We thank the reviewers for their detailed comments. Please see our response below.

R1: ○ *"... common implementation of weight decay [1] will usually multiply the amount of weight decay by the learning rate."* The same holds in our setup: We have an L_2 regularization term in the loss. In the gradient descent update equations, this results in a weight decay term where the L_2 coefficient is multiplied by the learning rate, as in [1]. Perhaps this is not immediately obvious in Theorem 2 where we consider gradient flow rather than gradient descent, but the statement holds in that case as well.

○ *"How do different learning rate schedules affect the conclusion?"*: We address LR schedule questions below.

○ *"It would be great if the authors can provide more experiments on ... AUTOL2"* We ran additional experiments training Wide ResNets and ConvNets on CIFAR-100 and SVHN, with similar conclusions: AutoL2 either beats or matches the performance and training speed of a tuned, constant L_2 parameter. As an example, for CIFAR-100 with a fixed lr and evolved for 500 epochs, the optimal L_2 parameter gives a test accuracy of 0.76 while AutoL2 gives 0.79.

R2: ○ *"((1)) If I could have access to the test set..."*. We reject the claim that our submission "violates the ethics of machine learning research". The method we propose for finding the optimal L_2 parameter does not rely on a validation or test set — it can be implemented using training data alone! Our baseline is to compare against tuning on the test set, as is commonly done in the deep learning literature. If our setup introduces any bias (compared with having a separate validation set), it is in favor of the baseline.

○ *"((2)) I have concerns on comparing AutoL2..."*. Indeed, the experiment of Fig. 1c does not include lr decay (as discussed in the text). Experiments with lr decay and AutoL2 are presented in the SM. Please see below for additional discussion.

○ *"((3))) The practically of the proposed work... The most interesting plots might be Test Acc vs Lambda in Fig. 2c and 2f, but the authors didn't report multiple runs of the same configuration..."* Please see Fig. 1a, 2b, and 2e for many additional runs illustrating the same effect. We made an effort in the paper to understand the settings under which our conclusions hold; these are summarized in the Discussion and in SM.D.

R3: ○ *"... more insights on the relation between learning rate scheduler and AutoL2..."* We address this point in the learning decay discussion at the bottom of the page.

○ *"... the lambda update refractory period is not detailed ..."* The refractory period lasts for $\frac{\lambda(t=0)}{\lambda(t)}$ steps. This is explained in the SM, and we will clarify it in the main text. The results are not sensitive to the choice of $\lambda(t=0)$ and we pick 0.1 in our experiments.

○ *"It would be interesting to see on the same graph, training with learning rate scheduler..."* In the SM we have the training curves for the models trained with a learning rate schedule (see figure S10).

○ *"In Figure 1a and 1b, how is the best test accuracy determined?..."* In Figs. 1a,2,3, the model is trained for a specified number of epochs and we report the best test accuracy during training. In Fig. 1b, we have multiple runs with different L_2 and the 'optimal L_2' corresponds to the value which achieved the best test accuracy during training.

R4: ○ *"Default test accuracy of WRN of 0.92 seems a low ..."* The accuracies of Figs. 1a,1b exceed 0.96 for some choices of the L_2 parameter. For figure 1c, we limited the training time to 200 epochs and used a fixed lr, which explains why the accuracy is lower.

○ *"It would be great to have a better description on the relationship of learning rate schedules..."* Please see LR schedule comments below.

○ *"An imagenet training with an overparameterized (resnet50) would boost..."* We agree. We have run several additional experiments on CIFAR-100 and SVHN and the results show consistent improvement when using AutoL2. It is not clear whether ResNet50 on ImageNet (when trained with data augmentation) is sufficiently overparameterized for our purposes — we comment on this in SM.D.

Comments on learning rate schedules. Here we address reviewers' questions regarding learning rate schedules, and we will revise the paper to include these points. Our empirical observations apply in the presence of learning rate schedules, as illustrated in the paper. Figure 1a shows that the test accuracy remains roughly the same if we scale the training time as $1/\lambda$. As to our proposed algorithms, we found that our predicted optimal L_2 value is within an order of magnitude of the tuned value when using learning rate schedules, when accounting for the schedule by weighing each step according to the learning rate. Therefore, as in the case of fixed learning rate, our prediction is beneficial for hyperparameter tuning. For the AutoL2 scheduler, in the presence of learning rate decay we find that it performs similarly to a tuned, fixed L_2 parameter but does not exceed it. This is still beneficial, in that it saves the need to tune the L_2 parameter.