We thank the reviewers for the valuable feedback. There were two common concerns: lack of a complex benchmark and unclear terminology in the experiments. We address these first and then follow with responses to each reviewer.

Other Benchmarks: We didn't run experiments on a more complex dataset because even on split-omniglot, existing continual learning methods perform extremely poorly. It is fair question if the conclusions extend to other datasets. We therefore ran our method on Mini-Imagenet and report the results in Figure 1. We incrementally learn a 20-way classifier using 30 samples per class for both train and test. The results support the same conclusions. Note that we go over the training trajectory *only once*, one sample at a time. We will include these results to provide further evidence for MRCL and that the strategy scales to more complex settings.

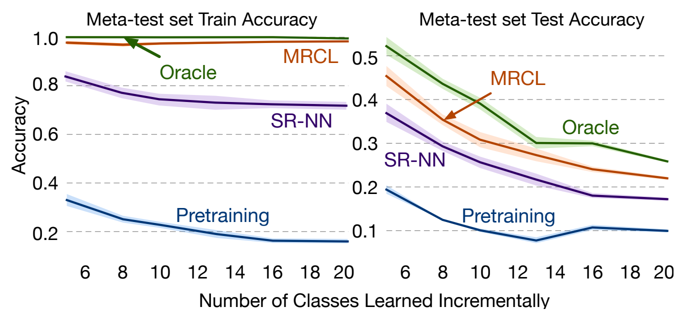

Figure 1: Reproducing the classification results on Mini-Imagenet.

To clarify the experiment protocol we have decided to use updated terminology in the paper. Our method is divided into two phases : (1) meta-training phase and (2) meta-testing phase. The meta-training phase involves optimizing the OML objective for learning a representation and the meta-testing phase involves training the **TLN** on a highly correlated trajectory ***in a single pass***. There is no overlap between data used in meta-training and meta-testing. All results are reported in the meta-testing phase; we also do not use IID sampling or multiple epochs for MRCL in any of the reported results. The first figure of Figure 4 is very meaningful because it shows MRCL—which learns incrementally in class order—is almost as effective as the Oracle—which learns using IID data. This highlights how much a reprsentation trained for online updating can help mitigate interference.

Reviewer 1: Writing and clarity: Thank you for pointing out issues with the writing and L73; we will fix those and move the algorithm to the main paper. In Appendix L379-380, we meant that a meta-learned initialization alone can not solve the interference problem; it is important to transform the input into a representation with non-interfering solution manifolds.

Improvements: ... increase your score? (1) The intuition behind MRCL is that instead of using sparsity as a proxy for good representations for continual learning, we directly measure interference caused by highly correlated updates over a finite horizon, and minimize this interference to learn a representation. We assume that a representation that minimizes interference for k correlated updates would also reduce interference in the long run. In the incremental sine experiment, k is actually equal to length of the complete trajectory whereas for Omniglot, k is much smaller than the complete trajectory. Empirical results support that in both cases, OML can recover a good non-interfering representation. One explanation is that as long as k is large enough to cause measurable interference, minimizing OML will result in a good representation.

(3): It's not clear how to compare our method with the three suggested approaches. One of the three approaches, TADAM, is specific to few-shot-learning - a different problem setting than ours. The remaining two approaches improve on gradient based meta-learning in general and **are complementary to our work** i.e. they can be combined with our objective function to potentially further improve the results. To better clarify this, we will extend the related work section of our paper to explain why a comparison with these approaches is tangential to our contributions.

Reviewer 2: Quality and Clarity: (1) Yes. The model is trained on the meta-training set using iid sampling, and online learning is done on meta-test set in a single pass. **(2)** Random batch is sampled from the entire meta-training set. Meta-training involves revisiting the data, whereas training during meta-tesing involves a single pass through data. **(3)** We agree that since class label is the same as class id, class id is implicitly available. However, our method does not exploit it in anyway, and learns a single classifier over all classes across task ids. We will nonetheless fix the inaccuracy in our claim. **(4)** The first figure in Figure 4 is the ***training error during meta-testing*** and does not involve IID sampling or multiple epochs. It measures degree of forgetting without taking into account the generalization error, and does in-fact perform very close to Oracle.

Limitations: (2) We fully agree that a fixed representation can not solve continual learning. We addressed this limitation in L273-L275 by suggesting one strategy which can be used to continuously update the representation. For the purpose of this paper, we focused on demonstrating that an effective representation can greatly reduce interference. We are currently extending this work using this proposed strategy, with a slowly changing representation updated using the OML objective.

Reviewer 3: Please... sparse SR-NN method works: SR-NN regularizes the activations across a mini-batch to be instance sparse where a feature is x% instance sparse if it is non-zero for x% of examples in a batch/mini-batch of data. We will add a more detailed description of SR-NN, and a precise definition of instance sparsity in the appendix.

Equation (3) The expectation is taken with respect to all possible length k trajectories, starting from $X_t = x$: over all the random variables $\{(X_{t+i}, Y_{t+i})\}$. The outer integral is an expectation over X_t, according to distribution μ.

Diagram and pseudocode We will move the pseudocode and diagram to the main paper.

Could ... How do these related to Figure 4? The results on the left of Table 1 (One class per task, 50 tasks) correspond to x=50 in Figure 4. Online + MRCL correspond to the MRCL line at x=50 whereas Online + Pretraining corresponds to Pretraining line at x=50.

...EWC are surprisingly low .. why? This is a great question. There are two reasons. (1) EWC tends to do extremely poorly on incremental classification tasks. (2) It does poorly when using a single pass through the trajectory, because the model does not necessarily converge on a task in a single pass. Our results are consistent with those reported by Riemer *et.al* 2019, Chaudhry et.al 2019 and others.