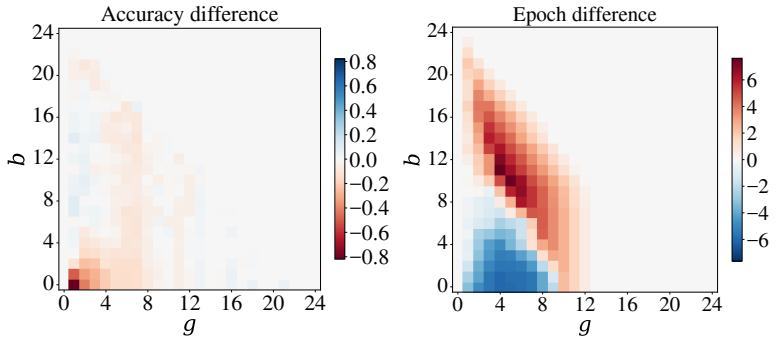
'unrolled' version of this heat map, making it possible to count how many parameter combinations result in early stops while also increasing accuracy, for example. The heatmaps, by contrast, make it possible to compare the behaviour of the two measures with respect to each parameter combination. Finally, the bottom row of every plot shows how many times each measure was triggered for every parameter combination. We consider a measure to be triggered if its stopping condition is satisfied prior to the last training epoch. Due to the way the parameter grid is set up, no configuration above the diagonal can stop, because $b + g$ would be larger than the total number of training epochs. This permits us to compare the 'slopes' of cells for each measure. Ideally, each measure should consist of a dark green triangle, as this would indicate that *parameter* configuration stops all the time.

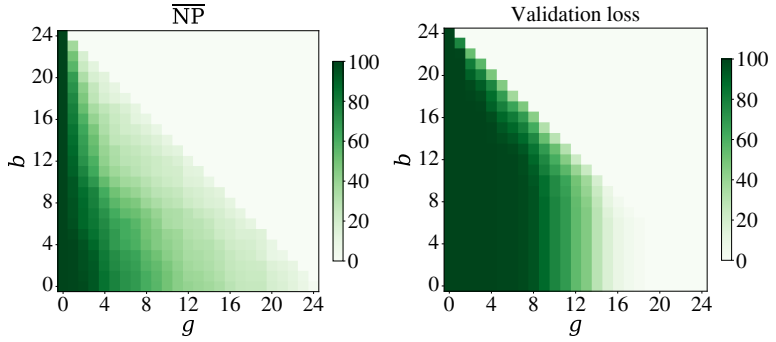
MNIST Please refer to Figures A.2 and A.3. The colours in the difference matrix of the top row are slightly skewed because in a certain configuration, our measure loses 0.8% of accuracy when stopping. However, there are many other configurations in which virtually no accuracy is lost and in which we are able to stop more than four epochs earlier. The heatmaps in the bottom row again indicate that neural persistence is capable of stopping for more parameter combinations in general. We do not trigger as often for some of them, though.

CIFAR-10 Please refer to Figure A.4. In general, we observe that this data set is more sensitive with respect to the parameters for early stopping. While there are several configurations in which neural persistence stops with an increase of almost 10% in accuracy, there are also scenarios in which we cannot stop training earlier, or have to train longer (up to 15 epochs out of 80 epochs in total). The second row of plots shows our measure triggers reliably for more configurations than validation loss. Overall, the scatterplot of all scenarios (Figure A.5) shows that most practical configurations are again located in Q_2 and Q_3. While we may thus find certain configurations in which we reliably outperform validation loss as an early stopping criterion, we also want to point out that our measures behaves correctly for many practical configurations. Points in Q_1, where we train *longer* and achieve a *higher* accuracy, are characterized by a high patience g of approximately 40 epochs and a low burn-in rate b, or vice versa. This is caused by the training for CIFAR-10, which does not reliably converge for FCNs. Figure A.6 demonstrates this by showing loss curves and the mean normalized neural persistence curves of five runs over training (loss curves have been averaged over all runs; standard deviations are shown in grey; we show the first half of the training to highlight the behaviour for practical early stopping conditions). For 'Fashion-MNIST', we observe that $\widetilde{\mathrm{NP}}$ exhibits clear change points during the training process, which can be exploited for early stopping. For 'CIFAR-10', we observe a rather incremental growth for some runs (with no clearly-defined maximum), making it harder to derive a generic early stopping criterion that does not depend on fine-tuned parameters. Hence, we hypothesize that neural persistence cannot be used reliably in scenarios where the architecture is incapable of learning the data set. In the future, we plan to experiment with deliberately selected 'bad' and 'good' architectures in order to evaluate to what extent our topological measure is capable of assessing their suitability for training, but this is beyond the scope of this paper.

IMDB Please refer to Figure A.7. For this data set, we observe that most parameter configurations result in *earlier* stopping (up to two epochs earlier than validation loss), with accuracy increases of up to 0.10%. This is also shown in the scatterplot A.8. Only a single configuration, viz. $g = 1$ and $b = 0$, results in a severe loss of accuracy; we removed it from the scatterplot for reasons of clarity, as its accuracy difference of -21% would skew the display of the remaining configurations too much (this is also why the legends do not include this outlier).



(a) Accuracy and epoch differences



(b) Number of triggers

Figure A.2: Additional visualizations for the 'MNIST' data set.

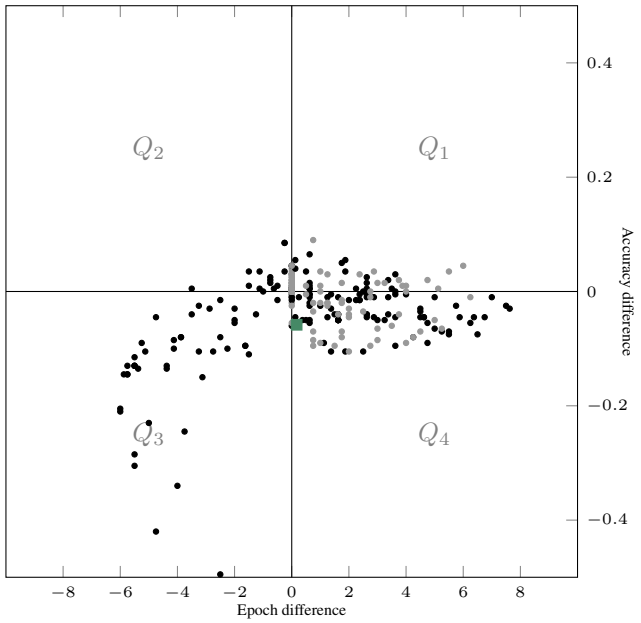
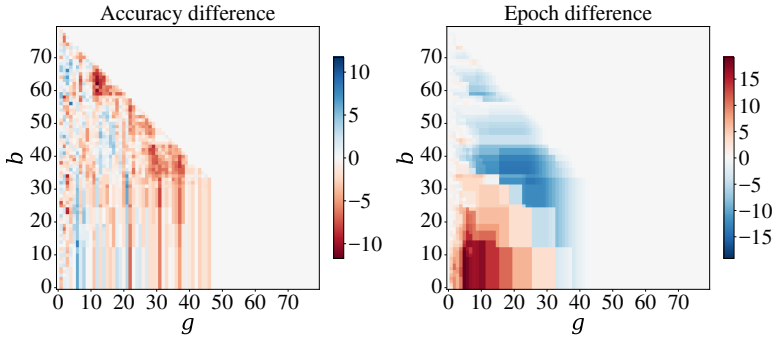
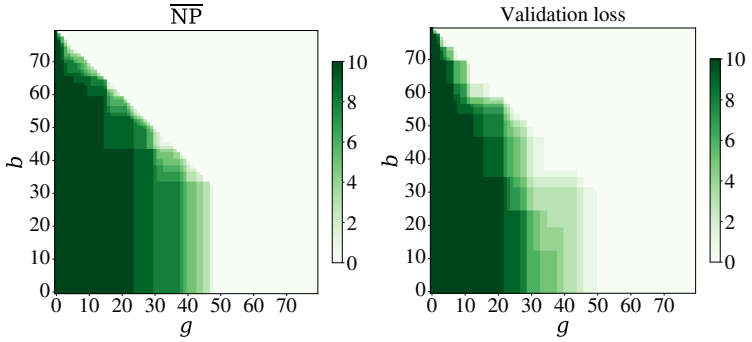


Figure A.3: Scatterplot of epoch and accuracy differences for 'MNIST'.



(a) Accuracy and epoch differences



(b) Number of triggers

Figure A.4: Additional visualizations for the 'CIFAR-10' data set.

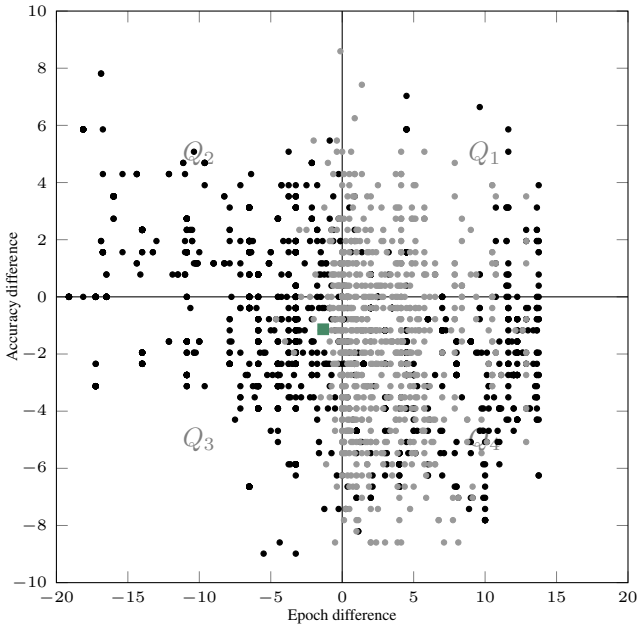


Figure A.5: Scatterplot of epoch and accuracy differences for 'CIFAR-10'.