Explicit Regularisation, Sharpness and Calibration

Israel Mason-Williams[*†] **Fredrik Ekholm**[*] **Ferenc Huszár**
Department of Computer Science and Technology
University of Cambridge
`{ifm24,fwe21}@cantab.ac.uk` `h277@cam.ac.uk`

Abstract

We probe the relation between flatness, generalisation and calibration in neural networks, using explicit regularisation as a control variable. Our findings indicate that the range of flatness metrics surveyed fail to positively correlate with variation in generalisation or calibration. In fact, the correlation is often opposite to what has been hypothesized or claimed in prior work, with calibrated models typically existing at sharper minima compared to relative baselines, this relation exists across model classes and dataset complexities.

1 Introduction

The learning process of neural networks around and up to the convergence point has been studied from the perspective of the loss landscape [1; 2]. Some work posits that flatness in the loss landscape, i.e the change of loss under perturbations of model parameters, is a predictor of generalisation [3; 4]. Others have pointed out that some metrics of flatness are not reparametrization invariant, meaning that flat landscapes could be arbitrarily sharpened [5].

Loss landscape flatness is often studied in the context of generalisation, for example investigating the impact of implicit regularisation, such as from stochastic gradient descent and residual connections [6]. But generalisation is only one way to measure the usefulness of the trained model. We further ask whether flatness correlates with other desirable properties of the model. For example, Liu et al. [7] claim that flatness correlates with the "transferability" of the learnt representation in self-supervised learning. Another desirable property of the trained model is calibration; calibration is a measure of the alignment between true and predicted class probabilities of the model, often used to highlight overconfidence in model predictions [8]. To the best of our knowledge, the relationship between flatness and calibration has not been studied directly, we extend our investigation in this direction. It would be reasonable to expect more calibrated models to sit at flatter minima. This is because poorly calibrated models typically fail by being overconfident, and in this case the variance of loss across samples is higher than for calibrated models, suggesting a more complex loss landscape.

In particular, we seek to answer the following two questions.

1. If using an explicit regulariser improves generalisation, does it also lead to finding a flatter minima (under any measures of flatness in the literature)?
2. Further, are well-calibrated models typically at flatter minima?

To this end, we train VGG-19 [9], ResNet-20 [10] and ViT [11] networks on the CIFAR [12] image classification tasks, and use the explicit regularisers dropout [13], weight decay [14], data augmentation and early stopping as control variables. There is no universal agreed-upon exact

[*] Equal contribution of authors.

[†] Work completed at Cambridge, now at UKRI Safe and Trusted AI.

Workshop on Scientific Methods for Understanding Deep Learning at Neural Information Processing Systems (NeurIPS 2024).

definition of model sharpness. Thus, we compute five different measures of sharpness from the literature, namely weight norm, Fisher-Rao norm [15], relative flatness [16], SAM-sharpness [4] and information geometric sharpness [17]. These are briefly introduced in section 2. Our measure of calibration will be expected calibration error (ECE) [8]. We provide mathematical formulations of these measures and more detail of our experimental setup in Appendix B. In the main body of this paper we will focus on our analysis of the results from training the VGG-19 network. We provide results for the other tasks and network architectures in Appendix C-E.

2 Sharpness Metrics

We detail five different sharpness measures from the literature: l^2-norm, *Fisher-Rao* norm, *SAM-Sharpness*, *Relative Flatness* and *IGS*. l^2-norm and Fisher-Rao norm might be considered capacity measures rather than sharpness measures, but have connections to loss landscape curvature [15] and has been hypothesised to predict generalisation [18]. We will refer to all five as *sharpness measures* for simplicity.

In previous work, nonreparametrisation-invariance have been raised as an issue of proposed sharpness metrics [16; 17]. This refers to scenarios where a reparametrisation of the model, i.e. a modification to the model weights without affecting the model function, ends up changing the sharpness. Scale-invariance is a weaker type of reparametrisation-invariance, where the sharpness is only invariant to scaling weights of the same layers by the same constant. In the measures we use, Relative Flatness and Fisher Rao norm are scale-invariant, while IGS is fully reparametrisation-invariant.

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. [4].

Relative Flatness Petzka et al. [16] 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.

IGS *Information Geometric Sharpness (IGS)* [17] 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:

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

The 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.

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 [15].

3 Results

Table 1: Results for VGG-19 on CIFAR-10. Lower numbers indicate 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 ($\times 10^{-2}$)	log IGS
`baseline`	$0.9_{0.5}$	$14.4_{0.1}$	$11.0_{0.2}$	$49.3_{4.1}$	$0.8_{0.3}$	$2.2_{1.0}$	$1.2_{0.7}$	$-1.9_{0.5}$
`+ temp scaling`	$1.8_{0.8}$		$7.5_{0.6}$		$0.7_{0.2}$	$3.8_{1.7}$	$1.3_{0.6}$	$-0.6_{0.3}$
`augmentation`	$7.0_{2.4}$	$11.1_{0.7}$	$6.6_{0.6}$	$51.5_{1.6}$	$2.3_{0.6}$	$12.1_{1.5}$	$6.6_{4.3}$	$0.9_{0.3}$
`+ temp scaling`	$8.6_{1.8}$		$2.7_{0.5}$		$1.8_{0.3}$	$13.2_{0.6}$	$4.6_{2.3}$	$0.8_{0.2}$
`dropout`	$0.8_{0.2}$	$13.9_{0.2}$	$11.0_{0.2}$	$39.1_{0.3}$	$0.8_{0.1}$	$1.2_{0.2}$	$2.8_{2.8}$	$-2.5_{0.4}$
`+ temp scaling`	$1.8_{0.3}$		$7.6_{0.2}$		$0.7_{0.1}$	$2.9_{0.4}$	$1.7_{1.6}$	$-0.9_{0.2}$
`weight decay`	$18.0_{3.1}$	$18.2_{1.3}$	$10.8_{1.3}$	$33.2_{0.9}$	$3.9_{0.5}$	$4.5_{0.2}$	$9.2_{0.4}$	$1.5_{0.1}$
`+ temp scaling`	$21.4_{2.0}$		$2.9_{1.3}$		$2.9_{0.1}$	$6.2_{0.3}$	$6.5_{2.0}$	$1.3_{0.0}$

Table 2: Results for VGG-19 on CIFAR-100. Lower numbers indicate 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`	$0.3_{0.3}$	$49.9_{1.1}$	$38.4_{0.7}$	$768.3_{17.7}$	$0.6_{0.2}$	$4.4_{3.4}$	$0.4_{0.4}$
`+ temp scaling`	$2.8_{1.0}$		$27.1_{0.4}$		$1.0_{0.2}$	$19.7_{4.8}$	$0.9_{0.8}$
`augmentation`	$36.0_{5.5}$	$39.5_{0.7}$	$20.0_{0.8}$	$693.6_{11.7}$	$5.4_{0.3}$	$78.4_{8.9}$	$6.5_{2.2}$
`+ temp scaling`	$45.7_{6.7}$		$6.1_{1.5}$		$4.5_{0.4}$	$71.7_{9.1}$	$8.0_{2.3}$
`dropout`	$1.1_{0.7}$	$47.4_{0.6}$	$36.2_{0.6}$	$700.2_{9.8}$	$1.1_{0.1}$	$6.9_{0.7}$	$1.2_{0.3}$
`+ temp scaling`	$3.6_{0.5}$		$24.9_{0.5}$		$1.2_{0.1}$	$17.0_{0.5}$	$1.4_{0.6}$
`weight decay`	$57.5_{3.6}$	$49.8_{0.3}$	$25.0_{0.2}$	$170.3_{2.8}$	$6.7_{0.3}$	$28.2_{0.7}$	$25.0_{6.8}$
`+ temp scaling`	$71.0_{2.6}$		$5.9_{0.3}$		$6.5_{0.0}$	$32.3_{0.6}$	$16.7_{1.5}$

Augmentation and dropout improve generalisation. The control with the largest error reduction is augmentation, error drops 23% from the baseline on CIFAR-10, and for CIFAR-100 it drops 21%. We find that dropout results in a modest reduction, with error dropping 4% and 5% on the respective datasets. Weight decay, overall has a negative effect on generalisation on CIFAR-10 (26% increase in error). On CIFAR-100, this network is numerically near-identical from baseline.

Explicit regularisation does not lead to flatter minima. For augmentation and weight decay, sharpness measures are larger than the baseline, indicating a sharper loss landscape. We find no consistent correlation between sharpness measures and generalisation across controls and experiments.

We visualise the loss landscape using the tools released by Li et al. [6]. From Appendix Figure 4 it is evident that the `augmentation` landscape is again more complex with a tighter minima. For `weight decay` we see a shallow loss region compared to the baseline and `dropout` models for the train and test, Figure 6, landscape. When task complexity is increased, corresponding to tighter decision boundaries, both `weight decay` and `augmentation` relate to more complex landscapes (Figure 5).

Better calibrated models may exist at sharper loss landscape. Our results show that `augmentation` improves generalisation and calibration the most with ECE 40% better on CIFAR-10 and 48% better on CIFAR-100 compared to the baseline.

We now add a tuned temperature parameter, T, to the baseline and controls, as is common in practice in calibration experiments [19]. T is a single parameter scaling the predicted probabilities before softmax to smooth prediction probabilities. Employing temperature scaling dramatically reduces ECE, improving calibration. In practice, temperature is $1 < T < 1.6$ for the trained models. In particular, we find that the reduction in ECE when using temperature scaling for `dropout` is similar to 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%. While augmentation leads to the best generalisation and creates a calibrated model, weight decay, which does not improve generalisation, also results in have a calibrated model.

In Appendix C.2 we show reliability plots for VGG-19, with and without temperature scaling. The figures confirm our findings that using `augmentation` and `weight decay` enable much better calibration, compared to using `dropout` or no explicit regularizer. They also again portray the overconfidence the of the networks across controls pre-calibration.

In summary, our baseline models without explicit regularisation cannot be tuned well with temperature-scaling. But we find 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. Suggesting that these regularisers are able to navigate to tighter decision boundaries that are more receptive to temperature scaling.

Calibration is negatively correlated with a sharpness One reasonable hypothesis is that a more confident model (one that is likely to predict higher probability) would show a flatter loss landscape than a less confident model. In the extreme case, a highly confident image classification model will always predict 100% on some class while predicting 0% in all others, with the class getting 100% depending on the input image. This means that a perturbation the input can cause the model to quickly change from 0% to 100% in some class, causing a large change of loss on that sample. As both input and model weights affect intermediate activations, a perturbation of model weights might reasonably also cause a sharp shift in output probabilities, leading to a large change in loss. Small weight perturbations causing large changes to the loss means the loss landscape is sharp.

In the other extreme case, a totally under-confident model will always predict a probability of $\frac{1}{C}$ for every class, with C being the number of classes. Hence its output does not depend on the input at all, meaning the output probably depends less on intermediate activations from earlier layers, and small weight perturbations should affect the loss less.

The above reasoning suggests that under-confident models should show a flatter landscape than overconfident models. In our experiments, we found that the less calibrated models tend to fail by being overconfident, and thus, we would expect the more calibrated models to be less confident in general and thus flatter. In particular, when performing temperature scaling with $T > 1$, the predicted probabilities will be closer to uniform and less dependent on the pre-softmax logits, hence the landscape is expected to be flatter.

Our results show that temperature scaling with $T > 1$ does in fact often decrease the sharpness for the SAM-sharpness and Fisher-Rao measures. Surprisingly however, relative flatness consistently shows a sharper landscape after temperature scaling.

Looking at the effect of the regularisers, we see that adding `augmentation` or `weight decay` give significantly reduced ECE after temperature scaling, 2.7% and 2.9% respectively compared to 7.5% and 7.6% for the baseline and dropout models. These two regularises are also the ones that show the sharpest loss landscapes across all sharpness metrics. This results in an overall negative trend between calibration and sharpness, refuting the hypothesis.

Overall, we report two trends here. The first is a trend that well-calibrated models sit at sharper minima. The second is that temperature-scaling makes this correlation stronger for initially less calibrated models. This finding goes against the above presented hypothesis.

4 Discussion

We have seen that neither `augmentation` nor `weight decay` lead to a flatter minima, and instead lead to sharper minima. We posit two possible explanations for these unexpected results.

Correlation with training loss. Both data `augmentation` and `weight decay` lead to notably higher training set loss than the baseline ($8\times$ and $20\times$ increase respectively), while for `dropout`, the training loss is similar to the baseline. The same two regularisers also consistently give a sharper landscape. Given this, we suspect there is a connection between ending up at sharper minima and ending up at minima that have higher training loss. In , we visualise this relationship by plotting train loss against IGS, SAM-sharpness and Relative Flatness. Both SAM-sharpness and IGS increase with gradient norm. Therefore, one intuitive argument for this relationship is that the smaller the loss is, the closer to a local minima the model is, and at a minima the gradient goes to zero.

Altered loss function makes sharpness comparison meaningless. 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 different part of the loss landscape with different sharpness characteristics, like weight decay pushing the model towards a region of the landscape with lower weight norm. As the previously established correlation between generalisation and sharpness has not focused on regularisation, the correlation might only hold when comparing models from the restricted sub-region of the loss landscape that non-regularised models tend to converge to.

Calibration and sharpness. Our findings show that sharper regions of the landscape, where tighter decision boundaries exist, can be useful for calibration. A possible explanation is that `weight decay` and `augmentation` reduce simplicity bias, especially for CIFAR-100, causing the increased sharpness of the loss landscape. We hypothesise that this could emerge due to more complex and entangled boundaries in calibrated models, which are less resilient to perturbation potentially explaining the relation between impacts on calibration.

Nascent understandings of landscape geometry In summary, we find the following answers to our research questions:

1. While the use of explicit regularisers can improve generalisation, it does not result in a flatter landscape; more often than not, the landscape is sharper.
2. The intuitive hypothesis of calibrated models having a flatter loss landscape is not true in general. In fact, calibrated models often exist at sharper minima, increasing in sharpness as dataset complexity increases across the explored architectures.

5 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 and slightly decreased in some architectures showing that being too sharp is also an issues for calibration. 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] J. Schmidhuber and S. Hochreiter, “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

[2] S. Hochreiter and J. Schmidhuber, “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

[3] 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

[4] 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

[5] 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

[6] 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

[7] H. Liu, S. M. Xie, Z. Li, and T. Ma, “Same pre-training loss, better downstream: Implicit bias matters for language models,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 22 188–22 214. [Online]. Available: https://proceedings.mlr.press/v202/liu23ao.html

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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_

[17] 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

[18] 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

[19] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic, "Revisiting the calibration of modern neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 682–15 694, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/8420d359404024567b5aefda1231af24-Abstract.html

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," *Advances in neural information processing systems*, vol. 32, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/hash/36ad8b5f42db492827016448975cc22d-Abstract.html

[26] 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

[27] 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

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106

[29] I. Loshchilov and F. Hutter, “Fixing weight decay regularization in adam,” 2018. [Online]. Available: https://openreview.net/forum?id=rk6qdGgCZ

[30] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, “Do vision transformers see like convolutional neural networks?” in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS ’21. Red Hook, NY, USA: Curran Associates Inc., 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/652cf38361a209088302ba2b8b7f51e0-Paper.pdf

A Related Work

Previous work [18; 16; 17], have shown the negative sharpness-generalisation correlation empirically by varying parameters of the model (such as number of layers) and training process (such as learning rate and batch size). However, to our knowledge this correlation has not been extensively studied with explicit regularisation as a control variable, even though it is a common way of improving generalisation.

The disparity in the literature regarding the notion of flatness [2] and sharpness [5] in loss landscapes and their relative merits leads to uncertainty when reasoning their impacts. Huang et al. [20] 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. [3] 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.

Many studies posit a correlation between generalisation and measures of sharpness of the loss landscape [e.g. 18; 21]. Mini-batch SGD's generalisation qualities have been aligned with the notion that its implicit regularisation favours convergence towards flatter minima [22; 23]. 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 [4; 24; 16; 17].

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 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 [16].

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.

Another line of work has shown that neural networks tend to be overconfident [8]. In other words, they are often poorly calibrated and predictions are correct less frequently than high confidences would suggest. Studies into calibration suggest that some explicit regularisers such as augmentation [25] and weight decay can improve calibration, however, increasing model capacity through implicit regularisers such as increased depth and width can harm calibration [8]. Although, other experiments have determined that architectural features between model classes can greatly improve calibration [19].

B Experimental Details

B.1 Expected Calibration Error

Naeini et al. [26] 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.

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

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

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

One approach shown to reduce ECE of a neural network, in turn improving calibration, is to temperature-scale the output class logits before softmax [8]. 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 $T = 1$.

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

B.2 Experimental Setup

Datasets We use the CIFAR-10 and CIFAR-100 image classification datasets [12]. 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 [9] 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).

For each of these models, we further follow Guo et al. [8] 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`. It should be noted that for the VGG-19 and ViT architectures it was too computationally expensive to calculate IGS on CIFAR-100 and, therefore, it is not recorded for this dataset on these architectures.

Metrics, measures and visualisations Our metric for generalisation is accuracy on the test set. For calibration, we use ECE. For sharpness we report the metrics described in Section 2: 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).

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). 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 [17] 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. [6], 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.

C Further VGG-19 Results

This section reports the results of the CIFAR datasets without early stopping, correlation plots for train loss and sharpness metrics, reliability plots, loss landscape visualisations and generalisation correlation plots.

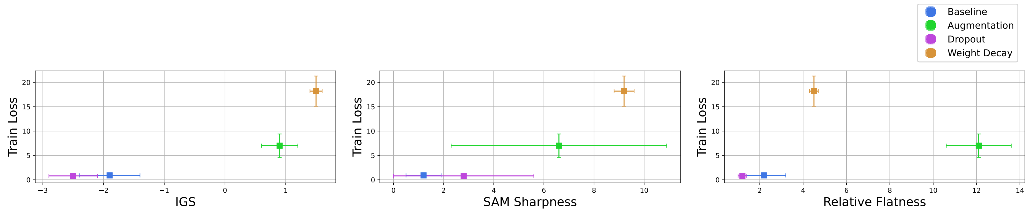
Table 3: Results for VGG-19 on CIFAR-10 without early stopping. Lower numbers indicate lower loss/error, more calibrated, or flatter.

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

Table 4: Results for VGG-19 on CIFAR-100 without early stopping. Lower numbers indicate 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	$0.6_{0.8}$	$49.5_{0.8}$	$37.6_{0.1}$	$752.3_{29.9}$	$0.8_{0.5}$	$4.4_{3.4}$	$0.8_{1.0}$
+ temp scaling	$3.7_{2.1}$	-	$26.0_{0.6}$	$752.3_{29.9}$	$1.1_{0.4}$	$19.7_{4.8}$	$1.0_{0.7}$
Augmentation	$38.5_{7.7}$	$38.7_{0.3}$	$18.0_{2.1}$	$661.5_{45.7}$	$5.2_{0.3}$	$78.3_{9.1}$	$6.3_{4.0}$
+ temp scaling	$49.6_{8.8}$	-	$4.6_{2.1}$	$661.5_{45.7}$	$4.6_{0.4}$	$71.7_{8.7}$	$7.6_{4.2}$
Dropout	$1.4_{1.2}$	$47.1_{0.3}$	$35.5_{0.3}$	$681.3_{17.7}$	$1.1_{0.3}$	$6.9_{0.7}$	$1.1_{0.4}$
+ temp scaling	$4.2_{1.4}$	-	$24.0_{0.6}$	$681.3_{17.7}$	$1.3_{0.2}$	$17.0_{0.4}$	$1.3_{0.8}$
Weight decay	$55.1_{3.5}$	$49.5_{0.2}$	$23.6_{1.3}$	$181.7_{8.8}$	$6.3_{0.2}$	$28.2_{0.7}$	$24.5_{9.2}$
+ temp scaling	$70.3_{2.8}$	-	$5.1_{0.9}$	$181.7_{8.8}$	$6.3_{0.1}$	$32.3_{0.7}$	$20.7_{7.5}$

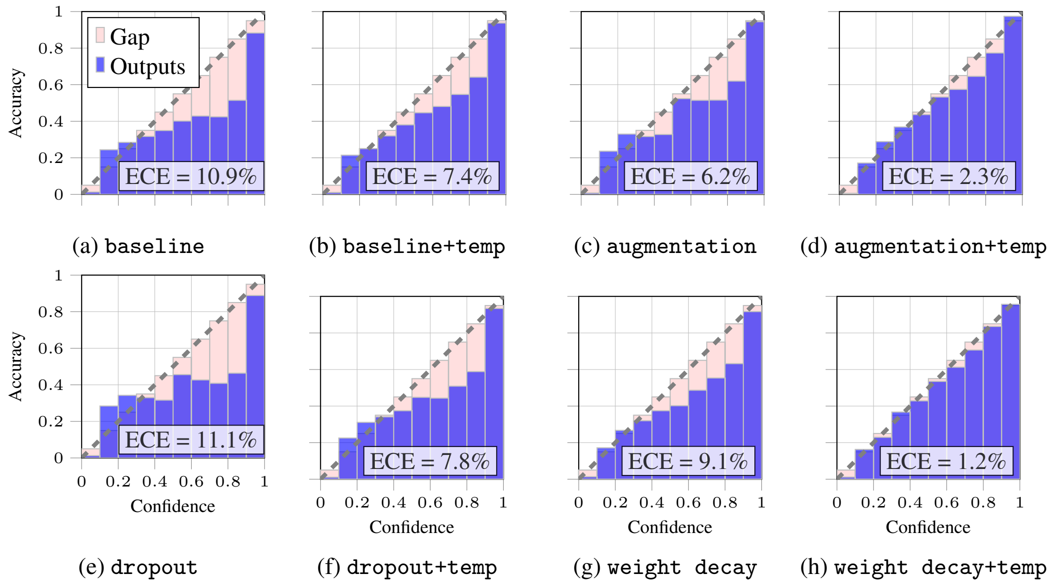
C.1 Correlation between Train Loss Sharpness



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

Figure 1: Scatter plot of train loss vs IGS (left), SAM-sharpness (middle) and Relative Flatness (right), for models on CIFAR-10 for early stopping (□).

C.2 Reliability Plots



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

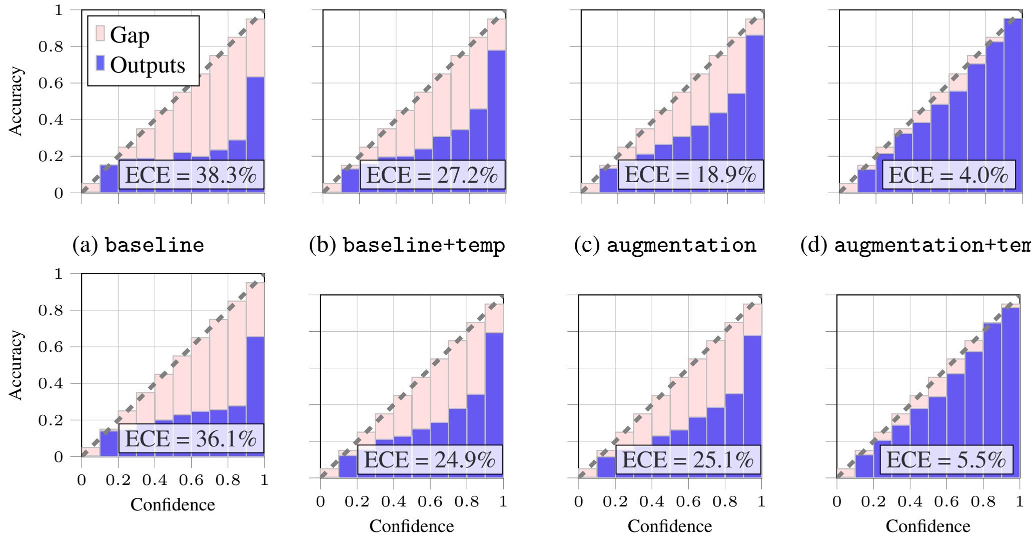
(e) `dropout` (f) `dropout+temp` (g) `weight decay` (h) `weight decay+temp`

Figure 2: Reliability plots for CIFAR-10. Model with median test accuracy among seeds used.

The reliability plots for CIFAR-100, Figure 3, are in line with our results from the plots from CIFAR-10 as seen in Figure 2, discussed in the main body of the report. The consistency suggests that regardless of task complexity different regularisers elicit the same sensitivity to temperature scaling. When considering the hypothesis we posit for augmentation and weight decay navigating to tighter loss landscape regions these results align as they would be more receptive to landscape modifications caused by scaling.

C.3 Train and Test Loss Landscapes

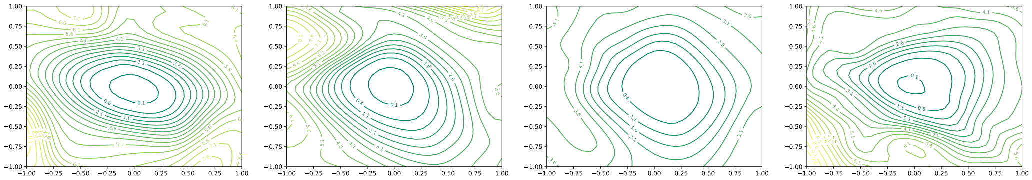
Here we see that the test loss landscape mimics that of the train loss landscape provided in Figures 4 and 5. We find that for both train and test the use of augmentation leads to a more complex loss landscape compared to the `baseline` and `dropout`. Considering the gains in ECE witnessed for these models, this could again be due to these explicit regularisers reducing the simplicity bias of the neural network [27], 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) `baseline+temp` (c) `augmentation` (d) `augmentation+temp`

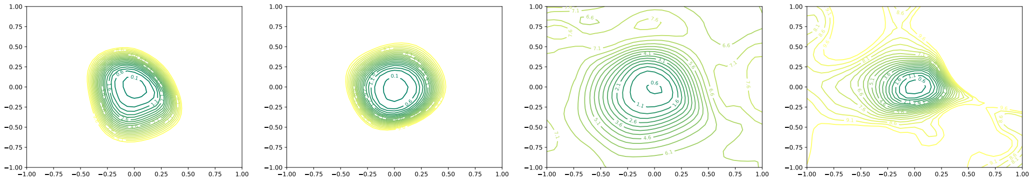
(e) `dropout` (f) `dropout+temp` (g) `weight decay` (h) `weight decay+temp`

Figure 3: Reliability plots for CIFAR-100. Model with median test accuracy among seeds used.



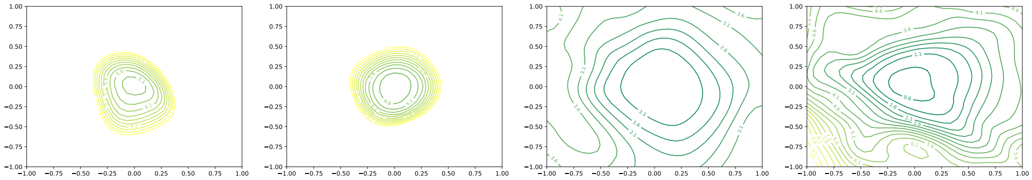
(a) `baseline` (b) `dropout` (c) `weight decay` (d) `augmentation`

Figure 4: Train contour loss landscapes for CIFAR-10 using 20% of training dataset for early stopped models.



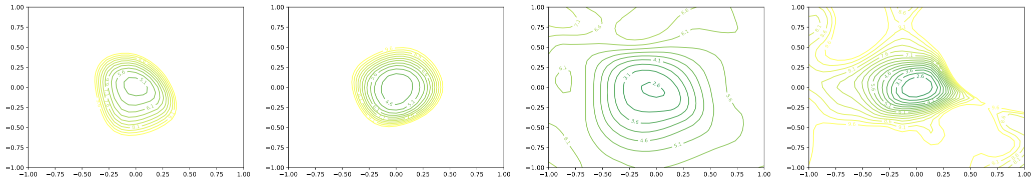
(a) `baseline` (b) `dropout` (c) `weight decay` (d) `augmentation`

Figure 5: Train loss landscapes for CIFAR-100 using 20% of training dataset for the early stopped models.



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

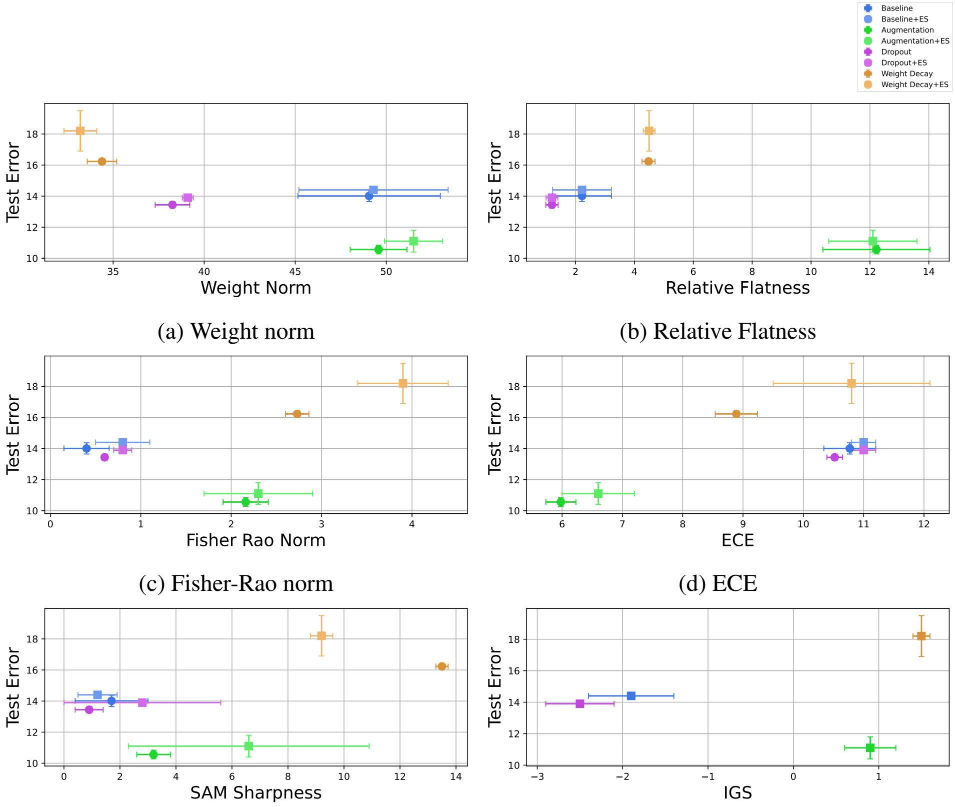
Figure 6: Test loss landscapes for CIFAR-100, loss landscapes generated 100% of the test dataset for the early stopped models.



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

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

C.4 Generalisation Correlation



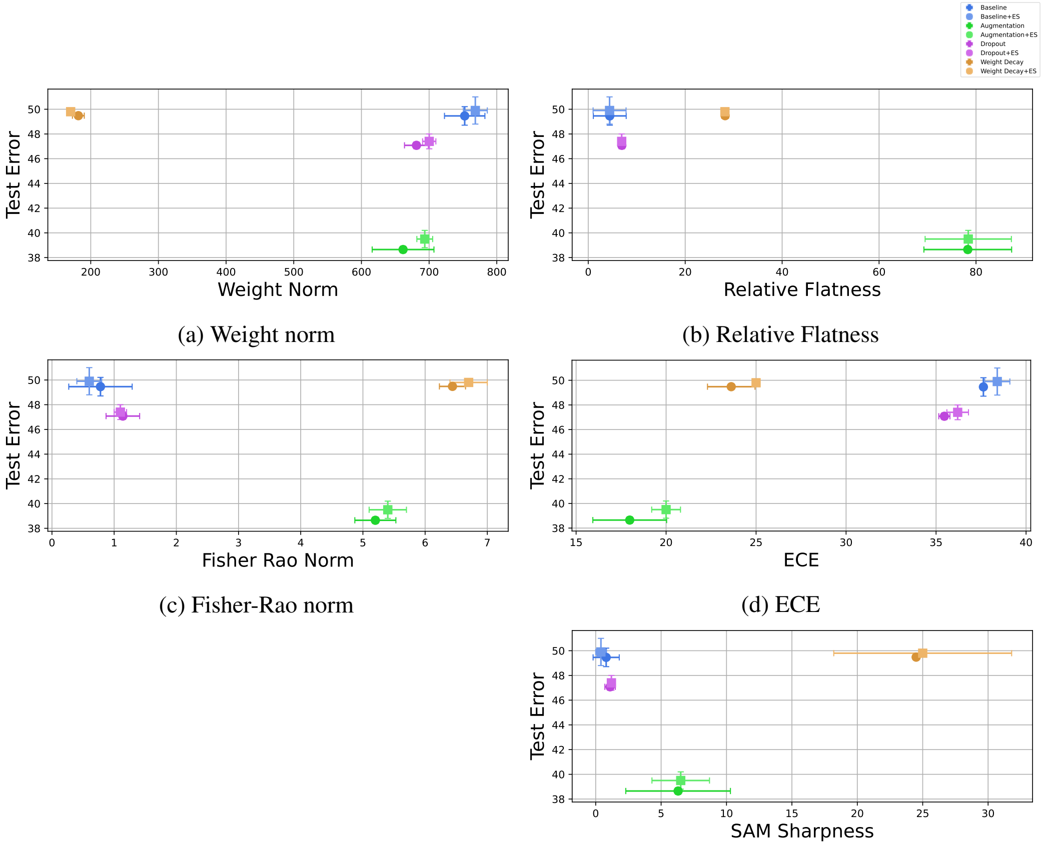
(a) Weight norm (b) Relative Flatness

(c) Fisher-Rao norm (d) ECE

(e) SAM sharpness (f) IGS

Figure 8: 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.

The generalisation correlation plots for the CIFAR datasets on the VGG-19 show no congruent relationship between generalisation and flatness across the sharpness metrics explored. Additionally, it appears that there is no particular requirement for flatness when considering models with comparatively low test error.



(a) Weight norm

(b) Relative Flatness

(c) Fisher-Rao norm

(d) ECE

(e) SAM Sharpness

Figure 9: Generalisation Correlation with CIFAR-100. Results here include both early stopping (□) and non-early stopping (○).

D ResNet-20 Results

The ResNet architecture is trained using the same setup as before, with no modifications. Please refer to Section B.2 for complete details.

The results collected on the ResNet architecture mostly mirror the results presented in the main body of the paper. The only disparities are that the effect of Dropout on the calibration error is different as it is reduced and, in line with the presented findings, the sharpness of the resulting model is increased. Also, there is an instance with weight decay on CIFAR-10 where the calibration error is high, but the landscape is also sharp; this provides more context as it shows that there are instances where sharpness and low calibration can emerge together but that this occurs only on CIFAR-10 and irregularly for the ResNet. This may suggest that the landscape can be too sharp, but with temperature scaling, the sharpness is reduced slightly, and a better calibration is recorded. Once again, the controls' impacts on the calibration of the model and the sharpness are accentuated when looking at the increased task complexity of CIFAR-100.

D.1 CIFAR-10

Table 5: Results for ResNet on CIFAR-10 with early stopping. Lower numbers indicate lower loss/error, more calibrated, or flatter.

Regularizer	Error (%)	ECE (%)	Weight norm	Fisher rao norm	Relative flatness	SAM sharpness ($\times10^{-2}$)	log IGS
`Baseline`	$17.37_{0.45}$	$13.53_{0.40}$	$155.24_{5.79}$	$0.01_{0.00}$	$1.24_{0.55}$	$0.00_{0.00}$	$-6.73_{-7.27}$
`+ temp scaling`	-	$10.04_{0.35}$	$155.24_{5.79}$	$0.17_{0.01}$	$24.31_{4.69}$	$0.01_{0.00}$	$-3.58_{-5.39}$
`Augmentation`	$12.32_{0.32}$	$7.39_{0.10}$	$167.98_{1.99}$	$2.64_{0.22}$	$304.30_{28.59}$	$0.86_{2.56}$	$1.34_{-1.52}$
`+ temp scaling`	-	$3.43_{0.05}$	$167.98_{1.99}$	$1.90_{0.11}$	$231.97_{13.94}$	$4.91_{0.80}$	$0.64_{-1.67}$
`Dropout`	$39.36_{1.84}$	$9.39_{0.57}$	$83.62_{0.95}$	$7.58_{0.32}$	$267.73_{4.62}$	$1.77_{0.59}$	$1.52_{-1.83}$
`+ temp scaling`	-	$1.72_{0.13}$	$83.62_{0.95}$	$5.20_{0.14}$	$195.33_{3.83}$	$1.62_{0.30}$	$1.24_{-2.75}$
`Weight decay`	$25.59_{0.47}$	$15.39_{0.43}$	$65.81_{0.71}$	$5.94_{0.40}$	$333.67_{4.43}$	$5.20_{0.86}$	$1.70_{-0.98}$
`+ temp scaling`	-	$7.42_{0.50}$	$65.81_{0.71}$	$3.54_{0.20}$	$234.90_{2.63}$	$3.46_{1.63}$	$1.02_{-1.85}$

Table 6: Results for ResNet on CIFAR-10 without early stopping. Lower numbers indicate lower loss/error, more calibrated, or flatter.

Regularizer	Error (%)	ECE (%)	Weight norm	Fisher rao norm	Relative flatness	SAM sharpness ($\times10^{-2}$)	log IGS
`Baseline`	$17.26_{0.49}$	$13.32_{0.49}$	$152.73_{7.16}$	$0.01_{0.00}$	$1.64_{0.98}$	$0.00_{0.00}$	$-6.72_{-7.27}$
`+ temp scaling`	-	$9.71_{0.53}$	$152.73_{7.16}$	$0.21_{0.03}$	$22.28_{3.51}$	$0.01_{0.00}$	$-3.60_{-5.51}$
`Augmentation`	$11.58_{0.29}$	$6.30_{0.70}$	$151.24_{15.29}$	$2.42_{0.17}$	$312.16_{28.66}$	$5.84_{1.91}$	$1.35_{-0.69}$
`+ temp scaling`	-	$2.44_{0.55}$	$151.24_{15.29}$	$1.88_{0.12}$	$231.03_{13.30}$	$2.61_{0.92}$	$0.66_{-1.73}$
`Dropout`	$34.56_{1.48}$	$2.16_{0.77}$	$74.89_{4.00}$	$5.81_{0.39}$	$270.88_{1.66}$	$2.75_{0.88}$	$1.52_{-1.79}$
`+ temp scaling`	-	$4.95_{0.17}$	$74.89_{4.00}$	$4.66_{0.19}$	$214.11_{3.22}$	$2.54_{0.84}$	$1.31_{-2.08}$
`Weight decay`	$19.53_{0.76}$	$10.46_{0.55}$	$66.60_{0.20}$	$2.81_{0.25}$	$346.84_{5.67}$	$4.16_{1.62}$	$1.70_{-1.14}$
`+ temp scaling`	-	$3.83_{0.57}$	$66.60_{0.20}$	$2.27_{0.12}$	$232.20_{3.27}$	$4.40_{0.85}$	$1.06_{-1.64}$

D.2 CIFAR-100

Table 7: Results for ResNet on CIFAR-100 for with early stopping. Lower numbers indicate lower loss/error, more calibrated, or flatter.

Regularizer	Error (%)	ECE (%)	Weight norm	Fisher rao norm	Relative flatness	SAM sharpness ($\times10^{-2}$)
`Baseline`	$49.90_{0.21}$	$35.92_{0.16}$	$1570.34_{46.33}$	$0.25_{0.02}$	$122.69_{10.86}$	$0.00_{0.00}$
`+ temp scaling`	-	$24.00_{0.13}$	$1570.34_{46.33}$	$0.98_{0.04}$	$1066.77_{94.27}$	$0.07_{0.02}$
`Augmentation`	$40.57_{0.12}$	$17.23_{0.24}$	$1497.57_{6.34}$	$6.31_{0.16}$	$5330.82_{44.94}$	$7.22_{0.88}$
`+ temp scaling`	-	$3.90_{0.24}$	$1497.57_{6.34}$	$5.32_{0.05}$	$3568.26_{40.12}$	$4.51_{1.64}$
`Dropout`	$73.84_{2.41}$	$6.51_{1.73}$	$656.01_{6.27}$	$10.62_{0.83}$	$1326.59_{34.86}$	$2.89_{2.47}$
`+ temp scaling`	-	$4.25_{0.72}$	$656.01_{6.27}$	$7.58_{0.14}$	$818.17_{66.65}$	$1.66_{0.57}$
`Weight decay`	$56.83_{1.19}$	$24.79_{0.41}$	$463.53_{2.87}$	$8.18_{0.70}$	$3307.63_{97.74}$	$2.72_{0.91}$
`+ temp scaling`	-	$6.96_{0.16}$	$463.53_{2.87}$	$6.13_{0.26}$	$2223.95_{80.46}$	$2.83_{0.43}$

Table 8: Results for ResNet on CIFAR-100 without early stopping. Lower numbers indicate lower loss/error, more calibrated, or flatter.

Regularizer	Error (%)	ECE (%)	Weight norm	Fisher rao norm	Relative flatness	SAM sharpness ($\times 10^{-2}$)
Baseline	$49.69_{0.16}$	$35.27_{0.30}$	$1544.91_{61.75}$	$0.33_{0.06}$	$871.98_{273.70}$	$0.00_{0.00}$
+ temp scaling	-	$23.03_{0.60}$	$1544.91_{61.75}$	$1.21_{0.20}$	$761.94_{24.76}$	$0.05_{0.01}$
Augmentation	$39.17_{0.10}$	$15.51_{0.42}$	$1468.30_{23.17}$	$5.84_{0.12}$	$5593.09_{99.42}$	$13.25_{2.26}$
+ temp scaling	-	$2.51_{0.48}$	$1468.30_{23.17}$	$5.30_{0.07}$	$3501.34_{37.05}$	$5.36_{1.81}$
Dropout	$72.08_{0.75}$	$4.75_{0.33}$	$611.74_{36.89}$	$9.80_{0.41}$	$1328.65_{45.74}$	$2.00_{1.05}$
+ temp scaling	-	$4.68_{0.23}$	$611.74_{36.89}$	$7.49_{0.13}$	$872.07_{24.14}$	$1.69_{1.02}$
Weight decay	$51.66_{0.56}$	$17.38_{0.35}$	$436.12_{3.26}$	$6.11_{0.33}$	$3659.94_{86.14}$	$3.73_{1.49}$
+ temp scaling	-	$1.59_{0.30}$	$436.12_{3.26}$	$5.87_{0.06}$	$2186.82_{92.02}$	$1.95_{0.20}$

E ViT Results

Pre-trained Vision Transformers, typically using Adam [28] or AdamW [29] optimisers, can perform sufficiently on the CIFAR datasets [11]. However, transformers struggle to perform well when trained on small datasets; it has been suggested that they struggle to incorporate locality in lower layers, typically improved via pre-training [30]. As a result, complex training setups employ weight decay, dropout and augmentation to improve accuracy on smaller datasets without pre-training. Provided that we use each of these explicit regularises as controls, our results do not represent competitive accuracy on the CIFAR datasets; we believe this is due to using SDG with a fixed learning rate and omitting cumulative explicit regularisers. Albeit, to ensure the continuity of the experimental setup and corresponding analysis, we only modified the training setup presented in Section B.2 to reduce the learning rate to 0.01 and provided 85 epochs of training. Our results further our understanding of expected calibration error and sharpness of loss landscapes by showing similar trends to the other explored model classes. It is of note that while the relationship holds that the use of explicit regularisers increases the sharpness of the landscape on the ViT architecture comparative to the baseline model, for weight norm and SAM sharpness these affects are less pronounced than on the previously presented architectures.

E.1 CIFAR-10

Table 9: Results for ViT on CIFAR-10 with early stopping. Lower numbers indicate lower loss/error, more calibrated, or flatter.

Regularizer	Error (%)	ECE	Weight norm	Fisher rao norm	Relative flatness	SAM sharpness ($\times 10^{-2}$)	log IGS
Baseline	$43.58_{0.47}$	$36.26_{0.52}$	$34.34_{1.01}$	$0.01_{0.00}$	$1.44_{0.02}$	$0.00_{0.00}$	$-7.69_{-11.54}$
+ temp scaling	-	$29.73_{0.51}$	$34.34_{1.01}$	$0.17_{0.00}$	$30.06_{1.54}$	$0.00_{0.00}$	$-3.59_{-6.88}$
Augmentation	$37.33_{0.44}$	$11.75_{1.01}$	$13.36_{0.46}$	$4.63_{0.13}$	$563.79_{28.09}$	$9.89_{4.71}$	$1.07_{-1.44}$
+ temp scaling	-	$1.37_{0.33}$	$13.36_{0.46}$	$3.73_{0.05}$	$393.03_{20.62}$	$4.68_{6.23}$	$1.00_{-2.55}$
Dropout	$49.32_{0.68}$	$13.46_{0.33}$	$10.31_{1.65}$	$8.37_{0.28}$	$147.68_{2.62}$	$0.07_{0.03}$	$1.40_{-1.09}$
+ temp scaling	-	$2.22_{0.57}$	$10.31_{1.65}$	$5.18_{0.10}$	$101.69_{0.77}$	$0.03_{0.01}$	$1.24_{-1.69}$
Weight decay	$45.47_{0.79}$	$32.39_{0.87}$	$42.68_{1.12}$	$2.02_{0.45}$	$392.46_{60.76}$	$0.06_{0.03}$	$0.72_{-0.44}$
+ temp scaling	-	$21.64_{0.89}$	$42.68_{1.12}$	$1.90_{0.13}$	$346.98_{25.15}$	$0.04_{0.02}$	$0.64_{-1.04}$

Table 10: Results for ViT on CIFAR-10 without early stopping. Lower numbers indicate lower loss/error, more calibrated, or flatter.

Regularizer	Error (%)	ECE	Weight norm	Fisher rao norm	Relative flatness	SAM sharpness ($\times 10^{-2}$)	log IGS
Baseline	$43.60_{0.42}$	$36.08_{0.75}$	$34.20_{1.18}$	$0.03_{0.04}$	$1.48_{0.01}$	$0.00_{0.00}$	$-7.70_{-11.45}$
+ temp scaling	-	$29.34_{1.02}$	$34.20_{1.18}$	$0.23_{0.09}$	$29.44_{1.68}$	$0.00_{0.00}$	$-3.62_{-6.65}$
Augmentation	$36.86_{0.40}$	$11.31_{1.24}$	$12.84_{0.75}$	$4.87_{0.12}$	$568.98_{27.12}$	$11.70_{6.55}$	$1.12_{-1.23}$
+ temp scaling	-	$1.35_{0.39}$	$12.84_{0.75}$	$3.83_{0.10}$	$389.99_{19.45}$	$6.95_{4.15}$	$0.99_{-2.15}$
Dropout	$49.04_{0.46}$	$12.95_{0.54}$	$10.11_{1.80}$	$8.50_{0.25}$	$149.39_{1.36}$	$0.08_{0.03}$	$1.44_{-0.95}$
+ temp scaling	-	$1.81_{0.44}$	$10.11_{1.80}$	$5.25_{0.11}$	$100.10_{2.22}$	$0.03_{0.01}$	$1.26_{-1.44}$
Weight decay	$44.15_{0.59}$	$32.00_{1.15}$	$38.54_{3.93}$	$1.66_{0.75}$	$397.58_{61.67}$	$0.03_{0.02}$	$0.79_{0.19}$
+ temp scaling	-	$21.77_{1.59}$	$38.54_{3.93}$	$1.68_{0.39}$	$342.32_{24.70}$	$0.07_{0.06}$	$0.62_{-1.14}$

E.2 CIFAR-100 Results

Table 11: Results for ViT on CIFAR-100 for with early stopping. Lower numbers indicate lower loss/error, more calibrated, or flatter.

Regularizer	Error (%)	ECE (%)	Weight norm	Fisher rao norm	Relative flatness	SAM sharpness ($\times 10^{-2}$)
Baseline	$69.52_{\pm 0.22}$	$44.42_{\pm 0.38}$	$309.62_{\pm 3.01}$	$0.14_{\pm 0.01}$	$43.92_{\pm 2.68}$	$0.00_{\pm 0.00}$
+ temp scaling	$69.52_{0.22}$	$25.36_{0.45}$	$309.62_{3.01}$	$1.00_{0.03}$	$763.68_{35.39}$	$0.00_{0.00}$
Augmentation	$64.62_{\pm 0.42}$	$36.41_{\pm 0.51}$	$603.70_{\pm 4.31}$	$2.38_{\pm 0.11}$	$8762.55_{\pm 497.09}$	$4.61_{\pm 3.35}$
+ temp scaling	$64.62_{0.42}$	$16.98_{0.58}$	$603.70_{4.31}$	$4.40_{0.07}$	$8818.26_{335.74}$	$2.07_{2.94}$
Dropout	$73.48_{\pm 0.12}$	$7.21_{\pm 0.51}$	$213.88_{\pm 0.44}$	$9.09_{\pm 0.21}$	$3347.07_{\pm 50.70}$	$0.05_{\pm 0.02}$
+ temp scaling	$73.48_{0.12}$	$4.05_{0.20}$	$213.88_{0.44}$	$7.11_{0.02}$	$2023.82_{11.40}$	$0.07_{0.02}$
Weight decay	$70.90_{\pm 1.41}$	$36.53_{\pm 0.65}$	$250.33_{\pm 12.59}$	$1.21_{\pm 1.06}$	$1353.63_{\pm 1474.40}$	$0.03_{\pm 0.03}$
+ temp scaling	$70.90_{1.41}$	$14.92_{0.71}$	$250.33_{12.59}$	$4.28_{1.31}$	$2060.67_{900.54}$	$0.04_{0.03}$

Table 12: Results for ViT on CIFAR-100 without early stopping. Lower numbers indicate lower loss/error, more calibrated, or flatter.

Regularizer	Error (%)	ECE (%)	Weight norm	Fisher rao norm	Relative flatness	SAM sharpness ($\times 10^{-2}$)
Baseline	$69.46_{\pm 0.22}$	$41.98_{\pm 0.20}$	$302.15_{\pm 3.72}$	$0.32_{\pm 0.03}$	$44.14_{\pm 2.69}$	$0.00_{\pm 0.00}$
+ temp scaling	$69.46_{0.22}$	$22.17_{0.22}$	$302.15_{3.72}$	$1.61_{0.04}$	$732.39_{37.02}$	$0.01_{0.00}$
Augmentation	$64.68_{\pm 0.47}$	$35.98_{\pm 0.34}$	$596.31_{\pm 12.92}$	$2.48_{\pm 0.10}$	$8785.91_{\pm 548.93}$	$3.28_{\pm 0.17}$
+ temp scaling	$64.68_{0.47}$	$16.38_{0.42}$	$596.31_{12.92}$	$4.57_{0.17}$	$8800.96_{341.38}$	$4.54_{2.40}$
Dropout	$73.00_{\pm 0.31}$	$7.40_{\pm 0.37}$	$210.46_{\pm 0.53}$	$9.16_{\pm 0.06}$	$3340.97_{\pm 50.33}$	$0.03_{\pm 0.01}$
+ temp scaling	$73.00_{0.31}$	$4.21_{0.18}$	$210.46_{0.53}$	$7.12_{0.03}$	$2008.89_{32.04}$	$0.04_{0.01}$
Weight decay	$69.29_{\pm 0.44}$	$37.19_{\pm 0.88}$	$256.23_{\pm 8.13}$	$0.43_{\pm 0.06}$	$1354.42_{\pm 1475.19}$	$0.00_{\pm 0.00}$
+ temp scaling	$69.29_{0.44}$	$16.10_{0.94}$	$256.23_{8.13}$	$2.89_{0.21}$	$2063.03_{888.47}$	$0.02_{0.00}$