We thank the reviewers for their thoughtful and positive feedback. We are encouraged they found our work well-argued **(R1)** and with substantive contributions **(R2)** . We are pleased **(R3)** finds our work consistently motivated by theoretical insights and **(R4)** identifies the idea as novel and insightful. We address reviewers comments below and we make sure to incorporate all feedbacks.

@**(R1)** - **Baselines not tuned properly**: We used the grid for hyper-params set by their authors. For example, for A-GEM and ER-Reservoir, we had the best hyper-params reported in the appendix of the original papers in our grid.

@**(R1)** - **Emphasize on two distinct contributions**: We agree and we'll update the introduction and conclusion section.
@**(R2)** - **Motivate your new forgetting measure**: We also believe this definition needs more context. Regarding the forgetting measure in experiments we agree it's not new and we will provide additional context to reduce the confusion.

@**(R2)** - **Early stopping as an important tool for CL**: Yes. Early stopping would prevent going further from the previous minima. However, there will be a trade-off since if the parameters do not change enough, the new tasks can not be learned. This is very related to the discussion in l204 as they both refer to the distance from the previous minima.

@**(R3)** - **Experimental setting**: Task labels are provided for CIFAR-100, only to have a similar implementation with other baselines (e.g., A-GEM, ER-Reservoir). Nevertheless, we agree that this deserves explicit discussion and we will update the paper. Regarding the number of epochs for CIFAR experiment, we did so to be compatible with all the benchmarks as explained in the paper. However, we will add the results for more training epochs.

@**(R3)** - **Reproducibility**: We apologize if you had difficulty to reproduce the results. We uploaded the polished and cleaned version of the code since we wanted to make sure the reviewers can investigate the code. However, there were some minor typos in the uploaded code and we sincerely apologize for that. We wish we could upload a new version of the code or share an anonymous link to our experiments to make it easy to verify the results. Unfortunately, according to the NeurIPS guidelines, we can not share any external links. However, as a proof of concept, you can verify experiment 1's rotation MNIST benchmark using the following command which should give an average accuracy of 92.4% with all accuracy metrics above 90%.

```
python -m stable_sgd.main --dataset rot-mnist --tasks 5 --epochs-per-task 5 --lr 0.15 --hiddens 100 --batch-size 16 --gamma 0.25  --dropout 0.25 --seed 1234
```

In addition, you can use the the following command to replicate experiment 2 on CIFAR-100:

```
python -m stable_sgd.main --dataset cifar100 --tasks 20 --epochs-per-task 1 --lr 0.2  --gamma 0.85 --batch-size 10 --dropout 0.1 --seed 2345
```

To ensure all results are reproducible we'll add external links to each experiment in the revision.

@**(R4)** - **Extension with high-dimensional input**: We acknowledge that our results may be affected by the curse of dimensionality. However, we discussed this issue in lines 106-116 and especially in 129-133. Moreover, recent work by Fort & Ganguli ("Emergent properties of the local geometry of neural loss landscapes") suggests that: (1) The Hessian eigenspectrum is composed of a bulk plus C outlier eigenvalues where C is the number of classes, (2) Gradient aligns with this tiny Hessian subspace which implies that most of the descent directions lie along extremely low dimensional subspaces of high local positive curvature. Hence, the bound would be still tight enough for our sake as also demonstrated empirically in Fig2-c,d on high dimensional models.

@**(R4)** - **Other CL metrics**: As noted by the reviewer, the focus of this work was on forgetting. However, we have included results for forward transfer in Fig. 1 & 2 on rotation-MNIST. We see the forgetting for stable net is much less by compromising a negligible amount of forward transfer. This is in line with our discussion on stability-plasticity dilemma. We will include this result in the paper. Thanks for pointing out.

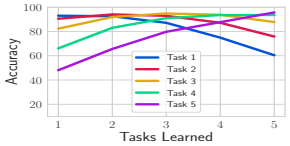

Figure 1: Plastic Net

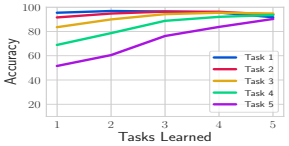

Figure 2: Stable Net

Method	Task 1	Task 2	Task 3	Task 4	Task 5
OGD	75.6	86.6	91.7	94.3	93.4
Stable-SGD	91.5	92.1	95.02	94.2	90.9

Table 1: OGD versus stable-SGD

Method	Average Acc
SGD	53.9 ($\pm$ 4.2)
Stable-SGD	65.5 ($\pm$ 3.7)

Table 2: CORe50 result

@**(R4)** - **Lack comparisons with similar approaches**: We compare our method with OGD. The validation accuracy for each task at the end of CL experience on rotation MNIST with five tasks is reported in Table 1. We will add discussions on OGD [Farajtabar et al] and its extensions [Bennani et al] and other papers (EWC, SI, etc) in the revision.

@**(R4)** - **Benchmarks - low-resolution data with supervised image classification tasks, which are a bit simple**: While we agree with the reviewer that these datasets are simple, we want to highlight that they are difficult for *sequential multitask* problems. For instance, the best result in our paper on rotation MNIST dataset with 20 tasks yields 78% accuracy which leaves a lot of room for improvement. Moreover, to demonstrate the feasibility of the proposed work on other applications we have included a preliminary result for five runs over NI setting of CORe50 dataset with Resnet18 network in Table 2. Full results will be reported in the revision.

@**(R4)** - **Computational metrics**: Computing the Hessian in the code is just for reproducing eigenvalues in Fig 3 and empirically showing that minima found by stable-SGD is wider. It is not a part of the algorithm. Hence, stable-SGD is as computationally cheap as normal SGD, which we believe is an advantage [over many alternatives].