First of all I would like to thank all reviewers for their valuable and unexpectedly detailed reviews!

R1,R2,R4: Figure 1 shows in addition to Figure 2 of the paper that the **parabolic property is also valid on further datasets and at the end of the training**. It fits best for MNIST and worsens for more complex tasks.

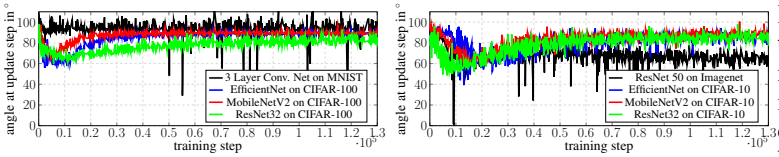

Figure 1: Direction angles at update step position over training steps.

R1: I used a common **weight decay** factor 10^{-4} for all experiments as described in appendix D.5.3. I considered searching for an optimal weight decay factor, but omitted it, since 1494 networks already had to be trained for the evaluation. The **smoothness of the learning rate curves** originates from averaging over epochs which is a standard procedure for plots against epochs. Figure 2 exemplarily shows that on a step wise scale the variance of the learning rate is larger. EfficientNet and MobileNet **contradict the parabolic assumption** only for about the first 500 steps. A **Hyperparameter study** is given in Appendix D.4. My **Box plots** are standard, the circles show outlier outside the interquartile range.

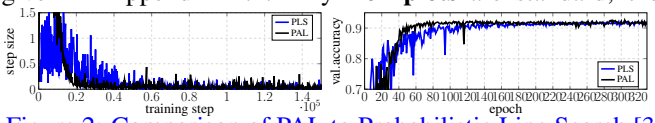

Figure 2: Comparison of PAL to Probabilistic Line Search [36].

R2: There is no easy to use implementation for **PLS [36]**. The sum of squared gradients has to be derived manually for each layer, which is a considerable amount of work for modern architectures but after 2 weeks o work I managed to do it for a ResNet32 see (Figure 2). Comparisons with **second-order methods** are not applicable since the 11GB memory of my graphic cards are not enough to save the Hessian or one of its approximations for the networks investigated.

A comparison of **CPU time** is given in Appendix D.2. The "conjugate" gradient factor has minor influence and most of the best results are achieved without it. Therefore, pure **SGD using PAL's schedule** would mostly exactly train like PAL. In general, however, the optimal schedule depends on the path an optimizer takes on the loss landscape, thus **an optimal schedule** has to be directly inferred during the training process and should generally **not be transferable**. Figure 3 shows that μ **has also a low sensitivity for other networks**.

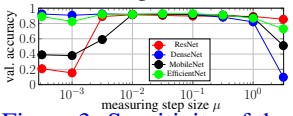

Figure 3: Sensitivity of the measuring step size μ of PAL for several Models on CIFAR-10.

R1,R2,R4: Although the best possible **validation accuracy** to be achieved is an important property, I want to stress that in practice the **robustness** of the hyperparameters is even more important. Thus, please focus your attention on the Box plots on the right of Figure 4,10,11,12. These show that for commonly used hyperparameters, **PAL has a better median for the validation accuracy than SGD** . This is important for scenarios where one has to train very long and cannot try lots of hyperparameter combinations. Note, that for robustness evaluations one should handle the datasets as if they were unknown and one should perform an evaluation with default hyperparameters that a common researcher would chose. Furthermore, several networks should be considered. **This robustness evaluation is one ot the strengths of my work and rarely seen in such detail**. Out of this perspective using ImageNet to check the transfer-ability of hyperparameters is a valid approach. Note that Adam, RMSP and SLS completely fail in this scenario (**R1**).

R1,R2,R4: The **clearly inappropriate name "conjugate" gradient** is a leftover, since I used approaches such as Fletcher Reeves in the beginning of my research to obtain a new direction which fulfills the conjugate condition if one assumes a quadratic. This approach, however, works as well as a fixed β, but is much more expensive. We will change this name. We can assure that the "parabolic property" is also valid in those "adapted" gradient directions. The "conjugate" gradient direction does have minor influence and most of the best results are achieved without it, thus it is **not a confounding factor** in the comparison between PAL and SLS (**R4**).

R4 (also interesting for R2): We have already **incorporated** your important and valuable suggestions about the **clarity of our formulations**, which was of minor effort. However, we have a **conflict of viewpoints and interests** here. This work is one of the rather **rare empirical works in optimization** for deep learning which tries to measure information from **real-world loss functions** to exploit these for optimization instead of starting form **theoretical assumptions, that are never fully valid** in practice (e.g. convexity, over-parameterization, lipschitz continuous gradient). **Thus, I consider my approach as equally valuable as the theoretical approaches** and therefore see no problem in having a weak theoretical part and a strong empirical part instead. **Naturally, the empirical findings of papers with a theoretical approach are relatively weak from a practical point of view due to restrictive assumptions**. My weak results for the SLS approach [54] are a perfect example of this. **I have to emphasize**, that if one comes from empirical results, it is **not necessarily given** that one can do an **in-depth theoretical evaluation**. In my case, the parabolic assumption, which is clearly useful empirically, limits the theoretical analysis because it does not satisfy the lipschitz continuous gradient assumption. Hence, we had to use a simpler theoretical model which still provides convergence guarantees and is likely more valid locally than globally. **I want to encourage you to reconsider your grading again since you mostly focused on the "weakest" part of the paper and mostly ignoring my strong empirical contributions. R1,R2:I also hope to have provided enough evidence for the other authors to rethink their grading.**