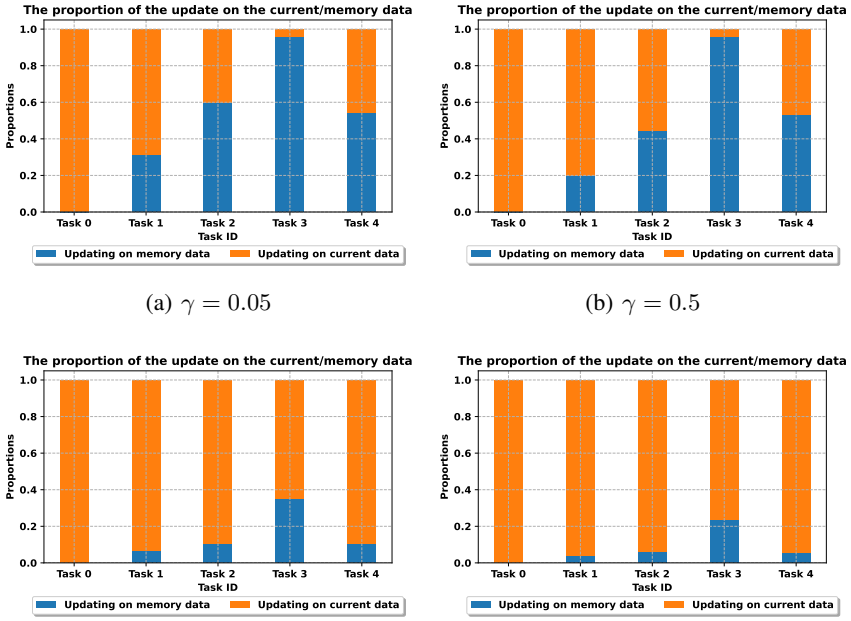
5. Synthetic Experiment for the counterexample



(a) The value of the $\theta^{(1)}$ vs. iterations. (b) The forgetting vs. iterations. (c) Loss on task 2 vs. iterations.

Figure 6. Synthetic experiment for the counterexample. (a), (b), (c) show the evolution of the value of $\theta^{(1)}$, the forgetting on task 1 and the loss on task 2. STREAM can find the optimal $\theta^{(1)}$ and achieve minimal forgetting and loss on the new task. But A-GEM and OGD fail to update their parameter $\theta^{(1)}$ toward the optimum throughout the training process, thus cannot minimize the loss on the new task. ER minimizes the loss functions f_1 and f_2 jointly (Assume that memory is large enough, it can visit f_1 as it needs). ER cannot find the the optimal value of $\theta^{(1)}$, thus can not achieve the low loss on task2. ER performs well in terms of forgetting, but STREAM still exhibits good forgetting metric (forgetting = 0).

6. The Frequency Statistics of Model Update on Current/Memory Data



(a) $\gamma = 0.05$ (b) $\gamma = 0.5$

(c) $\gamma = 0.7$ (d) $\gamma = 1.0$

Figure 7. The frequency statistics of model update on the current/memory data.

7. Hyperparameter Tuning

Table 4. Results vs. memory size (m is the memory size) on Multiple Dataset.

Methods	$m = 64$		$m = 128$		$m = 256$	
	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)
EWC	26.50±1.66	0.377±0.013	26.67±1.40	0.245±0.017	26.59±0.11	0.223±0.008
MAS	32.46±3.83	0.538±0.181	33.54±1.47	0.538±0.041	34.15±1.63	0.536±0.039
AGEM	33.47±1.61	0.541±0.179	36.04±0.94	0.480±0.005	41.90±1.52	0.424±0.026
OGD	27.88±1.23	0.375±0.015	32.24±1.34	0.544±0.051	33.63±1.24	0.303±0.022
DER	59.25±4.52	0.273±0.106	59.78±0.94	0.225±0.015	60.53±2.67	0.240±0.028
GDumb	53.03±0.34	**0.032±0.026**	58.64±0.42	**0.000±0.006**	69.11±2.83	**0.000±0.018**
MEGA	63.36±3.51	0.210±0.109	66.44±2.03	0.292±0.015	67.24±2.38	0.214±0.003
STREAM	**72.08±1.40**	0.152±0.035	**74.14±2.01**	0.145±0.016	**74.07±0.95**	0.025±0.048

Table 5. Results vs. memory size (m is the memory size) Split Tiny ImageNet.

Methods	$m = 256$		$m = 512$		$m = 1024$	
	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)
EWC	16.79±0.74	0.311±0.010	17.15±0.44	0.306±0.002	17.05±0.10	0.324±0.003
MAS	15.51±0.89	0.312±0.010	17.03±0.08	0.332±0.008	18.56±0.52	0.316±0.000
AGEM	26.22±0.36	0.210±0.003	27.95±0.58	0.217±0.004	32.97±0.15	0.181±0.004
OGD	15.21±0.53	0.325±0.008	16.53±0.67	0.357±0.007	17.15±0.52	0.347±0.005
DER	18.00±0.76	0.262±0.010	21.46±1.13	0.236±0.005	23.10±0.51	0.214±0.007
GDumb	27.15±0.83	**0.035±0.001**	32.81±0.25	**0.040±0.004**	33.51±0.14	**0.021±0.002**
MEGA	29.30±0.38	0.148±0.002	32.87±0.63	0.137±0.007	33.20±0.44	0.104±0.001
STREAM	**31.36±0.71**	0.121±0.008	**33.02±0.53**	0.115±0.003	**34.67±0.14**	0.104±0.002

Table 6. Results vs. the number of tasks (T denotes the number of tasks) on Split CIFAR-100.

Methods	$T = 10$		$T = 20$		$T = 25$	
	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)
EWC	27.98±1.57	0.403±0.013	48.56±2.16	0.473±0.014	47.43±3.23,	0.530±0.033
MAS	31.40±2.34	0.380±0.021	34.90±1.54	0.438±0.009	31.68±2.16	0.496±0.022
AGEM	39.72±2.43	0.292±0.019	54.02±1.61	0.251±0.013	48.19±1.65	0.345±0.019
OGD	29.13±1.52	0.262±0.025	34.19±1.57	0.451±0.019	36.35±1.43	0.353±0.023
DER	41.99±1.80	0.264±0.018	46.05±1.29	0.324±0.011	49.65±0.90	0.334±0.008
GDumb	38.80±1.08	**0.058±0.002**	55.85±0.46	**0.036±0.006**	67.56±0.53	**0.008±0.009**
MEGA	48.03±1.44	0.176±0.003	61.74±0.77	0.156±0.010	68.55±0.57	0.094±0.007
STREAM	**50.33±0.66**	0.167±0.008	**64.06±0.86**	0.132±0.010	**69.70±0.37**	0.080±0.010

8. New Advanced Baselines

Table 7. Results of task-incremental experiments on Multiple Dataset and Split Tiny-Imagenet (p denotes the noise rate).

	Methods	$p = 0.0$		$p = 0.2$	
		ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)
Multiple Dataset	OCS	55.65±2.26	**0.062±0.001**	45.03±4.16	**0.049±0.012**
	MetaSP	57.14±1.10	0.113±0.042	47.14±1.66	0.081±0.027
	STREAM	**72.08±1.40**	0.152±0.035	**73.50±2.25**	0.130±0.037
Split Tiny-Imagenet	OCS	41.29±0.09	**0.112±0.001**	35.36±0.94	**0.061±0.005**
	MetaSP	43.33±0.32	0.127±0.002	37.18±0.76	0.068±0.007
	STREAM	**47.92±0.45**	0.153±0.018	**43.32±0.02**	0.367±0.013