Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?
Appendix

A. Effect of Sampling Strategies

In Figure 4 of the manuscript, we present the baseline ER for two sampling strategies, namely FIFO and uniform. FIFO selects the latest seen samples to train on, whereas Uniform simply randomly and uniformly samples a set of previously seen samples to train on. Mixed sampling is a mix of both FIFO and Uniform, where half of the batch is constructed using FIFO sampling, and the other half using Uniform sampling.

A.1. CGLM

The results for various samplers for both Online Accuracy (Figure 1) and Near-Future Accuracy (Figure 2) on CGLM dataset are summarized below:

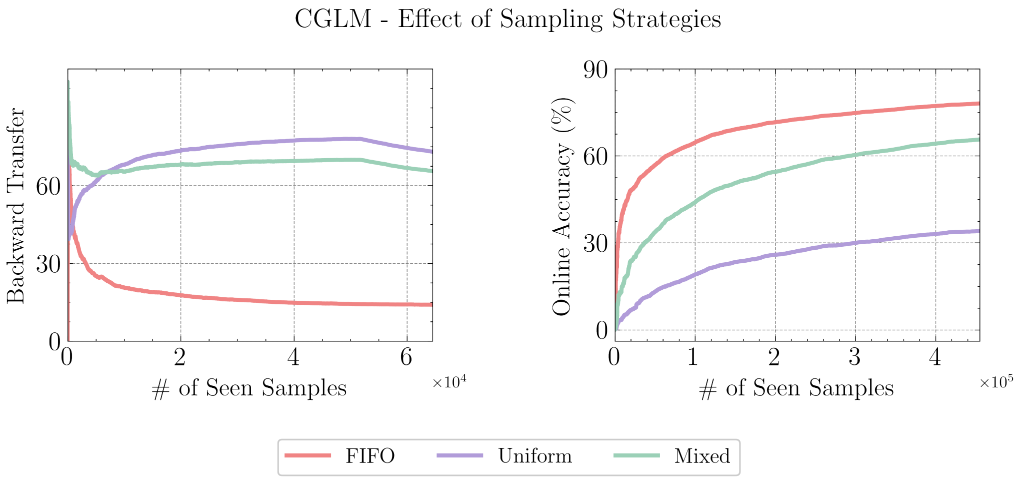

Figure 1. **Effect of Sampling Strategy on Online Accuracy.** In terms of online accuracy, FIFO sampling which focuses on the latest samples performs best where uniform performs the worst and mixed performs somewhere in the middle. In terms of retetion however, the order is reversed.

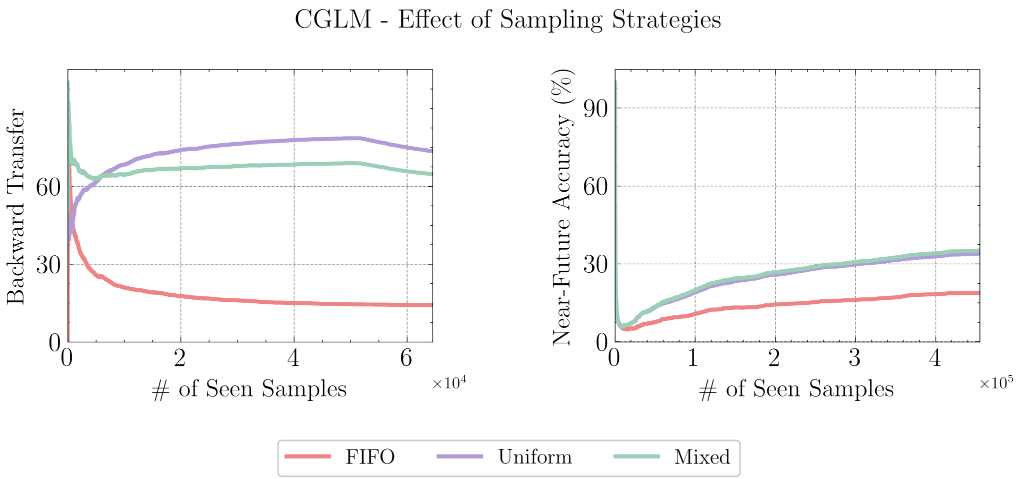

Figure 2. **Effect of Sampling Strategy on Near-Future Accuracy.** In terms of near-future accuracy, Uniform and Mixed sampling perform almost the same, however FIFO sampling is no where close to them. In terms of retention, Uniform still takes the lead.

Conclusion. Interestingly, mixed sampling is competitive with uniform sampling in near future accuracy, but performs worse in information retention.

A.2. CLOC

The results for various samplers for both Online Accuracy (Figure 3) and near-future accuracy (Figure 4) on CLOC dataset are summarized below.

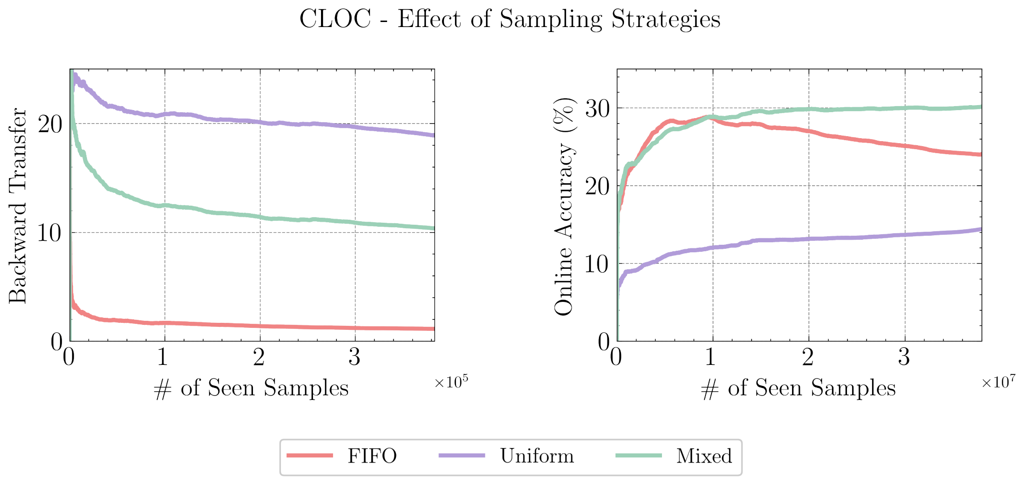

Figure 3. **Effect of Sampling Strategy on Online Accuracy.** Unlike what was observed for CLGM, for CLOC, the mixed sampling performs the best in terms of online accuracy followed followed by FIFO and then Uniform. However, Uniform still performs the best in terms of information retention with a large gap with a large margin.

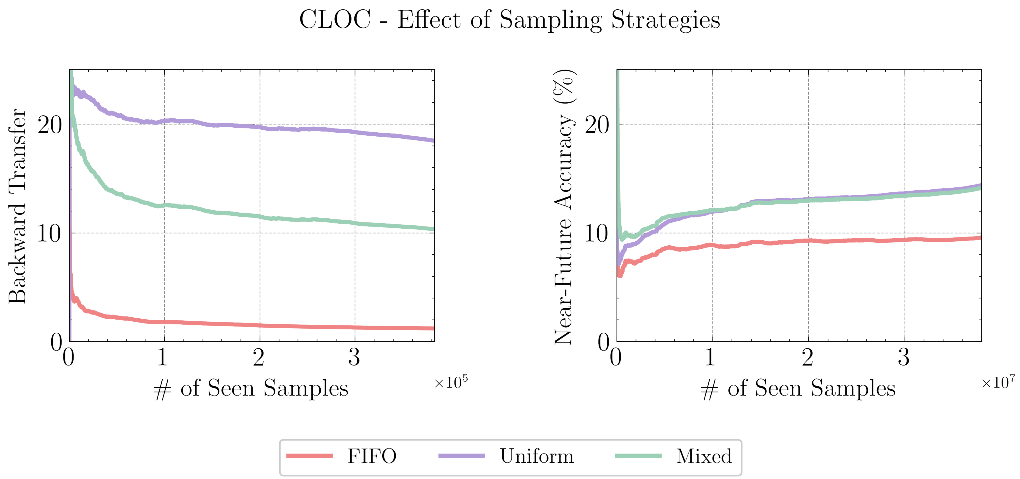

Figure 4. **Effect of Sampling Strategy on Near-Future Accuracy** In terms of near-future accuracy, mixed outperforms other samplers by a significant margin. However, it achieves significantly worse performance compared to uniform sampling in terms of information retention.

Conclusion. Mixed sampling is competitive with uniform sampling in near future accuracy, however, its information retention capabilities are half that of the ER baseline.

B. Sensitivity Analyis: Learning Rate and Weight Decay

In OCL, changing hyperparameters across datasets is uncertain as the stream might have a significant distribution shift from the pretrain. Hence, we use the hyperparameters for our model from Ghunaim *et al.* [1] for all our experiments. However, how do the selected hyperparameters transfer to CGLM is an interesting question. We demonstrate the sensitivity of the hyperparameters: learning rate and weight decay below:

B.1. Sensitivity to Weight Decay

Results and Conclusion. We present our results in Figure 5. We conclude that weight decay has minimal effect on the performance of the ER (Replay Only) method.

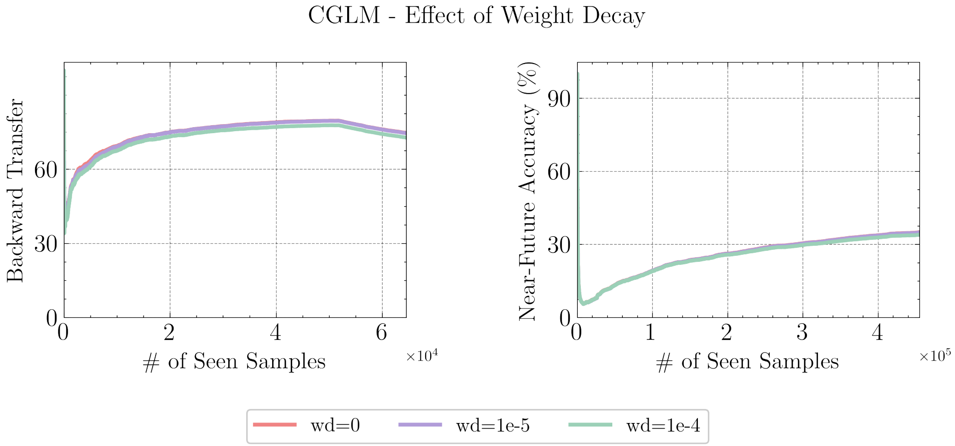

Figure 5. **Sensitivity to Weight Decay on ER (Replay Only).** Weight decay seems to have minimal effect on both near-future accuracy and backward transfer.

B.2. Sensitivity to Learning Rate

Results and Conclusion. We present our results in Figure 6. An order magnitude change in learning rate in either direction leads to a decrease in both information retention performance and near-future accuracy. The learning rate of 0.005 transfers well to CGLM dataset in terms of both near-future accuracy and information retention (Backward Transfer).

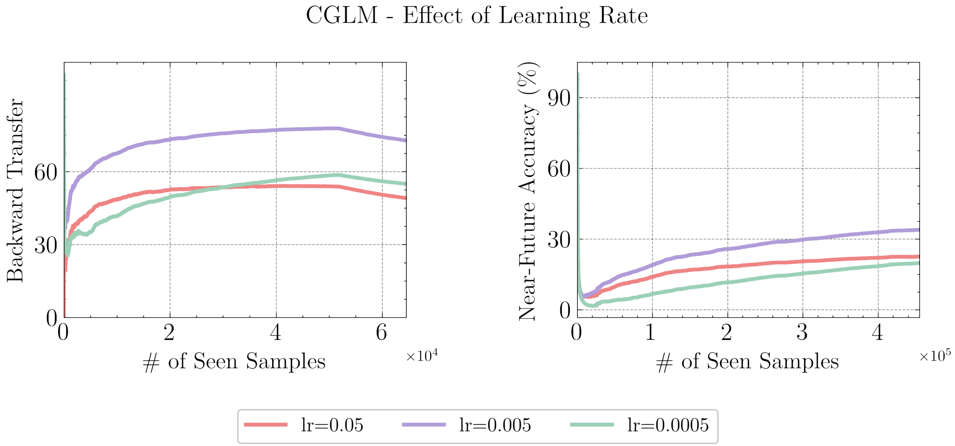

Figure 6. **Effect of Learning Rate on the ER Baseline on CGLM.** Both near future-accuracy and backward transfer are highly sensitive to the selection of the learning rate.

C. Fixed-Feature Extractor Based Methods: NCM, SDLA, and ACM

In this section, we compare ACM with other popular fixed-feature extractor based methods like NCM [4, 5, 3] and SLDA [2]. The results for running NCM, SLDA, and ACM on CGLM dataset are shown in Figure 7 where we see that ACM performs best in terms of backward transfer, online accuracy, and near-future accuracy.

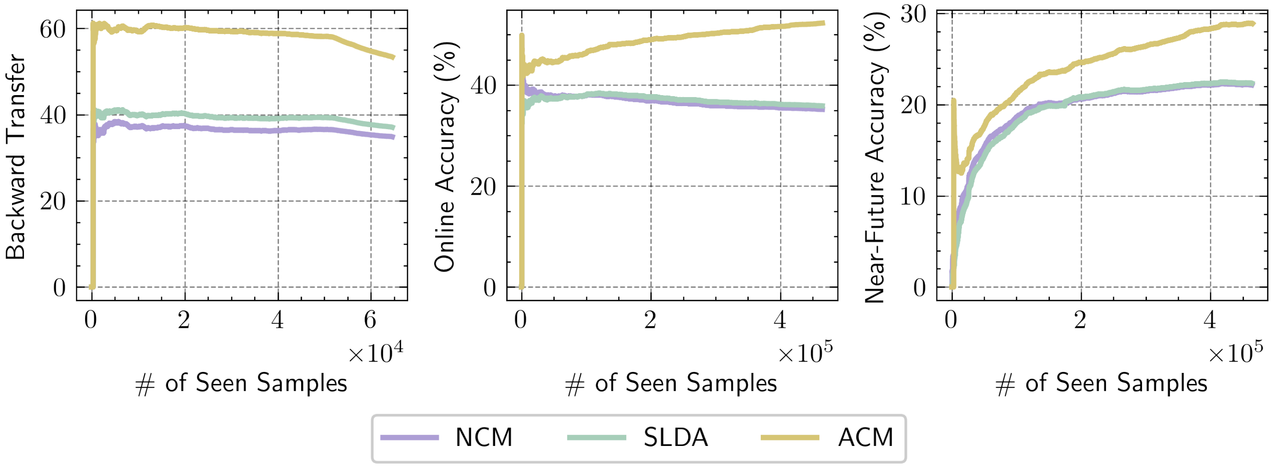

Figure 7. **Fixed-Feature Extractor Based Methods.** When compared to NCM and SLDA, ACM performs the best in terms of Backward Transfer, Online Accuracy and Near-Future Accuracy metrics.

D. Limitations & Future Directions

Limitations. While our work discovers interesting phenomena in OCL, it has important limitations:

- **Is near-future accuracy definitively measuring adaptation?** It is unclear why our proposed evaluation approach definitively measures rapid adaptation, and further investigation is needed to determine the problem exists in future OCL scenarios.
- **Dependency on shift $\mathcal{S}$.** Our proposed metric, near-future accuracy, depends on the calculated value (fixed) for the shift, $\mathcal{S}$, however in reality the stream label correlations could be dynamically changing with the stream and an adaptive value of S would be required in that case.
- **Mitigating Monitoring Costs.** In our experiments, we did not explicitly store the model itself as we were able to access future samples solely for evaluation purposes. In practice, a larger memory allocation would be necessary. To reduce storage costs we recommend performing this evaluation periodically instead of on every incoming sample to reduce storage requirements.
- **Why use a new metric instead of changing the data stream?** A question that might arise is why use a new metric if we could simply remove the correlated samples? The data streams used in our work are *naturally* ordered by timestamps, therefore changing the datasets would make it less natural or even complete unnatural if the correlation is long enough. Therefore, in our work we propose a new metric which could address the label correlations without modifying the datasets.

Future Directions. Our work leads to some interesting questions and problems to be explored:

- **Beyond Label Correlations.** We currently only measure and remove correlations from labels $p(y)$, while covariate correlations $p(X)$ are harder to reliably isolate, requiring further investigation.
- **Why aren't information retention and adaptation at odds?** Our results suggest that there is still a lot of room for improvement in OCL methods, as we are far from the pareto frontier where information retention and rapid adaptation are at odds.

References

[1] Yasir Ghunaim, Adel Bibi, Kumail Alhamoud, Motasem Alfarra, Hasan Abed Al Kader Hammoud, Ameya Prabhu, Philip HS Torr, and Bernard Ghanem. Real-time evaluation in online continual learning: A new paradigm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[2] Tyler L. Hayes and Christopher Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, 2020. 5

[3] Paul Janson, Wenxuan Zhang, Rahaf Aljundi, and Mohamed Elhoseiny. A simple baseline that questions the use of pretrained-models in continual learning. *arXiv preprint arXiv:2210.04428*, 2022. 5

[4] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *TPAMI*, 2013. 5

[5] Oleksiy Ostapenko, Timothee Lesort, Pau Rodríguez, Md Rifat Arefin, Arthur Douillard, Irina Rish, and Laurent Charlin. Continual learning with foundation models: An empirical study of latent replay. In *CoLLAs*, 2022. 5