HOW CURRICULUM LEARNING IMPACTS MODEL CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the significant progress made on deep learning models, concerns yet exist when a trained model is deployed to real-world applications. Model *calibration* is a key consideration that has recently attracted more attention—a learned model should not only achieve high predictive performance but also attain that with a proper level of confidence—a mismatch between predictive performance and confidence creates miscalibration and hence raises concerns about trusting a (miscalibrated) model. Even with the importance of the problem and many recent research efforts, calibration has not been fully understood yet, particularly when it faces the common challenges that deep learning models struggle with: specifically limited training resources and noisy data. In this paper, we study calibration emphasizing these scenarios. We particularly investigate the effect of curriculum learning, which, inspired by human curricula, leverages a guided learning regime to improve model generalization and has been found to improve predictive performance in the aforementioned cases. Specifically, we provide an empirical understanding on the impact of curriculum learning on model calibration under a variety of general contexts. Our studies suggest the following: most of the time curriculum learning has a negligible effect on calibration, but in certain cases under the context of limited training time and noisy data, curriculum learning can substantially reduce calibration error in a manner that cannot be explained by dynamically sampling the dataset. Second, curriculum and anti-curriculum learning appear to have nearly identical effects on model calibration. Lastly, the choice of pacing function and its parameters in curriculum learning can significantly impact model calibration, indicating that extra care should be taken to minimize the risk of severe model miscalibration. We hope the empirical insights will help us better understand calibration and guide the utilization of curriculum learning in practice.

1 INTRODUCTION

Deep learning has achieved state-of-the-art performance in a wide range of problems (LeCun et al., 2015; Bengio et al., 2021). Nevertheless, concerns exist when trained models are deployed for real-world applications (Kelly et al., 2019). Calibration is one of the key considerations that has recently attracted serious attention (Guo et al., 2017; Minderer et al., 2021). For example, in many safety critical applications (Jiang et al., 2012), it is crucial for a classifier to not only achieve high predictive performance but also attain that with a proper level of confidence—both underconfidence or overconfidence should be avoided, and any mismatch as such creates miscalibration and raises concerns about being able to trust model predictions. Research in the field of calibration has focused on recalibration methods (Guo et al., 2017), calibration metrics (Nixon et al., 2020), and how out-of-distribution (OOD) data affects confidence scores (Lee et al., 2018).

Many open questions related to deep learning models pertain the limited training resources (e.g., training data and training time) and noise in data. Such factors also affect the calibration of deep neural networks in a negative manner (Zhao et al., 2020), and a wide variety of techniques have been proposed to help overcome the challenges. Building on existing work on model calibration, in this work we investigate *curriculum learning* for calibration, which has become a popular paradigm in machine learning in general, and in particular for coping with limited training resources and data noise. Curriculum learning has been widely used in various problems in supervised learning (Hacohen & Weinshall, 2019), semi-supervised learning (Gong et al., 2016), and reinforcement

learning (Narvekar et al., 2020). Inspired by an intuitive notion of how curricula affect human learning, whereby humans typically learn better starting from easy problems and working their way to harder problems, curriculum learning in a machine learning context refers to presenting easier training samples earlier in training and gradually adding more difficult samples as training proceeds (Elman, 1993; Bengio et al., 2009). An alternate type of ordering called anti-curriculum learning is also widely used which learns in the opposite order—ergo hardest first. Benefits have been observed using both orderings in denoising (Jiang et al., 2018), faster generalization, and smoother gradients (Bengio et al., 2009; Weinshall et al., 2018), although results range widely between the best choice of ordering and consistent recommendations for use are hard to obtain (Wang et al., 2020). In recent work, it has been shown that for standard benchmark datasets, curriculum learning has significant benefits in improving model accuracy with limited training time and noisy data (Wu et al., 2021; Ovadia et al., 2019). Despite these revelations of the benefits of curriculum learning, the question of whether the same positive relationship exists with calibration remains unexplored.

When considering what theoretical effects curriculum learning can have on calibration, we note that many calibration methods work by modifying how models are trained (Kumar et al., 2018; Kong et al., 2020; Müller et al., 2019). Thus, it is reasonable to assume that curriculum learning, which affects the nature of training by altering the optimization strategy, can have an influence on calibration. Particularly curriculum learning's purported benefits in generalization and regularizing training towards better regions in parameter space by optimizing a smoother version of the training objective (Wang et al., 2020) can theoretically punish overconfidence. It has been observed that the increase in a neural network's confidence over the course of training is one of the key causes of miscalibration (Mukhoti et al., 2020), and as a result being exposed to the most difficult samples far into training can mean there is less of a chance of a model becoming overconfident on data that it is the poorest at classifying. Furthermore, when applying curriculum learning, at every training step the algorithm is deciding which subset of the training set to learn on. Depending on the initial subset used and the function used to add more samples for training each step, the degree of miscalibration over time can be strongly affected by the biases in that subset. Hence not only do we examine the overall effect on calibration, but also the propagation of error over the course of training.

We closely follow the approach taken by Wu et al. (2021) to conduct our experiments. We build on top of their work by collecting information with regards to the calibration error of models trained under different types of orders and pacing functions, using a number of metrics to validate our findings. For our experiments we do not apply any recalibration techniques to our models. Rather we wish to see 1) if there is any inherent advantage to curriculum learning in terms of reducing calibration, particularly, in the case of limited training time and noisy data where curriculum learning is found to have the most benefit; 2) if there are any pitfalls in certain curriculum learning or anti-curriculum learning configurations that would discourage their use in settings where good classifier calibration is critical. These factors are important to analyze as curriculum learning is deployed in situations where calibration is a major concern.

To the best of our knowledge, our research is the first attempt to investigate the influence of curriculum learning on calibration. Our key observations are four-fold. First, curriculum learning does have an influence over calibration, but we demonstrate that this influence is inconsistent in general. In many cases there is no statistically significant benefit over standard training, especially for full time training. However, there are specific cases where its benefits can be observed. Second, the significance of curriculum learning for calibration is most notable in limited training time and noisy data where substantial reductions in calibration error are observed compared to standard training, that are not seen when training using a dynamic-curriculum. This shows that curriculum learning is providing benefits that cannot be explained by dynamically sampling the dataset. Third, curriculum and anti-curriculum learning appear to have nearly identical trends regarding model calibration. Lastly, the choice of pacing function and its parameters can markedly impact calibration, necessitating extra care to be taken in selecting these hyperparameters to minimize risk of serious model miscalibration.

2 BACKGROUND

2.1 BASIC NOTATION AND DEFINITION OF CALIBRATION

Let $X = \{x_1, x_2, ..., x_N\}$ be a set of N feature vectors where each element $x \in \mathbb{R}^d$ has dimensionality d. Let N be defined as the size of the test set being evaluated. Let $Y = \{y_1, y_2, ..., y_N\}$ be the corresponding true labels where $y \in \{1, 2, ..., K\}$ and where K is the number of classes. We define the classifier, in this case a neural network, as a function $\mathcal{F} \rightarrow f(x) = (\hat{Y}, \hat{P})$ that takes an input datapoint and outputs a predicted class $\hat{Y}$ and corresponding predicted probability distribution over the K classes $\hat{P} = \{\hat{p}_1, \hat{p}_2, ..., \hat{p}_K\}$. In a typical supervised neural network, this probability vector is often produced after the softmax is taken in the final output layer where $\sum_{i=1}^{K} \hat{p}_i = 1$. The confidence score is taken as the probability of the predicted class. A few different notions of calibration exist. The strongest view, multiclass-calibration, is defined in the following equation for all input datapoints $(x_n, y_n) \in \mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ in dataset $\mathcal{D}$:

$$\mathbb{P}(y_n = k \mid \hat{P}_k(x_n) = p) = p, \ \forall \, k \in \{1, 2, .., K\}, \tag{1}$$

where p is a prediction vector $p = \{p_1, p_2, p_k\} \mid \sum_{i=1}^{K} p_i = 1$. In this view, a network is considered well calibrated when it predicts a probability distribution over all the classes and the probability that the model predicts the correct labels matches the probabilities from its predicted distribution over the classes. Any mismatch between these, the left and right hand sides of the equation, creates miscalibration (calibration error).

A less stringent and more commonly used notion in various metrics is called classwise calibration (Kull et al., 2019). Here for all input datapoints $(x_n, y_n) \in \mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ in dataset $\mathcal{D}$:

$$\mathbb{P}(y_n = \hat{Y}(x_n) \mid \hat{p}_k(x_n) = p_k) = p_k, \tag{2}$$

where p_k is the confidence score for predicted class k. Here only the predicted (or a specific class of choice) is considered. It is important to emphasize that good calibration and accuracy are both desirable properties, thus it is key to attempt to optimize both simultaneously and not improve one at the expense of the other (Krishnan & Tickoo, 2020).

2.2 CALIBRATION METRICS

Equation 2 is an idealized representation of calibration, which, however, is impractical to calculate in practice as it features a continuous function that requires infinite datapoints to compute the true value. Numerous metrics have instead been developed to approximate calibration error, with their own advantages and drawbacks. To provide a complete picture in our research we use three metrics.

Binning-based metrics remain the most popular method of approximating calibration error, of which expected calibration error (ECE) (Pakdaman Naeini et al., 2015) is widely used in research and the primary one we use for this study. To calculate ECE, the confidence scores on the predicted classes are binned into *M* number evenly spaced bins, and the weighted sum over the differences between the average confidence score and accuracy in each bin constitutes the expectation of the calibration error of the model. This can be seen in the equation for ECE:

$$ECE = \sum_{m=1}^{M} \frac{n_m}{N} |acc(B_m) - conf(B_m)|, \tag{3}$$

where B_m are the data points in the m^{th} bin, and n_m is the number of data points in the bin. ECE and other metrics relying on binning have documented flaws and should not be used as the only measurements of calibration error. The concerns include being sensitive to the number of bins chosen and not being a proper scoring rule (Ovadia et al., 2019), leading to situations where there can be zero or minimal calibration error despite severe miscalibration (Nixon et al., 2020). An additional recently proposed metric, Kolmogorov-Smirnov Calibration Error (KS error), is binning-free and rectifies some of these aforementioned flaws (Gupta et al., 2021). We use it alongside ECE as one of our primary metrics.

Central to the KS error is leveraging the Kolmogorov-Smirnov statistical test for comparing the equality of two distributions. The cumulative probability distributions of the confidence scores and

Table 1: Expressions for the six different types of pacing functions we test.

Pacing Function Type	Expression for ratio of training set used at step t
Logarithmic:	$Nb + N(1-b)(1 + 0.1log(\frac{t}{aT} + e^{-10}))$
Exponential:	$Nb + \frac{N(1-b)}{e^{10}-1}(e^{\frac{10t}{aT}} - 1)$
Step:	$Nb + N\lfloor\frac{t}{aT}\rfloor$
Root:	$Nb + \frac{N-b}{\sqrt{aT}}\sqrt{t}$
Linear:	$Nb + \frac{N-b}{aT}t$
Quadratic:	$Nb + \frac{N-b}{(aT)^2}t^2$

labels are compared and the maximum difference between them is calculated and taken as the error. First the predictions are sorted according to the confidence score on class k, i.e., $\hat{p}_k$:

$$\begin{aligned} \text{KS error} &= \max_i |h_i - \tilde{h}_i|, \\ \text{where, } h_0 &= \tilde{h}_0 = 0, \\ h_i &= h_{i-1} + \mathbf{1}(y_i = k)/N, \\ \tilde{h}_i &= \tilde{h}_{i-1} + p_k(x_i)/N. \end{aligned} \tag{4}$$

Lastly, we use an additional metric called contraharmonic expected calibration error (ECE) (Obadinma et al., 2021). This metric was originally introduced for imbalanced data classification by taking the contraharmonic mean over the ECE calculated for each class when binning each class individually. We find that since ECE values for individual classes vary widely when there are many classes, we use CECE to be able to judge if any individual classes become miscalibrated compared to what is suggested by the ECE value for the whole data. It is undesirable for there to be severe miscalibration on certain classes that are harder to predict for example. CECE is defined as follows for class-wise ECE_i calculated using only datapoints belonging to each class k.

$$CECE = \frac{ECE_1^2 + ECE_2^2 + ... ECE_K^2}{ECE_1 + ECE_2 + ... ECE_K}. \tag{5}$$

2.3 Curriculum Learning Components

In curriculum learning an explicit curriculum has to be defined that alters the order a model is exposed to the training data. The paradigm we follow has been widely used (Bengio et al., 2009; Hacohen & Weinshall, 2019) where a curricula is defined as having two necessary components, a scoring function S and a pacing function g_θ. The scoring function $S(x, y)$ creates a mapping for each training datapoint to a scalar score based on the difficulty of the sample. Scoring functions are typically loss based and scoring can be done as the model is being trained by dynamically updating scores as in the case of self-paced learning (Kumar et al., 2010). They can also be static and pre-calculated using a separate model or pre-defined method (Wu et al., 2021). We do not focus on the influence of the scoring function on calibration but it is a worthy topic to examine further. Instead, we use the estimated c-score loss scoring function used in Wu et al. (2021) that was found to perform the best. Originally proposed for detecting of regularities in data (Jiang et al., 2020), here the scoring function is defined as $S(x_n, y_n) = \mathbb{E}^r_{D \overset{x}{\sim} \hat{D} \setminus \{x_n, y_n\}}[l(x_n, y_n)|D]$ where D is a training set sampled from the full training set but without datapoint (x_n, y_n). $l(x_n, y_n)$ is the loss over the datapoint. Essentially, the score for a sample captures how consistently a group of models trained on random training sets of various sizes excluding the datapoint can predict the correct label, to ensure that the regularity of a given datapoint is represented.

The pacing function g_θ, at training step t out of the total training steps T, determines the mini-batches available for training based on the scoring functions which are sampled uniformly for that step. In the case of curriculum learning, the data is sorted from the lowest score to the highest using the scoring function. The pacing function selects the lowest scoring samples/mini-batches for training at each step. Pacing functions are parameterized by $\theta = (a, b)$ where a is what fraction of training time it takes the pacing function to begin sampling from the entire training set, and b denotes what fraction of the training set the pacing function exposes to the model at the start of training. Both can have dramatic effects on how the model trains, and due to this are an important topic for this study.

Theoretically, a can range from 0 to infinity, though any value above 1 means the full training data is not utilized. b ranges from 0 to 1. There are different types of pacing functions corresponding to common function families that alter at what rate the training set is gradually increased, with different functions affecting whether the size of the set increases fast at the start of training or vice-versa. As in Wu et al. (2021), we test the 6 function families: logarithmic, exponential, step, linear, quadratic, and root, and their definitions can be seen in Table 1.

After selecting the scoring function and pacing function the final step is determining the order. Curriculum learning orders samples from lowest scoring to highest scoring, anti-curriculum learning goes from highest to lowest scoring. Random-curriculum randomly samples data not according to the scoring function in a given step, and the pacing function serves to dynamically increase the amount of data the model is exposed to in a given step.

3 Experimental Setup

We base our experiments on the work of Wu et al. (2021). In addition, we add on top of their work our analysis of the calibration of the models using the metrics detailed in Section 2.2. We focused on one standard model for image classification, ResNet-18 (He et al., 2016), on the benchmark datasets CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009). The same settings specified in Wu et al. (2021) are used in our research, which will be briefly summarized in this section. Random horizontal flip and normalization are used as data augmentation methods. The hyperparameters for the model include: a batch size of 128, a learning rate set at an initial value of 0.1 with a cosine decay learning rate scheduler, 0.9 momentum, 5e-4 weight decay, and a Stochastic Gradient Descent (SGD) optimizer. We use these same settings for standard training, curriculum, anti-curriculum, and random ordering for a fair comparison. We use 45,000 images for training and 5,000 for validation. Note that for all the models we present the calibration results for the split with the best validation accuracy. We use the pre-generated orders based on c-scores for everything apart from standard training. As brought up in Wu et al. (2021), for the curricula approaches the class balance during training is still preserved, and the learning rate still decays to 0 even with the reduced training times. The number of iterations per epoch varies during training since the size of the dataset increases. As a result, the number of actual epochs changes depending on the pacing function. Thus, it is more accurate to refer to the total number of iterations/steps through the data during training, which is calculated using the formula $(training\ set\ size/batch\ size) * number\ of\ epochs$. Lastly for both ECE and CECE we use 15 bins to estimate the error.

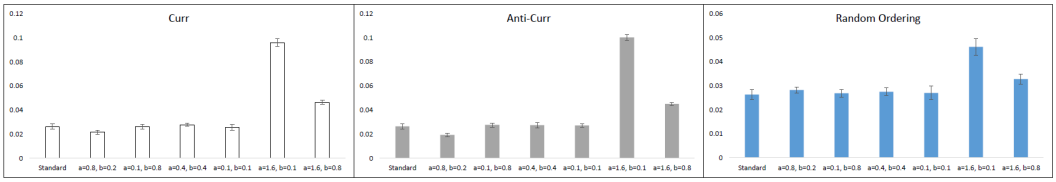


Figure 1: Bar plots comparing expected calibration error (ECE) on CIFAR-10 at 25 epochs (8800 steps) for standard training and 6 different combinations of pacing function parameters. We compare results on curriculum ordering (left), anti-curriculum ordering (center), and random-curriculum (right). A linear pacing function is used for all trials. The error bars show the standard deviation of the measurements. It can be seen that one configuration ($a = 0.8, b = 0.2$) produces lower ECE than standard training considering the error. The same effect is not seen in the random-curriculum where the same parameters performer poorer than standard training.

4 Observations on Order and Pacing Function Parameters

The training order is crucial for curriculum learning. We compare the performance between curriculum learning, anti-curriculum learning, and random-curriculum. In the literature, the preference over these types of ordering is largely unsolved, though in general Wang et al. (2020) suggest that curriculum learning tends to be helpful in noisy and difficult settings, while a preference for anti-curriculum may be considered for cleaner datasets. To facilitate the comparison, we test a few configuration of the hyperparameters *a* and *b* for each type of curriculum and compare the results

to standard training. We test over three different training durations (35,200, 8,800, and 1,760 iterations) corresponding to 100, 25, and 5 epochs through the data, respectively, in order to capture a good range of various training times. We base the range of possible values from Wu et al. (2021) where they have $b \in \{0.0025, 0.1, 0.2, 0.4, 0.8\}$ and $a \in \{0.01, 0.1, 0.2, 0.4, 0.8, 1.6\}$. Due to the computational cost of testing every possible combination, we choose only a few different representative combinations, in order to capture combinations that have a fair range of low, medium, and high values. Specifically we choose, $a = 0.8, b = 0.2$; $a = 0.1, b = 0.8$; $a = 0.4, b = 0.4$; $a = 0.1, b = 0.1$; $a = 1.6, b = 0.1$; $a = 1.6, b = 0.8$. For each different configuration we test (e.g. curriculum learning $a = 0.8, b = 0.2$ with a linear pacing function), we average the results over three randomly initialized trials for 100 and 25 epochs and calculate the standard deviation. For five epochs we run five trials for every configuration due to the lower computational cost. Full tabular data can be seen in Section A.3 in the Appendix.

Through our analysis, we determine that there is a great similarity in the calibration error produced by models training using curriculum and anti-curriculum learning considering all metrics, and the trends are near identical to both. Figures 3a and 3b both show near identical trends between curriculum learning and anti-curriculum learning in all tested configurations when considering change in calibration during training, but a notable contrast can be seen with random ordering. Note that the choice of a and b has a larger influence over calibration than choosing between curriculum or anti-curriculum learning. Values of the parameter a greater than 1 have increased calibration error compared to the other configurations. This can be seen in the configurations with *a=1.6*, with *a=1.6, b=0.1* being particularly poor parameters for the pacing function creating the highest miscalibration in all tested scenarios, likely due to poorer generalizability as a result of the decreased training set. The random-curriculum is interesting in that there is no case where it outperforms the standard training scenario conclusively, and it largely performs in line or slightly worse than standard training. In any case, curriculum parameters have to be carefully tuned to have a noticeable effect and most of the time the effect on calibration is negligible, and we note the beneficial scenarios below. We also want to point out that the calibration metrics generally agree with one another.

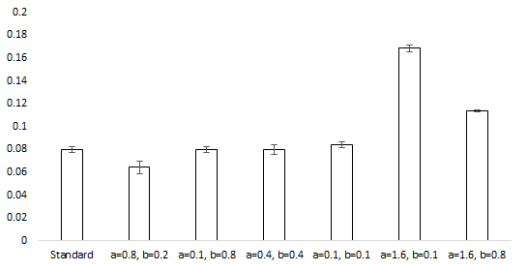
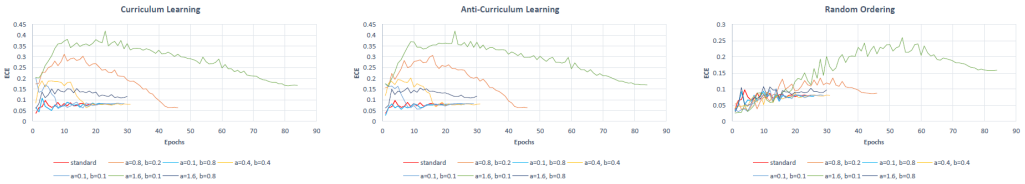


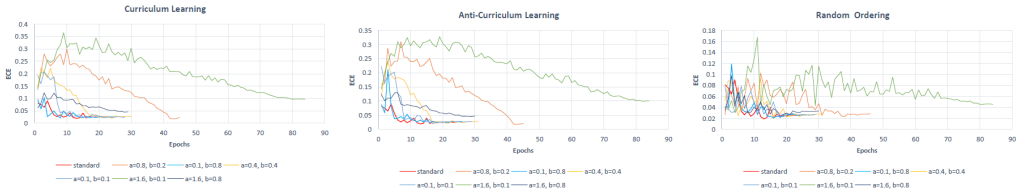
Figure 2: Bar plot comparing the ECE on the test set of CIFAR-100 at 25 epochs (8,800 steps) using curriculum learning ordering models. We compare the same 6 combinations of pacing function parameters with the standard training model (leftmost). Similar to CIFAR-10, $a = 0.8, b = 0.2$ produces lower ECE than standard training.

We observed that for the full 100 epochs there is no benefit to curricula approaches. Rather, in most cases the error becomes slightly higher. As such, we do not think curriculum learning is of much interest in this scenario. When considering the more limited training time scenario of 25 epochs, the benefits of curriculum learning can be noted. As can be seen in Figure 1 and 2, for both curriculum learning and anti-curriculum learning with $a = 0.8, b = 0.2$, the calibration error drops below the standard training conclusively when considering the standard deviation. This is true for both CIFAR-10 and CIFAR-100. This effect is not observed for the random order. This signals that curriculum learning has the potential to help model calibration. The only issue is the decrease in accuracy compared to standard training, creating a trade-off. Most of the other configurations are about equal with standard training or poorer signifying that tuning is critical to obtain benefits.

There is great variability in the results at 5 epochs that make it difficult to conclusively determine whether curricula are beneficial. The models are not trained for a long enough duration to fit the data. Nevertheless, some configurations can still be conclusively proven worse, namely those with an a value of 1.6. Despite averaging better accuracy at $a = 1.6, b = 0.8$, it is not recommended as it produces higher calibration error compared to standard training. Furthermore, $a = 0.8, b = 0.2$ actually performs quite poorly here at 5 epochs compared to 25 epochs, signalling that each training time appears to have its own ideal pacing function parameters. Regardless, we can see that the significance of curriculum learning for calibration is most notable in limited training time. In this case we see substantial reductions in calibration error compared to standard training that are not present when simply dynamically sampling the dataset.



(a)



(b)

Figure 3: Graphs capturing the change in calibration over time by tracking the change in ECE on test set per training epoch for CIFAR-10 (a) and CIFAR-100 (b) at 25 epochs where the curricula approaches are found to have the most benefit. Three different orderings are shown: curriculum (left), anti-curriculum (centre), and random (right). The trends for standard training can be seen in the bright red line. We test six configurations of pacing function parameters using a linear pacing function. The most prominent observations are: (1) curriculum and anti-curriculum learning have near identical trends; (2) they both differ from random ordering; (3) the curricula-based approaches are prone to severe miscalibration early on during training that gradually improves to finish even below the error of standard training.

5 Effect of the Choice of Pacing Functions

In this section, we provide empirical evidence for understanding the effect of varying the choice of pacing functions on calibration error.

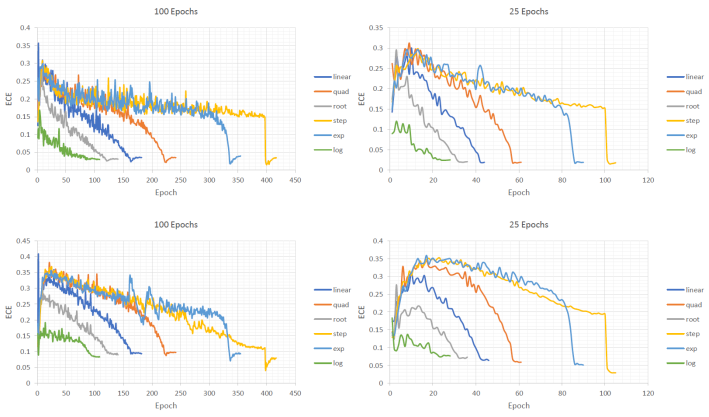


Figure 4: Top row: CIFAR-10 results. Bottom row: CIFAR-100 results. Graphs comparing the progression of test ECE over the course of training. Left graph is for 100 epochs (35,200 training steps) and right is 25 epochs (8,800 training steps). Note the difference in how different types of pacing functions converge to their final calibration error despite the same parameters and especially the sudden improvement in calibration seen with the step function.

Given the similarity in results we observed between curriculum learning and anti-curriculum learning in the previous section, we choose to only test the variability in pacing functions using curriculum learning. We keep the parameters a and b constant at 0.8 and 0.2, respectively, except for five epochs where we use $a = 0.1, b = 0.8$ for CIFAR-10 and $a = 0.1, b = 0.1$ for CIFAR-100, since they have better performance at this training time. We alter the pacing functions among the choices detailed in section 2. We test at the three different training times at the 35,200, 8,800, and 1,760 steps. Full results are in Section A.3 of the Appendix. Overall, we demonstrate that the empirical study demonstrates that pacing functions have a substantial effect on model calibration and that it is key to consider them. Differences

between pacing functions can be large despite the same configuration of pacing function parameter. The best performing pacing function is strongly case dependant and varies by dataset and training time as we detail in the following paragraphs.

At 100 epochs (35,200 steps), the logarithmic pacing function performs the best, even more so than standard training in the case of CIFAR-10. In contrast, the step pacing function performs very well for CIFAR-100, yet calibration error is average among the options for CIFAR-10. Most of the functions offer negligible difference among each other and no statistically significant benefits can be observed. Numerous differences are seen for limited training under 25 epochs (8,800 training steps). The step pacing function performs the best for CIFAR-10 along with the quadratic function, however the accuracy for step is notably worse than for quadratic despite a minor difference in ECE. The logarithmic pacing function performs the poorest among the options. For CIFAR-100, step is by far the best in terms of calibration and gets an extreme reduction in ECE and KS error. Log is again found to perform worst and calibration error is notably higher than linear and root, the next highest. At five epochs, four pacing functions (linear, quadratic, root and log) have very similar calibration error values. The step function and especially the exponential function perform very poorly; the exponential error is almost double in CIFAR-10 compared to the average. For CIFAR-100 the step and exponential pacing functions are the best in terms of ECE, with step excelling particularly when considering the KS error. Linear, quadratic, root, and log are similar in terms of ECE, but vary widely in terms of the KS error, making it difficult to evaluate their performance. Once again, due to inconsistency, it is not clear which choice of pacing function creates any conclusive benefit for this highly limited training time.

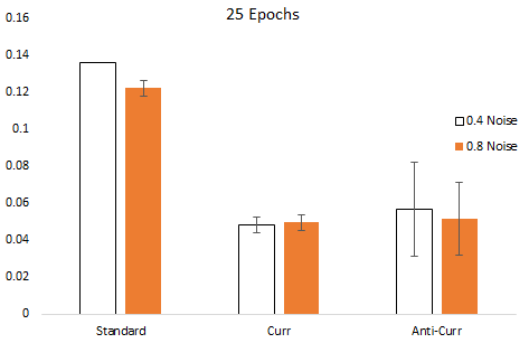


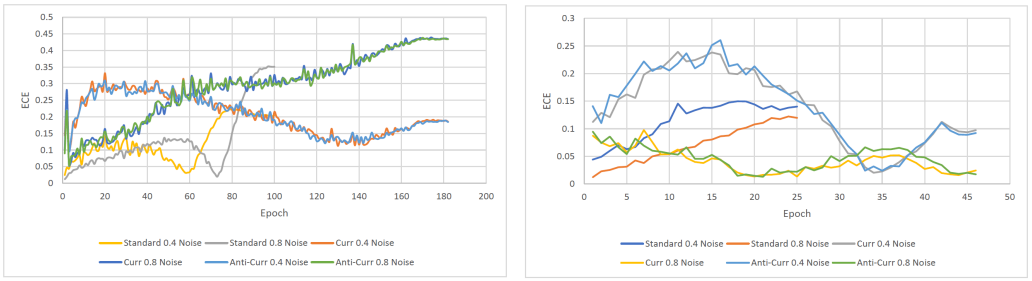
Figure 5: Bar plot showing the high effectiveness of curriculum learning and anti-curriculum learning at reducing calibration error on the test set compared to regular training with 40% and 80% label noise at 25 epochs on CIFAR-100.

6 Impact of Noise

In this section we show that it is under noisy data that we observe the largest benefits for using curricula to reduce calibration error.

We keep the parameters a and b constant at 0.8 and 0.2, except for CIFAR-10 at 5 epochs where its $a = 0.1, b = 0.8$. The standard training, curriculum learning, and anti-curriculum learning are compared. We forgo random ordering since we want to understand whether the previously observed benefits of curriculum learning occur for noisy data. We test at two different noise levels, 40% and 80% label corruption following the approach of Hendrycks & Dietterich (2019).

There is no benefit to using curriculum learning when training for the full 35,200 training steps and it appears to hurt the calibration of the model slightly in the case of CIFAR-10 and heavily in the case of CIFAR-100 for high levels of label noise. The accuracy, however, does improve as was observed in Wu et al. (2021). Figure 6 shows that standard training is able to hit a lower minimum error at its best epoch than curriculum learning at the same noise threshold. At 8,800 training steps there are significant benefits to be gained by using curriculum learning. Both CIFAR-10 and CIFAR-100 have high reductions under both curriculum learning and anti-curriculum learning, as can be seen in Figure 5. This decrease is far starker than observed in any other scenario and the benefits are significant enough to recommend using curriculum learning as a calibration method. Observing the ECE progression in Figure 5, the curricula approaches manage to calibrate better over time while standard training only gets more miscalibrated. Both curriculum learning and anti-curriculum learning have near identical trends. At 1,760 training steps, results are mixed. For CIFAR-10 the calibration error is about the same. For CIFAR-100 there appears to be a disadvantage to using curriculum learning, especially when there is 80% label noise. In contrast to the results at 25 epochs, we cannot recommend curriculum learning for a training time this low.



(a) (b)

Figure 6: Graphs showing the change in model calibration over time under noisy data while comparing standard training, curriculum learning, and anti-curriculum learning. The curricula share the same linear pacing function at $a = 0.8, b = 0.2$ We track the change in ECE on test set per training epoch for CIFAR-100 at 40% and 80% label noise. (a) shows results at 100 epochs and (b) at 25 epochs. There is a stark difference in the trends between the two training times that favours standard training for 100 epochs and the curricula approaches for 25 epochs. The noise level strongly affects the trends in a difficult to predict manner.

7 Related Work

Recent studies have been conducted into the calibration of specific types of model architectures, including convolutional based architectures such as ResNet and DenseNet (Guo et al., 2017; Minderer et al., 2021), pre-trained transformer-based models (Desai & Durrett, 2020), and ReLU-type neural networks (Hein et al., 2019). In addition, there have been investigations into the effectiveness of specific calibration methods like temperature scaling (Kumar et al., 2019). Similar to these approaches, we seek to evaluate the calibration of models trained with curriculum learning based architectures.

We would also like to highlight the work by Sakaridis et al. (2019), where they use a guided curriculum adaptation to reach state-of-the-art performance on a semantic nighttime image segmentation task under an uncertainty aware metric that rewards predictions with confidence consistent with human annotators. This study utilizes curriculum learning to improve confidence scores, but unlike our work they do not provide a deeper examination into calibration specifically.

For a fuller examination of related work in the fields of calibration and curriculum learning, please see Appendix A.1.

8 Conclusions and Future Work

Along with the recent surge of interest in curriculum learning, we provide the first empirical study on this promising technique for calibration. Our research here answers the following question: does curriculum learning have a significant impact on model calibration similar to its great benefits in improving model accuracy? Through extensive experiments, we contributed the following main insights. First, under the context of limited training time and noisy data, curriculum learning can substantially reduce the miscalibration error in certain cases, which cannot be explained by dynamically sampling the dataset. Second, curriculum and anti-curriculum learning appear to have nearly identical effects on model calibration. Last, the choice of pacing function and its parameters in curriculum learning can significantly impact model calibration, indicating that extra care should be taken to minimize the risk of severe model miscalibration. Based on the empirical observations presented in this study, we are interested in establishing a theoretical framework that can systematically analyze the interplay between curriculum learning and model calibration in the future.

9 Reproducibility Statement

Our work is easily reproducible by following our experimental protocol in Section 3. Section 2.2 shows how we calculate the calibration error metrics. In terms of the curricula parameters that we use, Section 4 lists the pacing function parameter combinations, Section 2.3 shows the pacing

function types, and Section 6 shows the settings for the amount of label corruption. Since we follow the approach by (Wu et al., 2021), the Github for their experiments can be referenced and used to replicate our experiments. We will make our modifications to their code publicly available upon the acceptance of the paper.

REFERENCES

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for ai. *Commun. ACM*, 64(7): 58–65, June 2021. ISSN 0001-0782. doi: 10.1145/3448250. URL https://doi.org/10.1145/3448250.

Shrey Desai and Greg Durrett. Calibration of Pre-trained Transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Jeffrey L. Elman. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99, 1993. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(93)90058-4. URL https://www.sciencedirect.com/science/article/pii/0010027793900584.

Chen Gong, Dacheng Tao, Stephen J. Maybank, Wei Liu, Guoliang Kang, and Jie Yang. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7):3249–3260, 2016. doi: 10.1109/TIP.2016.2563981.

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1311–1320. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/graves17a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1321–1330. JMLR.org, 2017.

Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *European Conference on Computer Vision (ECCV)*, September 2018.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2021.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2535–2544. PMLR, 2019. URL http://proceedings.mlr.press/v97/hacohen19a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41–50, 2019.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association : JAMIA*, 19:263–74, 2012.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C. Mozer. Exploring the memorization-generalization continuum in deep learning. *CoRR*, abs/2002.03206, 2020. URL https://arxiv.org/abs/2002.03206.

Christopher Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17, 12 2019. doi: 10.1186/s12916-019-1426-2.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1326–1340, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.102. URL https://www.aclweb.org/anthology/2020.emnlp-main.102.

Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 2020.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

Meelis Kull, Miquel Perelló-Nieto, Markus Kängsepp, Telmo de Menezes e Silva Filho, Hao Song, and Peter A. Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *NeurIPS*, 2019.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *NeurIPS*, 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2805–2814, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/kumar18a.html.

M. Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/e57c6b956a6521b28495f2886ca0977a-Paper.pdf.

Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015. doi: 10.1038/nature14539.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryiAv2xAZ.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Ann Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *ArXiv*, abs/2106.07998, 2021.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *International Conference on Machine Learning*, 2020.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. 2020.

R. Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *NeurIPS*, 2019.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *J. Mach. Learn. Res.*, 21:181:1–181:50, 2020.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning, 2020.

Stephen Obadinma, Hongyu Guo, and Xiaodan Zhu. Class-wise calibration: a case study on covid-19 hate speech. In *The 34th Canadian Conference on Artificial Intelligence*, Vancouver, 2021. Canadian Artificial Intelligence Association.

Yaniv Ovadia, Emily Fertig, J. Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:2901–2907, 04 2015.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pp. 61–74. MIT Press, 1999.

Rahul Rahaman and Alexandre H. Thiery. Uncertainty quantification and deep ensembles, 2020.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7373–7382, 2019.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/36ad8b5f42db492827016448975cc22d-Paper.pdf.

X. Wang, Y. Chen, and W. Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, mar 2020. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3069908.

Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks, 2018.

Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. When do curricula work? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tW4QEInpni.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, pp. 609–616, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *KDD*, pp. 694–699, 2002. URL https://doi.org/10.1145/775047.775151.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Yuan Zhao, Jiasi Chen, and Samet Oymak. On the role of dataset quality and heterogeneity in model confidence, 2020.

Tianyi Zhou, Shengjie Wang, and J. Bilmes. Robust curriculum learning: from clean label detection to noisy label self-correction. In *ICLR*, 2021.

A Appendix

A.1 Additional Related Work On Calibration and Curriculum Learning

Here we provide a wider review of related work that has been conducted in the fields of calibration and curriculum learning.

Research on the calibration of machine learning classifiers typically focuses on improving metrics for calibration and creating novel calibration methods. Moreover, extensive studies have been conducted in related topics focusing on confidence scores, such as out-of-distribution detection (Lee et al., 2018) and uncertainty quantification (Rahaman & Thiery, 2020). Calibration methods in particular have received much attention. These methods are divided into post-calibration methods and training-based methods. Post-calibration methods take an existing model and modify the predictions at test time to be better calibrated, usually using validation data to tune a recalibration function. The most popular recalibration method is temperature scaling (Guo et al., 2017), which scales the logits to have higher entropy by dividing them by a temperature parameter before feeding them into the soft-max function. Other established methods include Platt scaling (Platt, 1999), isotonic regression (Zadrozny & Elkan, 2002), and histogram binning (Zadrozny & Elkan, 2001). More recently developed methods that achieve top performance include using splines for recalibration (Gupta et al., 2021).

The second class of methods involve modifying the training regime to improve regularization and punish overconfidence. Techniques that train using soft-labels like label smoothing (Müller et al., 2019) have been found to improve model calibration by focusing neural networks to output less confident predictions due to the smoothing parameter. Other techniques use data augmentation for regularization. Methods of this type include Mix-Up (Zhang et al., 2018), which convexly combines random pairs of images and their labels and helps calibration due to soft-labels (Thulasidasan et al., 2019), and manifold smoothing (Kong et al., 2020), a method which combines on-manifold and off-manifold regularization by creating pseudo-samples that are used as additional training data to improve calibration. Other methods choose to modify the loss function to explicitly bias the model towards learning to output better calibrated probabilities. MMCE is a RKHS kernel-based measure of calibration that is optimized alongside negative likelihood loss (Kumar et al., 2018) and is able to minimize calibration error without heavily punishing rightful high confidence predictions. Mukhoti et al. (2020) find that using focal loss, rather than cross entropy loss, in conjunction with temperature scaling creates models that are very well calibrated and attain state-of-the-art results. Moon et al. (2020) use a novel loss function called Correctness Ranking Loss, which regularizes class probabilities explicitly. The diverse range of these training-based calibration methods show that modifying how the model trains can have a significant effect on calibration, and that methods that provide regularization, like curriculum learning, warrant examination as we do in this study.

Curriculum learning has been used in a wide range of contexts, and prominent surveys like that of (Wang et al., 2020) exist that describe the diverse landscape of research in this domain. In terms of works pertinent to our analysis, we wish to highlight the studies on curriculum learning's effectiveness with noisy data and convergence speedup. The effectiveness of curriculum learning approaches with noisy data has been well established by research (Zhou et al., 2021). MentorNet (Jiang et al., 2018) is an algorithm for jointly optimizing deep CNNs on large-scale data using a data-driven curriculum created by a neural network, and was found to improve the generalization performance on corrupted labels. There is widespread use of curriculum learning in the weakly supervised domain as a method of regularization (Guo et al., 2018; Gong et al., 2016), where curriculum learning has been able to reduce the negative effects of the inherently noisy datasets to achieve state-of-the-art predictive performance. Regarding performance under limited training time, curriculum learning has been found to accelerate training (Bengio et al., 2009; Hacohen & Weinshall, 2019), and (Graves et al., 2017) found that a curriculum learning-based approach reaches satisfactory performing models half

the time with LSTM. Studies like these provide justification to see whether the improvements in early convergence carry on to calibration.

A.2 Progression of Model Calibration During Training

We discussed how calibration error changes over the course of training briefly in our main analysis, and in this section we provide further details and insights. As we trained the curriculum learning, anti-curriculum learning, and random order models, we measured the ECE on the test set at the end of each dynamic epoch to witness how these approaches affect calibration convergence. We averaged the measurements over multiple trials to remove any bias from an individual run. We previously mentioned that both curriculum learning and anti-curriculum learning follow a similar trend largely distinct from random ordering, as can be seen in Figures 3 and 7. Model calibration error is initially higher than standard training in all cases before it rises and peaks early before gradually decreasing over time, with the parameters *a* and *b* determining the peak and over how many dynamic epochs it takes to reach its minimum value. The peak is particularly interesting as many combinations significantly hurt model calibration early on during training creating model snapshots that are severely miscalibrated compared to standard training. Standard training remains relatively low and contained throughout training and does not rise to exceedingly high values. Despite this, all of the curriculum learning configurations gradually lower to reach minimum values comparable to standard training with the exception of $a = 1.6, b = 0.1$. Random order yields a more even distribution where it takes longer to reach the point in training where the model produces its maximum calibration error. The same rise-peak-decay pattern can be seen, however the miscalibration does not become nearly as severe and remains relatively close to standard training even in the worst combinations of *a* and *b*. The severe miscalibration early in training yields us to believe that learning from a limited subset of data continuously early on leads to the model being unable to generalize well. As the model gets exposed to more data it begins to learn to output confidence scores that take into account the entire data distribution, rendering better calibration. This explains the difference between curriculum/anti-curriculum learning and the random curriculum since random ordering is not necessarily training on the same subset, providing less bias. Overall, curriculum learning does alter how a model's calibration changes over time, and we discover that curriculum learning approaches learn to recover from initially severe levels of miscalibration over the course of training.

A.2.1 Effect of the type of pacing function on model calibration over time

One aspect of interest is the progression in calibration error the model produces on the test set over the course of training using different pacing functions. Even for the same pacing function parameters the trends between the different function families are markedly different in how they converge to their optimal level of calibration. Examining Figure 4, a trend that can be seen, most notably for the step and exponential pacing functions, is that at a certain point towards the end of training the calibration error that the model produces plummets dramatically in only a few dynamic epochs. This is in contrast with the other pacing functions that largely have model calibration steadily improving over time in a nearly logarithmic fashion. This indicates that the severe miscalibration is being self-corrected in a staggeringly short period of time over a few iterations through the whole data. The graphs show results using the pacing function parameters $a = 0.8, b = 0.2$, and this effect occurs close to the end of training when classifiers have a tendency to have low entropy for their outputted probabilities, particularly for the longer training time. This means that these pacing functions are able to suddenly shift the distribution of confidence scores despite being at the point in training where deep neural networks trained using NLL tend to be highly confident on most data. In any case, this phenomena is difficult to explain and warrants further investigation as it can offer insight into the factors that can make model calibration improve dramatically.

A.3 Full Tabular Results

We present our full tabular data in this section for each of our experiments in Sections 4, 5, and 6.

Table 2: Table comparing the accuracy and calibration performance using ECE, CECE, and KS error on the test set using a curriculum learning approach on CIFAR-10. We compare standard training to 6 combinations of pacing function parameters using a linear pacing function for 100, 25, and 5 epochs. We show the average and standard deviation from three different runs for 100 and 25 epochs, and 5 for 5 epochs.

			Standard		a=0.8, b=0.2		a=0.1, b=0.8		a=0.4, b=0.4		a=0.1, b=0.1		a=1.6, b=0.1		a=1.6, b=0.8	
			μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
	100 epochs	Best Epoch	95.7	1.5	176.0	2.6	94.7	5.0	115.7	2.5	111.3	3.5	328.7	4.2	114.7	1.5
		ECE	0.0296	0.0012	0.0347	0.0016	0.0319	0.0023	0.0325	0.0018	0.0315	0.0012	0.0867	0.0018	0.0400	0.0018
		CECE	0.0417	0.0020	0.0517	0.0027	0.0492	0.0050	0.0476	0.0044	0.0453	0.0036	0.0991	0.0031	0.0529	0.0020
		KS	0.0294	0.0014	0.0346	0.0015	0.0318	0.0022	0.0321	0.0018	0.0306	0.0014	0.0865	0.0019	0.0395	0.0016
		Accuracy	0.9470	0.0017	0.9324	0.0033	0.9456	0.0032	0.9444	0.0021	0.9466	0.0011	0.8694	0.0015	0.9341	0.0020
	25 epochs	Best Epoch	23.0	1.7	44.3	0.6	24.0	1.0	29.0	1.0	27.0	1.0	82.7	0.6	28.7	0.6
		ECE	0.0263	0.0020	0.0215	0.0018	0.0260	0.0018	0.0275	0.0013	0.0256	0.0023	0.0958	0.0032	0.0461	0.0018
Curr		CECE	0.0426	0.0023	0.0358	0.0023	0.0407	0.0017	0.0423	0.0027	0.0380	0.0039	0.1083	0.0030	0.0595	0.0020
		KS	0.0255	0.0015	0.0212	0.0013	0.0257	0.0020	0.0272	0.0016	0.0252	0.0026	0.0958	0.0033	0.0461	0.0018
		Accuracy	0.9218	0.0014	0.9088	0.0010	0.9225	0.0009	0.9224	0.0011	0.9221	0.0012	0.8400	0.0030	0.9081	0.0006
	5 epochs	Best Epoch	4	0	9	0	4	0	5.8	0.4	5	0	15.8	0.4	5	0
		ECE	0.0247	0.0036	0.0318	0.0052	0.0225	0.0040	0.0231	0.0042	0.0231	0.0052	0.1540	0.0041	0.0579	0.0143
		CECE	0.0681	0.0081	0.0603	0.0047	0.0635	0.0097	0.0713	0.0105	0.0668	0.0072	0.1645	0.0057	0.0840	0.0065
		KS	0.0231	0.0041	0.0313	0.0059	0.0208	0.0045	0.0217	0.0054	0.0224	0.0050	0.1539	0.0040	0.0577	0.0142
		Accuracy	0.7267	0.0235	0.7310	0.0341	0.7375	0.0139	0.7047	0.0443	0.7394	0.0251	0.6932	0.0179	0.7317	0.0592

Table 3: Table comparing the accuracy and calibration performance using ECE, CECE, and KS error on the test set using an anti-curriculum learning approach on CIFAR-10. Presentation is the same as in Table 2

			Standard		a=0.8, b=0.2		a=0.1, b=0.8		a=0.4, b=0.4		a=0.1, b=0.1		a=1.6, b=0.1		a=1.6, b=0.8	
			μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
	100 epochs	Best Epoch	95.7	1.5	177.3	1.5	93.7	2.1	116.3	4.5	110.7	4.5	329.7	3.8	110.7	5.1
		ECE	0.0296	0.0012	0.0320	0.0009	0.0310	0.0025	0.0323	0.0005	0.0303	0.0009	0.0842	0.0030	0.0405	0.0015
		CECE	0.0417	0.0020	0.0452	0.0026	0.0457	0.0044	0.0478	0.0017	0.0458	0.0003	0.0963	0.0032	0.0533	0.0023
		KS	0.0294	0.0014	0.0319	0.0008	0.0308	0.0023	0.0322	0.0004	0.0299	0.0004	0.0841	0.0030	0.0403	0.0015
		Accuracy	0.9470	0.0017	0.9358	0.0015	0.9461	0.0031	0.9445	0.0006	0.9471	0.0008	0.8714	0.0041	0.9323	0.0010
	25 epochs	Best Epoch	23.0	1.7	44.3	0.6	23.7	1.5	29.0	1.0	27.7	0.6	82.0	1.0	27.7	0.6
		ECE	0.0263	0.0020	0.0192	0.0016	0.0274	0.0017	0.0273	0.0021	0.0270	0.0012	0.1002	0.0023	0.0448	0.0011
Anti-Curr		CECE	0.0426	0.0023	0.0352	0.0033	0.0432	0.0025	0.0398	0.0022	0.0454	0.0031	0.1129	0.0025	0.0558	0.0004
		KS	0.0255	0.0015	0.0182	0.0028	0.0267	0.0013	0.0265	0.0019	0.0262	0.0017	0.0999	0.0024	0.0447	0.0011
		Accuracy	0.9218	0.0014	0.9075	0.0020	0.9192	0.0070	0.9237	0.0013	0.9254	0.0010	0.8351	0.0023	0.9118	0.0041
	5 epochs	Best Epoch	4	0	9	0	4.2	0.4	6	0	5	0	15.4	0.5	5	0
		ECE	0.0247	0.0036	0.0360	0.0059	0.0185	0.0072	0.0235	0.0037	0.0175	0.0033	0.1551	0.0033	0.0579	0.0055
		CECE	0.0681	0.0081	0.0632	0.0163	0.0690	0.0117	0.0695	0.0215	0.0663	0.0058	0.1662	0.0041	0.0802	0.0044
		KS	0.0231	0.0041	0.0357	0.0055	0.0169	0.0081	0.0210	0.0044	0.0153	0.0037	0.1550	0.0032	0.0577	0.0054
		Accuracy	0.7267	0.0235	0.7199	0.0447	0.7001	0.0231	0.7216	0.0411	0.7092	0.0203	0.7115	0.0209	0.7448	0.0350

Table 4: Table comparing the accuracy and calibration performance using ECE, CECE, and KS error on the test set using random ordering on CIFAR-10. Presentation is the same as in Table 2

			Standard		a=0.8, b=0.2		a=0.1, b=0.8		a=0.4, b=0.4		a=0.1, b=0.1		a=1.6, b=0.1		a=1.6, b=0.8	
			μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
	100 epochs	Best Epoch	95.7	1.5	171.0	1.7	97.7	2.5	117.7	2.5	109.0	3.5	328.0	2.6	107.7	3.2
		ECE	0.0296	0.0012	0.0314	0.0004	0.0305	0.0013	0.0313	0.0016	0.0293	0.0014	0.0440	0.0028	0.0323	0.0006
		CECE	0.0417	0.0020	0.0457	0.0011	0.0442	0.0045	0.0461	0.0023	0.0438	0.0033	0.0605	0.0021	0.0493	0.0022
		KS	0.0294	0.0014	0.0310	0.0004	0.0305	0.0013	0.0309	0.0021	0.0292	0.0014	0.0439	0.0028	0.0321	0.0007
		Accuracy	0.9470	0.0017	0.9429	0.0006	0.9467	0.0014	0.9455	0.0023	0.9479	0.0023	0.9170	0.0036	0.9425	0.0018
	25 epochs	Best Epoch	23.0	1.7	44.0	1.0	25.0	0.0	28.7	0.6	26.7	0.6	81.7	1.5	27	0
		ECE	0.0263	0.0020	0.0282	0.0013	0.0267	0.0016	0.0275	0.0016	0.0269	0.0028	0.0461	0.0035	0.0327	0.0021
Random		CECE	0.0426	0.0023	0.0444	0.0044	0.0429	0.0054	0.0429	0.0007	0.0403	0.0015	0.0610	0.0030	0.0491	0.0035
		KS	0.0255	0.0015	0.0279	0.0015	0.0264	0.0019	0.0274	0.0017	0.0270	0.0028	0.0452	0.0041	0.0319	0.0028
		Accuracy	0.9218	0.0014	0.9200	0.0034	0.9224	0.0003	0.9219	0.0047	0.9217	0.0026	0.9007	0.0044	0.9193	0.0034
	5 epochs	Best Epoch	4	0	9	0	4	0	5.8	0.4	5	0	15.8	0.4	5	0
		ECE	0.0247	0.0036	0.0217	0.0059	0.0244	0.0044	0.0259	0.0053	0.0183	0.0083	0.0267	0.0032	0.0228	0.0070
		CECE	0.0681	0.0081	0.0645	0.0111	0.0741	0.0158	0.0714	0.0130	0.0696	0.0075	0.0790	0.0124	0.0768	0.0137
		KS	0.0231	0.0041	0.0198	0.0074	0.0225	0.0038	0.0245	0.0064	0.0167	0.0086	0.0253	0.0039	0.0216	0.0074
		Accuracy	0.7267	0.0235	0.7148	0.0340	0.7048	0.0344	0.7074	0.0105	0.6993	0.0405	0.7027	0.0170	0.6901	0.0463

Table 5: Table comparing the accuracy and calibration performance using ECE, CECE, and KS error on the test set using a curriculum learning approach on CIFAR-100. Presentation is the same as in Table 2

			Standard		a=0.8, b=0.2		a=0.1, b=0.8		a=0.4, b=0.4		a=0.1, b=0.1		a=1.6, b=0.1		a=1.6, b=0.8	
			μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
	100 epochs	Best Epoch	93.7	1.2	176.0	4.6	95.7	2.9	114.3	4.9	111.7	1.5	331.3	0.6	113.3	1.2
		ECE	0.0790	0.0033	0.0935	0.0015	0.0808	0.0012	0.0822	0.0030	0.0785	0.0025	0.1222	0.0017	0.0852	0.0020
		CECE	0.1202	0.0027	0.1516	0.0068	0.1256	0.0044	0.1248	0.0024	0.1355	0.0046	0.1606	0.0082	0.1325	0.0192
		KS	0.0784	0.0041	0.0935	0.0015	0.0803	0.0015	0.0818	0.0032	0.0781	0.0024	0.1218	0.0017	0.0848	0.0014
		Accuracy	0.7678	0.0036	0.7397	0.0017	0.7652	0.0027	0.7600	0.0032	0.7642	0.0042	0.6853	0.0023	0.7541	0.0022
	25 epochs	Best Epoch	23.0	1.0	44.0	1.7	23.3	1.5	29.7	0.6	26.0	1.0	80.7	1.5	28	1
		ECE	0.0799	0.0025	0.0643	0.0054	0.0801	0.0027	0.0796	0.0043	0.0842	0.0023	0.1687	0.0031	0.1139	0.0007
Curr		CECE	0.1197	0.0066	0.1123	0.0096	0.1166	0.0123	0.1249	0.0070	0.1308	0.0156	0.2031	0.0070	0.1456	0.0044
		KS	0.0796	0.0025	0.0642	0.0057	0.0800	0.0029	0.0796	0.0043	0.0841	0.0023	0.1686	0.0031	0.1139	0.0006
		Accuracy	0.7244	0.0055	0.7108	0.0089	0.7242	0.0041	0.7271	0.0043	0.7202	0.0065	0.6345	0.0011	0.7115	0.0032
	5 epochs	Best Epoch	4	0	9	0	4	0	6	0	5	0	15.6	0.5	5	0
		ECE	0.0145	0.0023	0.0405	0.0043	0.0138	0.0020	0.0166	0.0050	0.0127	0.0027	0.1628	0.0109	0.0477	0.0019
		CECE	0.1377	0.0103	0.1305	0.0114	0.1373	0.0086	0.1340	0.0082	0.1297	0.0099	0.1854	0.0105	0.1357	0.0114
		KS	0.0088	0.0025	0.0407	0.0043	0.0083	0.0013	0.0127	0.0045	0.0054	0.0017	0.1628	0.0109	0.0478	0.0019
		Accuracy	0.4540	0.0147	0.4881	0.0097	0.4540	0.0123	0.4552	0.0261	0.4379	0.0401	0.4601	0.0139	0.4734	0.0105

Table 6: Table comparing the accuracy and calibration performance using ECE, CECE, and KS error on the test set using an anti-curriculum learning approach on CIFAR-100. Presentation is the same as in Table 2

			Standard		a=0.8, b=0.2		a=0.1, b=0.8		a=0.4, b=0.4		a=0.1, b=0.1		a=1.6, b=0.1		a=1.6, b=0.8	
			μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Anti-Curr	100 epochs	Best Epoch	93.7	1.2	176.0	1.7	98.3	2.1	118.3	2.1	110.7	2.5	329.7	2.1	110.3	5.5
		ECE	0.0790	0.0033	0.0929	0.0067	0.0808	0.0045	0.0815	0.0035	0.0805	0.0005	0.1260	0.0023	0.0893	0.0035
		CECE	0.1202	0.0027	0.1277	0.0086	0.1240	0.0114	0.1262	0.0070	0.1298	0.0108	0.1584	0.0072	0.1312	0.0076
		KS	0.0784	0.0041	0.0927	0.0066	0.0800	0.0038	0.0804	0.0036	0.0801	0.0005	0.1251	0.0016	0.0884	0.0034
		Accuracy	0.7678	0.0036	0.7409	0.0039	0.7647	0.0014	0.7615	0.0027	0.7622	0.0023	0.6808	0.0016	0.7511	0.0021
	25 epochs	Best Epoch	23.0	1.0	44.3	0.6	22.3	0.6	29.0	1.0	26.7	1.5	81.7	1.5	26	1
		ECE	0.0799	0.0025	0.0656	0.0060	0.0809	0.0041	0.0785	0.0025	0.0840	0.0004	0.1713	0.0069	0.1112	0.0028
		CECE	0.1197	0.0066	0.1184	0.0030	0.1172	0.0140	0.1171	0.0051	0.1265	0.0038	0.2083	0.0133	0.1411	0.0074
		KS	0.0796	0.0025	0.0654	0.0063	0.0809	0.0041	0.0785	0.0025	0.0838	0.0008	0.1712	0.0070	0.1111	0.0030
		Accuracy	0.7244	0.0055	0.7078	0.0052	0.7190	0.0015	0.7255	0.0012	0.7201	0.0037	0.6274	0.0098	0.7126	0.0045
	5 epochs	Best Epoch	4	0	9	0	4	0	5.8	0.4	5	0	15.4	0.5	5	0
		ECE	0.0145	0.0023	0.0433	0.0064	0.0158	0.0020	0.0130	0.0020	0.0157	0.0057	0.1730	0.0054	0.0453	0.0029
		CECE	0.1377	0.0103	0.1398	0.0056	0.1389	0.0022	0.1337	0.0078	0.1359	0.0096	0.1882	0.0067	0.1462	0.0112
		KS	0.0088	0.0025	0.0433	0.0065	0.0080	0.0017	0.0077	0.0033	0.0084	0.0044	0.1729	0.0055	0.0454	0.0030
		Accuracy	0.4540	0.0147	0.4920	0.0183	0.4441	0.0066	0.4565	0.0185	0.4566	0.0165	0.4607	0.0083	0.4711	0.0093

Table 7: Table comparing the accuracy and calibration performance using ECE, CECE, and KS error on the test set using random order on CIFAR-10. Presentation is the same as in Table 2

			Standard		a=0.8, b=0.2		a=0.1, b=0.8		a=0.4, b=0.4		a=0.1, b=0.1		a=1.6, b=0.1		a=1.6, b=0.8	
			μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Random	100 epochs	Best Epoch	93.7	1.2	177.3	2.1	96.7	2.5	116.3	3.2	106.3	3.2	331.7	0.6	110.3	1.5
		ECE	0.0790	0.0033	0.1016	0.0042	0.0796	0.0024	0.0787	0.0032	0.0778	0.0010	0.1210	0.0037	0.0800	0.0026
		CECE	0.1202	0.0027	0.1561	0.0048	0.1345	0.0058	0.1167	0.0179	0.1213	0.0083	0.1718	0.0074	0.1326	0.0099
		KS	0.0784	0.0041	0.1013	0.0042	0.0783	0.0035	0.0778	0.0037	0.0773	0.0007	0.1204	0.0041	0.0791	0.0019
		Accuracy	0.7678	0.0036	0.7299	0.0066	0.7694	0.0013	0.7644	0.0058	0.7680	0.0016	0.6563	0.0078	0.7566	0.0013
	25 epochs	Best Epoch	23.0	1.0	44.7	0.6	24.3	0.6	28.7	0.6	26.7	1.2	80.7	0.6	28.3	0.6
		ECE	0.0799	0.0025	0.0877	0.0010	0.0780	0.0022	0.0809	0.0016	0.0796	0.0025	0.1577	0.0011	0.0926	0.0101
		CECE	0.1197	0.0066	0.1274	0.0131	0.1130	0.0101	0.1191	0.0052	0.1138	0.0192	0.1961	0.0153	0.1364	0.0150
		KS	0.0796	0.0025	0.0876	0.0010	0.0779	0.0023	0.0809	0.0016	0.0793	0.0028	0.1577	0.0011	0.0923	0.0103
		Accuracy	0.7244	0.0055	0.7177	0.0046	0.7268	0.0037	0.7239	0.0046	0.7250	0.0041	0.6195	0.0036	0.7203	0.0059
	5 epochs	Best Epoch	4	0	9	0	4.2	0.4	6	0	5	0	15.8	0.4	5	0
		ECE	0.0145	0.0023	0.0120	0.0014	0.0146	0.0026	0.0135	0.0055	0.0136	0.0021	0.0175	0.0051	0.0126	0.0028
		CECE	0.1377	0.0103	0.1409	0.0101	0.1359	0.0100	0.1415	0.0129	0.1351	0.0102	0.1295	0.0032	0.1416	0.0074
		KS	0.0088	0.0025	0.0065	0.0031	0.0068	0.0030	0.0082	0.0037	0.0056	0.0015	0.0156	0.0064	0.0062	0.0012
		Accuracy	0.4540	0.0147	0.4485	0.0211	0.4470	0.0070	0.4404	0.0142	0.4453	0.0160	0.4437	0.0150	0.4406	0.0104

Table 8: Table comparing the accuracy and calibration performance using ECE, CECE, and KS error on the test set using 6 different pacing functions on CIFAR-10. Here we only use a curriculum learning approach at the same combination of a and b. Again we compare three different training times of 100, 25, and 5 epochs. For 100 and 25 epochs we use $a = 0.8, b = 0.2$. For 5 epochs we use $a = 0.1, b = 0.8$ since its performance is not as bad as the aforementioned combination. We show the average and standard deviation from three different runs for 100 and 25 epochs, and 5 for 5 epochs.

		linear		quad		root		step		exp		log	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
100 epochs	Best Epoch	176.0	2.6	237.3	1.5	136.3	4.7	415.7	0.6	350.3	1.5	104.7	1.5
	ECE	0.0347	0.0016	0.0350	0.0020	0.0320	0.0001	0.0336	0.0014	0.0364	0.0028	0.0285	0.0006
	CECE	0.0517	0.0027	0.0497	0.0033	0.0468	0.0022	0.0504	0.0022	0.0526	0.0034	0.0439	0.0012
	KS	0.0346	0.0015	0.0345	0.0023	0.0319	0.0002	0.0331	0.0013	0.0364	0.0027	0.0283	0.0006
	Accuracy	0.9324	0.0033	0.9305	0.0026	0.9403	0.0021	0.9216	0.0014	0.9221	0.0024	0.9456	0.0002
25 epochs	Best Epoch	44.3	0.6	59.3	0.6	32.7	0.6	104.0	0.0	88.3	1.2	26.7	0.6
	ECE	0.0215	0.0018	0.0186	0.0005	0.0191	0.0017	0.0178	0.0026	0.0195	0.0015	0.0252	0.0029
	CECE	0.0358	0.0023	0.0376	0.0032	0.0354	0.0035	0.0411	0.0059	0.0407	0.0012	0.0360	0.0024
	KS	0.0212	0.0013	0.0182	0.0003	0.0184	0.0011	0.0152	0.0042	0.0179	0.0029	0.0251	0.0030
	Accuracy	0.9088	0.0010	0.9007	0.0002	0.9170	0.0015	0.8626	0.0018	0.8790	0.0064	0.9201	0.0040
5 epochs	Best Epoch	9	0	12	0	7	0	20	0	18	0	6	0
	ECE	0.0318	0.0052	0.0328	0.0038	0.0321	0.0034	0.0482	0.0048	0.0635	0.0052	0.0289	0.0038
	CECE	0.0603	0.0047	0.0578	0.0084	0.0611	0.0057	0.0696	0.0029	0.0776	0.0039	0.0604	0.0029
	KS	0.0313	0.0059	0.0326	0.0039	0.0315	0.0037	0.0482	0.0048	0.0630	0.0050	0.0281	0.0041
	Accuracy	0.7310	0.0341	0.7173	0.0664	0.7472	0.0198	0.6857	0.0083	0.7063	0.0088	0.7430	0.0201

Table 9: Table comparing the accuracy and calibration performance using ECE, CECE, and KS error on the test set using 6 different pacing functions on CIFAR-100. Experimental setup is the same as in table 8 except that $a = 0.1, b = 0.1$ is used for 5 epochs instead.

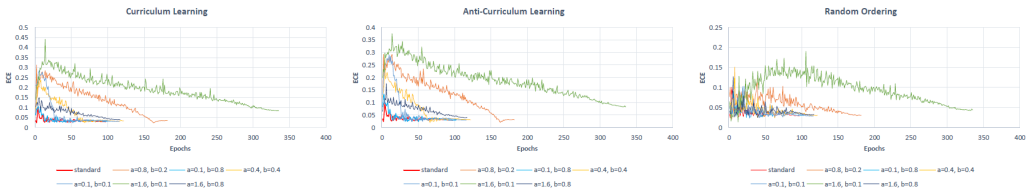
		linear		quad		root		step		exp		log	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
100 epochs	Best Epoch	176.0	4.6	239.0	1.0	135.7	5.9	413.7	2.1	351.0	1.0	103.3	3.5
	ECE	0.0935	0.0015	0.0964	0.0044	0.0910	0.0031	0.0787	0.0013	0.0927	0.0030	0.0833	0.0004
	CECE	0.1516	0.0068	0.1409	0.0047	0.1317	0.0122	0.1292	0.0116	0.1387	0.0136	0.1169	0.0136
	KS	0.0935	0.0015	0.0962	0.0045	0.0908	0.0032	0.0786	0.0011	0.0921	0.0024	0.0818	0.0006
	Accuracy	0.7397	0.0017	0.7306	0.0054	0.7467	0.0023	0.7242	0.0012	0.7232	0.0034	0.7611	0.0012
25 epochs	Best Epoch	44.0	1.7	59.0	1.0	34.3	0.6	103.3	0.6	88.3	0.6	26.0	1.0
	ECE	0.0643	0.0054	0.0593	0.0022	0.0704	0.0011	0.0279	0.0040	0.0509	0.0040	0.0761	0.0009
	CECE	0.1123	0.0096	0.1267	0.0102	0.1153	0.0113	0.1170	0.0129	0.1239	0.0084	0.1270	0.0077
	KS	0.0642	0.0057	0.0591	0.0024	0.0703	0.0010	0.0262	0.0045	0.0507	0.0042	0.0760	0.0010
	Accuracy	0.7108	0.0089	0.6974	0.0051	0.7171	0.0002	0.6370	0.0096	0.6671	0.0049	0.7255	0.0042
5 epochs	Best Epoch	5.0	0.0	6.0	0.0	4.3	0.6	8.0	0.0	7.0	0.0	4.0	0.0
	ECE	0.0144	0.0015	0.0135	0.0042	0.0142	0.0010	0.0100	0.0008	0.0130	0.0007	0.0158	0.0020
	CECE	0.1328	0.0119	0.1277	0.0106	0.1369	0.0095	0.1438	0.0117	0.1341	0.0117	0.1403	0.0028
	KS	0.0060	0.0021	0.0095	0.0070	0.0080	0.0007	0.0035	0.0012	0.0080	0.0008	0.0070	0.0014
	Accuracy	0.4287	0.0503	0.4537	0.0074	0.4479	0.0207	0.4307	0.0161	0.4461	0.0307	0.4469	0.0188

Table 10: Table comparing the accuracy and calibration performance using ECE, CECE, and KS error on the test set under noisy labels on CIFAR-10. We compare standard training to curriculum and anti-curriculum learning under 40% and 80% label noise. Again we compare three different training times of 100, 25, and 5 epochs. For 100 and 25 epochs we keep the pacing function parameters constant at $a = 0.8, b = 0.2$. For 5 epochs we use $a = 0.1, b = 0.8$. We show the average and standard deviation from three different runs for 100 and 25 epochs, and 5 for 5 epochs.

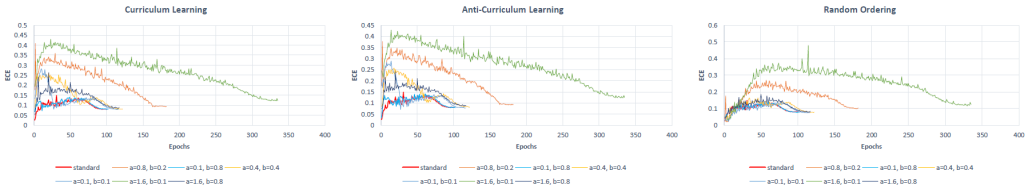
		Standard				Curr				Anti-Curr			
		0.4 Noise		0.8 Noise		0.4 Noise		0.8 Noise		0.4 Noise		0.8 Noise	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
100 epochs	Best Epoch	97.3	0.6	89.00	3.46410	177.3	3.1	180.3	1.2	176.3	0.6	177.7	4.0
	ECE	0.0291	0.0023	0.0316	0.0016	0.0337	0.0011	0.0356	0.0014	0.0326	0.0024	0.0317	0.0009
	CECE	0.0444	0.0048	0.0467	0.0011	0.0493	0.0024	0.0529	0.0031	0.0467	0.0047	0.0475	0.0011
	KS	0.0289	0.0025	0.0314	0.0015	0.0335	0.0011	0.0355	0.0012	0.0324	0.0023	0.0316	0.0009
	Accuracy	0.9483	0.0016	0.9449	0.0024	0.9357	0.0009	0.9335	0.0023	0.9367	0.0031	0.9364	0.0020
25 epochs	Best Epoch	23.0	1.0	23.7	0.6	45.0	0.0	44.3	0.6	43.0	1.0	44.3	0.6
	ECE	0.0301	0.0015	0.0260	0.0013	0.0198	0.0019	0.0181	0.0017	0.0178	0.0007	0.0184	0.0005
	CECE	0.0458	0.0037	0.0438	0.0036	0.0349	0.0012	0.0366	0.0038	0.0336	0.0016	0.0340	0.0030
	KS	0.0301	0.0015	0.0253	0.0012	0.0173	0.0010	0.0169	0.0026	0.0173	0.0005	0.0182	0.0007
	Accuracy	0.9216	0.0026	0.9217	0.0049	0.9122	0.0023	0.9099	0.0023	0.9099	0.0057	0.9114	0.0026
5 epochs	Best Epoch	4	0	4	0	4	0	4	0	4	0	4	0
	ECE	0.0255	0.0054	0.0208	0.0036	0.0205	0.0039	0.0180	0.0047	0.0243	0.0052	0.0235	0.0086
	CECE	0.0742	0.0221	0.0630	0.0134	0.0745	0.0121	0.0805	0.0117	0.0695	0.0152	0.0714	0.0098
	KS	0.0250	0.0050	0.0191	0.0053	0.0195	0.0050	0.0159	0.0055	0.0235	0.0057	0.0223	0.0094
	Accuracy	0.7259	0.0344	0.7276	0.0385	0.7109	0.0202	0.6900	0.0269	0.7401	0.0209	0.7156	0.0261

Table 11: Table comparing the accuracy and calibration performance using ECE, CECE, and KS error on the test set under noisy labels on CIFAR-10. We compare standard training to curriculum and anti-curriculum learning under 40% and 80% label noise. Experimental setup is the same as in table 10 except that $a = 0.8, b = 0.2$ is used for 5 epochs instead.

		Standard				Curr				Anti-Curr			
		0.4 Noise		0.8 Noise		0.4 Noise		0.8 Noise		0.4 Noise		0.8 Noise	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
100 epochs	Best Epoch	48.7	4.7	62.67	4.61880	109.7	5.1	29.3	7.4	118.0	3.6	35.0	3.0
	ECE	0.0767	0.0343	0.1102	0.0252	0.1658	0.0158	0.1354	0.0174	0.1323	0.0163	0.1652	0.0178
	CECE	0.2015	0.0273	0.2098	0.0804	0.2231	0.0315	0.1962	0.0204	0.2100	0.0317	0.2048	0.0067
	KS	0.0745	0.0362	0.1093	0.0266	0.1658	0.0158	0.1354	0.0174	0.1323	0.0163	0.1652	0.0178
	Accuracy	0.5235	0.0051	0.2353	0.0142	0.5603	0.0064	0.2325	0.0056	0.5636	0.0118	0.2263	0.0039
25 epochs	Best Epoch	23.0	1.0	23.3	0.6	36.3	2.5	33.7	2.1	37.7	1.5	29.3	3.8
	ECE	0.1364	0.0042	0.1225	0.0043	0.0482	0.0254	0.0497	0.0044	0.0570	0.0225	0.0519	0.0197
	CECE	0.1941	0.0025	0.1857	0.0307	0.1532	0.0161	0.1415	0.0207	0.1479	0.0073	0.1552	0.0157
	KS	0.1365	0.0042	0.1225	0.0043	0.0443	0.0306	0.0495	0.0046	0.0564	0.0242	0.0518	0.0196
	Accuracy	0.5709	0.0025	0.1675	0.0053	0.5836	0.0087	0.2241	0.0028	0.5816	0.0022	0.2309	0.0065
5 epochs	Best Epoch	4	0	4	0	5	0	4.4	0.547723	5	0	5	0
	ECE	0.1063	0.0033	0.0454	0.0098	0.1176	0.0082	0.0872	0.0058	0.1191	0.0097	0.0838	0.0123
	CECE	0.1783	0.0183	0.0255	0.0070	0.1907	0.0158	0.1023	0.0328	0.1845	0.0208	0.1414	0.0427
	KS	0.1062	0.0032	0.0449	0.0100	0.1174	0.0084	0.0871	0.0058	0.1191	0.0097	0.0834	0.0124
	Accuracy	0.2361	0.0093	0.0665	0.0113	0.2730	0.0121	0.1152	0.0077	0.2710	0.0257	0.1122	0.0139



(a)



(b)

Figure 7: Graphs showing the progression of ECE on the test set calculated at every dynamic epoch during training for CIFAR-10 (top) and CIFAR-100 (bottom) at 100 epochs to measure the calibration of the model over the course of training. We compare 3 different orderings: curriculum (left), anti-curriculum (centre) and random (right) at the same 6 combinations of a and b discussed in section 4. We also include the trend of standard training in bright red.