Explicit Regularisation, Sharpness and Calibration

Israel F. Mason-Williams
ifm24@cam.ac.uk

Igor Sterner
is473@cam.ac.uk

Fredrik Ekholm
fwe21@cam.ac.uk

Abstract

We aim to investigate how using explicit regularisation impacts the loss landscape and calibration of neural networks. We gather a range of sharpness measures from the literature and investigate how they correlate with generalisation and calibration. Our experiments using VGG-19 on the CIFAR datasets show that the sharpness measures on the training loss landscape do not consistently correlate with generalisation, for different regularisers. This finding contradicts other claims in the literature. For these regularisers, we instead find consistent correlation between better calibration and a sharper landscape. We believe this introduces a new perspective to the study of loss landscapes, calibration and explicit regularisation, wherein sharpness is a desired quality for well-calibrated models.

1 Introduction

Explicit regularisation is a machine-learning-practitioner's tool to improve generalisation [1; 2]. When to use which regulariser can, however, in general, only be answered with empirical experimentation [3; 4; 5]. Meanwhile, loss landscape analysis has been used to better understand the mechanisms that govern the learning process of neural networks around and up to the convergence point [6; 7]. A conflict exists in literature between how different features of the loss landscape are predictors of generalisation [6], with some work positing the virtues of flatness [8; 9] while other work identifies that flat landscapes can be arbitrarily sharpened due to their complex geometry [10]. Implicit regularisers such as that from stochastic gradient descent (SGD) and skip connections [11] have been studied from the loss landscape perspective. Meanwhile similar studies, to our knowledge, are lacking for explicit regularisers.

We seek to gain insight into how explicit regularisers improve generalisation by exploring how they impact the loss landscape. In particular, we explore dropout, weight decay and data augmentation, combined with early stopping as often used in practice. In tandem, we explore the impact of explicit regularisers on the calibration of the neural networks [12]. Since neural networks are generally over-confident [12], our line of enquiry is to discover how and if these explicit regularisers impact the calibration of the resulting model and if geometric properties align.

We seek to answer the following questions:

1. Which measures of sharpness from the literature have claimed correlation with generalisation?
2. If using an explicit regulariser improves generalisation, does it also lead to finding a flatter minima? Do trends hold for tasks of different complexity?
3. Are there geometric properties associated with well-calibrated models and how do explicit regularisers affect this?

The contributions of this work are as follows.

- We investigate wide range of geometric measures on the loss landscape from the literature. We release our code which includes a unified framework for researchers to experiment with any or all of them on any PyTorch model[1].
- Our results suggest that none of these measures are predictors of generalisation for different regularisers, prompting further research into suitable generalisation measures.
- Instead, we find a significant correlation between more calibrated models and sharper minima, again for different regularisers.

2 Background

2.1 Regularisation

Explicit Regularisation Explicit regularisation techniques are architecture-independent and designed to constrain the effective capacity of a neural network [13]. Dropout [14] and weight decay [15] are commonly used explicit regularisation techniques. Data augmentation has been referenced in the literature as both an explicit [16] and implicit [13] regulariser. We treat it as an explicit regulariser. A combination of both implicit and explicit regularisers in neural networks has proven effective in improving model generalisation. Empirical studies suggest the reason for this is that both implicit and explicit regularisers impact the simplicity bias [17] of the network [18].

Implicit Regularisation Implicit regularisation defines how architecture adaptations and optimization constraints impact a model's effective capacity [16]. There have been numerous works studying implicit regularisation in neural network training, often with respect to SGD [19] due to its ability to avoid overfitting and generalising well. An example of an implicit regulariser is slo skip connections for ResNets [20]. It is recognised that implicit regularisation such as skip connections have an impact on the otherwise non-convex landscapes causing them to become more convex, which offers a potential geometric explanation for its effectiveness [11].

2.2 Loss Landscapes

The disparity in the literature regarding the notion of flatness [7] and sharpness [10] in loss landscapes and their relative merits leads to uncertainty when reasoning their impacts. Huang et al. [21] explored a line of enquiry regarding the decision boundaries of flat and sharp minima. They observe that flatter minima have wider decision boundaries, and, therefore, are more resilient to weight perturbation. As a result, they posit that the complexity of the decision boundary of the data, the distribution itself, rather than flatness is more important when considering generalisation. Kaddour et al. [8] extend this line of enquiry and show that when considering the flatness of a loss landscape, datasets matter, architectures matter, and flat-minima optimizers offer asymmetric payoffs.

2.3 Sharpness of the loss landscape

Many studies posit a correlation between generalisation and measures of sharpness of the loss landscape [e.g. 22; 23]. Mini-batch SGD's generalisation qualities have been aligned with the notion that its implicit regularisation favours convergence towards flatter minima [24; 19]. Motivated by this, various measures of sharpness have been proposed. They have also been explored to explicitly include sharpness measures in the optimization process, guiding convergence towards minima that generalise better [9; 25; 26; 27] .

2.4 Reparametrisation Invariance

For some sharpness measures, it is possible to alter the weights of an already trained network such that it models the exact same function while the sharpness measure changes. In such cases, the measure is not *reparametrisation invariant*. Reparametrisation invariance can be said to be a desirable feature of a sharpness measure, as the generalising ability of a network does not change with reparameterization. In fact, some argue that a measure that is not reparametrisation invariant can not be a suitable sharpness measure, as networks can be constructed wherein the correspondence with generalisation is

[1] Our code is available here: `https://github.com/IFMW01/R252_Group_Project`

contradicted. Alternative views are that such reparametrisation is a pathological case of networks that don't arise in practice, and thus non-invariant measures may also be useful. In fact, in some cases non-invariant measures have shown comparable correlation to generalisation as invariant measures [26].

A weaker form of reparametrisation invariance is scale invariance: a measure which is not dependent on linear re-scalings of the network weights. This offers a middle ground between reparametrisation invariance and strict reparametrisation invariance.

2.5 Measures

We will now detail five different sharpness measures from the literature: l^2-norm, *Fisher-Rao* norm, *SAM-Sharpness*, *Relative Flatness* and *IGS*. l^2-norm nad Fisher-Rao norm might be considered capacity measures rather than sharpness measures, but have connections to loss landscape curvature [28]. We will refer to all five as *sharpness measures* for simplicity.

In the following, θ is the vector of model parameters, $L(\theta)$ is the average loss over the dataset and $L_i(\theta)$ is the loss for the i-th sample.

l^2-Norm The l^2-Norm of the weights is a common explicit regulariser, but has also been touted a predictor of generalisation under the theory that a lower norm network should model a smoother function.

SAM-Sharpness We define *SAM-sharpness* as the difference $L^{SAM}(\theta) - L(\theta)$, where $L^{SAM}(\theta) = \max_{||\epsilon||_2 < \rho} L(\theta + \epsilon)$ as the loss function introduced by Foret at al. [9]. It is not a reparametrisation invariant measure.

Relative Flatness Petzka et al. [26] introduce *Relative Flatness*. They consider the decomposition of general neural networks into a feature extractor and a single layer classification model, and calculate sharpness only for this final classification layer. Relative Flatness is calculated using the trace of the hessian for each pair of neurons in the last layer of a model, scaled by the inner product of the weights of the pair of neurons. This is a scale-invariant sharpness measure.

IGS *Information Geometric Sharpness (IGS)* [27] was introduced to achieve full reparametrisation invariance. In IGS, the average magnitude of gradients are calculated using the pseudo-norm induced by the Fisher Information Matrix (FIM). The FIM is defined as

$$F(\theta) = \mathbb{E}_{p(x,y;\theta)}[\nabla_\theta L(f_\theta(x), y) \nabla_\theta L(f_\theta(x), y)^\top]$$

Where $L(f_\theta(x), y)$ is the loss for input x and label y. We consider the model-FIM, where x is distributed according to the dataset, and y is distributed according to the categorical probabilities given by the model. Information Geometric Sharpness is then defined as:

$$IGS(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial L_i(\theta)}{\partial \theta} \right) F(\theta)^\dagger \left(\frac{\partial L_i(\theta)}{\partial \theta} \right)^\top$$

Here, the FIM serves as a metric to measure the distance between parametric probability density functions. The FIM is closely related to the KL-divergence, in that

$$\mathrm{KL}\left(p(x, y; \theta + d\theta) || p(x, y; \theta)\right) \approx \frac{1}{2} d\theta^\top F(\theta) d\theta$$

for infinitesimal $d\theta$. Therefore, IGS can be viewed as calculating the gradient norm in function space instead of in Euclidean parameter space, and thus yielding a reparametrisation invariant metric. As the FIM is expensive to calculate, the original paper provides an approximation method based on power iteration to find the largest eigenvalues of the FIM (see appendix of [27]).

Fisher-Rao Similar to IGS, the FIM can be used to calculate a weight norm where each parameter of the network weights is scaled by its impact on the probability density function described by the network: $|\theta|_{fr} = \theta F(\theta)\theta^\top$. This gives a scale invariant measure called Fisher-Rao norm [28]. The original work by Liang et al. [28] also provides a direct formula for classification networks (Equation B.3 in their Appendix).

2.6 Model Calibration

Neural Networks have empirically been found to be overconfident [12]. In other words, they are often poorly calibrated and predictions are correct less frequently than high confidences would suggest. Calibration is particularly relevant when considering the application of neural networks in the real world. Naeini et al. [29] propose a metric for how calibrated a classifier is: expected calibration error (ECE). This metric consists of taking a weighted average of the difference between accuracy (Equation 1) and confidence (Equation 2) across equally spaced bins B, as per Equation 3.

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}(\hat{y}_i = y_i) \tag{1}$$

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{q}_i \tag{2}$$

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \big|\text{acc}(B_m) - \text{conf}(B_m)\big| \tag{3}$$

One approach shown to reduce ECE of a neural network, in turn improving calibration, is to temperature-scale the output class logits before softmax [12]. Temperature-scaling involves multipying the logits by the inverse of a learnt scalar T. Equation 4 shows this form of calibration, where $\mathbf{z_i}$ is the logit vector before scaling and $\mathbf{q_i}$ is it after. A model without temperature scaling has fixed $T = 1$.

$$\mathbf{q_i} = \frac{\exp\left(\mathbf{z_i}/T\right)}{\sum_j \exp\left(\mathbf{z_i}/T\right)} \tag{4}$$

3 Experimental Setup

Datasets We use the CIFAR-10 and CIFAR-100 for image classification datasets [30]. In both, there are 50,000 training and 10,000 evaluation (image, class) pairs. For evaluation we use the first 5,000 for validation and the rest for test. We do not normalize the images using the (μ, σ) of Imagenet, as is standard practice to obtain SoTA accuracies on CIFAR. We hypothesize that the difference between results with different explicit regularisation will be more pronounced without the impact of normalisation.

Systems and training We train the VGG-19 architecture [31] which contains approximately 144 million parameters for image classification. While our setup is modular in the architecture used, we limit our experiments to this architecture due to computational budget.

We optimize cross-entropy loss on the train set, specifically using SGD with momentum=0.9, learning rate=0.05 and batch size=256. We train for 50 and 75 epochs for CIFAR-10 and CIFAR-100 respectively, with early stopping on the validation set. Our `baseline` model is free from all explicit regularisation. We train models with augmentation, dropout and weight decay separately added. We set dropout to 0.5 for our `dropout` models, weight decay to 5e-4 for our `weight decay` models and augment the input images with random horizontal flips + random crops for our `augmentation` models. Our goal is not to replicate SoTA accuracies, but to report the isolated and controlled effect of each of these explicit regularizers. Each of these regularisers are used in practice (despite use of `weight decay` waning).

Table 1: Results for VGG-19 on CIFAR-10. ↓ indicates lower loss/error, more calibrated, or flatter. Note that train loss here is not the final loss achieved in training, but rather the non-regularised loss function evaluated on the train set

Regularizer	Train loss ($\times 10^{-2}$) ↓	Test error (%) ↓	ECE(%) ↓	Weight norm ↓	Fisher-Rao norm ↓	Relative Flatness↓	SAM-sharpness ↓ ($\times 10^{-2}$)	log IGS ↓
`baseline`	$0.9_{\pm 0.5}$	$14.4_{\pm 0.1}$	$11.0_{\pm 0.2}$	$49.3_{\pm 4.1}$	$0.8_{\pm 0.3}$	$2.2_{\pm 1.0}$	$1.2_{\pm 0.7}$	$-1.9_{\pm 0.5}$
`+ temp`	$1.8_{\pm 0.8}$		$7.5_{\pm 0.6}$		$0.7_{\pm 0.2}$	$3.8_{\pm 1.7}$	$1.3_{\pm 0.6}$	$-0.6_{\pm 0.3}$
`augmentation`	$7.0_{\pm 2.4}$	$11.1_{\pm 0.7}$	$6.6_{\pm 0.6}$	$51.5_{\pm 1.6}$	$2.3_{\pm 0.6}$	$12.1_{\pm 1.5}$	$6.6_{\pm 4.3}$	$0.9_{\pm 0.3}$
`+ temp`	$8.6_{\pm 1.8}$		$2.7_{\pm 0.5}$		$1.8_{\pm 0.3}$	$13.2_{\pm 0.6}$	$4.6_{\pm 2.3}$	$0.8_{\pm 0.2}$
`dropout`	$0.8_{\pm 0.2}$	$13.9_{\pm 0.2}$	$11.0_{\pm 0.2}$	$39.1_{\pm 0.3}$	$0.8_{\pm 0.1}$	$1.2_{\pm 0.2}$	$2.8_{\pm 2.8}$	$-2.5_{\pm 0.4}$
`+ temp`	$1.8_{\pm 0.3}$		$7.6_{\pm 0.2}$		$0.7_{\pm 0.1}$	$2.9_{\pm 0.4}$	$1.7_{\pm 1.6}$	$-0.9_{\pm 0.2}$
`weight decay`	$18.0_{\pm 3.1}$	$18.2_{\pm 1.3}$	$10.8_{\pm 1.3}$	$33.2_{\pm 0.9}$	$3.9_{\pm 0.5}$	$4.5_{\pm 0.2}$	$9.2_{\pm 0.4}$	$1.5_{\pm 0.1}$
`+ temp`	$21.4_{\pm 2.0}$		$2.9_{\pm 1.3}$		$2.9_{\pm 0.1}$	$6.2_{\pm 0.3}$	$6.5_{\pm 2.0}$	$1.3_{\pm 0.0}$

For each of these models, we further follow Guo et al. [12] by tuning a temperature parameter to scale the logits. This is our intervention to explicitly calibrate the models. We use LBFGS optimization with respect to the negative log likelihood loss (on the temperature–softmax scaled logits), with learning rate=0.01, for 50 iterations. We call the temperature-scaled models for each regularizer `+temp`.

Metrics, measures and visualisations Our metric for generalisation is accuracy on the test set. For calibration, we use ECE as discussed in Section 2.6. For sharpness we report the metrics described in Section 2.5: l^2-norm, Fisher-Rao norm, Relative Flatness, SAM-sharpness and IGS. All sharpness measures are evaluated on the train set. The Fisher-Rao norm is calculated using the direct formula for classification models (see Section 2.5).

Relative Flatness is calculated by explicitly by computing the hessian for each pair of neurons in the last layer. Due to the computational intensity of this operation, we compute the Relative Flatness over a random subset of 7,500 training datapoints.

SAM-sharpness is approximated by randomly sampling 20 weight vectors of a distance $\rho = 0.05$ from the original model, and calculating the loss over the whole dataset.

Finally, for IGS we adapt the reference implementation of the power iteration approximation algorithm (see Section 2.5). The code was modified to allow for the IGS to be calculated on any arbitrary model and dataset in multiple batches. Due to our relatively large model size and dataset, we do the following to reduce the computational load. First we only estimate the 20 largest eigenvectors of the FIM, as opposed to 100 in the original implementation. We calculate the IGS as the average over batches of size 64. This means we are effectively calculating an m-sharpness [27] measure with $m = 64$. We limit the calculation to 20 batches. And finally we only compute IGS on CIFAR-10, excluding CIFAR-100 for this measure.

To visualise loss landscapes we adapted the implementation of Li et al. [11], which uses scale invariant filter-wise normalized directions to plot the loss region.

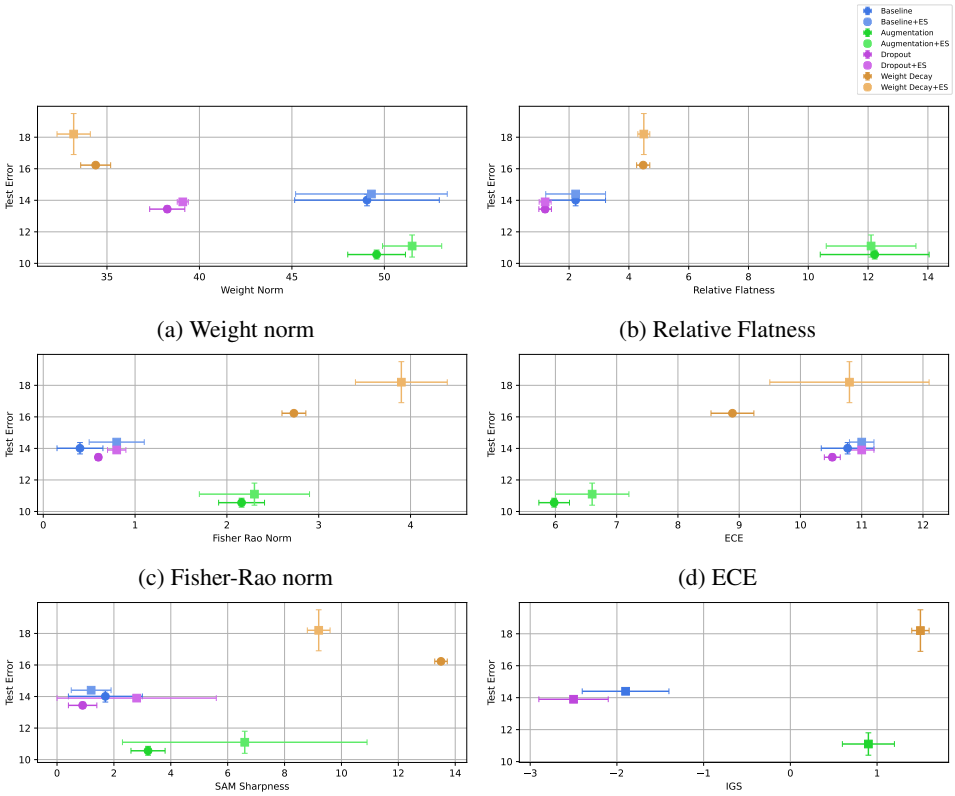
For all models we report the mean of three runs with different random model seed, plus or minus one standard deviation. For statistical measurements of correlation, we compute Spearman correlation coefficients, and use significance threshold $\alpha = 0.05$.

4 Results and discussion

Numerical results for CIFAR-10 are given in Table 1 and for CIFAR-100 in Table 2. Subsequent analysis and plots will focus on the results from CIFAR-10, but we present corresponding plots for CIFAR-100 in the Appendix. Results for the final model without early stopping are also in the Appendix.

Table 2: Results for VGG-19 on CIFAR-100. ↓ indicates lower loss/error, more calibrated, or flatter

Regularizer	Train loss ($\times 10^{-2}$) ↓	Test Error ↓ (%)	ECE ↓ (%)	Weight norm ↓	Fisher-Rao norm ↓	Relative Flatness↓	SAM-sharpness↓ ($\times 10^{-2}$)
`baseline`	$0.3_{\pm 0.3}$	$49.9_{\pm 1.1}$	$38.4_{\pm 0.7}$	$768.3_{\pm 17.7}$	$0.6_{\pm 0.2}$	$4.4_{\pm 3.4}$	$0.4_{\pm 0.4}$
`+ temp`	$2.8_{\pm 1.0}$		$27.1_{\pm 0.4}$		$1.0_{\pm 0.2}$	$19.7_{\pm 4.8}$	$0.9_{\pm 0.8}$
`augmentation`	$36.0_{\pm 5.5}$	$39.5_{\pm 0.7}$	$20.0_{\pm 0.8}$	$693.6_{\pm 11.7}$	$5.4_{\pm 0.3}$	$78.4_{\pm 8.9}$	$6.5_{\pm 2.2}$
`+ temp`	$45.7_{\pm 6.7}$		$6.1_{\pm 1.5}$		$4.5_{\pm 0.4}$	$71.7_{\pm 9.1}$	$8.0_{\pm 2.3}$
`dropout`	$1.1_{\pm 0.7}$	$47.4_{\pm 0.6}$	$36.2_{\pm 0.6}$	$700.2_{\pm 9.8}$	$1.1_{\pm 0.1}$	$6.9_{\pm 0.7}$	$1.2_{\pm 0.3}$
`+ temp`	$3.6_{\pm 0.5}$		$24.9_{\pm 0.5}$		$1.2_{\pm 0.1}$	$17.0_{\pm 0.5}$	$1.4_{\pm 0.6}$
`weight decay`	$57.5_{\pm 3.6}$	$49.8_{\pm 0.3}$	$25.0_{\pm 0.2}$	$170.3_{\pm 2.8}$	$6.7_{\pm 0.3}$	$28.2_{\pm 0.7}$	$25.0_{\pm 6.8}$
`+ temp`	$71.0_{\pm 2.6}$		$5.9_{\pm 0.3}$		$6.5_{\pm 0.0}$	$32.3_{\pm 0.6}$	$16.7_{\pm 1.5}$



(a) Weight norm

(b) Relative Flatness

(c) Fisher-Rao norm

(d) ECE

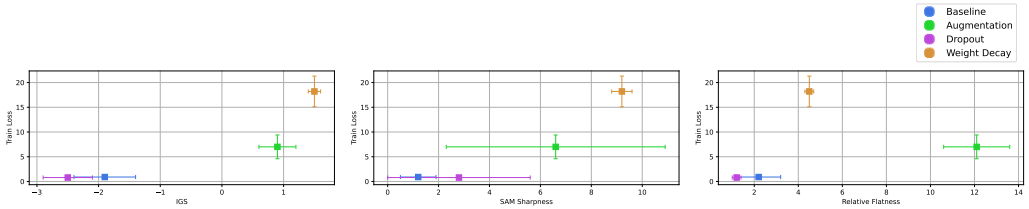
(e) SAM sharpness

(f) IGS

Figure 1: Generalisation correlation on CIFAR-10 for sharpness metrics and ECE. Results here include both early stopping (□) and non-early stopping (○). Error bars indicate one standard deviation.

4.1 Regularisation

The explicit regularizer with the largest boost to accuracy is augmentation. For CIFAR-10, error drops 23% from the baseline, and for CIFAR-100 it drops 21%. we find that dropout results in a modest improvement, with error dropping 4% and 5% on the datasets respectively. Finally, weight decay has a negative effect on generalisation on CIFAR-10 (26% worse) and is indistinguishable from baseline on CIFAR-100.



(a) IGS (b) SAM-Sharpness (c) Relative Flatness

Figure 2: Scatter plot of train loss vs IGS (left), SAM-sharpness (middle) and Relative Flatness (right), for models on CIFAR-10 for early stopping (□).

4.2 Sharpness

4.2.1 Explicit regularisation and sharpness

In our second research question, we suggested adding explicit regularisers might make the model converge to a flatter minima. Our experimental results provide evidence that this is not true in general. For both data augmentation and weight decay, all sharpness measures are larger than the baseline, indicating that the loss landscape is sharper.

We do not find a consistent correlation between sharpness measures and generalisation across our experiments. In Figure 1 we plot test error against the various measures/metrics. Only for Weight Norm we find a statistically significant negative correlation ($\rho = -0.74$, $p = 0.03$). Relative Flatness shows a weak, but not significant, negative correlation ($\rho = -0.2$, $p = 0.6$). All other metrics show a slight positive (ρ between 0.3 and 0.4) but not significant ($p > 0.4$) correlation.

4.2.2 Potential Explanations

We have seen that neither `augmentation` nor `weight decay` follow the hypothesis of leading to a flatter minima. We posit two possible explanations for these unexpected results.

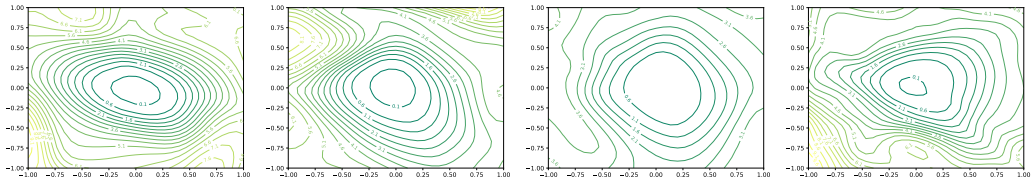
Correlation with training loss Both data `augmentation` and `weight decay` lead to a notably higher train set loss than the baseline ($8\times$ and $20\times$ increase respectively). For `dropout`, the train loss is similar to the baseline. Given this, we suspect there is a connection between augmentation and weight decay finding sharper minima and minima that have higher train loss.

In Figure 2 we plot train loss against IGS, SAM-sharpness and Relative Flatness. We find a strong correlation for all of them ($\rho = 1.0, 0.86, 0.67, p < 0.01, p = 0.01, p = 0.07$ respectively). For our limited number of runs, the correlation is significant for SAM-sharpness and IGS, and Relative Flatness is approaching significance. The strongest correlation is for IGS. An intuitive argument for this relationship could be as follows. The smaller the loss function is, the closer it is to a minima, where the gradient goes to zero. Therefore, we could expect a smaller loss to relate to a smaller loss gradient. As the IGS is calculated as the sum of FIM-metrics of per sample loss gradient, this result is plausible.

For other definitions of sharpness, we might also expect the sharpness to increase with the gradient norm. For instance, when the gradient of the loss function is larger, the highest point on the loss landscape at a distance ρ from the original weights will be higher, when ρ is small enough that the local landscape is well approximated by a linear function. SAM-sharpness accordingly also shows strong correlation with train loss.

In order to conclusively determine such a correlation between sharpness and loss, we would need to investigate a much larger number of models, with varying hyperparameters, training setups and regularisations. If such a correlation holds in the regime our experiments are run in, then we can explain the higher sharpness of data `augmentation` and `weight decay` as simply being a consequence from their training loss being higher.

We can also note that it is reasonable to expect the original training loss to be higher when applying data augmentation or weight decay, as both these regularisation methods change the loss function



(a) `baseline` (b) `dropout` (c) `weight decay` (d) `augmentation`

Figure 3: Train contour loss landscapes for CIFAR-10 using 20% of training dataset on early stopping models

that SGD is minimising. Augmentation does this by changing the dataset the loss is calculated over and weight decay does this by adding a term to the expression for the loss.

Altered loss function makes sharpness comparison meaningless As mentioned above, both `augmentation` and `weight decay` alter the loss function that is being optimised, while for `dropout` the loss function stays the same. It could be that when the loss function is explicitly altered, it becomes meaningless to compare sharpness values on the original train dataset. In these cases, it is not certain that the model converges towards a minimum in the original train loss, which means that a different loss landscape geometry should be expected. We could also say that the explicit regularisers might bring the model into a new region of the loss landscape, where the correlation between generalisation and sharpness does not hold when comparing models from different regions. For instance, weight decay will bring the model to a region with lower weight norm. Previous work [22; 26; 27], have shown the sharpness-generalisation correlation empirically within the same region, as they focus on varying parameters of the model and training process that don't alter the loss function, but haven't shown correlation between different regions.

A counterpoint to this is that features of the training setup that typically have been considered in the sharpness-generalisation correlation, such as learning rate and batch size, actually affect the implicit regularisation of the network. Hence they also alter the loss landscape. Regardless, these results highlight a need of further investigation in sharpness measures and their comparability across explicit regularisers.

4.2.3 Sharpness and task complexity

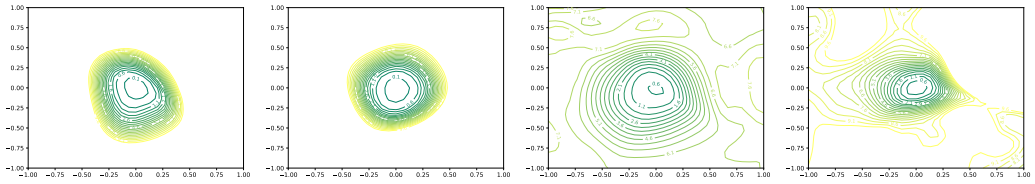
The impacts of the explicit regularisers on sharpness are exacerbated when the complexity of the dataset is increased from CIFAR-10 to CIFAR-100. For the weight norm, Relative Flatness and SAM sharpness measures, the `baseline` and `dropout` models increase in sharpness is less severe. But for `augmentation` and `weight decay` there is an increase in excess of 100% in sharpness between CIFAR-10 and CIFAR-100.

Visualising the train landscape for CIFAR-10 (Figure 3) it is evident that the `augmentation` landscape is more complex with the minima also being smaller. For `weight decay` we see a shallower loss region when compared to the baseline and `dropout` models in the train landscape. However, when the dataset complexity is increased, corresponding to tighter decision boundaries, both `weight decay` and `augmentation` relate to more complex landscapes (Figure 4).

For the VGG-19 architecture on a the harder classification task of CIFAR-100, we see sharper landscapes emerge, moving towards tighter decision boundaries. For the sharpest models, we often see better calibration and sometimes improved accuracy in the case of `augmentation`. This suggests that sharper regions of the landscape where tighter decision boundaries exist can be useful for generalisation and calibration. A possible explanation is that `weight decay` and `augmentation` reduce the simplicity bias, especially for CIFAR-100 (Figure 4), causing the increased sharpness of the loss landscape and, in turn, potentially explaining impacts on calibration. We will now discuss calibration in more detail.

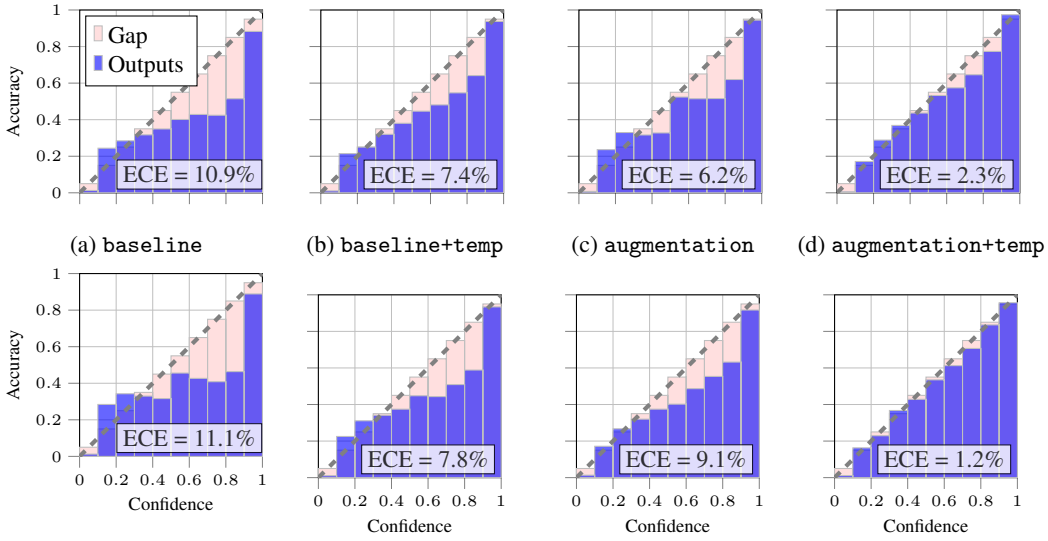
4.3 Calibration

Regularisers and calibration Our results show that `augmentation` is not only the best regularizer to improve generalisation, but also the most calibrated (ECE 40% better on CIFAR-10 and 48% better



(a) `baseline` (b) `dropout` (c) `weight decay` (d) `augmentation`

Figure 4: Train contour loss landscapes for CIFAR-100 using 20% of training dataset on early stopping models



(a) `baseline` (b) `baseline+temp` (c) `augmentation` (d) `augmentation+temp`

(e) `dropout` (f) `dropout+temp` (g) `weight decay` (h) `weight decay+temp`

Figure 5: Reliability plots for CIFAR-10. Model with median test accuracy among seeds used.

on CIFAR-100 than baseline). Meanwhile, the other regularizers have ECE indistinguishable from baseline on CIFAR-10. On CIFAR-100, `dropout` leads to a modest improvement to calibration; `weight decay` substantially improves ECE, despite indistinguishable error.

Temperature and calibration Adding the tuned temperature parameter dramatically reduces ECE: models are better calibrated across the board with temperature scaling. In practice, temperature T is always greater than 1 and around 1.6 for all of the models. This represents calibration that makes the model less confident on average. In other words, the networks begin over-confident, in line with expectation. The improvement is, however, greater for some regularizers than others. In particular, we find that the reduction in ECE for `dropout` is almost the same as for the baseline model (reduction of 29-31% for baseline and `dropout` models in both datasets). In contrast, the reduction in ECE for `augmentation` and `weight decay` is much larger, between 58-77%. Our results show that augmentation leads to the best generalisation; it also enables very good calibration. In contrast, weight decay did not improve generalisation, but does have better calibration than baseline.

In terms of correlations: Before temperature-scaling, there is a moderate ($\rho = 0.54$) but not significant ($p = 0.17$) correlation between how calibrated the final models are (ECE) and test error. After scaling, any notion of this correlation is gone: we find correlation drops to $\rho = 0.24$. Our results for weight decay are an obvious empirical explanation for this: generalisation doesn't improve but the model is much better calibrated as discussed. Figures 5 and Appendix Figure 9 show reliability plots of the models in CIFAR-10 and CIFAR-100 respectively, with and without temperature scaling. These visualize our findings that using `augmentation` and `weight decay` enables much better calibration, compared to using `dropout` or no explicit regularizer. They also again portray how over-confident the networks all are pre-calibration (as indicated by the relative size of the last bar).

In summary, our baseline models without explicit regularisation cannot be tuned well with temperature-scaling. Adding dropout does not help this problem. In contrast, using data augmentation or weight decay both improves the calibration of the model immediately after training and further enables temperature-scaling to be used to near-perfectly calibrate the model on the validation set.

Calibration and sharpness Geometric properties of the loss landscape may not correlate with accuracy, but our results suggest they do with calibration.

For all models where temperature scaling was performed, $T > 1$ was found to be optimal, indicating that the model was previously overconfident. With lower temperature, the output probabilities can quickly shift from high probability in one class to high probability in another from only a small change in the logits. With higher temperature, the probabilities and therefore the loss is less sensitive to the logits. With this, one could expect temperature scaling to give a flatter loss landscape. In Table 1 we see that this is not the case in general. For the Relative Flatness measure, a consistently sharper landscape is reported before and after temperature scaling.

We find consistent negative correlations between ECE and the sharpness measures. We find strong and significant negative correlation between ECE and relative flatness both before ($\rho = -0.75, p = 0.03$) and after ($\rho = -0.78, p = 0.02$) temperature scaling. This correlation also holds for IGS (before: $\rho = -0.74$, after: $\rho = -0.80$), but now they are no longer significant ($p = 0.26, 0.20$). In general we report two trends here. The first is a consistent (albeit not always significant) correlation whereby better calibrated models sit at sharper minima. The second is that temperature-scaling only makes this correlation stronger.

5 Limitations and Future Work

To further interrogate our findings, it would be necessary to set up and run our setup on a range of architectures: other variants of the VGG architecture, ResNets [20] and vision transformers [32] would be starting candidates. Few of our correlation coefficients were significant, almost certainly because we simply had too few datapoints. Such further investigations would enable us to discover which correlations are significant.

Additionally, we trained our VGG-19 model for up to 75 epochs. In particular for CIFAR-100, other literature would suggest that longer training with an adaptive learning rate can reach better accuracies. While our decision to train for less epochs with a higher learning rate was as a result of our budget constraints, we would have liked to observe the trends over longer training durations to higher accuracies closer to SoTA.

Finally, we were only able to compute the IGS metric for CIFAR-10 with early stopping. In practice, computing these numbers took as long as training all the other models combined on an A100 GPU. And that was only for 20 batches of size 64 (less than 3% of the training set). To compute IGS on CIFAR 100, we would have needed to reduce batch size to 4, which would take too long under our budget constraints.

6 Conclusions

In this work, we have compared the effects of different regularisers on loss landscape sharpness and calibration. In contradiction to some claims in the literature, our results provide empirical evidence that better-generalising models (via explicit regularisers) tend to have sharper loss landscapes, or there is no trend at all. We provided two explanations for why this could be: a potential correlation between train loss and sharpness, and an argument centered around regularisation changing the loss function. Our results either suggest that sharpness is not a good predictor of generalisation across different regularisation methods, or that current sharpness measures are not sufficient descriptors of sharpness.

Furthermore, we have shown that better calibrated models often have sharper landscapes. Indeed, temperature scaling can lead to increased sharpness. Consequently we posit that to achieve a well-calibrated and accurate model, it may be necessary to navigate to sharper regions of the landscape where tighter decision boundaries exist. We also discovered that weight decay seems to enable

temperature-scaling to be an effective mode of calibration, even if it does not improve generalisation directly.

Overall, this work highlights a need for further investigation, both empirical and theoretical, on the interplay between loss landscape geometry, calibration, and how different forms of explicit regularisation lead to better generalisation.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[2] J. Kukačka, V. Golkov, and D. Cremers, "Regularization for deep learning: A taxonomy," 2017.

[3] Y. Tian and Y. Zhang, "A comprehensive survey on regularization strategies in machine learning," *Information Fusion*, vol. 80, pp. 146–166, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156625352100230X

[4] R. Moradi, R. Berangi, and B. Minaei, "A survey of regularization strategies for deep models," *Artif. Intell. Rev.*, vol. 53, no. 6, p. 3947–3986, aug 2020. [Online]. Available: https://doi.org/10.1007/s10462-019-09784-7

[5] C. F. G. D. Santos and J. a. P. Papa, "Avoiding overfitting: A survey on regularization methods for convolutional neural networks," *ACM Comput. Surv.*, vol. 54, no. 10s, sep 2022. [Online]. Available: https://doi.org/10.1145/3510413

[6] S. Hochreiter and J. Schmidhuber, "Simplifying neural nets by discovering flat minima," *Advances in neural information processing systems*, vol. 7, 1994. [Online]. Available: https://proceedings.neurips.cc/paper/1994/file/01882513d5fa7c329e940dda99b12147-Paper.pdf

[7] ——, "Flat minima," *Neural computation*, vol. 9, no. 1, pp. 1–42, 1997. [Online]. Available: https://direct.mit.edu/neco/article-abstract/9/1/1/6027/Flat-Minima?redirectedFrom=fulltext

[8] J. Kaddour, L. Liu, R. Silva, and M. J. Kusner, "When do flat minima optimizers work?" *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 577–16 595, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/69b5534586d6c035a96b49c86dbeece8-Abstract-Conference.html

[9] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=6Tm1mposlrM

[10] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio, "Sharp minima can generalize for deep nets," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1019–1028. [Online]. Available: https://proceedings.mlr.press/v70/dinh17b

[11] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," *Advances in neural information processing systems*, vol. 31, 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330. [Online]. Available: http://proceedings.mlr.press/v70/guo17a.html

[13] A. Hernández-García and P. König, "Data augmentation instead of explicit regularization," *arXiv preprint arXiv:1806.03852*, 2018. [Online]. Available: https://arxiv.org/abs/1806.03852

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[15] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Proceedings of the 4th International Conference on Neural Information Processing Systems*, ser. NIPS'91. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1991, p. 950–957. [Online]. Available: https://dl.acm.org/doi/10.5555/2986916.2987033

[16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Sy8gdB9xx

[17] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli, "The pitfalls of simplicity bias in neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9573–9585, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6cfe0e6127fa25df2a0ef2ae1067d915-Abstract.html

[18] D. Zhao, "Combining explicit and implicit regularization for efficient learning in deep networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3024–3038, 2022. [Online]. Available: https://arxiv.org/pdf/2306.00342.pdf

[19] S. L. Smith, B. Dherin, D. Barrett, and S. De, "On the origin of implicit regularization in stochastic gradient descent," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=rq_Qr0c1Hyo

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[21] W. R. Huang, Z. Emam, M. Goldblum, L. Fowl, J. K. Terry, F. Huang, and T. Goldstein, "Understanding generalization through visualizations," in *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, ser. Proceedings of Machine Learning Research, J. Zosa Forde, F. Ruiz, M. F. Pradier, and A. Schein, Eds., vol. 137. PMLR, 12 Dec 2020, pp. 87–97. [Online]. Available: https://proceedings.mlr.press/v70/dinh17b

[22] Y. Jiang*, B. Neyshabur*, H. Mobahi, D. Krishnan, and S. Bengio, "Fantastic generalization measures and where to find them," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SJgIPJBFvH

[23] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro, "Exploring generalization in deep learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/10ce03a1ed01077e3e289f3e53c72813-Paper.pdf

[24] Z. Yao, A. Gholami, Q. Lei, K. Keutzer, and M. W. Mahoney, "Hessian-based analysis of large batch training and robustness to adversaries," *Advances in Neural Information Processing Systems*, vol. 31, 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/102f0bb6efb3a6128a3c750dd16729be-Abstract.html

[25] J. Kwon, J. Kim, H. Park, and I. K. Choi, "Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 5905–5914. [Online]. Available: https://proceedings.mlr.press/v139/kwon21b.html

[26] H. Petzka, M. Kamp, L. Adilova, C. Sminchisescu, and M. Boley, "Relative flatness and generalization," *Advances in neural information processing systems*, vol. 34, pp. 18 420–18 432, 2021. [Online]. Available: https://openreview.net/forum?id=sygvo7ctb_

[27] C. Jang, S. Lee, F. C. Park, and Y.-K. Noh, "A reparametrization-invariant sharpness measure based on information geometry," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=AVh_HTC76u

[28] T. Liang, T. Poggio, A. Rakhlin, and J. Stokes, "Fisher-rao metric, geometry, and complexity of neural networks," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 888–896. [Online]. Available: https://proceedings.mlr.press/v89/liang19a.html

[29] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 29, no. 1, 2015. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/9602

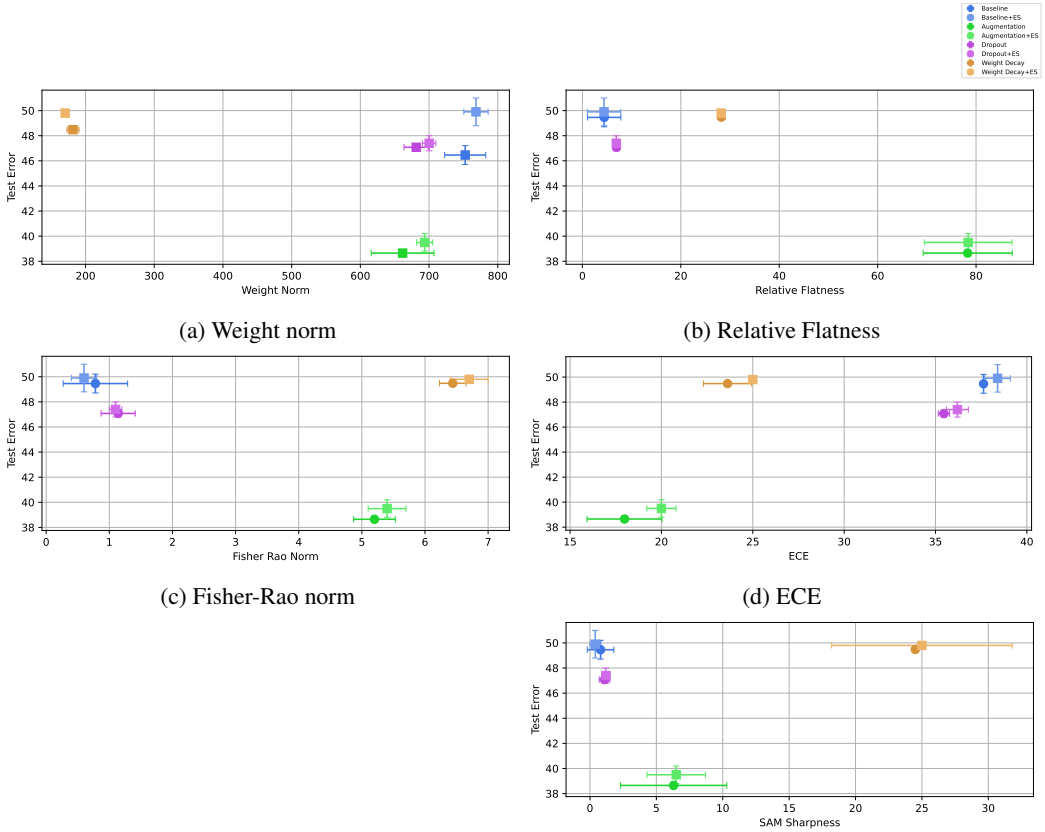
[30] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

A CIFAR-100 Generalisation Correlation

Generalisation correlation of CIFAR-100 is in line with results from CIFAR-10. One exception is that weight norm no longer has a strongly negative correlation with the test error.



(a) Weight norm

(b) Relative Flatness

(c) Fisher-Rao norm

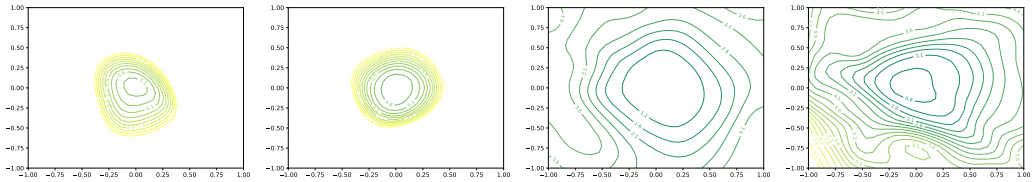
(d) ECE

(e) SAM Sharpness

Figure 6: Generalisation Correlation with CIFAR-100. Results here include both early stopping (□) and non-early stopping (○).

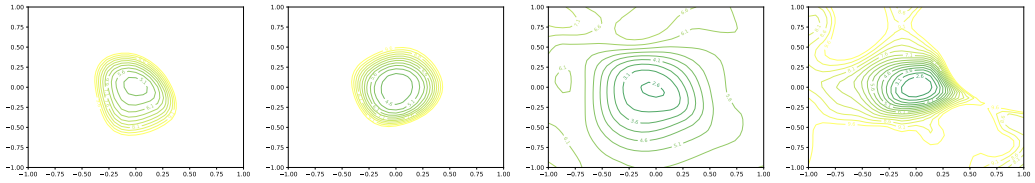
B Test Loss Landscapes

Here we see that the test loss landscape mimics that of the train loss landscape provided in the report in Figures 3 and 4. Once again we find the use of augmentation leads to a more complex loss landscape compared to the `baseline` and `dropout` loss landscape. Considering the gains in ECE that we see for these models, this could again be due to these explicit regularisers reducing the simplicity bias of the neural network [17], therefore causing sharper points in the landscape to be reached. There is also an increase in the complexity of the loss landscape in line with task complexity as we note in the main body of this report.



(a) `baseline` (b) `dropout` (c) `weight decay` (d) `augmentation`

Figure 7: Test loss landscapes for CIFAR-100, loss landscapes generated 100% of the test dataset for the early stopped models.

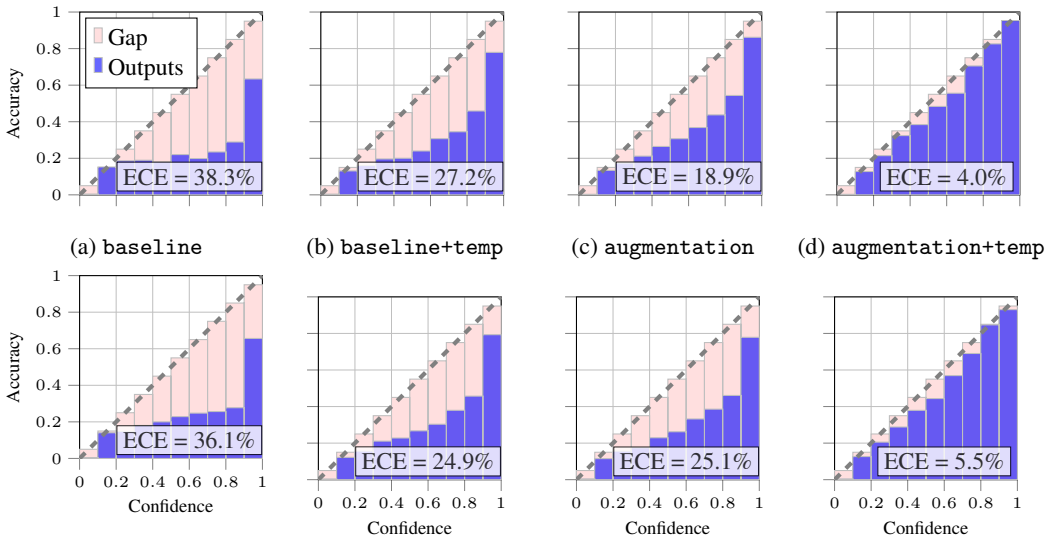


(a) `baseline` (b) `dropout` (c) `weight decay` (d) `augmentation`

Figure 8: Test loss landscapes for CIFAR-100, loss landscapes generated 100% of the test dataset for the early stopped models.

C Reliability Plots

The reliability plots for CIFAR-100 are in line with our results from the plots from CIFAR-10.



(a) `baseline` (b) `baseline+temp` (c) `augmentation` (d) `augmentation+temp`

(e) `dropout` (f) `dropout+temp` (g) `weight decay` (h) `weight decay+temp`

Figure 9: Reliability plots for CIFAR-100

D Results without early stopping

Regularizer	Train loss ($\times 10^{-2}$) ↓	Test error (%) ↓	ECE(%) ↓	Weight norm ↓	Fisher-Rao norm ↓	Relative Flatness ↓	SAM-sharpness ↓ ($\times 10^{-2}$)
`Baseline`	$0.5_{\pm 0.4}$	$14.0_{\pm 0.4}$	$10.8_{\pm 0.4}$	$49.1_{\pm 3.9}$	$0.4_{\pm 0.3}$	$2.2_{\pm 1.0}$	$1.7_{\pm 1.3}$
`+ temp`	$1.2_{\pm 0.7}$	$14.0_{\pm 0.4}$	$7.4_{\pm 0.6}$	$49.1_{\pm 3.9}$	$0.5_{\pm 0.2}$	$3.8_{\pm 1.7}$	$1.3_{\pm 0.6}$
`Augmentation`	$6.4_{\pm 1.4}$	$10.6_{\pm 0.3}$	$6.0_{\pm 0.3}$	$49.6_{\pm 1.6}$	$2.2_{\pm 0.3}$	$12.2_{\pm 1.8}$	$3.2_{\pm 0.6}$
`+ temp`	$8.3_{\pm 1.36}$	$10.6_{\pm 0.3}$	$2.2_{\pm 0.3}$	$49.6_{\pm 1.6}$	$1.8_{\pm 0.1}$	$13.3_{\pm 0.6}$	$4.0_{\pm 1.0}$
`Dropout`	$0.6_{\pm 0.2}$	$13.4_{\pm 0.2}$	$10.5_{\pm 0.1}$	$38.3_{\pm 1.0}$	$0.6_{\pm 0.0}$	$1.2_{\pm 0.2}$	$0.9_{\pm 0.5}$
`+ temp`	$1.5_{\pm 0.3}$	$13.4_{\pm 0.2}$	$7.1_{\pm 0.3}$	$38.3_{\pm 1.0}$	$0.6_{\pm 0.1}$	$2.8_{\pm 0.4}$	$0.7_{\pm 0.3}$
`Weight decay`	$11.2_{\pm 1.4}$	$16.2_{\pm 0.0}$	$8.9_{\pm 0.4}$	$34.4_{\pm 0.8}$	$2.7_{\pm 0.1}$	$4.5_{\pm 0.2}$	$13.5_{\pm 3.0}$
`+ temp`	$15.8_{\pm 1.2}$	$16.2_{\pm 0.0}$	$2.4_{\pm 0.6}$	$34.4_{\pm 0.8}$	$2.4_{\pm 0.1}$	$6.1_{\pm 0.3}$	$9.6_{\pm 1.4}$

Table 3: Final results (without early stopping) for VGG-19 on CIFAR-10

Regularizer	Train loss ($\times 10^{-2}$) ↓	Test error (%) ↓	ECE(%) ↓	Weight norm ↓	Fisher-Rao norm ↓	Relative Flatness ↓	SAM-sharpness ↓($\times 10^{-2}$)
`Baseline`	$0.6_{\pm 0.8}$	$49.5_{\pm 0.8}$	$37.6_{\pm 0.1}$	$752.3_{\pm 29.9}$	$0.8_{\pm 0.5}$	$4.4_{\pm 3.4}$	$0.8_{\pm 1.0}$
`+ temp`	$3.7_{\pm 2.1}$	$49.5_{\pm 0.8}$	$26.0_{\pm 0.6}$	$752.3_{\pm 29.9}$	$1.1_{\pm 0.4}$	$19.7_{\pm 4.8}$	$1.0_{\pm 0.7}$
`Augmentation`	$38.5_{\pm 7.7}$	$38.7_{\pm 0.3}$	$18.0_{\pm 2.1}$	$661.5_{\pm 45.7}$	$5.2_{\pm 0.3}$	$78.3_{\pm 9.1}$	$6.3_{\pm 4.0}$
`+ temp`	$49.6_{\pm 8.8}$	$38.7_{\pm 0.3}$	$4.6_{\pm 2.1}$	$661.5_{\pm 45.7}$	$4.6_{\pm 0.4}$	$71.7_{\pm 8.7}$	$7.6_{\pm 4.2}$
`Dropout`	$1.4_{\pm 1.2}$	$47.1_{\pm 0.3}$	$35.5_{\pm 0.3}$	$681.3_{\pm 17.7}$	$1.1_{\pm 0.3}$	$6.9_{\pm 0.7}$	$1.1_{\pm 0.4}$
`+ temp`	$4.2_{\pm 1.4}$	$47.1_{\pm 0.3}$	$24.0_{\pm 0.6}$	$681.3_{\pm 17.7}$	$1.3_{\pm 0.2}$	$17.0_{\pm 0.4}$	$1.3_{\pm 0.8}$
`Weight decay`	$55.1_{\pm 3.5}$	$49.5_{\pm 0.2}$	$23.6_{\pm 1.3}$	$181.7_{\pm 8.8}$	$6.3_{\pm 0.2}$	$28.2_{\pm 0.7}$	$24.5_{\pm 9.2}$
`+ temp`	$70.3_{\pm 2.8}$	$49.5_{\pm 0.2}$	$5.1_{\pm 0.9}$	$181.7_{\pm 8.8}$	$6.3_{\pm 0.1}$	$32.3_{\pm 0.7}$	$20.7_{\pm 7.5}$

Table 4: Final results (without early stopping) for VGG-19 on CIFAR-100