Appendix

A. Task-Aware Accuracy

Task-aware accuracy is a metric used in continual learning to evaluate model performance when task identity is known during inference. Task-aware accuracy requires the model only to distinguish classes within-task, as opposed to task-agnostic accuracy, which requires both within-task class separation and correct task classification. Therefore, task-aware accuracy is considered to be an easier setting. In this section, we show task-aware results corresponding to experiments from Sec. 4.2 and Sec. 4.3, and show that even when we evaluate task-specific linear head it performs worse than NMC, empowering our previous claims. The results can be seen in Fig. 11, Fig. 12, and Fig. 13.

B. Continual evaluation metrics

The stability gap (SG) is a per-task metric, making it particularly useful for analyzing the dynamics of a model's behavior across individual tasks during sequential learning. Similarly, worst-case accuracy (WC-ACC) and average minimum accuracy (min-ACC) are metrics that can be evaluated on a per-iteration basis, providing further insights into the model's performance during training. Visualizing these metrics for all tasks on the plots is more informative than considering only the final scores, as it highlights how the model evolves and adapts over time. We present the plots of those metrics in Fig. 14.

C. Other approaches

We expand our investigation to regularization-based (LwF [22], SS-IL [1]), and parameter-isolation (EWC [16]) methods. We use a constant memory buffer for all the methods (2000 exemplars). We also present continual evaluation metrics on the whole training on CIFAR100 (see Fig. 18, Fig. 19, and Fig. 20).

C.1. Task-Agnostic results

More advanced methods, which usually perform better on CIL setups, provide better knowledge transfer while reducing forgetting and ultimately better representations, but still suffer from the linear multi-head classifier. Methods based on modified loss functions provide better latent representations, and NMC additionally mitigates the problems associated with the classifier head (see Tab. 4, Fig. 15, and Fig. 17).

C.2. Latest-Task Prediction Bias

Experiment analogous to the one in Fig. 9. We observe that NMC reduces the LTB even in approaches that try to improve the linear heads of previous tasks, e.g. by knowledge distillation or other methods, more than simple finetuning with exemplars (see results in Fig. 17).

Table 4. NMC improves the stability metrics and final accuracy of different CIL methods.

	WC-ACC (↑)	min-ACC (↑)	SG(↓)	ACC (↑)
	CIFAR100/10			
LwF	10.44±4.32	4.12±4.67	81.36±15.08	21.55±0.55
+NMC	**25.33**±4.00	**23.08**±4.60	**27.79**±8.83	**30.97**±0.36
EWC	6.92±0.10	0.71±0.04	95.38±0.89	17.87±0.27
+NMC	**15.95**±0.83	**13.87**±1.07	**49.88**±3.97	**25.84**±0.26
SS-IL	22.72±0.45	21.75±0.45	36.09±1.10	**31.81**±0.39
+NMC	**26.74**±0.23	**25.83**±0.31	**21.11**±1.34	31.05±0.53
	ImageNet100/10			
LwF	25.78±0.36	19.66±0.36	41.24±1.06	31.36±0.30
+NMC	**43.09**±0.13	**42.31**±0.08	**10.67**±0.42	**45.66**±0.46
EWC	17.9±0.88	10.12±0.89	68.87±2.64	31.46±0.39
+NMC	**36.01**±0.99	**30.17**±1.04	**34.46**±1.97	**44.58**±0.30
SS-IL	37.42±0.45	35.43±0.81	21.53±0.74	46.05±0.67
+NMC	**38.75**±0.53	**36.89**±0.61	**9.15**±0.94	**46.52**±0.42

D. Different CNN architectures

To further investigate the influence of the non-parametric NMC classifier, in addition to experiments with ResNet18, we test it with other standard convolutional neural networks. We evaluate finetuning with constant memory (2000 exemplars) on linear multi-head and NMC with MobileNetV2 [34], ResNet50 [12], EfficientNet-B4 [38], and VGG11 [37] as backbones. We use CIFAR100 split into 10 equally sized tasks and train the networks as described in Sec. 3.1. We still notice that using NMC constantly improves the results, regardless of the architecture we use.

Table 5. CIFAR100 split into 10 disjoint tasks.

Network	WC-ACC (↑)	min-ACC (↑)	SG(↓)	ACC (↑)
ResNet18	13.94±0.39	7.58±0.32	63.04±2.57	21.04±0.31
+NMC	**27.27**±0.50	**24.01**±0.65	**25.35**±2.66	**31.14**±0.38
ResNet50	7.43±1.79	1.51±2.12	89.23±15.16	14.92±0.62
+NMC	**11.49**±8.43	**7.26**±9.08	**69.76**±29.29	**23.36**±1.31
VGG11	16.58±0.40	10.61±0.45	57.44±0.71	25.14±0.23
+NMC	**24.75**±0.61	**20.15**±0.75	**31.31**±2.13	**28.81**±0.39
MobileNetV2	11.49±0.76	5.56±0.94	65.56±7.53	17.56±0.91
+NMC	**22.72**±0.53	**19.43**±0.59	**24.93**±0.96	**25.44**±0.55
EfficientNet-B4	6.94±0.70	0.43±0.68	97.65±3.50	17.79±2.16
+NMC	**14.02**±5.71	**9.6**±6.23	**63.77**±23.88	**27.10**±0.50

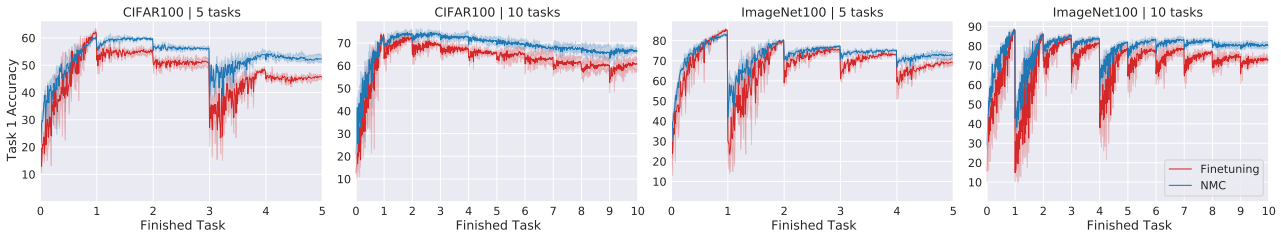

Figure 11. Task-aware results demonstrate that even when using a task-specific linear head, performance is lower than with NMC, reinforcing our previous claims.

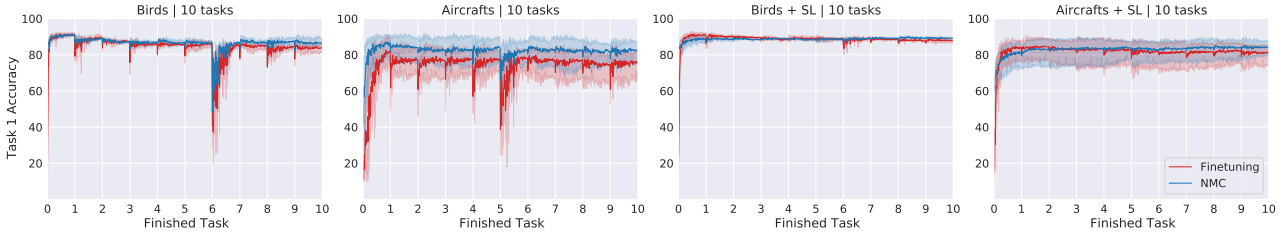

Figure 12. Task 1 task-aware accuracy during training on finegrained datasets.

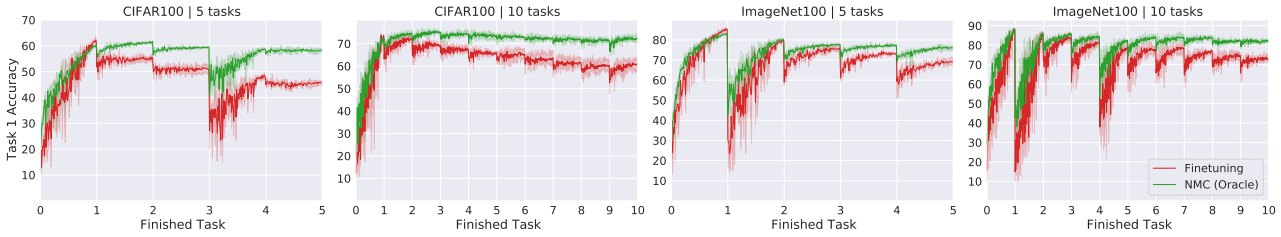

Figure 13. Oracle NMC has the best prototype estimates so it further improves TAw accuracy.

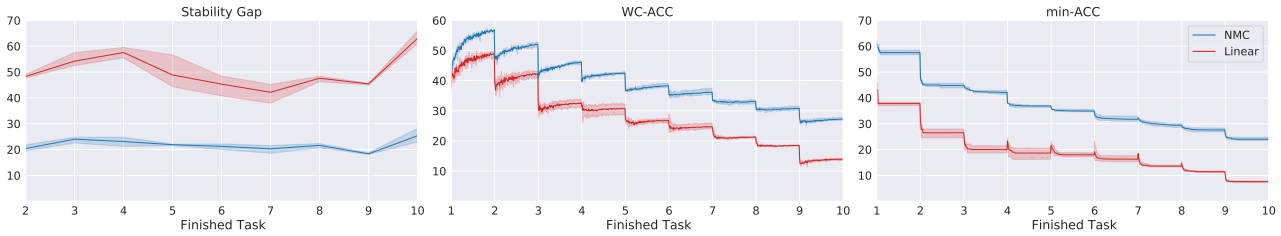

Figure 14. Fine-tuning on CIFAR100 (10 tasks). Continual evaluation metrics through full training of 10 tasks.

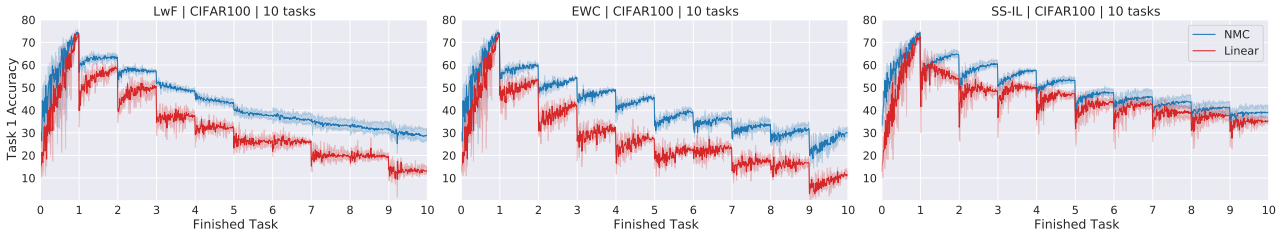

Figure 15. TAg accuracy. More advanced methods yield improved latent representations, while NMC further alleviates issues related to the classifier head. Observations from main experiments scale to other approaches.

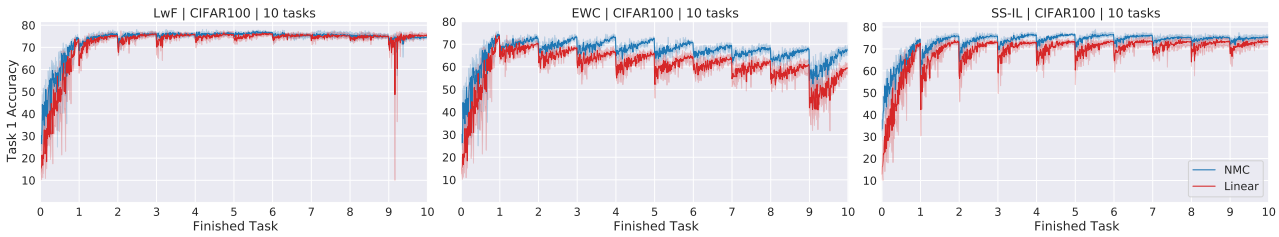

Figure 16. For completeness we also present TAw accuracy scores.

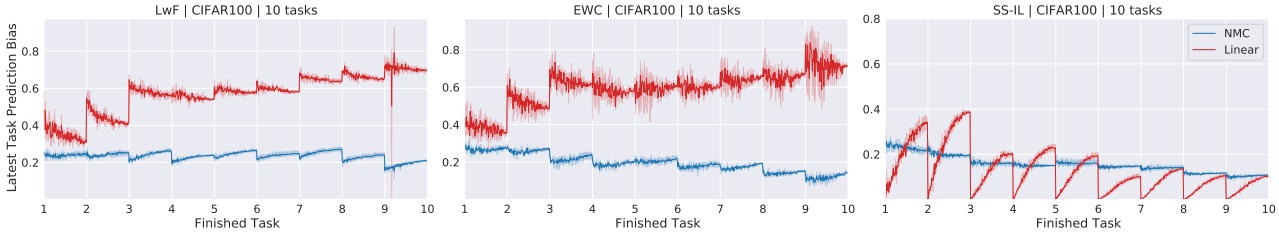

Figure 17. LwF and EWC suffer from latest task prediciton bias, but it can be reduced with NMC. SS-IL has internal mechanism to mitigate task-recency bias, but still its LTB is comparable to NMC's.

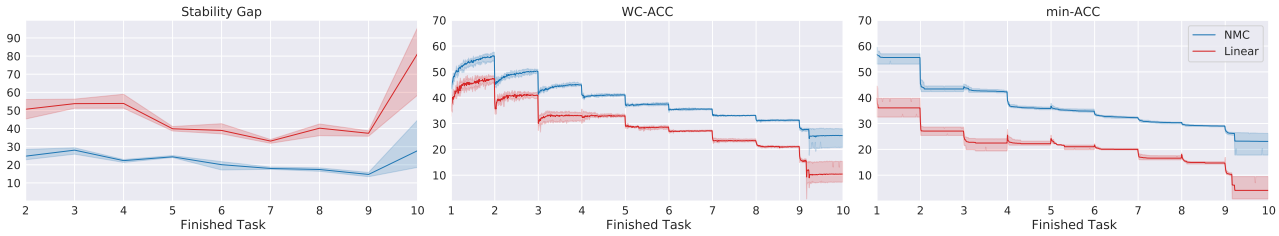

Figure 18. LwF on CIFAR100 (10 tasks). Continual evaluation metrics.

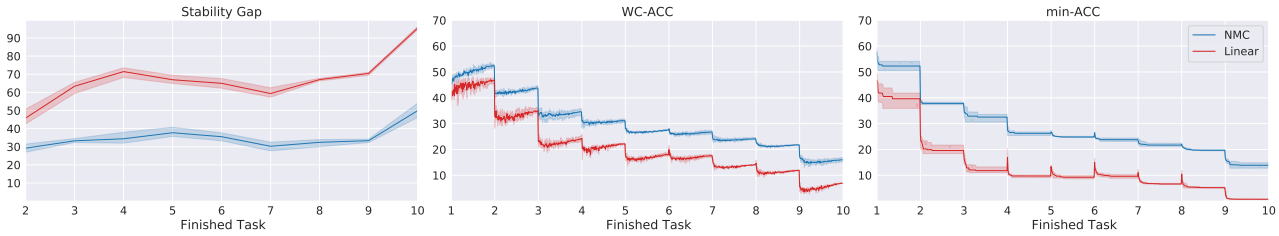

Figure 19. EWC on CIFAR100 (10 tasks). Continual evaluation metrics.

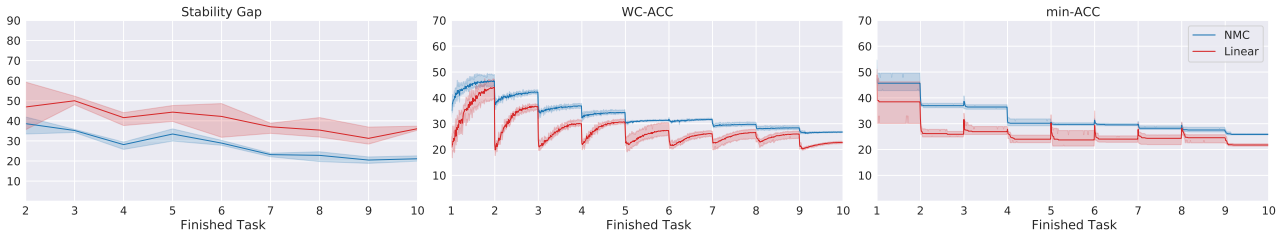

Figure 20. SS-IL on CIFAR100 (10 tasks). Continual evaluation metrics.