Analysis of Class Incremental Learning: Role of Learning Rate, Number of Classes, and Dataset Size

Aditya Somasundaram* V Sai Vignesh *

April 2025

Abstract

We present a systematic empirical study of cyclic class-incremental learning (CIL) on the MNIST-1D benchmark, examining how learning rate, per-class sample size, and total number of classes jointly shape learning dynamics, catastrophic forgetting, and the recently observed "latching" phenomenon. By training a simple two-layer perceptron for 5,000 class-wise cycles under a range of hyperparameter settings, we (1) validate theoretical bounds on forgetting in cyclic regimes, (2) quantify how higher learning rates can compensate for fewer samples (and vice versa), and (3) uncover persistent, non-zero forgetting even after extensive training. Moreover, we derive practical scaling laws that relate optimal learning rates and sample sizes to the number of classes, and we demonstrate a pronounced last-in-first-out bias: early classes maintain $> 90\%$ accuracy while later classes fluctuate by 20–40% across cycles. Notably, our framework achieves a test accuracy of 86% on the MNIST-1D dataset, significantly outperforming both our recalculated baseline (72%) and the previously reported baseline (68%). These findings offer concrete guidelines for hyperparameter selection in CIL and reveal new avenues for closing the remaining gap to zero forgetting.

1 Introduction

Class-Incremental Learning (CIL) addresses a critical challenge in modern machine learning: enabling models to learn new concepts incrementally while retaining knowledge of previously encountered classes. Unlike traditional machine learning paradigms that assume static datasets, CIL operates in dynamic environments where data arrives sequentially, often with novel classes emerging over time. This approach mirrors how humans learn throughout their lives, making it crucial for developing truly adaptive intelligent systems.

*Department of Electrical Engineering, Columbia University

1.1 Core Challenge: Catastrophic Forgetting

The defining obstacle in CIL is catastrophic forgetting—the tendency of neural networks to abruptly lose performance on earlier tasks when trained on new data [13, 14, 51]. This phenomenon occurs because neural parameters optimized for new classes overwrite those critical for recognizing previous ones [50]. For example, a model trained to distinguish dogs and cats might lose this ability when later taught to recognize birds, unless specifically designed to prevent forgetting.

Class incremental learning differs from related paradigms through three fundamental properties.

- Expanding Label Space: Each training phase introduces entirely new classes, requiring the model to progressively increase its discrimination capacity.
- Data Constraints: Only current class data and a limited exemplar set from previous classes are available during training.
- Unified Evaluation: The model must maintain performance across all classes seen during inference, not just the most recent task.

These characteristics create a tension between **plasticity** (learning new concepts) and **stability** (preserving old knowledge) [25]. Modern approaches balance these through techniques such as replaying experience, regularization, and architectural modifications.

Class incremental learning has another caveat: the loss landscape for each class is different. Since each class is separate from others and the model does not see all the classes at once it becomes nearly impossible to find a minima which encompasses all the classes. Figure 1 provides an intuitive visual.

2 Related Works

Numerous techniques have been proposed to overcome the problem of catastrophic forgetting. But there is a little work on theoretical work on solving the underlying problem. Adding regularization to minimize change in parameters is a common approach [22, 3, 48, 38, 41, 46, 4, 20, 7, 28, 1, 30, 52]. Another classic method is memorization of some old examples (*replay*) [6, 19, 45, 39, 2, 36, 29, 5, 47, 44, 40, 31, 37, 16]. Finding an optimal balance between plasticity and stability has been a subfield of study in continual learning as well [9, 23, 29, 34, 35, 37, 47]. Recent research has proposed Generalized Class Incremental Learning (GCIL), which adopts a more realistic framework by removing some hard limitations. [42, 32, 11].

2.1 Theoretical Backgrounds

The theoretical foundations of CIL [26, 49, 43] are deeply rooted in computational learning theory and recursion-theoretic inductive inference. Early work established

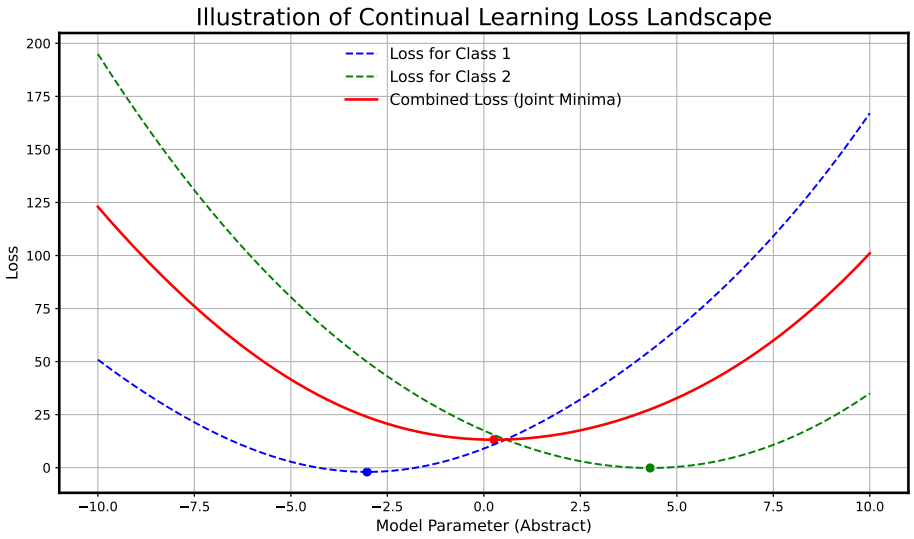


Figure 1: The loss landscape varies marginally for each class. It is difficult to find the optima for all (shown in red), while only seeing each class iteratively (blue and green).

critical distinctions between informationally incremental and operationally incremental learning paradigms. Informationally incremental algorithms process data sequentially without revisiting historical information, while operationally incremental systems may access prior data, but face constraints on effective utilization. These concepts were formalized through containment decision list learning frameworks, demonstrating fundamental limitations in universal CIL solvability due to intrinsic memory constraints.

Recent theoretical advances in CIL have focused on analyzing catastrophic forgetting under cyclic task orderings and developing frameworks for repeated class exposure. [10] establishes worst-case bounds on forgetting in overparameterized linear models under cyclic task sequences. For T tasks in d dimensions repeated for n cycles, they prove:

1. Cyclic Order: Forgetting bounded by $T^2 \min\{\frac{1}{\sqrt{n}}, \frac{d}{n}\}$
2. Random Order: Forgetting bounded by $O(\frac{d}{n})$

While most prevalent CIL methods can be categorized into either regularization-based or dynamic structure-based methods, ours follows neither. We analyze the process of naive training and retraining of classes in a cyclic process, devoid of any memory module or constraints in the default algorithm.

2.2 Optimization Theory

Understanding the relationship between local and global minima is fundamental to optimization theory. For a function $f : S \to \mathbb{R}$, a point $\mathbf{x}^* \in S$ is a local mini-

mum if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all points in some neighborhood of $\mathbf{x}^*$. In contrast, $\mathbf{x}^*$ is a global minimum if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in S$. For convex functions, any local minimum is also a global minimum. However, for non-convex functions (common in deep learning), there may be multiple local minima that are not global minima. This multiplicity of minima creates challenges for optimization algorithms, which may get trapped in suboptimal solutions.

Remarkably, recent theoretical work has shown that under certain conditions in deep learning, "every local minimum achieves the globally optimal value of the perturbable gradient basis model at any differentiable point" [24]. Analysis of the Hessian eigenspectrum of neural network loss landscapes reveals that for classification tasks, the loss landscape exhibits C highly curved directions (where C is the number of classes) [12], while remaining much flatter in the vastly larger number of remaining directions in weight space.

Figure 1 shows an example of how continual learning algorithms fall into suboptimal points. [27] proposes interpolating weights to better consolidate network capabilities before and after learning new tasks. Their work emphasizes that "The impact of loss landscape properties on continual learning is a very important, yet largely unexplored area", citing [33, 21].

3 Methods

3.1 Problem Setup

In this subsection, we clearly define the class incremental learning paradigm. The system learns a sequence of classes $\{(\mathbf{X_k}, \mathbf{k})\}_{k=1,\cdots,T}$, where $\mathbf{X_k}$ corresponds to class $\mathbf{k}$, with $\mathbf{k}$ being the label. Once the model sees all of class $\mathbf{i}$, class $\mathbf{i+1}$ is shown. Each cycle consists of T sequential training phases, where phase k uses only data from class k, with no revisiting of previous classes.

[10] showcase that in linear regression, such cyclic training would result in some forgetting (assuming realizability, forgetting is defined as the average loss). This metric tends to 0 with an increase in cycles. Our aim is to analyze the evolution of such a forgetting nature in a simple multilayer perceptron.

3.2 Dataset selection

Our work aims to analyze cyclic class incremental learning on the recently introduced MNIST-1D dataset [15]. MNIST-1D has a reduced dimensionality (from 784 to 40), which allows for a deep analysis of the network working and increases the feasibility for a large number of experiments.

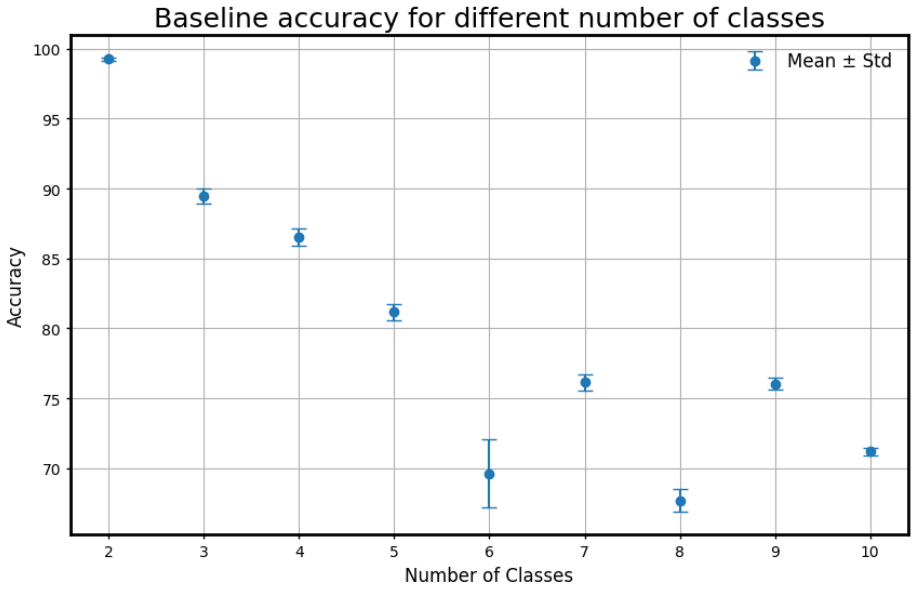


Figure 2: Baseline accuracy vs number of classes. Note how while choosing the first 6 or first 8 classes, the model is unable to perform well, and are outliers to the monotonically decreasing trend. This could be an anomaly within the dataset.

4 Experiments

The simplicity of the MNIST-1D dataset allows for fine grained experiments.

4.1 Baseline

To understand the effects of the class wise training scheme, learning rate and the number of classes used during the class incremental training, we required a robust baseline. For consistency across all experiments, we employed a uniform model architecture—a simple neural network comprising two fully connected layers. This architecture was deemed sufficient given the relative simplicity of the dataset, while also facilitating straightforward experimentation and analysis. The baseline model was trained using a learning rate of 10^{-4} over 1000 epochs, with early stopping applied using a patience of 5 epochs. We trained nine variants of this model: the first on only two classes, and each subsequent variant on an incrementally larger number of classes. Evaluation was conducted on a held-out test set. The training followed a conventional supervised learning protocol, with shuffled input data fed in mini-batches of 256 samples. The resulting accuracy metrics are presented in Figure 2.

Our model demonstrates a clear pattern when trained using standard stochastic gradient descent with a fixed learning rate. When distinguishing between just 2 classes, the model achieves near-perfect accuracy (99.4%). However, performance decreases substantially as the number of classes increases, dropping to 72.5% with 10 classes. Notably, the decline is not strictly monotonic: we observe significant fluctuations after 5 classes. These baseline results will serve as our control when

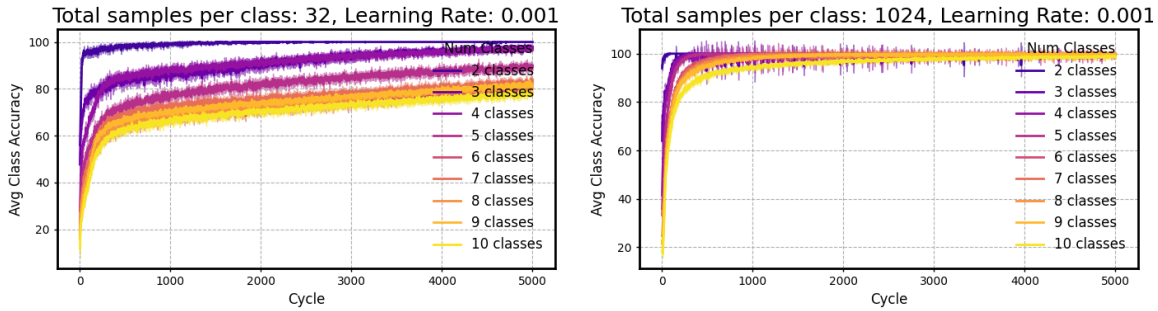


Figure 3: Accuracy versus training cycle is shown above for (sample size, learning rate) pairs of (32, 10^{-3}) and (1024, 10^{-3}). Notably, experiments with fewer total classes converge faster and reach baseline accuracy more quickly.

evaluating the effectiveness of various incremental learning approaches.

4.2 Setup

We perform experiments systematically by varying 3 key factors: total number of classes $\{2, \cdots, 10\}$, total samples in each class $\{2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}\}$, and learning rates $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. Our model has fixed size: $40 \rightarrow 128 \rightarrow 10$, and fixed batch size of 32. We train for a total of 5000 cycles. We tested the model after every cycle on every class, providing us with insight into the class wise accuracy. This test is performed on 100 randomly sampled data points from each class (data points are sampled after every cycle).

4.3 Performance Analysis with Varying Classes

In this section, we analyze how varying the number of classes affects the learning process. It is intuitive that having a larger number of classes would lead to higher difficulty in learning, and our experiments make this clear. We see ease of learning when there are fewer classes in all graphs (see Figure 3 and Appendix A). The difficulty in learning is also affected by choice of hyper parameters. We analyze this in depth in Section 4.4.3. We begin by discussing the key challenge in continual learning: *catastrophic forgetting*.

4.4 Catastrophic Forgetting

Reiterating, it is difficult for a network to maintain knowledge learned in the past while simultaneously acquiring new world information (Section 2). As mentioned earlier, an optimal balance between stability and plasticity is necessary. The large number of experiments allow for a detailed analysis of effects of learning rate, total number of classes, and sample size on catastrophic forgetting. Throughout our experiments, we observe this "noise" in accuracy, which indicated forgetting. We have a select few results shown below.

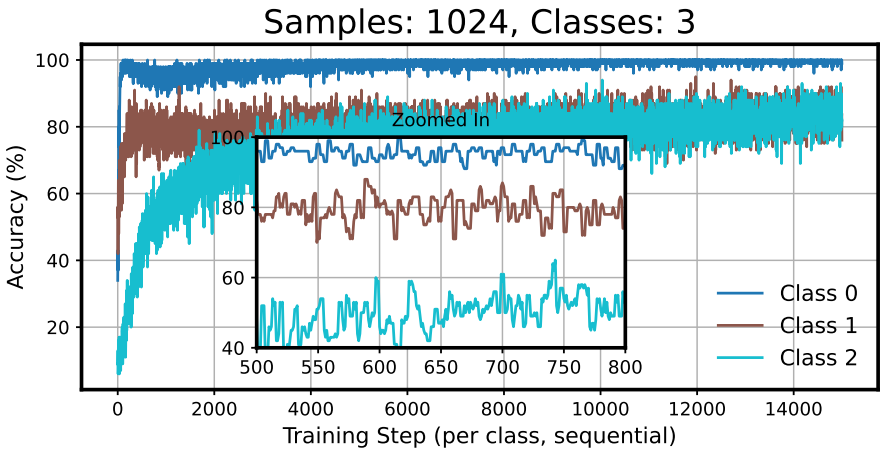


Figure 4: Zoomed in plot of class wise accuracy noted after training on each class in the cyclic training scheme. Note how the accuracies are very noisy with training, indicating repeated forgetting and relearning of data.

4.4.1 Effect of Learning Rate

The strides across the loss landscape is crucial to finding the right set of parameters. We experiment with 4 learning rates and find that moving from 10^{-5} to 10^{-6} breaks the combined learning process (see Figure 5). The noise observed in the Figure 5 is the drop in accuracy, indicating forgetting.

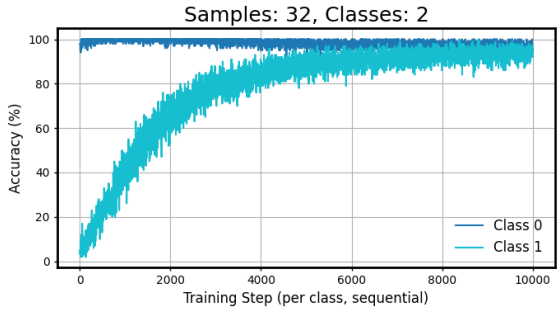
4.4.2 Combined Effect of Learning Rate and Total Number of Samples

We theorize that a high learning rate with a few samples is equivalent to a low learning rate with a high number of samples. We assume that the loss function for this dataset is simpler than most, as the dataset itself is simple. Taking one long stride in the landscape might be equivalent to taking multiple short strides. Figure 6 empirically proves this. But this need not necessarily result in the best performance (see Table 1).

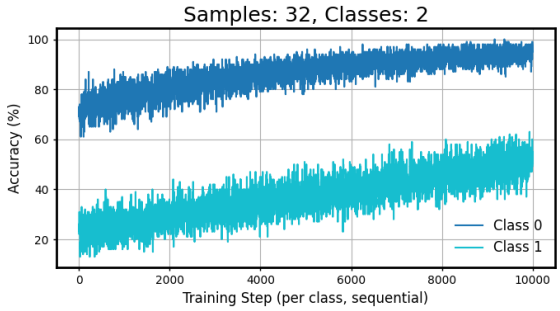
4.4.3 Changing the Total Number of Classes

It is intuitive that having a higher number of classes would make learning increasingly cumbersome. Here, we aim to understand the workings of the learning rate and total number of samples on the forgetting process. Reiterating, the noise observed here is drop in accuracy, indicating forgetting. It is both intuitive and interesting that this "noise" is larger in the case of 10 classes (Figure 7) as compared to 2 (Figure 6)

In an ordinary learning paradigm, having a large sample size is always better. But here, that need not be that case. Size of dataset is closely linked with the learning rate and changing each changes the learning process drastically (see Figure 8). When we have a high sample size and a high learning rate, the network fully learns



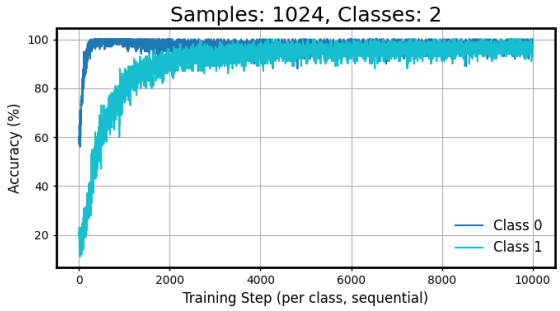
(a) Learning rate of 10^{-5}



(b) Learning rate of 10^{-6}

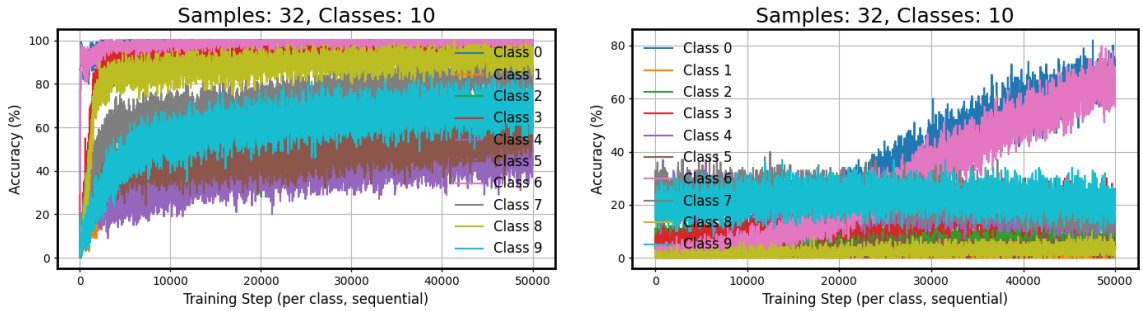
Figure 5: We have the accuracies of across classes after training on a particular class shown here. Since the number of samples is 32 (which is the batch size), only a single step is taken during training on a class. Note how a lower learning rate (on the right) affects learning for both classes.

(a) Learning rate of 10^{-4}



(b) Learning rate of 10^{-6}

Figure 6: In contrast with Figure 5b, Figure 6b shows better performance with training. This shows that training with a larger number of samples with a low learning rate does well. Also note how Figures 6b and 6a appear similar, indicating that training few samples with a large learning rate is equivalent to training with large samples and a smaller learning rate.

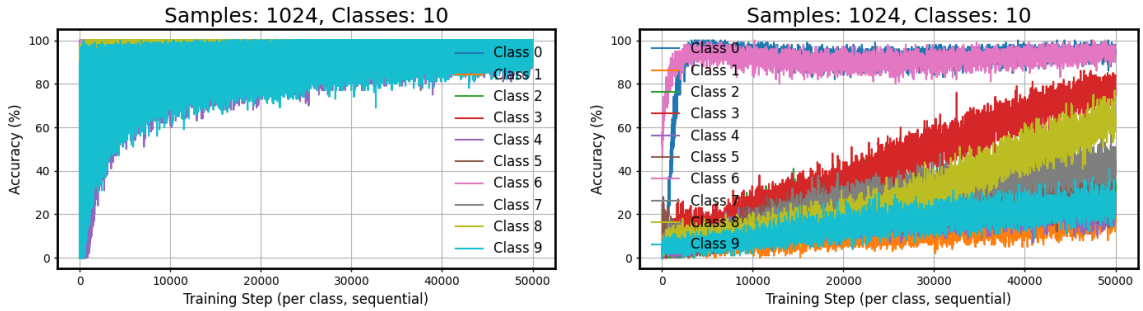


(a) Learning rate of 10^{-3}

(b) Learning rate of 10^{-6}

Figure 7: An increase in number of classes does increase difficult in learning. But it also depends on learning rates and sample size. The effect of having a low learning rate is shown here. Similar to observations from section 4.4.2, having a low learning rate does not facilitate learning.

at every iteration. This means that the network is more susceptible to forgetting. This susceptibility is shown in Figure 8a experimentally. The forgetting is very high initially but decreases with training. It is also interesting how the drops in accuracy (forgetting) is symmetric across classes in this case (compared to Figure 8b).



(a) Learning rate of 10^{-3}

(b) Learning rate of 10^{-6}

Figure 8: Increasing number of samples plays a role here as well. Having a high number of samples and a low learning rate facilitates learning of some classes very well (unlike what is observed in Figure 5b!).

4.4.4 Can Forgetting hit Zero?

An important result is shown in Figure 9: forgetting does not go to zero even with a lot of training. The forgetting does decrease in the initial stages, but subsequently flattens. In all our experiments, forgetting does not go to zero (we observe accuracy loss spikes).

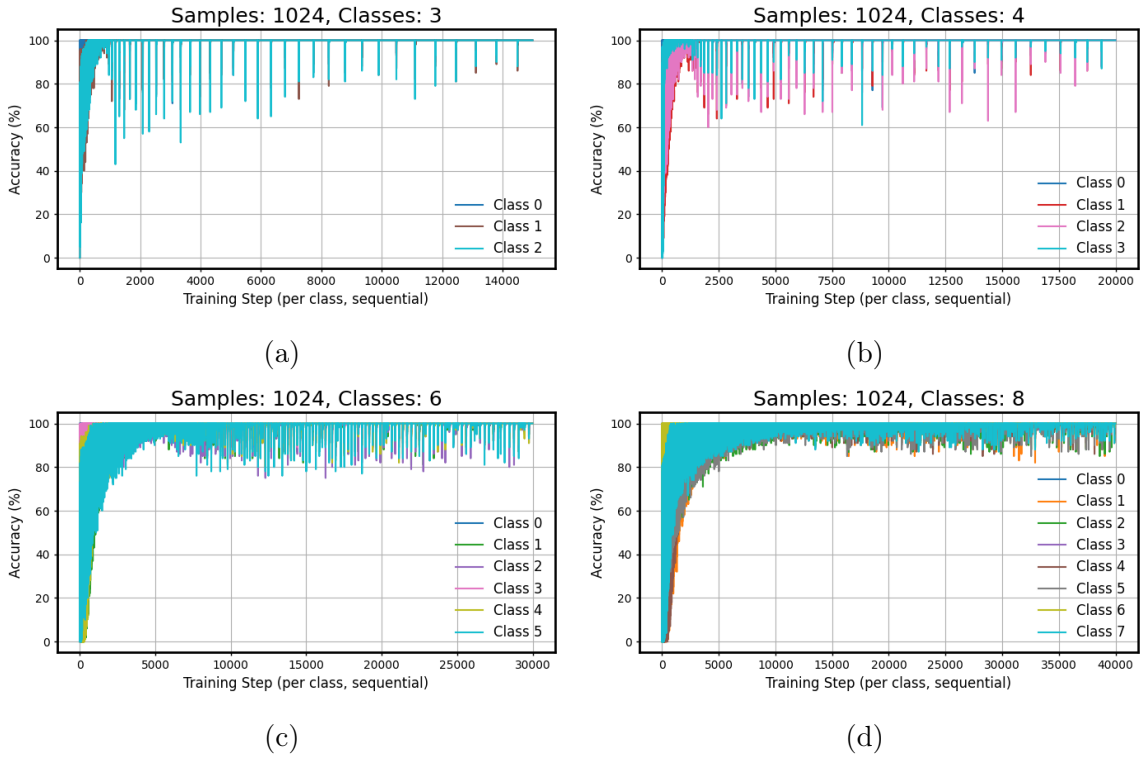


Figure 9: Forgetting analysis on 1024 samples and learning rate of 10^{-3} while varying number of classes. Note how after a series of training steps, there is a constant non zero forgetting!

4.4.5 Latching Behavior

Although the accuracies vary with the change in total classes, learning rate, and sample size, a consistent "latching" behavior is observed. The model learns the first class it sees and latches on to it (see Figures 8b, 7b, 6b, 5b). When a neural network first encounters data, it begins with unbiased weights and maximum plasticity. This clean slate state provides several advantages that become increasingly difficult to maintain as learning progresses: The first class a model encounters enjoys a privileged learning position for several reasons:

- Maximum Plasticity: The network has its full capacity for change and adaptation [8].
- No Competing Knowledge: Without previously learned patterns, there are no conflicting representations to impede learning [17].
- Unbiased Parameter Space: The random initialization provides an equal opportunity to form any representation necessary [18].

As a result, initial learning typically proceeds efficiently, establishing strong, stable representations for the first-encountered classes.

4.4.6 The Last-In-First-Out Pattern

Research confirms that in continual learning scenarios, examples learned first tend to be forgotten last, a phenomenon directly related to latching behavior [17]. This creates a "last-in-first-out" pattern where:

- Representations which are learned early form deeply entrenched pathways in the network
- Later learned classes are more susceptible to forgetting
- The model develops a persistent bias toward early training examples

As stated in one study: "Examples that are learned early are rarely forgotten, while those learned later are more susceptible to forgetting" [17]. This creates a fundamental imbalance in how knowledge is retained across the learning timeline.

4.5 Scaling Laws

We propose potential scaling laws for the cyclic continual learning framework. Specifically, given prior knowledge of the total number of classes the model is expected to learn, we investigate whether heuristic estimates for optimal learning rates and sample sizes can be derived. Our experiments systematically evaluate various configurations to identify those that yield the highest accuracies. The corresponding learning rates and sample sizes for the best-performing settings are summarized in Table 1.

Number of Classes	Top 3 (Learning Rate, Total Number of Samples) tuples
2	(0.001, 32), (0.0001, 64), (0.00001, 1024)
3	(0.001, 64), (0.001, 128), (0.0001, 512)
4	(0.001, 128), (0.0001, 1024), (0.001, 1024)
5	(0.001, 256), (0.001, 1024), (0.001, 128)
6	(0.001, 512), (0.001, 1024), (0.001, 256)
7	(0.001, 256), (0.001, 512), (0.001, 1024)
8	(0.001, 1024), (0.001, 512), (0.001, 256)
9	(0.001, 1024), (0.001, 512), (0.001, 256)
10	(0.001, 1024), (0.001, 512), (0.001, 256)

Table 1: We observe that the same (learning rate, number of samples) combinations performs the best at higher number of classes. When the number of classes is lower, there is more flexibility with choosing sample size and learning rate. This can be due to the relative simplicity in the distribution.

We observe that the same combinations of learning rate and number of samples work well when the number of classes increase. When the number of classes is lower, there is more flexibility in choosing these hyper parameters for the best performance.

4.6 Revisiting the Baseline

The anomaly identified in Section 4.1 warrants further investigation. In our cyclic Continual Incremental Learning (CIL) experiments, we observed that the model's performance exceeds the baseline by a substantial margin, although the reported test accuracies were initially computed on a subset of the test dataset. Figures 8a and 9 illustrate the notable increase in accuracy under the cyclic CIL setting. To better understand this discrepancy, we trained a network using the same cyclic protocol and evaluated it on the complete test dataset. Remarkably, the model **achieved an accuracy of 86**%, which significantly surpasses both the authors' originally reported performance (68%) and the results from our baseline (72%). Despite our efforts to optimize the baseline model, we were only able to achieve marginal improvements—limited to a few percentage points—over the accuracy reported by the original authors. The trained model weights are publicly available in our GitHub repository.

This observation raises an important question: can models trained extensively under the cyclic continual learning (CIL) paradigm consistently outperform traditional baselines? Addressing this question remains an open avenue for future research.

5 Discussion

Our systematic investigation of cyclic class-incremental learning (CIL) reveals three critical insights into catastrophic forgetting and learning dynamics.

- **Hyperparameter Interdependence:** Although the simpler experiments (with fewer total classes and total samples) suggest that learning rate and number of samples have a combined effect on learning (Section 4.4.2), the final scaling laws show otherwise. The learning rate can be fixed around 10^{-3}, but the number of samples per class needs to grow roughly exponentially with the total number of classes.
- **Latching Behavior:** First-learned classes maintain >90% accuracy across cycles due to maximum initial plasticity, while later classes show 20-40% fluctuation (Figure 8b), aligning with Last-In-First-Out forgetting patterns.
- **Non Zero Forgetting:** Forgetting never fully dissipates-even after 5,000 cycles, residual accuracy fluctuations of ±5% persist (Figure 9).

Our experiments give us insights as to how widely used architectures and learning algorithms behave when it comes to cyclic CIL. In the future, we aim to analyze this further by integrating regularization mechanisms mentioned in Section 2. We believe that this would aid the learning process. It would be interesting to see if our claims about scaling and hyper parameter independence holds when regularization is added into the system.

An intriguing outcome of our experiments is the apparent advantage of cyclic class incremental learning (CIL) models under certain configurations. While our baseline model closely aligns with previously reported results, the CIL-trained models, when sufficiently optimized, exhibit notably higher performance. This finding prompts a key question for future investigation: could models trained under CIL frameworks, given appropriate conditions, consistently surpass traditional methods? Further exploration is required to generalize these results and to understand the underlying factors contributing to such improvements.

References

[1] Hongjoon Ahn, Sungmin Cha, Donggyu Lee, and Taesup Moon. Uncertainty-based continual learning with adaptive regularization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[2] 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.

[3] Raffaello Camoriano, Giulia Pasquale, Carlo Ciliberto, Lorenzo Natale, Lorenzo Rosasco, and Giorgio Metta. Incremental robot learning of new objects with fixed update time. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3719–3726. IEEE, 2017.

[4] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 233–248, 2018.

[5] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 943–952, 2021.

[6] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2019.

[7] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5138–5146, 2019.

[8] Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632:768–774, 2024.

[9] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision (ECCV)*, pages 86–102. Springer, 2020.

[10] Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4028–4079. PMLR, 02–05 Jul 2022.

[11] Di Fang, Hong Zhuang, Zhi Lin, and Kar-Ann Toh. G-acil: Analytic learning for exemplar-free generalized class incremental learning. *arXiv preprint arXiv:2403.15706*, 2024.

[12] Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes. *arXiv preprint arXiv:1910.05929*, 2019.

[13] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

[14] Lukasz Golab and M Tamer Özsu. Issues in data stream management. *ACM SIGMOD Record*, 32(2):5–14, 2003.

[15] Sam Greydanus and Dmitry Kobak. Scaling down deep learning with MNIST-1D. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[16] Yiduo Guo, Bing Liu, and Dongyan Zhao. Online continual learning through mutual information maximization. In *International Conference on Machine Learning*, pages 8109–8126. PMLR, 2022.

[17] Guy Hacohen and Tinne Tuytelaars. Forgetting order of continual learning: Examples that are learned first are forgotten last. *arXiv preprint arXiv:2406.09935*, 2024.

[18] Md Yousuf Harun and Christopher Kanan. A good start matters: Enhancing continual learning with data-driven weight initialization. *arXiv preprint arXiv:2503.06385*, 2025.

[19] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019.

[20] Wenpeng Hu, Zhou Lin, Bing Liu, Chongyang Tao, Zhengwei Tao, Jinwen Ma, Dongyan Zhao, and Rui Yan. Overcoming catastrophic forgetting for continual learning via model adaptation. In *International Conference on Learning Representations*, 2019.

[21] Zhongzhan Huang, Mingfu Liang, Senwei Liang, and Wei He. Altersgd: Finding flat minima for continual learning by alternative training. *arXiv preprint arXiv:2107.05804*, 2021.

[22] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. *arXiv preprint arXiv:1607.00122*, 2016.

[23] Minsoo Kang, Jaeyoo Park, and Bohyung Han. Class-incremental learning by knowledge distillation with adaptive feature consolidation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3330–3339, 2022.

[24] Kenji Kawaguchi, Jiaoyang Huang, and Leslie Pack Kaelbling. Every local minimum value is the global minimum value of induced model in nonconvex machine learning. *Neural Computation*, 31(12):2293–2323, 12 2019.

[25] Dongwan Kim and Bohyung Han. On the stability-plasticity dilemma of class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20196–20204, 2023.

[26] Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. A theoretical study on solving continual learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5065–5079. Curran Associates, Inc., 2022.

[27] Jędrzej Kozal, Jan Wasilewski, Bartosz Krawczyk, and Michał Woźniak. Continual learning with weight interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4187–4195, 2024.

[28] Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. *arXiv preprint arXiv:1903.12648*, 2019.

[29] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Adaptive aggregation networks for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2544–2553, 2021.

[30] Yu Liu, Sarah Parisot, Gregory Slabaugh, Xu Jia, Ales Leonardis, and Tinne Tuytelaars. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. In *European Conference on Computer Vision*, pages 699–716. Springer, 2020.

[31] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 6467–6476, 2017.

[32] 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 (CVPRW)*, pages 128–129, 2020.

[33] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Dilan Görür, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. *arXiv preprint arXiv:2010.04495*, 2020.

[34] Dongmin Park, Seokil Hong, Bohyung Han, and Kyoung Mu Lee. Continual learning by asymmetric loss approximation with single-side overestimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3330–3339, 2019.

[35] Jaeyoo Park, Minsoo Kang, and Bohyung Han. Class-incremental learning for action recognition in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13698–13707, 2021.

[36] Jathushan Rajasegaran, Munawar Hayat, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Ming-Hsuan Yang. An adaptive random path selection approach for incremental learning. *arXiv preprint arXiv:2002.04750*, 2020.

[37] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[38] Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[39] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P Lillicrap, and Greg Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[40] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[41] Jonathan Schwarz, Jelena Luketina, Wojciech M Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. *arXiv preprint arXiv:1805.06370*, 2018.

[42] Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

[43] Fu-Yun Wang, Da-Wei Zhou, Liu Liu, Han-Jia Ye, Yatao Bian, De-Chuan Zhan, and Peilin Zhao. Beef: Bi-compatible class-incremental learning via energy-based expansion and fusion. In *The eleventh international conference on learning representations*, 2022.

[44] Liyuan Wang, Xingxing Zhang, Kuo Yang, Longhui Yu, Chongxuan Li, Lanqing Hong, Shifeng Zhang, Zhenguo Li, Yi Zhong, and Jun Zhu. Memory replay with data compression for continual learning. *arXiv preprint arXiv:2202.06592*, 2022.

[45] 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*, pages 374–382, 2019.

[46] Ju Xu and Zhanxing Zhu. Reinforced continual learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[47] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. *arXiv preprint arXiv:2103.16788*, 2021.

[48] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017.

[49] Zihan Zhang, Wenhao Zhan, Yuxin Chen, Simon S. Du, and Jason D. Lee. Optimal multi-distribution learning, 2024.

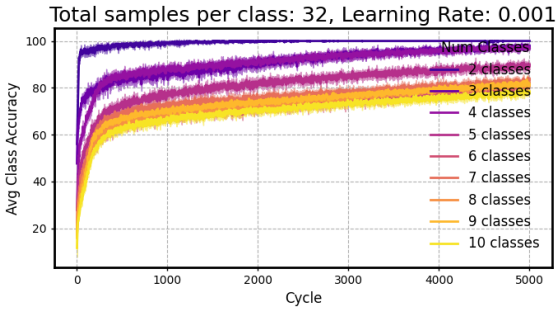
[50] Haiyan Zhao, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Does continual learning equally forget all parameters? In *International Conference on Machine Learning*, pages 42280–42303. PMLR, 2023.

[51] Da-Wei Zhou, Qi-Wei Wang, Han-Jia Ye, and De-Chuan Zhan. A model or 603 exemplars: Towards memory-efficient class-incremental learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

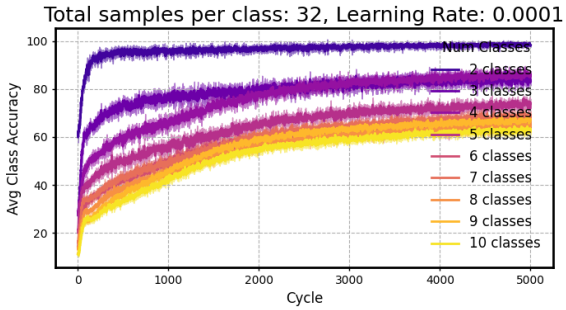
[52] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021.

A Accuracy Results

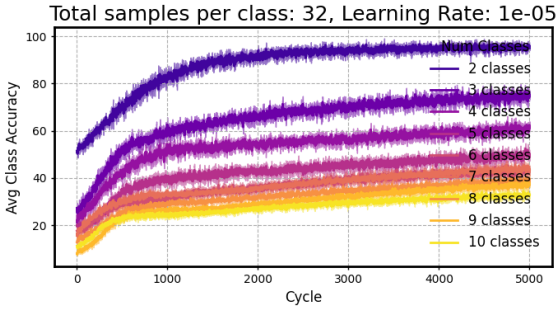
Below is the performance of the model with varying samples per class, learning rates, and number of classes selected. Each plot shown covers the accuracy while fixing the sample size of each class and learning rate fixed. Note that these plots, including Figure 3 are very slightly denoised using a smoothing filter for improved visuals.



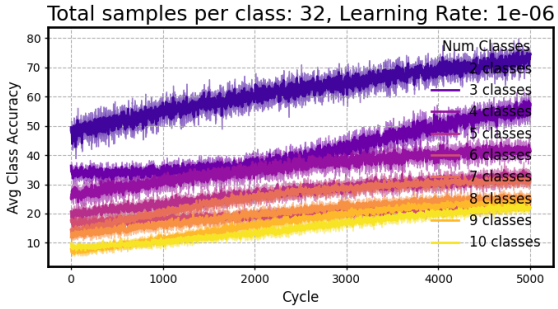
Samples = 32, LR = 1e-3



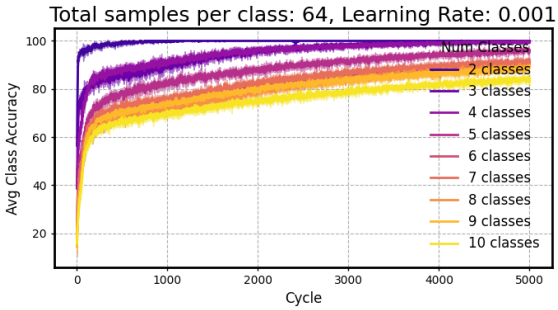
Samples = 32, LR = 1e-4



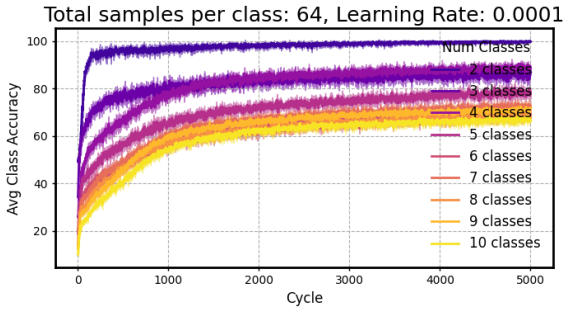
Samples = 32, LR = 1e-5



Samples = 32, LR = 1e-6

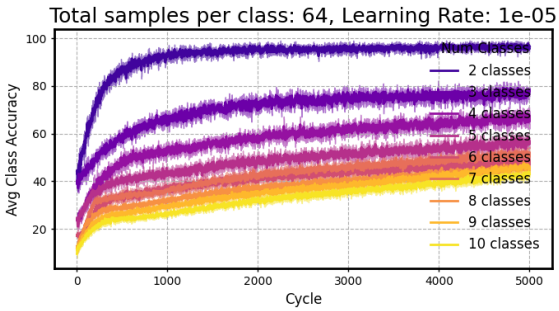


Samples = 64, LR = 1e-3

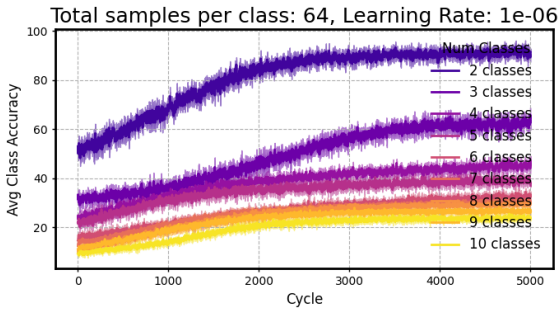


Samples = 64, LR = 1e-4

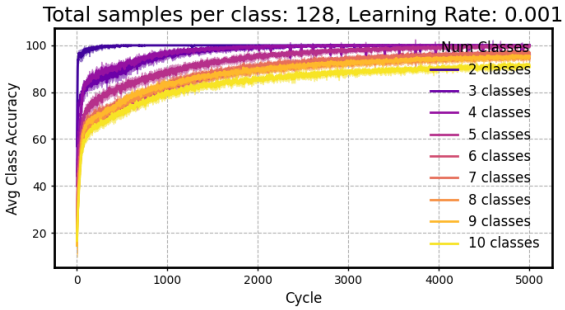
Figure 10: Plots for 32 and 64 sample sizes with learning rates from 1e-3 to 1e-6.



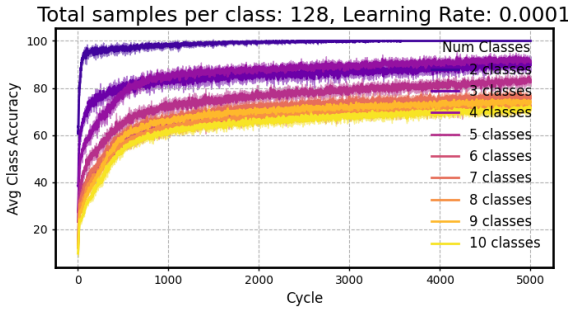
Samples = 64, LR = 1e-5



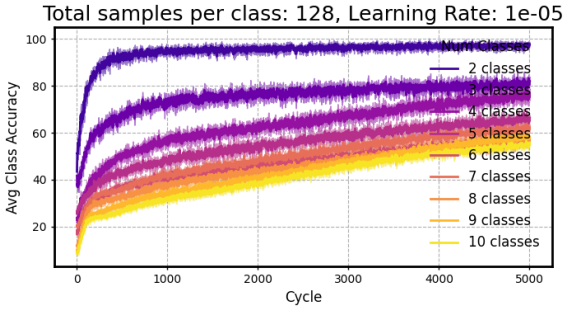
Samples = 64, LR = 1e-6



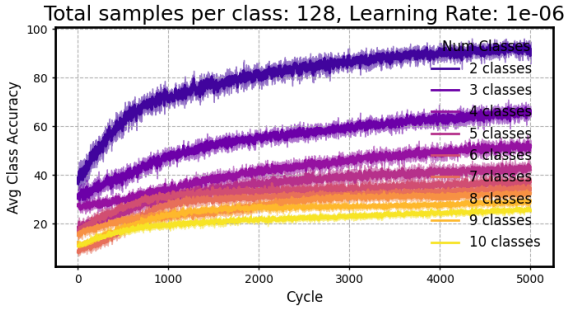
Samples = 128, LR = 1e-3



Samples = 128, LR = 1e-4

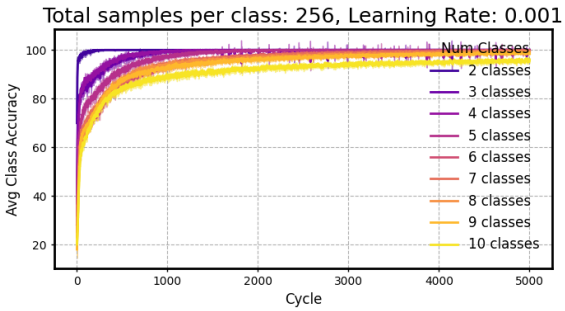


Samples = 128, LR = 1e-5

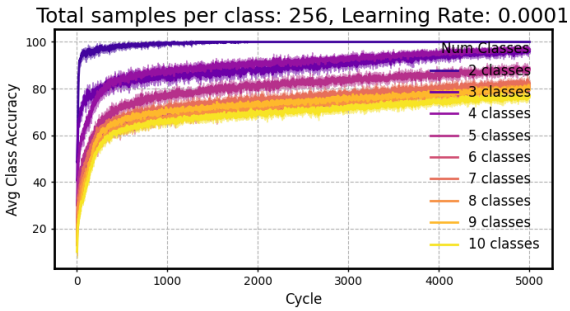


Samples = 128, LR = 1e-6

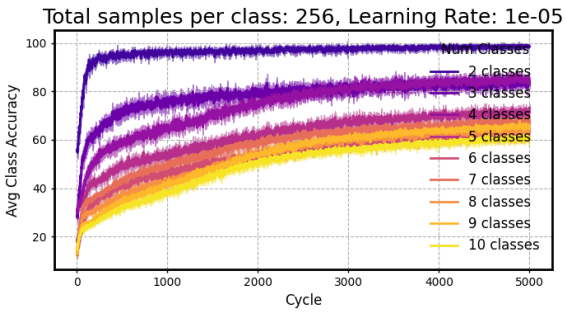
Figure 11: Plots for 64 and 128 sample sizes with learning rates from 1e-3 to 1e-6.



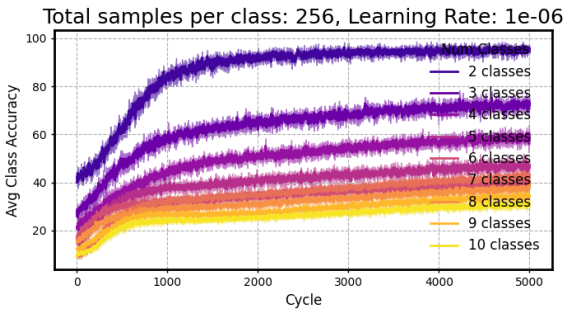
Samples = 256, LR = 1e-3



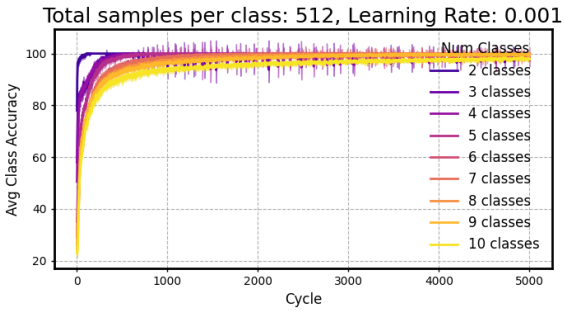
Samples = 256, LR = 1e-4



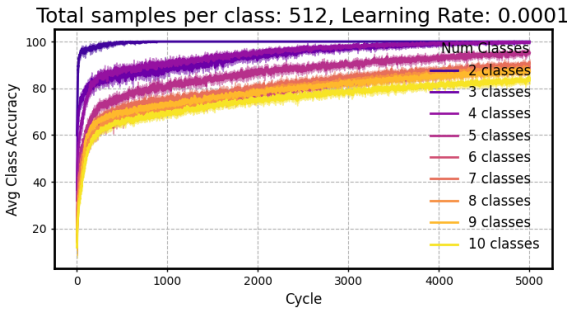
Samples = 256, LR = 1e-5



Samples = 256, LR = 1e-6

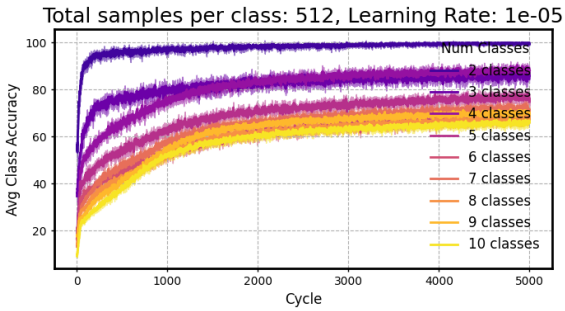


Samples = 512, LR = 1e-3

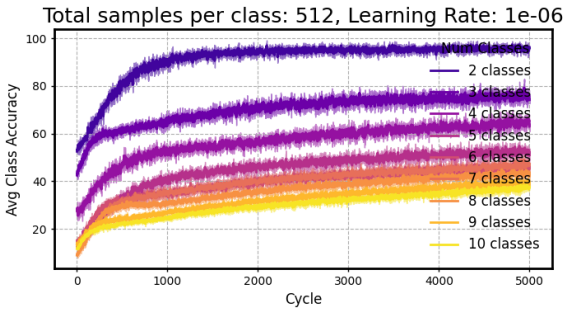


Samples = 512, LR = 1e-4

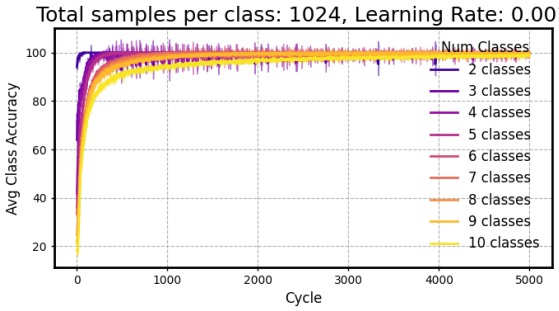
Figure 12: Plots for 256 and 512 sample sizes with learning rates from 1e-3 to 1e-6.



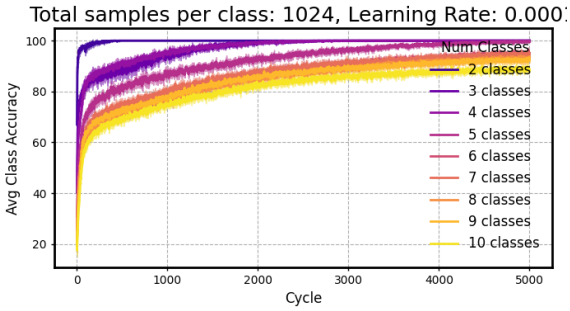
Samples = 512, LR = 1e-5



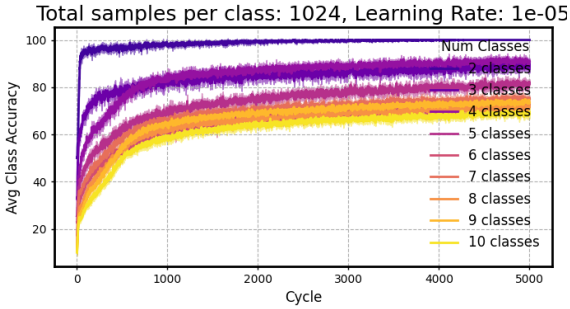
Samples = 512, LR = 1e-6



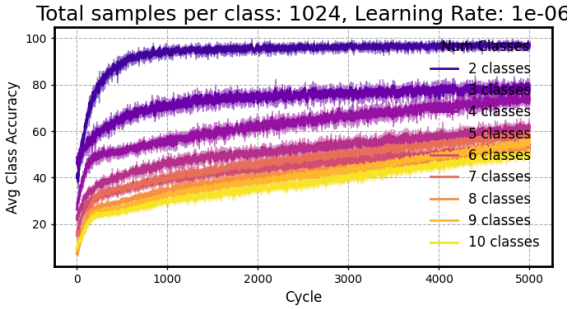
Samples = 1024, LR = 1e-3



Samples = 1024, LR = 1e-4



Samples = 1024, LR = 1e-5



Samples = 1024, LR = 1e-6

Figure 13: Plots for 512 and 1024 sample sizes with learning rates from 1e-3 to 1e-6.