Stability Gap in Continual Learning: The Role of Learning Rate

[1] Matthias De Lange, Gido van de Ven, and Tinne Tuytelaars. Continual evaluation for lifelong learning: Identifying the stability gap. arXiv preprint arXiv:2205.13452, 2022.

[2] Timm Hess, Tinne Tuytelaars, and Gido M van de Ven. Two complementary perspectives to continual learning: Ask not only what to optimize, but also how. arXiv preprint arXiv:2311.04898, 2023.

1. BACKGROUND

- **Continual learning** → training neural networks on tasks **sequentially**, without retraining the whole model
- Early challenge: **catastrophic forgetting** → network forgets earlier tasks completely
- Current methods reduce forgetting but still face the **stability gap** → sudden, short-lived drop in accuracy on earlier tasks after learning a new one
- **Key idea:** the issue may not be **what** we optimize, but **how** we optimize → is there a path in the parameter space that leads to good performance on all tasks without this drop?

2. RESEARCH QUESTION AND HYPOTHESES

RQ: How does the learning rate influence the stability gap in continual learning, and can it be reduced through scheduling?

Hypotheses

H1. Lower constant learning rates will reduce the stability gap.

H2. Well-tuned scheduled learning rates will further help reduce the stability gap.

3. EXPERIMENTAL SETUP

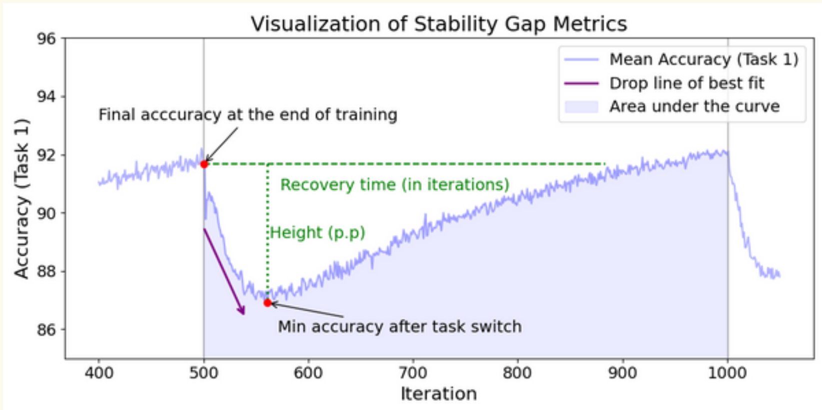
- **Baseline Model** → a simple **Multilayer Perceptron** model, optimizer → **SGD with no momentum**
- The model is trained sequentially on **rotated MNIST**
  - Task 1 → 0°
  - Task 2 → 80°
  - Task 3 → 160°

Task 1 Task 2 Task 3

- **PHASES:**
  - **Phase 1** → change the constant learning rate in the range [0.001, 1.5]
  - **Phase 2** → apply learning rate scheduling (CyclicLR and IncreaseLROnPlateau)

4.1 METHODOLOGY – CONSTANT LR

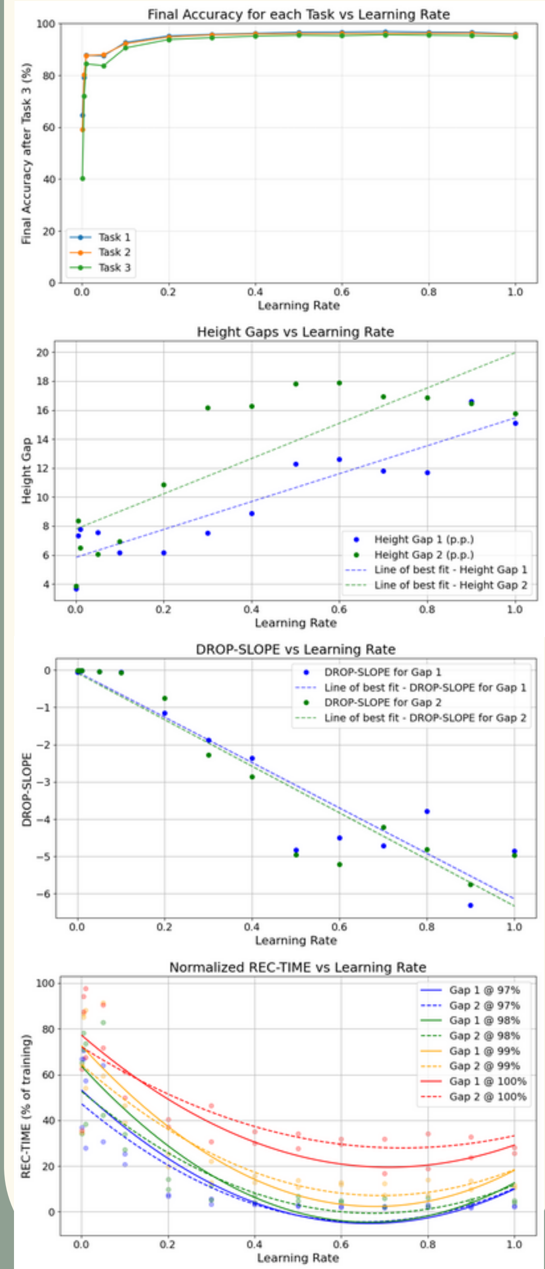
1. **Run the model** with each constant LR **20 times**
2. **Compute the metrics** shown in the figure below based on all runs
3. These metrics were developed to help discover **trends and patterns in Learning Rate vs Stability Gap Shape**



4.2 METHODOLOGY – SCHEDULED LR

- **Chosen schedulers:**
  a. **CyclicLR** → available as a standard option in PyTorch's torch.optim.lr_scheduler module
  b. **IncreaseLROnPlateau** → a **custom scheduler** developed for this study, inspired by ReduceLROnPlateau; works the same, except that it **increases the LR when the metric stops improving**
- **Experiment steps:**
  1. **Run grid search** to find the best configuration for each scheduler based on the objective function: **Objective = MeanFinalAcc – MeanHeight**, where:
     - **MeanFinalAcc** → mean final test accuracy of all tasks measured after training on Task 3
     - **MeanHeight** → mean height of all observed stability gaps
  2. Apply the same **objective function to all tested constant LR** to use the **best-performing one as a baseline**
  3. **Run the best configurations 20 times, then compare** all visually via plots and numerically via 2 defined metrics (MeanFinalAcc and MeanHeight)

5.1 RESULTS – CONSTANT LR



5.2 RESULTS – SCHEDULED LR

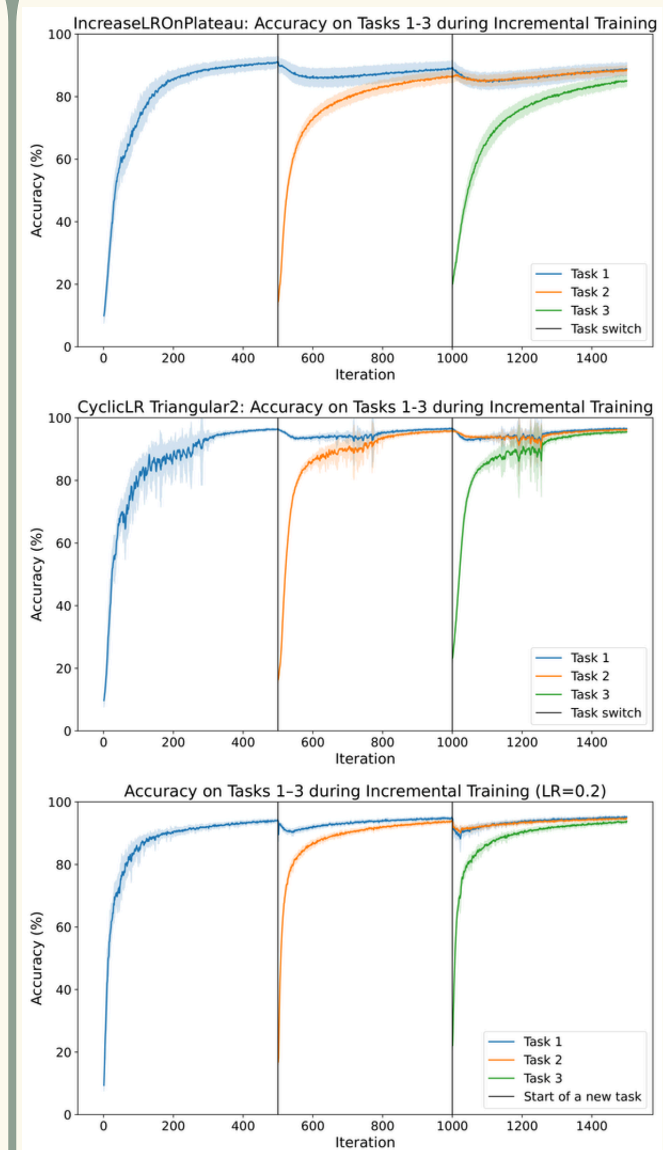


Table 1: Summary of the MeanFinalAcc defined in Equation 2 and MeanHeight defined in Equation 3 for each approach. For CyclicLR, the gap height is computed only over the first 100 iterations after each task switch, to better isolate the initial drop from the scheduler's further oscillations. Fluctuations occurring later are not included in the height value but are clearly visible in the accuracy plots. All values are reported as mean ± standard deviation across 20 runs. The best values are highlighted in bold.

Scheduler	MeanFinalAcc ↑	MeanHeight ↓
IncreaseLROnPlateau	87.54 ± 2.49	5.33 ± 0.70
CyclicLR Triangular2	**96.08 ± 0.57**	**4.39 ± 0.47**
Constant LR = 0.2	94.61 ± 0.72	7.75 ± 1.40

6. CONCLUSION

Key findings

- **Low LR** → smaller gap, slower recovery
- **High LR** → deeper gap, faster recovery
- **Schedulers might help only with careful tuning** (CyclicLR adds oscillations)

Implications

- **Safety-critical applications** → pick moderate/low LR for worst-case stability
- **Others** → depending on needs, might choose higher LR for faster recovery
- If resources are available, it might be beneficial to tune a scheduler for a more dynamic control

Author: Paulina Sobocińska ⋮ P.K.Sobocinska@student.tudelft.nl

Affiliations: EEMCS, Delft University of Technology, The Netherlands

Supervisor: Gido van de Ven

Responsible Professor: Tom Viering