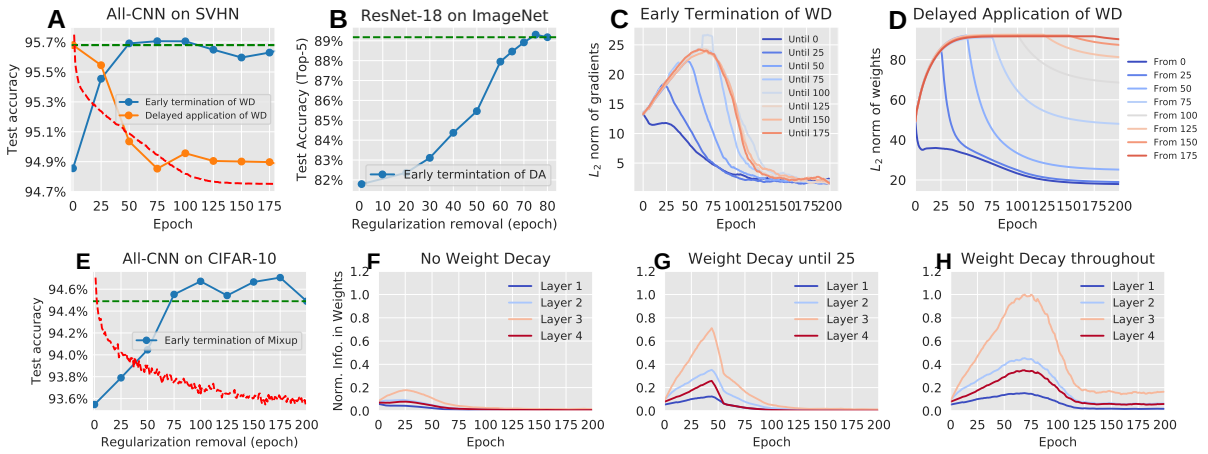

Figure 1: **(A)** Final test accuracy as a function of the weight decay termination/application epoch for All-CNN trained on SVHN dataset. Critical period for regularization occurs during the initial rapid decreasing phase of the training loss (red dotted line), which in this case is from epoch 0 to 75. **(B)** A ResNet-18 trained on ImageNet shows a critical period for data augmentation. **(C)** Plot of the L_2 norm of the gradients during training for ResNet-18 on CIFAR-10. **(D)** Plot of L_2 norm of the weights during training for ResNet-18 on CIFAR-10. The norm curves are almost parallel towards the end of the training, this suggests that further training will not reduce the weights significantly. **(E)** Plot of the test accuracy as a function of the regularization removal epoch for Mixup. **(F-H)** Layer-wise normalized Fisher Information of the weights as a function of training epochs for a ResNet-18 trained on CIFAR-10.

We thank the reviewers for their thoughtful analysis of the paper, and their suggestions, which we address below.

R1: Regularization may not necessarily bias the network toward a set equally good solutions, rather convergence to a single (flat) optimum can explain the results.

This is a viable hypothesis that is not in conflict with our interpretation. After the end of the critical period, the unregularized network may have converged to a (flat) local minimizer with sub-optimal generalization properties. Since adding weight decay at this point does not change the performance of the network, we conclude that weight decay does not (solely) help escaping bad local minimizers by reshaping the loss function. Rather, we claim that weight decay and other regularizers help generalization by biasing the network toward particular minimizers during the initial convergence phase. Note also that applying/removing regularization during the critical period (when the network is far from convergence to a minimum, see Fig. 1C) still leads to partial improvement/deficits.

R1: Delayed application of WD may need more time to reach convergence.

In Fig. 1D we plot the norm of the weights for an additional 25 epochs to confirm that the norm of the weights stabilises and would not improve further with additional training.

R1, R3: Testing on multiple datasets, architectures, and regularizers (*e.g.*, Mixup).

In the paper we verified our claims on a combination of several datasets, architectures, and regularizers, which is sufficient to establish the claim that *"applying regularization at different epochs of training can yield different outcomes."* Nonetheless, to assuage the reviewer's concern, we performed additional experiments on SVHN and on ImageNet (with a correspondingly larger architecture), that further corroborate our result (Fig. 1A-B). As suggested by R3, we also try the Mixup regularization on CIFAR-10 (Fig. 1E). In all cases we observe the same trends described in the paper.

R2: Lack of layer-wise analysis of Fisher. Another decomposition, less trivial to try, would be sample-wise.

See Fig. 1 (F-H) above for layer-wise analysis. Unlike [1] we do not observe changes in the relative ordering of the layers, which makes sense since unlike the experiments in [1], the underlying data distribution is not changing. We will work on sample-wise validation, thanks for the suggestion!

R3: Is the critical regularization period correlated with the training loss? It would be useful to plot the training loss curves. Is the critical regularization period data dependent?

As also observed by [1], critical periods for regularization occur during the initial training epochs when the training loss decreases rapidly, see Fig. 1 (A) and (E) (training loss is the red dotted line). Critical regularization periods do depend on the dataset as different tasks (datasets) have different different learning dynamics and different responses to the regularizers (some regularizer may be ineffective on some datasets), and thus different critical periods.