A. Algorithm

Algorithm 3: Adaptive Sampling Algorithm

Input: Replay Buffer: $\mathcal{X}$, Replay Buffer weights: W, Candidate Pool: $\mathcal{C}_t$, Task sample count: n_t, Entire sample count: n

Output: Data sample: {(x, y, z, w)}

Calculate probability: $p = \frac{n_t}{n}$

Sample a random number: $p_f = random([0, 1])$

if $p_f \leq p$ **then**

  Sample a index randomly: $i = random([1, |\mathcal{C}_t|])$

  $(x, y, z, w) = (\mathcal{C}_t[i], 1)$

else

  Sample a index randomly: $i = random([1, |\mathcal{X}|])$

  $(x, y, z, w) = (\mathcal{X}[i], W[i])$

B. Additional Results

B.1. Forgetting metric

In this section we present additional results for the experiments shown in Section 6.1 and 6.2. We report the *forgetting* metric (FRG) which shows how much the accuracy of learnt tasks over time as the continual models tries to learn subsequent tasks. The average forgetting across the tasks is reported in Table 6 (offline setting), and Table 7 (online streaming setting). It is worth noting that FRG should only be seen along with final accuracy to draw comparisons between two continual learning models. This is because FRG alone can be misleading–a model which does not learn any subsequent tasks throughout the training will give near to 0 forgetting but will give random final accuracy which is undesirable. A clear example is that of iCaRL [36] in the offline setting; we see that the method has poor overall accuracy (Table 1) but highly favorable forgetting metrics (Table 6).

B.2. OCS vs GCR

Table 8 compares published performance numbers from one setting, for the OCS algorithm [47], against our trained model in the same setting. Our approach shows better performance in the comparison, suggesting that our gradient approximation objective is superior to the gradient diversity-based selection objective of OCS. However, this comparison is incomplete; the authors of OCS have not made their code available for comparison, the paper's description of the algorithm was insufficient for reproduction, and they did not publish numbers on any of the other settings, datasets, buffer sizes we explored in our paper.

C. Generality of gradient-based coresets

We also examined whether the gains from our gradient approximation procedure for coreset selection (Section 4.1) were dependent on the specific loss function that we use for CL (Section 4.2). To evaluate this, we enhanced ER [37] (a simple replay-based continual learning procedure that does not use the distillation loss from Section 4.2) with our gradient approximation procedure. The results in Table 9 show that the gains from GCR are robust, and apply to other replay-based methods as well. Other baseline methods like iCARL, GSS, etc have specific replay buffer selection methods, unlike ER which uses random samples, and it was not clear how to add GCR on top of those methods. In any case, our results show that GCR beats those methods by significant margins.

Finally, we conducted experiments on the significantly more difficult S-Imagenet-1k dataset [40], comprising high-resolution Imagenet images broken down into 5 tasks of 200 categories each. Table 10 shows that GCR outperforms ER and DER++ by significant margin. Note, however, that all three methods have fairly low accuracy on the task overall; this is expected given that the task is significantly harder than S-Cifar100.

D. Implementation details

D.1. Hyperparameter Search

Table 11 shows the hyperparameter values selected from the grid search that were used in our experiments.

D.2. Hyperparameter Search Space

Table 12 shows the hyperparameter search space for offline and online setting on which grid search was done.

Setting	Method	S-Cifar-10 K=200	S-Cifar-10 K=500	S-Cifar-10 K=2000	S-Cifar-100 K=200	S-Cifar-100 K=500	S-Cifar-100 K=2000	S-TinyImageNet K=200	S-TinyImageNet K=500	S-TinyImageNet K=2000
Class-IL	ER	59.3±2.48	43.22±2.1	23.85±1.09	75.06±0.63	67.96±0.78	49.12±0.57	76.53±0.51	75.21±0.54	65.58±0.53
	GEM	80.36±5.25	78.93±6.53	82.33±5.83	77.4±1.09	71.34±0.78	55.27±1.37	-	-	-
	GSS	72.48±4.45	59.18±4.0	44.59±6.13	77.62±0.76	74.12±0.42	67.42±0.62	76.47±0.4	75.3±0.26	72.49±0.43
	iCARL	**23.52±1.27**	28.2±2.41	21.91±1.15	**47.2±1.23**	40.99±1.02	30.64±1.85	**31.06±1.91**	**37.3±1.42**	39.88±1.51
	DER	35.79±2.59	24.02±1.63	12.92±1.1	62.72±2.69	49.07±2.54	28.18±1.93	64.83±1.48	59.95±2.31	39.83±1.15
	GCR	32.75±2.67	**19.27±1.48**	**8.23±1.02**	57.65±2.48	**39.2±2.84**	**19.29±1.83**	65.29±1.73	56.4±1.08	**32.45±1.79**
Task-IL	ER	6.07±1.09	3.5±0.53	1.37±0.44	27.38±1.46	17.37±1.06	8.03±0.66	40.47±1.54	30.73±0.62	18.0±0.83
	GEM	9.57±2.05	5.6±0.96	2.95±0.81	29.59±1.66	20.44±1.13	9.5±0.73	-	-	-
	GSS	8.49±2.05	6.37±1.55	4.31±1.68	32.81±1.75	26.57±1.34	18.98±1.13	50.75±1.63	45.59±0.99	38.05±1.17
	iCARL	25.34±1.64	22.61±3.97	24.47±1.36	36.2±1.85	27.9±1.37	16.99±1.76	42.47±2.47	39.44±0.84	30.45±2.18
	DER	6.08±0.7	3.72±0.55	1.95±0.32	25.98±1.55	25.98±1.55	7.37±0.85	40.43±1.05	28.21±0.97	15.08±0.49
	GCR	7.38±1.02	**3.14±0.36**	**1.24±0.27**	**24.12±1.17**	**15.07±1.88**	**5.75±0.72**	40.36±1.08	**27.88±1.19**	**13.1±0.57**

Table 6. Forgetting metric in Offline Class-IL and Task-IL Continual Learning

Method	S-Cifar-10 K=200	S-Cifar-10 K=500	S-Cifar-10 K=2000	S-Cifar-100 K=200	S-Cifar-100 K=500	S-Cifar-100 K=2000	S-TinyImageNet K=200	S-TinyImageNet K=500	S-TinyImageNet K=2000
ER	47.01±6.63	38.72±7.94	31.96±8.93	30.16±0.69	26.29±1.31	16.42±2.17	27.86±1.69	32.53±1.18	27.91±1.41
GEM	73.63±3.96	73.07±6.58	53.27±10.93	32.94±2.88	27.15±3.78	29.97±7.12	-	-	-
GSS	48.8±6.56	40.62±6.74	40.67±5.75	33.06±1.05	25.37±1.93	19.56±1.64	36.91±1.44	32.67±1.36	23.63±1.18
iCARL	23.78±3.64	26.2±4.31	**22.11±4.61**	**9.53±0.57**	**9.15±0.49**	**8.9±0.49**	**6.95±0.5**	**7.22±0.38**	**6.89±0.37**
DER	34.12±7.04	29.05±8.59	27.5±8.69	26.84±1.7	22.92±2.73	13.72±2.03	31.68±1.46	27.09±0.79	14.97±1.28
GCR	26.7±8.37	**20.1±3.32**	**22.18±9.9**	21.86±1.77	19.46±1.72	17.91±2.3	34.19±1.07	27.47±0.8	22.31±1.35

Table 7. Forgetting metric in Online Continual Learning.

Method	OCS vs GCR S-Cifar 100 (20 Tasks)
OCS	60.5±0.55
GCR	60.86±3.53

Table 8. Comparing OCS with GCR for Task-IL setting of S-Cifar-100 (20 tasks) and buffer size of 100.

	S-Cifar-10				S-Cifar-100			
	Class-IL		Task-IL		Class-IL		Task-IL	
Method	K=500	K=2000	K=500	K=2000	K=500	K=2000	K=500	K=2000
ER	62.03±1.70	77.13±0.87	93.82±0.41	96.01±0.28	27.66±0.61	42.80±0.49	66.23±1.52	74.67±1.2
ER+GCR	**66.66±2.1**	**80.15±1.17**	94.17±0.46	**96.47±0.22**	**30.68±0.47**	**47.09±1.08**	**70.25±0.81**	**78.59±0.5**

Table 9. GCR coreset selection with ER method. Numbers represent mean ± SEM of model test accuracy over 15 runs. Best-performing models in each column are bolded (paired t-test, $p < 0.05$).

Method	Class-IL	Task-IL
ER [36]	7.43	15.32
DER++ [6]	10.22	17.79
GCR	**11.33**	**19.03**

Table 10. Scaling up to S-ImageNet1k (5 tasks, buffer size 1000)

Offline Class-IL				
Method	Buffer Size	S-Cifar10	S-Cifar100	S-Tinyimg
ER	200	lr: 0.1	lr: 0.1	lr: 0.03
	500	lr: 0.03	lr: 0.1	lr: 0.1
	2000	lr: 0.1	lr: 0.1	lr: 0.03
GEM	200	lr: 0.01 γ: 1.0	lr: 0.03 γ: 0.5	
	500	lr: 0.01 γ: 0.5	lr: 0.1 γ: 0.5	-
	2000	lr: 0.1 γ: 0.5	lr: 0.03 γ: 1.0	
GSS	200	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1
	500	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1
	2000	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1
iCARL	200	lr: 0.1 wd: 5e-5	lr: 0.1 wd: 5e-5	lr: 0.1 wd: 1e-5
	500	lr: 0.01 wd: 1e-5	lr: 0.1 wd: 5e-5	lr: 0.03 wd: 1e-5
	2000	lr: 0.1 wd: 1e-5	lr: 0.1 wd: 1e-5	lr: 0.03 wd: 1e-5
DER	200	lr: 0.03 α: 0.2 β: 1.0	lr: 0.03 α: 0.5 β: 0.1	lr: 0.03 α: 0.2 β: 0.1
	500	lr: 0.03 α: 0.1 β: 1.0	lr: 0.03 α: 0.5 β: 0.1	lr: 0.03 α: 0.2 β: 0.1
	2000	lr: 0.03 α: 0.2 β: 1.0	lr: 0.03 α: 0.2 β: 0.1	lr: 0.03 α: 0.1 β: 0.5
GCR	200	lr: 0.03 α: 0.5 β: 0.5 γ: 0.01	lr: 0.03 α: 0.2 β: 0.1 γ: 0.01	lr: 0.03 α: 0.5 β: 0.5 γ: 0.01
	500	lr: 0.03 α: 0.1 β: 0.1 γ: 0.1	lr: 0.03 α: 0.1 β: 0.1 γ: 0.01	lr: 0.03 α: 0.5 β: 0.1 γ: 0.01
	2000	lr: 0.03 α: 0.1 β: 1.0 γ: 0.1	lr: 0.03 α: 0.2 β: 0.1 γ: 0.1	lr: 0.03 α: 0.2 β: 1.0 γ: 0.01
Offline Task-IL				
Method	Buffer Size	S-Cifar10	S-Cifar100	S-Tinyimg
ER	200	lr: 0.01	lr: 0.03	lr: 0.1
	500	lr: 0.1	lr: 0.1	lr: 0.1
	2000	lr: 0.03	lr: 0.1	lr: 0.03
GEM	200	lr: 0.01 γ: 1.0	lr: 0.1 γ: 0.5	
	500	lr: 0.03 γ: 0.5	lr: 0.03 γ: 0.5	-
	2000	lr: 0.03 γ: 0.5	lr: 0.1 γ: 0.5	
GSS	200	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1
	500	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1
	2000	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1
iCARL	200	lr: 0.1 wd: 5e-5	lr: 0.1 wd: 5e-5	lr: 0.1 wd: 1e-5
	500	lr: 0.01 wd: 5e-5	lr: 0.1 wd: 5e-5	lr: 0.03 wd: 1e-5
	2000	lr: 0.01 wd: 5e-5	lr: 0.1 wd: 1e-5	lr: 0.03 wd: 1e-5
DER	200	lr: 0.03 α: 0.2 β: 0.1	lr: 0.03 α: 0.1 β: 0.1	lr: 0.03 α: 0.1 β: 0.1
	500	lr: 0.03 α: 0.2 β: 0.5	lr: 0.03 α: 0.1 β: 0.1	lr: 0.03 α: 0.1 β: 0.5
	2000	lr: 0.03 α: 0.2 β: 1.0	lr: 0.03 α: 0.1 β: 0.5	lr: 0.03 α: 0.1 β: 0.1
GCR	200	lr: 0.03 α: 0.1 β: 0.1 γ: 0.1	lr: 0.03 α: 0.1 β: 0.1 γ: 0.01	lr: 0.03 α: 0.1 β: 0.5 γ: 0.01
	500	lr: 0.03 α: 0.2 β: 0.5 γ: 0.01	lr: 0.03 α: 0.1 β: 0.1 γ: 0.05	lr: 0.03 α: 0.1 β: 1.0 γ: 0.01
	2000	lr: 0.03 α: 0.2 β: 0.5 γ: 0.05	lr: 0.03 α: 0.1 β: 0.1 γ: 0.1	lr: 0.03 α: 0.1 β: 1.0 γ: 0.01
Online Streaming				
Method	Buffer Size	S-Cifar10	S-Cifar100	S-Tinyimg
ER	200	lr: 0.03	lr: 0.01	lr: 0.1
	500	lr: 0.01	lr: 0.01	lr: 0.01
	2000	lr: 0.01	lr: 0.03	lr: 0.01
GEM	200	lr: 0.1 γ: 0.5	lr: 0.03 γ: 0.5	lr: 0.03 γ: 0.5
	500	lr: 0.03 γ: 1.0	lr: 0.1 γ: 0.5	lr: 0.03 γ: 1.0
	2000	lr: 0.1 γ: 1.0	lr: 0.03 γ: 1.0	lr: 0.03 γ: 1.0
GSS	200	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1	lr: 0.1 gmbs: 32 nb: 1
	500	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1	lr: 0.1 gmbs: 32 nb: 1
	2000	lr: 0.03 gmbs: 32 nb: 1	lr: 0.03 gmbs: 32 nb: 1	lr: 0.1 gmbs: 32 nb: 1
iCARL	200	lr: 0.1 wd: 1e-5	lr: 0.1 wd: 5e-5	lr: 0.1 wd: 5e-5
	500	lr: 0.1 wd: 1e-5	lr: 0.1 wd: 5e-5	lr: 0.1 wd: 5e-5
	2000	lr: 0.1 wd: 1e-5	lr: 0.1 wd: 5e-5	lr: 0.1 wd: 5e-5
MIR	200	lr: 0.03	lr: 0.03	lr: 0.03
	500	lr: 0.03	lr: 0.03	lr: 0.03
	2000	lr: 0.03	lr: 0.03	lr: 0.03
DER	200	lr: 0.03 α: 0.1 β: 1.0	lr: 0.03 α: 0.5 β: 0.5	lr: 0.03 α: 0.2 β: 2.0
	500	lr: 0.01 α: 0.2 β: 2.0	lr: 0.03 α: 1.0 β: 0.5	lr: 0.005 α: 1.0 β: 3.0
	2000	lr: 0.01 α: 0.2 β: 1.0	lr: 0.01 α: 1.0 β: 2.0	lr: 0.01 α: 1.0 β: 3.5
GCR	200	lr: 0.03 α: 0.2 β: 1.0 γ: 1.0	lr: 0.005 α: 1.0 β: 3.5 γ: 1.0	lr: 0.01 α: 1.0 β: 3.0 γ: 0.1
	500	lr: 0.005 α: 0.5 β: 3.5 γ: 1.0	lr: 0.01 α: 0.2 β: 3.0 γ: 1.5	lr: 0.005 α: 1.0 β: 3.5 γ: 1.0
	2000	lr: 0.03 α: 0.1 β: 0.5 γ: 1.5	lr: 0.01 α: 1.0 β: 2.0 γ: 1.0	lr: 0.01 α: 1.0 β: 3.0 γ: 1.5

Table 11. Hyperparameter values obtained from the grid search.

Method	Parameters	Offline	Online
ER	lr	[0.01, 0.03, 0.1]	[0.01, 0.03, 0.1]
GEM	lr	$[0.01, 0.03, 0.1]$	$[0.01, 0.03, 0.1]$
	γ	$[0.5, 1.0]$	$[0.5, 1.0]$
GSS	lr	$[0.01, 0.03]$	$[0.005, 0.01, 0.03, 0.1]$
iCARL	lr	$[0.01, 0.03, 0.1]$	$[0.01, 0.03, 0.1]$
	wd	$[1e-5, 5e-5]$	$[1e-5, 5e-5]$
MIR	lr	-	$[0.01, 0.03]$
DER	lr	[0.03]	$[0.005, 0.01, 0.03]$
	α	$[0.1, 0.2, 0.5, 1.0]$	$[0.1, 0.2, 0.5, 1.0]$
	β	$[0.1, 0.5, 1.0]$	$[0.1, 0.5, 1.0, 2.0, 3.0, 3.5]$
GCR	lr	$[0.03]$	$[0.005, 0.01, 0.03]$
	α	$[0.1, 0.2, 0.5, 1.0]$	$[0.1, 0.2, 0.5, 1.0]$
	β	$[0.1, 0.5, 1.0]$	$[0.1, 0.5, 1.0, 2.0, 3.0, 3.5]$
	γ	$[0, 0.01, 0.05, 0.1, 0.2]$	$[0, 0.1, 0.5, 1.0, 1.5]$

Table 12. Hyperparameter Search Space.