A Function-Centric Perspective on Flat and Sharp Minima

Anonymous authors
Paper under double-blind review

Abstract

Flat minima are strongly associated with improved generalisation in deep neural networks. However, this connection has proven nuanced in recent studies, with both theoretical counterexamples and empirical exceptions emerging in the literature. In this paper, we revisit the role of sharpness in model performance and examine sharpness as a function-dependent property rather than as a direct indicator of poor generalisation. We conduct extensive empirical studies ranging from single-objective optimisation, synthetic non-linear binary classification tasks, to modern image classification tasks. In single-objective optimisation, we show that flatness and sharpness are relative to the function being learned: equally optimal solutions can exhibit markedly different local geometry. In synthetic non-linear binary classification tasks, we show that increasing decision-boundary tightness can increase sharpness even when models generalise perfectly, suggesting that sharpness is not reducible to memorisation alone. Finally, in large-scale experiments, we find that sharper minima often emerge when models are regularised (e.g., via weight decay, data augmentation, or SAM), and can coincide with better generalisation, calibration, robustness, and functional consistency. Our findings are consistent with a function-centric interpretation of minima geometry: flatness should not be treated as a universal proxy for generalisation, and sharper minima can reflect desirable inductive biases associated with improved predictive behaviour, calling for a function-centric reappraisal of minima geometry.

1 Introduction

Neural network architectures with different implicit biases are known to exhibit distinct geometric properties, with flatness often associated with improved generalisation performance via reduced generalisation gaps (Li et al., 2018; Han et al., 2025; Petzka et al., 2021; Lee & Yoon, 2025; Cha et al., 2021; Zhao et al., 2022). This perspective is commonly associated with the idea that flat minima correspond to wide error margins and increased robustness, in line with Occam's Razor and Minimum Description Length (MDL) arguments (Hochreiter & Schmidhuber, 1994). Empirical and theoretical studies have reinforced this perspective (Kaddour et al., 2022; Foret et al., 2021; Petzka et al., 2021), supporting the view that flatter solutions lead to better generalisation. Flatness has also been associated with benefits such as improved representation transfer (Liu et al., 2023) and the effects of architectural choices such as residual connections (Li et al., 2018). Notably, optimisation methods such as Sharpness Aware Minimization (SAM) (Foret et al., 2021), which improve generalisation in the vision domain, were explicitly motivated by the idea of biasing training toward flatter minima. Yet generalisation is only one dimension of model quality.

At the same time, this connection has proven more nuanced than these accounts suggest. Dinh et al. (2017) showed that flat minima, under Hessian-based definitions, can be arbitrarily sharpened through reparameterisation, without changing the model's learned function or generalisation behaviour. This motivated the development of reparameterisation-invariant sharpness metrics, such as the Fisher-Rao norm (Liang et al., 2019) and Relative-Flatness (Petzka et al., 2021), which have been used to recover the association between flatness and generalisation. However, these developments leave open a central question: does sharpness, under reparametrisation invariant metrics, indicate poor generalisation, or does its interpretation depend on the function learned by the model and the training conditions under which that function is obtained?

This question becomes especially important once model quality is assessed beyond test accuracy alone. In practical settings, reliability-related properties such as calibration (Guo et al., 2017), robustness to common corruptions (Hendrycks & Dietterich, 2019), and functional consistency across runs (Wang et al., 2024) are also important for reliable deployment. However, their relationship to minima geometry remains underexplored. In particular, it is unclear whether flatter solutions consistently support better reliability, or whether higher-performing models on these dimensions may occupy sharper regions of the loss landscape.

In this paper, we revisit minima geometry through a function-centric lens. Specifically, we use the term *function complexity* to refer to properties of the learned function such as representation constraints, decision-boundary structure, or degree of representational precision, that may be reflected in minima geometry. Our hypothesis is that the geometry of a solution reflects the complexity of the learned function, rather than directly determining performance. Our central claim is not that sharpness is universally beneficial, nor that flatness is irrelevant. Rather, we argue that sharpness should be interpreted relative to the learned function and the model's inductive biases. From this perspective, sharper minima need not indicate overfitting or memorisation, but instead may reflect more structured or better-regularised solutions, particularly in high-dimensional learning tasks.

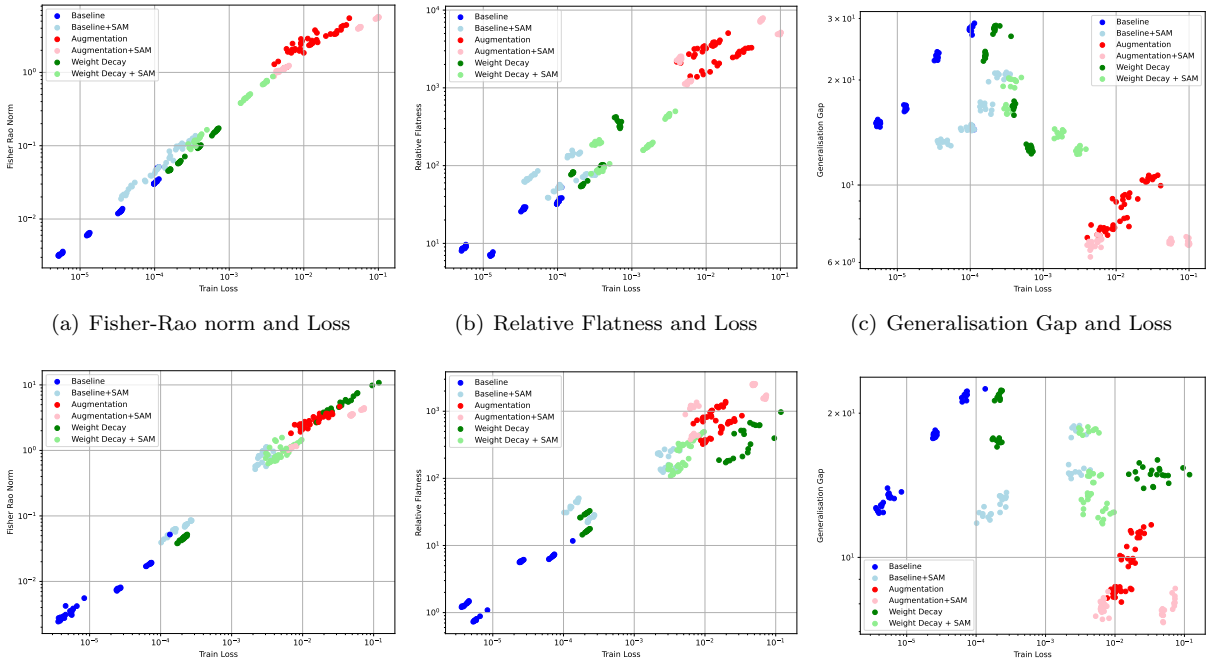
Our function-centric framing is motivated by Occam's Razor and the MDL perspective commonly used in the flatness literature, where flatter minima are often interpreted as simpler, lower-complexity solutions that require less precision to represent (Hochreiter & Schmidhuber, 1994). From this perspective, sharper minima can be viewed as corresponding to learned solutions that are more tightly constrained or that require more precise representation, and are therefore related to more complex decision-boundary structure in an Occam/MDL sense. We use function complexity as an interpretive lens for understanding how minima geometry relates to the learned function, inductive bias, and training regularisation.

We first study this in controlled synthetic settings. We begin with seven standard single-objective optimisation problems, where the global minima are known and their local geometry can be compared directly. These experiments illustrate that optimal solutions can be either sharp or flat, depending on the intrinsic complexity of the objective being fitted: some functions (e.g., Sphere) have flat global minima, while others (e.g., Rosenbrock) have inherently sharp global minima. This suggests that the geometry of the solution space is tied to function complexity, not optimality alone. We then turn to non-linear binary classification tasks. Such settings have been used to show that sharp minima have an implicit relationship with memorised solutions, arguing that sharp minima are an indicator of poorly generalising models (Huang et al., 2020). We revisit this relationship by explicitly manipulating decision-boundary tightness and showing that models can generalise perfectly while still exhibiting increased sharpness under tighter decision boundaries. This allows us to partially decouple sharpness from memorisation alone and to show that the way in which data are fitted can shape the geometry of the solution. Additionally, we show that learning tight decision boundaries requires increased computational budgets (with all other training conditions fixed), indicating that such functions may be harder for neural networks to fit.

We then scale our analysis to high-dimensional problems, and use the CIFAR (Krizhevsky & Hinton, 2009) and Tiny ImageNet (Le & Yang, 2015) data sets to train the ResNet (He et al., 2016), VGG (Simonyan & Zisserman, 2015) and ViT (Dosovitskiy et al., 2021) architectures under multiple matched-seed training controls. We compare baseline models to models trained with commonly used regularisation techniques (weight decay, data augmentation and SAM). Across these settings, we evaluate reparameterisation-invariant sharpness metrics, together with generalisation performance and reliability-related measures: expected calibration error, average-case perturbation robustness, and functional/prediction agreement.

Our findings provide empirical support for a function-centric interpretation of sharpness: models trained with regularisation often converge to sharper minima while also outperforming flatter, unregularised counterparts across reliability and generalisation metrics (Figure 1). This suggests that regularisation can give rise to learned solutions of greater functional complexity, yielding sharper yet more effective minima. At the same time, the relationship is not monotone or universal: distinct trends emerge across control conditions, with some favouring flatter solutions across seeds and others not, consistent with Simpson's paradox (Simpson, 1951). Minima geometry therefore requires a nuanced interpretation than a universal preference for flatness: while models without regularisation (blue in Figure 1) have a strong correlation with small generalisation

gaps and flatness, models using augmentation and SAM (pink in Figure 1) show no such preference. While SAM and related methods were originally motivated by the goal of encouraging flatness, we show that their empirical benefits frequently arise alongside increased sharpness. Taken together, these results call into question the assumption that flatness is inherently beneficial and support a reappraisal of sharpness through the lens of function complexity. We further show that these findings are robust to training hyperparameters such as learning rate and batch size.



(a) Fisher-Rao norm and Loss (b) Relative Flatness and Loss (c) Generalisation Gap and Loss

(d) Fisher-Rao norm and Loss (e) Relative Flatness and Loss (f) Generalisation Gap and Loss

Figure 1: Scatter plots of 240 converged minima for ResNet-18 (top) and VGG19 (bottom) on CIFAR-10 across batch size $128, 256$ and learning rate $10^{-3}, 10^{-2}$: (a) Fisher–Rao norm vs. train loss, (b) Relative Flatness vs. train loss, and (c) generalisation gap vs. train loss (log scale). Full results for ResNet and VGG in Appendix D.1 and E.1, respectively.

Concretely, we make the following contributions:

- We present a function-centric interpretation of sharpness, in which the geometry of minima reflects the complexity of the learned function rather than serving as a universal proxy for (poor) generalisation.
- We provide a broad empirical reassessment of minima geometry across controlled optimisation problems, synthetic classification tasks, and high-dimensional deep learning tasks, and find that sharper minima can coincide with statistically significant improvements in generalisation, calibration, robustness, and functional consistency.
- We find across architectures, data sets, hyperparameters and matched-seed training controls that widely used regularisation techniques (e.g., weight decay, augmentation and SAM) often induce statistically significant sharper minima than unregularised baselines, calling into question the assumption that regularisation generally promotes flatter solutions.
- We demonstrate that sharpness cannot be meaningfully interpreted across architectures or tasks without accounting for function complexity and inductive bias, cautioning against overgeneralised geometric claims.

Our findings can be summarised as follows:

1. Sharpness varies across global minima in single-objective optimisation and with decision-boundary proximity in non-linear binary classification tasks, reflecting function complexity rather than solution quality alone. In high-dimensional problems, regularised models typically converge to sharper minima, yet often achieve better generalisation, calibration, robustness, and functional consistency than flatter unregularised baselines.
2. We reconcile SAM's local robustness objective with increased global sharpness, showing that its empirical benefits need not arise from flatter minima and are consistent with a function-centric view of geometry.
3. Our results support a function-centric view of sharpness: solution geometry is shaped by the complexity of the learned function and the model's inductive biases. Crucially, we find no clear Goldilocks zone for sharpness across architectures and data sets, as the interpretation of sharpness dependents on the function being learned, an architectures implicit bias and the training conditions under which the model is optimised.

2 Related Work

Hochreiter & Schmidhuber (1997) presented seminal empirical evidence that neural networks can adhere to Occam's Razor. They showed that a flat-minima search algorithm using a second-order Hessian approximation could yield the smallest generalisation gap on two-class classification problems. As a result of the observed empirical relationship between flatness and generalisation, the antipodal sharp minima were often viewed as undesirable. The importance of flatness in more complex learning tasks was later reaffirmed by Li et al. (2018) who introduced landscape visualisation to study the geometry of deep networks. They argued that skip connections help prevent explosions of non-convexity, reducing the occurrence of chaotic plateaus often associated with sharp minima. Building on this, Sharpness Aware Minimisation (Foret et al., 2021) was proposed as an optimisation method (motivated in part by Hochreiter & Schmidhuber (1997)) that explicitly aims to reduce sharpness in the loss landscape. SAM has yielded strong empirical performance gains over traditional optimisation methods (Foret et al., 2021). However, some recent work has challenged this interpretation, arguing that SAM does not necessarily find flatter minima (Wen et al., 2023). At the same time, sharp minima are still commonly associated in the literature with memorisation and poor generalisation Huang et al. (2020).

The necessity of flatness for generalisation has also been questioned more fundamentally. Notably, Dinh et al. (2017) showed that sharpness can be arbitrarily increased through reparameterisation without affecting the learned function or its generalisation, casting doubt on the intrinsic value of flatness under naive Hessian-based definitions. In response, reparameterisation-invariant sharpness metrics were developed (Petzka et al., 2021) and have since been used to reaffirm the association between flatness and generalisation. Together, these developments highlight a conceptual tension: although sharpness was shown to be manipulable through reparameterisation, reparametrisation invariant metrics have reaffirmed a preference for flatness and, as such, flatness is still widely used as a desirable indicator of generalisation.

More recent work has continued to recover a relationship between flat minima and strong generalisation in neural network optimisation. Stutz et al. (2021) conducted a study which related average-case flatness to robustness under adversarial perturbation. Cha et al. (2021) introduced Stochastic Weight Averaging Densely (SWAD), which they argue improves domain generalisation by seeking flatter minima, supported by loss-landscape visualisations. Zhao et al. (2022) built on sharpness-aware minimisation through gradient-norm constraints designed to find flatter minima more efficiently and improve generalisation. In reinforcement learning, Lee & Yoon (2025) related the robustness of trained agents to flatter minima at the end of optimisation. Finally, flatness has also been explored as a necessary component for generalisation in grokking, where a phase of extreme memorisation is followed by generalisation (Han et al., 2025). In that setting, the authors showed that preventing Relative Flatness minimisation during training inhibited the escape from the memorisation phase of grokking. Taken together, there is a strong empirical and theoretical basis for favouring flat minima in deep learning.

In contrast to the view that sharpness necessarily signals poor generalisation, we find that sharper solutions can emerge precisely when models generalise better and exhibit improved reliability-related properties. We propose that sharpness should be interpreted through the complexity of the learned function and the model's inductive bias, rather than through a universal preference for flatness.

3 Sharpness, Generalisation and Reliability-Related Evaluations

In this section we define the sharpness metrics used in our empirical experiments (Section 3.1), and the reliability-related metrics used to assess model behaviour beyond test accuracy (Section 3.2). Our aim is to examine whether minima geometry aligns not only with generalisation, but also with broader properties relevant to reliable deployment.

3.1 Sharpness Metrics

We employ three established measures of sharpness from the literature, namely Fisher-Rao norm (Liang et al., 2019), Relative Flatness (Petzka et al., 2021), and average-case SAM-Sharpness (Foret et al., 2021). Hessian-based metrics, such as the top eigenvalue or the trace of the Hessian, were shown not to be reparameterisation invariant as they can be manipulated through linear reparameterisations. These reparameterisations do not change the function of the model but can make a minimum appear sharper Dinh et al. (2017), thereby undermining any direct relationship between such measures and generalisation. For this reason, we focus in particular on Fisher-Rao norm (Liang et al., 2019) and Relative Flatness (Petzka et al., 2021), both of which are reparametrisation invariant. This ensures that our study, and the conclusions drawn from it, are robust to such transformations.

Below we provide the formal definitions of the three sharpness metrics used throughout this work: Fisher-Rao norm, SAM-Sharpness and Relative Flatness. Information Geometric Sharpness (IGS) (Jang et al., 2022) is also a relevant candidate metric; however, we omitted it from this study because its computation is not feasible for the network sizes and data sets considered here. For Fisher-Rao and Relative Flatness we use the implementation provided by Petzka et al. (2021).[1]

Fisher-Rao: Fisher-Rao norm (Liang et al., 2019) uses information geometry to provide a norm-based measure of model complexity. In architectures satisfying the assumptions under which it is derived, it also provides a reparametrisation-invariant measure of minima sharpness, as verified by Petzka et al. (2021). In this paper, we use Fisher-Rao norm as an empirical geometry measure following Petzka et al. (2021). In line with Petzka et al. (2021), we use the analytical expression for cross-entropy loss from the appendix of Liang et al. (2019), reproduced in Equation equation 1, as the empirical estimator throughout this work.

In our notation, L denotes the number of hidden layers in the sense of Liang et al. (2019), so that $L+1$ is the number of weight matrices in the predictor. For practical implementation, we compute this quantity by counting the number of Linear, Conv1d, Conv2d, Conv3d, and Embedding layers in each network, and then matching this count to the $L+1$ factor in Equation equation 1. Under this convention, the corresponding counts are 21 for ResNet18, 18 for VGG19, and 26 for ViT.

$$\mathrm{FR}_{\mathrm{norm}} = \sqrt{(L+1)^2 \cdot \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial \mathcal{L}_i}{\partial \theta} \cdot \theta \right)} \tag{1}$$

SAM-Sharpness: When using the SAM optimiser (Foret et al., 2021) the ρ parameter determines the radius at which an adversarial weight-space perturbation is taken. A small ρ value indicates a reduced radius from the existing weight space. To study the loss regions in the weight spaces around the minima at the end of training we define SAM-sharpness as the average loss difference across 100 perturbed parameter locations at distance 0.005ρ from the original model, following Foret et al. (2021). This provides a local measure of how rapidly the loss changes under small parameter perturbations. A sharp minimum would therefore be

[1] Code base for Fisher-Rao norm and Relative Flatness from Petzka et al. (2021): `https://github.com/kampmichael/RelativeFlatnessAndGeneralization/blob/main/CorrelationFlatnessGeneralization/measure_comparison.py`

expected to exhibit a larger increase in loss from the baseline minimum under a new adversarial weight-space sampling, whereas for a flat minimum the loss would be expected to remain approximately constant.

$$S(\theta) = \frac{1}{N} \sum_{n=1}^{N} \left| \frac{\mathcal{L}(\theta + \Delta\theta_n) - \mathcal{L}(\theta)}{\rho} \right|. \tag{2}$$

Relative Flatness: Petzka et al. (2021) define Relative Flatness as a reparameterisation-invariant sharpness measure, and show that it has one of the strongest correlations with low generalisation gap among the measures they consider. In our setting, Relative Flatness sharpness is computed between the feature extraction layer and the classifier, and is computationally expensive because it requires the trace of Hessian blocks over the output matrices. We use the formulation shown in Equation equation 3.

$$\kappa^{\phi}_{\text{Trace}}(\mathbf{w}) \;:=\; \sum_{s,s'=1}^{d} \langle \mathbf{w}_s, \mathbf{w}_{s'} \rangle \cdot \text{Trace}(H_{s,s'}(\mathbf{w}, \phi(S))) \tag{3}$$

where $\mathbf{w}_s$ denotes the s-th row of $\mathbf{w}$, $\langle \cdot, \cdot \rangle$ is the scalar product, and $H_{s,s'}(\mathbf{w}, \phi(S))$ is the Hessian of the empirical loss with respect to $\mathbf{w}_s$ and $\mathbf{w}_{s'}$, evaluated at $\phi(S)$ (Han et al., 2025). We interpret Relative Flatness as an empirical geometry measure that is informative about the learned solution.

In line with the function-centric perspective of this paper, we interpret these sharpness measures, particularly Fisher-Rao norm, as capturing aspects of the geometry of the learned function, rather than treating them as universal scalar proxies for generalisation.

3.2 Reliability-Related Metrics

Beyond the generalisation properties of a neural network, there are several additional quantities of interest that are relevant to reliable and safe deployment. In this section, we provide the reliability-related metrics used in this study to examine how minima geometry relates to broader model behaviour.

Calibration: Calibration measures how well a model's predicted confidence aligns with its empirical likelihood of correctness. Deep neural networks, including ResNets, have been shown to be systematically overconfident (Guo et al., 2017), reducing trust in their predictions. We measure calibration using Expected Calibration Error (ECE) (Guo et al., 2017), where lower values indicate better calibration and higher trustworthiness. To compute Expected Calibration Error (ECE), we use the Lighting AI Pytorch Metrics implementation of Multiclass Calibration Error following Kumar et al. (2019).[2]

$$\text{ECE} = \sum_{i=1}^{B} b_i \, |p_i - c_i| \tag{4}$$

where p_i denotes the accuracy in bin i, and c_i the average confidence in that bin under uniform binning (Kumar et al., 2019).[2]

Functional Consistency: Functional diversity reflects how similar neural networks are in function or representation space (Wang et al., 2024; Mason-Williams et al., 2024; Mason-Williams, 2024). Prior work has linked diversity in function space to improved ensemble performance (Fort et al., 2020; Lu et al., 2024), while other work argues that representation convergence can also benefit ensembles (Wang et al., 2024). In this paper, we focus on the consistency of the learned functions across runs by measuring prediction disagreement on the test set, which captures how often two models disagree on their top-1 outputs. Lower disagreement implies greater agreement on individual predictions given the same training data, and we interpret this as a form of functional consistency under training stochasticity.

[2]Calibration error documentation from Lighting AI: `https://lightning.ai/docs/torchmetrics/stable/classification/calibration_error.html`

Concretely, we use **Prediction Disagreement**, defined in Equation equation 5 following Fort et al. (2020), where $f(x;\theta)$ denotes the top-1 predicted class for sample x under parameters θ. Lower Prediction Disagreement indicates that two models agree more frequently on their top-1 predictions.

$$\frac{1}{N}\sum_{n=1}^{N}\big[\, f(x_n;\theta_1) \neq f(x_n;\theta_2)\,\big]. \tag{5}$$

Corruption Robustness: Robustness assesses how well a model performs under distribution shift or input perturbations, which is important for reliable deployment. We evaluate robustness on CIFAR10-C , CIFAR100-C and Tiny ImageNet-C (Hendrycks & Dietterich, 2019), which include common corruptions such as impulse noise, JPEG compression, and contrast distortions. Performance is quantified via mean corruption accuracy, where higher values indicate greater robustness. These perturbations represent average-case corruption robustness, rather than worst-case adversarial robustness (Hendrycks & Dietterich, 2019). We employ these data sets to study how geometric properties relate to corruption robustness. Each corruption is provided at five levels of severity. The metric we use is **Corruption Accuracy**, defined as the average accuracy of a classifier f on the corrupted test set $\mathcal{D}_{\text{corruption}}$ across all corruption types and severity levels (Hendrycks & Dietterich, 2019).

$$\text{Corruption Accuracy} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{S}\sum_{s=1}^{S}\frac{1}{N_{c,s}}\sum_{n=1}^{N_{c,s}}\mathbf{1}(f(x_n^{(c,s)};\theta) = y_n). \tag{6}$$

where C is number of corruption types, S is the number of severity levels, $N_{c,s}$ the number of samples under corruption c and severity s, $f(x_n;\theta)$ the top-1 prediction for input x_n under parameters θ, and y_n the corresponding label.

4 Single-Objective Optimisation

We posit that the sharpness reached by a model depends on the geometric properties of the function it is trained to approximate, given that neural networks are universal function approximates. To illustrate that loss-landscape geometry can be tied to solution complexity / the function being fitted, we begin with a toy setting: single-objective optimisation. Toy settings have previously been used to study geometric properties of neural networks; for example Huang et al. (2020) used the Swiss Roll data set to explore generalisation and flat minima.

Consider Himmelblau's function in Equation equation 7 (visualised in Figure 2). It has four global minima whose local geometry differs markedly (Table 1), yet each achieves zero loss. Thus, no minimum is intrinsically preferable from an optimisation-objective standpoint. Under flatness-centric views, flatter minima would be deemed preferable; however, any network that represents the target function can plausibly converge to any of these minima. Flatness is therefore not a necessary criterion for optimality in this setting.

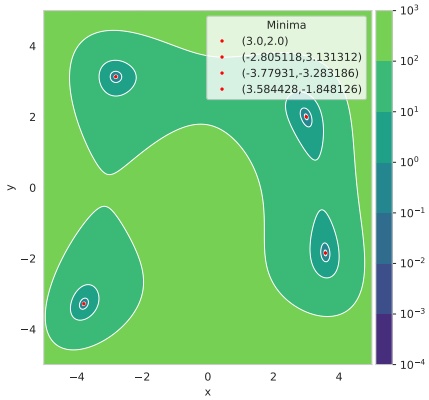


Figure 2: Himmelblau's function landscape.

$$f(\boldsymbol{x},\boldsymbol{y}) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2 \tag{7}$$

Global Minimum	Condition Number	Hessian Trace	Hessian determinant	Max Eigenvalue
(3.0, 2.0)	3.200	108.000	2116.000	82.284
(-2.805118, 3.131312)	1.242	145.39	5222.890	80.550
(-3.77931, -3.283186)	1.892	204.500	9460.560	133.786
(3.584428, -1.848126)	3.674	134.110	3024.540	105.419

Table 1: Local geometric properties at the four global minima of Himmelblau's function.

Moving beyond this example, we examine a set of single-objective problems with a single global minimum. Figure 3 visualises six such functions (definitions are provided in Appendix A). Each exhibits a distinct landscape, implying different local curvature at its global minimum. Table 2 reports sharpness statistics at the global minimum. For instance, the Sphere function is the flattest across metrics, whereas functions with more intricate landscapes (e.g., Rosenbrock, Beale, Booth) have sharper optima. Accurately representing these objectives therefore entails reaching minima with geometry commensurate with the complexity of the target function.

We next fit an MLP to each objective using the same initialisation and average over ten models. As shown in Figure 4, the sharpness of the local minima encountered during training reflects the sharpness of the global optimum: under a fixed training budget, model sharpness, training loss, and generalisation gap vary systematically with the complexity/structure of the target function (see Figure 3). Although absolute generalisation gaps differ across objectives, they exhibit similar relative reductions over training. Appendix A.2 further shows that matching the final loss across functions still yields different sharpness levels, as expected from their intrinsic geometry.

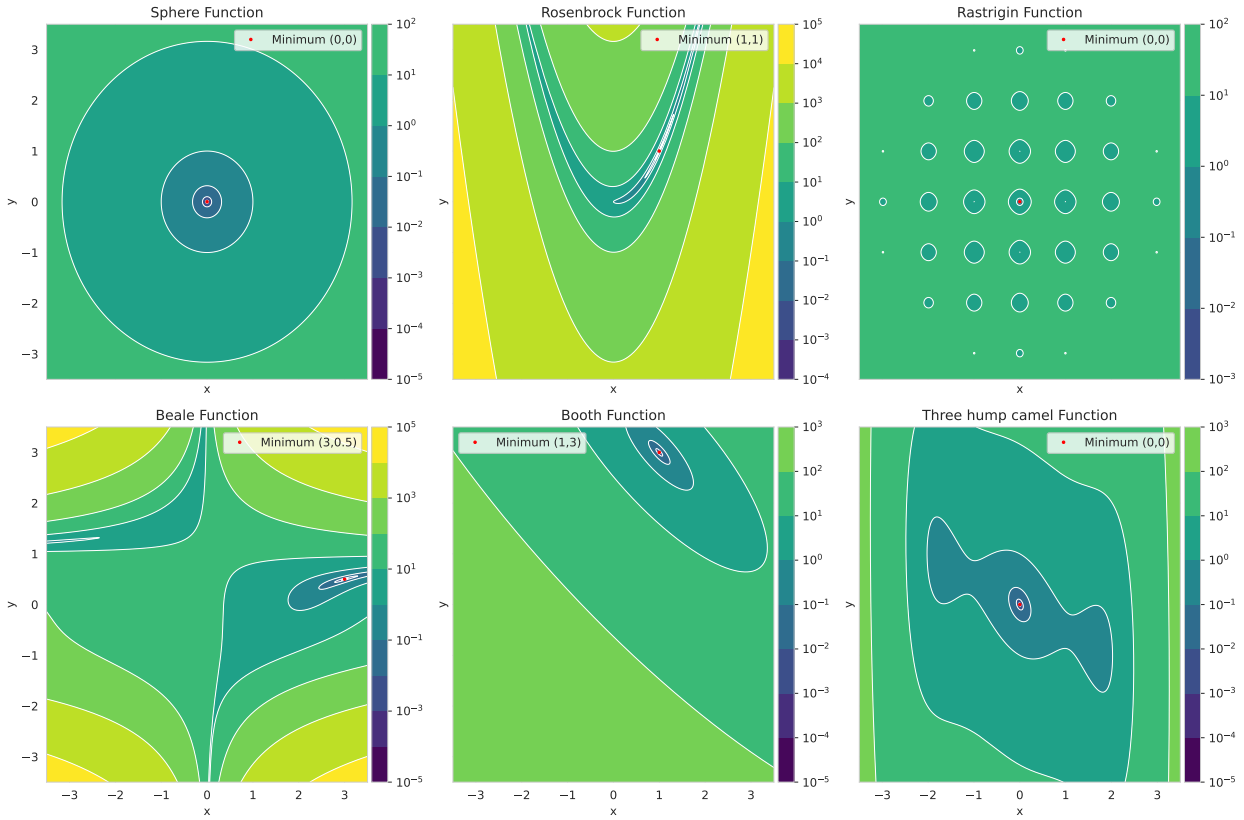


Figure 3: Landscapes for six single-objective functions.

From this perspective, flatness is desirable only when demanded by the target function (e.g., Sphere, Booth and Three hump camel). Seeking flat solutions for intrinsically sharper objectives (e.g., Rosenbrock) may therefore be suboptimal: their structure is consistent with the need for tighter solution geometry and thus sharper minima. It is important to note that this section is intended only as an illustration of how neural network minima geometry can relate to function complexity and the function being fitted. In the regression setting considered here, this analysis does not extend directly to measures such as Relative Flatness, which require locally constant labels Petzka et al. (2021). In the following section, we show how the same broader perspective extends to reparameterisation-invariant measures, including Fisher-Rao norm and Relative Flatness, in classification settings where the locally constant label condition holds.

Function	Condition Number	Hessian Trace	Hessian determinant	Max Eigenvalue
Sphere	1.000	4.000	4.000	2.000
Rosenbrock	2508.010	1002.000	400.000	1001.600
Rastrigin	1.000	793.568	157438.000	396.784
Beale	162.473	49.281	14.766	48.980
Booth	9.000	20.000	36.000	18.000
Three hump camel	2.784	6.000	7.000	4.414

Table 2: Sharpness at the global minimum for six single-objective optimisation functions.

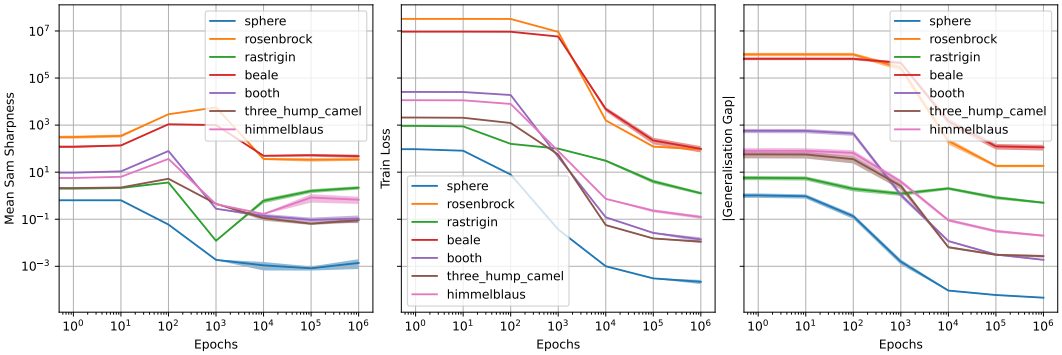


Figure 4: Training an MLP on single-objective problems over epochs: mean sharpness, training loss, and absolute generalisation gap (averaged over 10 runs).

Furthermore, in Section 7.3 we show that arbitrarily increasing the complexity of the training data in a classification setting results in models reaching sharper minima under Fisher-Rao norm and Relative Flatness. The findings from that experiment further support the relationship between function complexity and minima geometry illustrated in this toy setting.

5 Decision Boundaries, Memorisation and Sharpness

In this section, we explore and extend the analysis conducted by Huang et al. (2020), who studied sharp and flat minima through the lens of memorisation and generalisation. Huang et al. (2020) showed that neural networks that reach sharp minima, either through "bad" minima or through training on memorised data, exhibit sharper landscape visualisations than models trained on standard data. As such, they introduced the perspective that the geometric properties of a model relate to the decision boundaries of solution, with tighter decision boundaries corresponding to models that have memorised training points.

We first reproduce the results of Huang et al. (2020) on the make_circles data set, which is a non-linear binary classification task with 100 data points split evenly between 2 classes (Pedregosa et al., 2011).[3] We train a 3-layer MLP to classify the two classes. In the standard setting with no randomisation (random label probability 0.0), the model is trained with SGD using a learning rate of 0.01, momentum 0.9, batch size 5, for 300 epochs using cross-entropy loss. To show that increasing the randomisation of data points results in sharper minima, we scale the random label probability from 0.0 to 0.9. In the randomised setting, the MLP is trained for 20,000 epochs, with all other hyperparameters unchanged to allow convergence to 100% accuracy on the training data set. For both the standard ±0.0 and randomised conditions (0.1–0.9), we initialise 100 models across random seeds 0–99 and use the same training-data order for each model. The increased training budget / epochs required for the randomised data already suggests that such training data are harder for the MLP to fit.

[3]Sklearn make_circles data set: `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_circles.html`

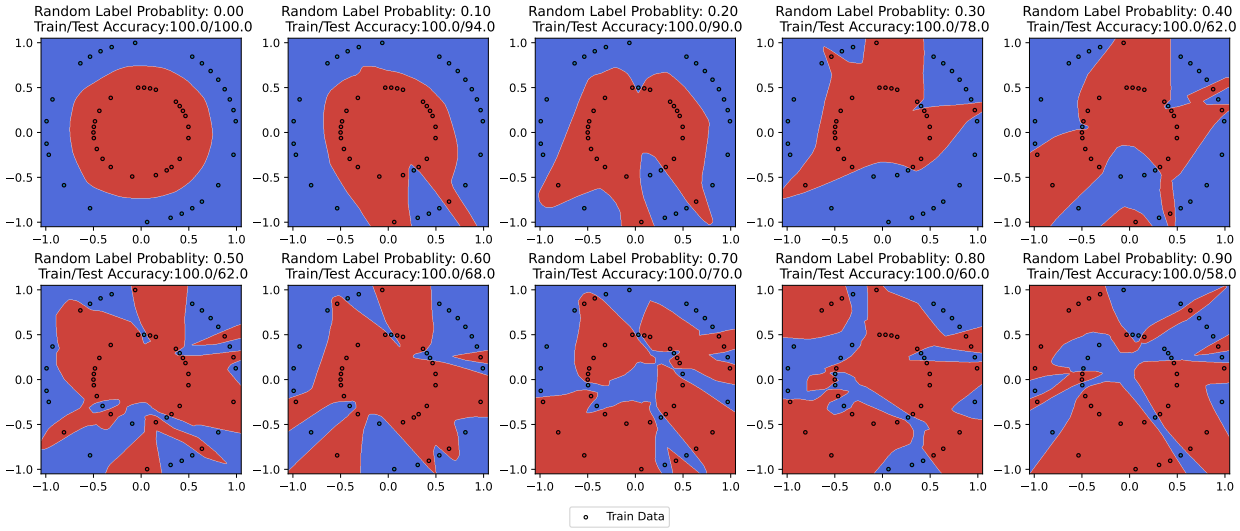


Figure 5: Visualising the decision boundaries of an MLP trained on the make_circles data set. The proportion of randomised data is increased in 0.1 intervals between 0.0 and 0.9. **Blue** indicates the decision boundary for class 1 and **red** indicates the decision boundary for class 2. As the data is increasingly randomised, the learned decision boundaries become more complex (irregular and tightly constrained).

Figure 5 shows the corresponding decision boundaries for seed 0, with the two class regions shown in red and blue. As the decision boundaries become tighter and more irregular/complex, we also observe a general reduction in test accuracy, although the decrease is not strictly monotonic, likely due to stochasticity in the data-randomisation process. It can be observed that as the proportion of randomised data increases (0.1-0.9), the decision boundaries in the visualisation become tighter and more irregular. In Figure 6, we show the mean sharpness of the minima found at the end of training (measured by Fisher-Rao norm and Relative Flatness), together with ± 1 Standard Error of the Mean (SEM) (Belia et al., 2005), as we increase the proportion of random data in the training data set. Under both reparametrisation-invariant sharpness metrics, the models trained on randomised data and exhibiting poor generalisation are substantially sharper than the baseline model trained without randomisation. Figure 5 also shows that, without randomisation, the decision boundaries are wide and the model generalises perfectly. This reproduces the central observation of Huang et al. (2020) that memorisation coincides with sharp minima and tight decision boundaries, and therefore that increased minima sharpness indicates memorisation.

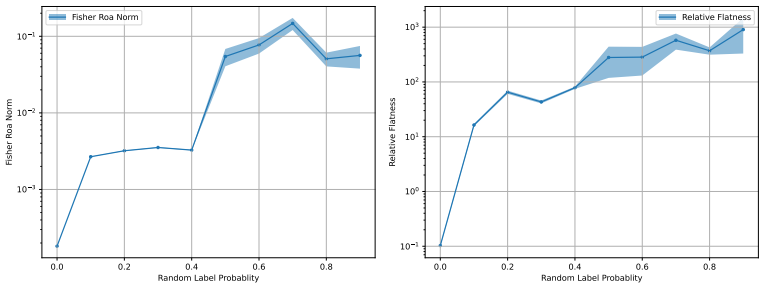


Figure 6: Fisher-Rao norm (Left) and Relative Flatness (Right) of the resulting minima after increasing the proportion of randomised data.

Building on our results in Single-Objective Optimisation discussed in Section 4, we now adopt a broader function-centric view. Rather than associating minima geometry only with memorisation or generalisation, we consider it to be tied more fundamentally to the function being learned and to the way in which the training data are fitted. To investigate this hypothesis under experimental conditions similar to those studied

by Huang et al. (2020), we modify the training data not by randomising labels, but by reducing the distance between the two classes while preserving perfect generalisation. Specifically, we vary the scale factor between the inner and outer circles, while controlling the experiment using the same 100 initialisations and data-order seeds. A higher scale factor produces tighter class boundaries, while a lower scale factor leads to wider separation between classes. Thus, the scale factor directly controls the margin of the decision boundary.

To probe this relationship, we train the same MLP architecture across scale factors from 0.0 to 0.9 using the training hyperparameters as described in the standard experiment above. Our objective is to determine how increasing the tightness of the decision boundary affects the recorded sharpness of the minima at the end of training, outside the memorisation regime. At every scale factor, all models perfectly fit the training data and generalise perfectly to the test data. In Figure 7, we observe that as the scale factor increases, the decision boundaries of the two classes move closer together. This is expected, since the scale factor explicitly reduces the separation between the two classes shown in red and blue. Under existing interpretations of flat and sharp minima, one might expect all of these models to be found at relatively flat minima, given that they all achieve 100% test accuracy and therefore generalise perfectly regardless of the scale factor. However, Figure 8 shows that increasing the scale factor is accompanied by increased sharpness under both Fisher-Rao norm and Relative Flatness. Because the scale factor explicitly modifies the decision boundaries while preserving perfect generalisation, this suggests that sharpness can provide an indication of tighter decision boundaries rather than memorisation alone. In this way, we can partially decouple the relationship between sharp minima and memorisation, and instead interpret increased sharpness, more appropriately, as reflecting tighter and more complex decision-boundary structure in the learned function.

It is also interesting to compare the sharpness values in the randomised-data and the scale-factor settings (Figures 6 and 8). The randomised models require a substantially larger training budget to fit the training data and reach sharpness values, particularly for Relative Flatness, that are orders of magnitude larger than those in the scale-factor setting. These results suggest that functions with extremely tight or irregular decision boundaries are harder for neural nets to fit than functions with wider or more structured decision boundaries, even when both are associated with increased sharpness. Our findings in this section therefore support a more nuanced relationship between sharp minima, memorisation, and decision-boundary structure:

While models that memorise training data may occupy sharp minima with tight decision boundaries, not all sharp minima with tight decision boundaries imply memorisation.

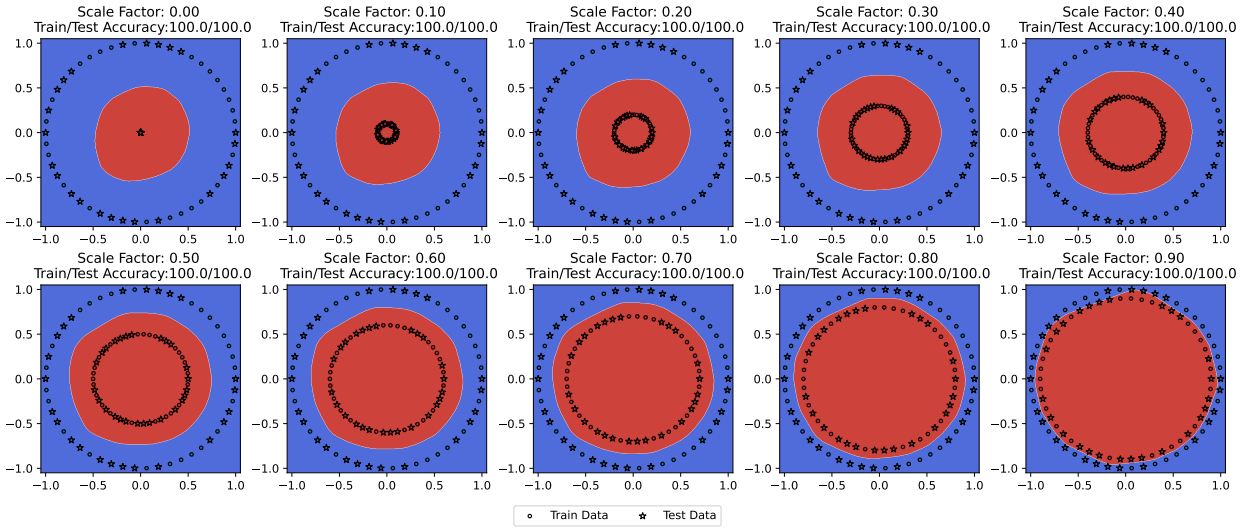


Figure 7: Visualising the decision boundaries of an MLP trained on the make_circles data set. The scale factor is increased in 0.1 intervals between 0.0 and 0.9. **Blue** indicates the decision boundary for class 1 and **red** indicates the decision boundary for class 2. As the Scale factor increases, the decision boundaries between class 1 and 2 become progressively tighter.

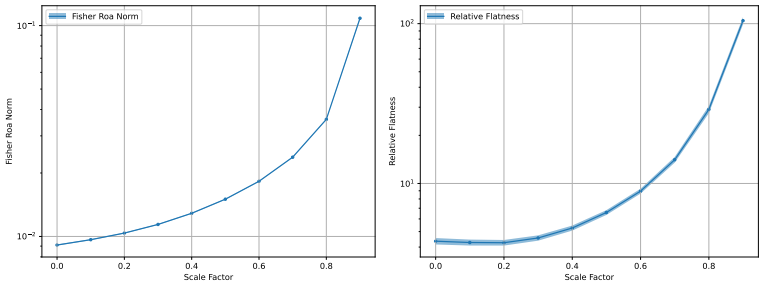


Figure 8: Fisher-Rao norm (Left) and Relative Flatness (Right) of the resulting minima after increasing the Scale factor of the data.

More broadly, these results strengthen the function-centric perspective developed throughout this paper: the manner in which data are learned can shape the geometry of the solution and, in doing so, reflect the complexity of the learned function rather than poor generalisation or memorisation alone.

6 High-Dimensional Optimisation Problems

Building on the view developed in the previous sections, namely that minima geometry reflects properties of the function being fit and the way in which that function is learned, we now extend our analysis to high-dimensional settings in the vision domain. In practice, deep vision neural networks are routinely trained with regularisation (Goodfellow et al., 2016; Kukačka et al., 2017), yet why specific regularisers improve generalisation remains only partially understood despite extensive prior work (Tian & Zhang, 2022; Moradi et al., 2020; Santos & Papa, 2022). This makes vision an ideal test-bed to study how geometry relates not only to generalisation, but also to reliability properties (calibration, robustness, and prediction agreement) at scale. Our contribution in this section is to examine these phenomena through the lens of learned solution (function-space) complexity, explicitly associating minima geometry with both generalisation and broader model behaviour. This function-centric perspective offers a complementary reading of flat and sharp minima.

Function Complexity: Occam's Razor, or the principle of parsimony, states that when two competing hypotheses, $\mathcal{H}$ and $\mathcal{H}'$ both adequately explain an event, preference should be given to the hypothesis that relies on fewer assumptions. Formally, if $\mathcal{H}$ is characterised by an assumption set $\{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_\mathcal{K}\}$ and $\mathcal{H}'$ by an assumption set $\{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_\mathcal{J}\}$, with $\mathcal{K} < \mathcal{J}$, then Occam's Razor favours $\mathcal{H}$ over $\mathcal{H}'$ (Good, 1977). This principle is commonly invoked in the flatness literature to motivate a preference for simpler learned solutions. In neural networks, this idea has often been linked to Minimum Description Length (MDL), where flatter minima are interpreted as requiring less precision to describe because the loss remains approximately constant under perturbation Hochreiter & Schmidhuber (1994).

From this perspective, sharper minima may be viewed as corresponding to learned solutions that are more tightly constrained, require more precise representation, or induce more complex decision-boundary structure. Through this lens, we study how sharpness (function complexity) is governed by regularisation to better understand generalisation and training dynamics. We extend this by connecting this study to reliability-related evaluations (robustness, calibration, and prediction consistency) as well as optimisation dynamics. Despite the arguments of Dinh et al. (2017), flat minima continue to be widely regarded as important for strong generalisation (Petzka et al., 2021; Cha et al., 2021; Zhao et al., 2022; Lee & Yoon, 2025; Han et al., 2025). However, the relationship between minima geometry, reliability-related metrics, and function complexity remains underexplored. Existing perspectives in the flatness literature suggest that neural networks with small generalisation gaps and, by extension, strong calibration or robustness, should be found at flatter minima. Our single-objective and non-linear binary classification analyses suggest a more nuanced picture: regularisation can yield sharper minima when the learned function is more structured, more tightly constrained, or requires tighter decision-boundary geometry. We therefore examine, in a controlled manner, how commonly used regularisers affect sharpness and the corresponding reliability evaluations across matched seeds.

6.1 Training Controls

We train neural networks (ResNets, VGGs and ViTs) across a range of data sets (CIFAR10, CIFAR100 and Tiny ImageNet) using $n = 10$ matched seeds. For each seed, all controls share the same initial weights and data order. This ensures that models trained under different controls begin from the same point in the loss landscape and, in principle, could traverse toward (and even reach) similar minima, enabling controlled geometric comparisons. Each control is applied independently, while all other training details (optimiser, learning-rate schedule, epochs, etc.) are held fixed within each architecture–data set setting, isolating the effect of each control on the resulting solution geometry.

Our objective is not to optimise for state-of-the-art performance, but to characterise, under controlled conditions, how different training controls affect minima geometry and reliability-related behaviour. We define the controls as follows:

Baseline: A model trained in a vanilla setting that has no extra regularisation terms applied – for each architecture and data set, we define how the baseline model is created in Appendix Section B. For each seed, the baseline provides a reference point for the geometric, generalisation, and reliability-related evaluations that should be expected for each architecture and dataset.

Weight Decay (WD), Augmentation (AUG) and Sharpness Aware Minimisation (SAM): We consider weight decay (5×10^{-4}) (WD), data augmentation (random rotation and crop) (AUG), and sharpness aware minimisation (SAM), applied individually or in combination, as defined in Subsection 6.2, as the regularisation controls used in this study. We record their effect on the sharpness metrics and the reliability evaluations under the matched-seed setup, allowing a controlled and precise causal insight into the exact impact of each regularisation control on the resulting geometric landscapes for each architecture–dataset pair.

6.2 Experimental Setup

More formally, given a training control (regulariser) c, we examine how it affects sharpness, and the corresponding reliability evaluations. Let the set of controls (training conditions) be $\mathcal{C} = \{\text{Baseline, Baseline+SAM, Aug, Aug+SAM, WD, WD+SAM}\}$. Let $\mathcal{M} = \{\text{SS}, \text{FR}, \text{RF}\}$ denote sharpness metrics (SAM sharpness, Fisher-Rao, Relative Flatness), $\mathcal{G}$ denote generalisation related metrics (generalisation gap, and test accuracy), $\mathcal{G} = \{\text{Gen}_{\text{gap}},\ \text{Test}_{\text{acc}}\}$ and let $\mathcal{R}$ denote reliability-related metrics (corruption-robust accuracy, calibration, prediction disagreement) where $\mathcal{R} = \{\text{Acc}_{\text{corr}},\ \text{ECE},\ \text{Disagree}\}$. We run seeds $i \in \{0, \ldots, 9\}$ with identical initialisation and data order across controls. For each control $c \in \mathcal{C}$ and seed i, we record $S_{i,m}^{(c)}\,(m \in \mathcal{M})$, $G_{i,g}^{(c)}\,(g \in \mathcal{G})$, $R_{i,r}^{(c)}\,(r \in \mathcal{R})$. We report per-control, per-metric summaries as means across seeds:

$$\bar{S}_m^{(c)} = \frac{1}{n}\sum_{i=0}^{n-1} S_{i,m}^{(c)}, \qquad \bar{G}_g^{(c)} = \frac{1}{n}\sum_{i=0}^{n-1} G_{i,g}^{(c)}, \qquad \bar{R}_r^{(c)} = \frac{1}{n}\sum_{i=0}^{n-1} R_{i,r}^{(c)},$$

and present mean ± SEM Belia et al. (2005) across ten seeds for each control.

Our hypothesis for the high-dimensional experiments is as follows:

- H_0: Regularisation controls ($\mathcal{C}$) reduce sharpness ($\mathcal{M}$) relative to the baseline, i.e., $\bar{S}_m^{(c)} \ll \bar{S}_m^{(baseline)}$, while improving evaluations in $\mathcal{G}$ and $\mathcal{R}$, i.e., reduced generalisation gap, higher accuracy-based metrics, and lower ECE and disagreement.
- H_a: Regularisation controls ($\mathcal{C}$) increase sharpness ($\mathcal{M}$) relative to the baseline, i.e., $\bar{S}_m^{(c)} \gg \bar{S}_m^{(baseline)}$, while still improving evaluations in $\mathcal{G}$ and $\mathcal{R}$, i.e., reduced generalisation gap, higher accuracy metrics, and lower ECE and disagreement.

This hypothesis states the experiment-level expectation in terms of per-control average behaviour across seeds. The actual statistical tests are conducted on seed-matched paired differences between each control

and the corresponding baseline run. Specifically, for each architecture–data set pair and metric, we use one-sided Wilcoxon signed-rank tests (Woolson, 2007) on the paired seed-level differences to test whether a control improves performance in the preferred direction relative to the baseline.

For each architecture–data set pair and metric, we compare each control to the *Baseline* model across the 10 matched runs ($\mathcal{C} = \{$*Baseline+SAM*, *Aug*, *Aug+SAM*, *WD*, *WD+SAM*$\}$) using one-sided Wilcoxon signed-rank tests (Woolson, 2007). Our scientific question is directional: for sharpness metrics in $\mathcal{M}$, the alternative tests whether a regularisation control yields sharper minima than the baseline; for $\mathcal{G}$, this means lower values for Gen_{gap} and higher Test_{acc}; for $\mathcal{R}$, this corresponds to higher values for Acc_{corr} and lower values for ECE and Disagree.

To control the expected false discovery rate across the reliability and sharpness metrics for each control, we apply Benjamini–Hochberg False Discovery Rate (FDR) correction (Yekutieli & Benjamini, 1999) and report significance after FDR for $\mathcal{M}$ and $\mathcal{R}$. This means that FDR is conducted independently for $\{\text{Acc}_{\text{corr}}$, ECE, Disagree$\}$ for $\mathcal{R}$ and for $\mathcal{M}$ $\{$*SS*,*FR*,*RF*$\}$. For $\mathcal{G}$ $\{\text{Gen}_{\text{gap}}, \text{Test}_{\text{acc}}\}$, we retain the significance threshold at 0.05, since Gen_{gap} is defined as $100 \times (\text{Train}_{\text{acc}} - \text{Test}_{\text{acc}})$ and is therefore algebraically dependent on Test_{acc}. We consider a regularisation control to be *statistically supported* if the corrected significance threshold is satisfied for the majority of reliability measures against the baseline. Similarly, a regularisation control leads to *statistically supported* increased sharpness if all sharpness values under correction reach the significance threshold against the baseline. We follow this experimental setup to promote reproducibility of our findings and conclusions (Mason-Williams & Mason-Williams, 2025).

7 High Dimensional Results and Discussion

Our central aim in this section is to test whether the common flatness narrative survives under controlled modern training settings. More specifically, we examine whether regularised models are consistently flatter than unregularised baselines, whether flatter models consistently achieve stronger downstream behaviour, and whether these relationships depend on the architecture and training control. The first two questions are addressed primarily in Section 7.1. The role of training control is examined further in Section 7.2, while Section 7.3 provides additional support for our function-centric interpretation by showing that artificially increasing task complexity can also lead to sharper minima. Evidence across data set complexities (CIFAR10, CIFAR100 and Tiny ImageNet) is provided in the main body for ResNet18, and in Appendix Sections E and F we present results for VGG and ViT, respectively.

We present results for ResNet18 trained on CIFAR10, CIFAR100, and Tiny ImageNet. For each control in $\mathcal{C}$, we report generalisation metrics ($\mathcal{G}$) together with reliability-related evaluations ($\mathcal{R}$) and sharpness metrics ($\mathcal{M}$) across 10 matched seeds. Appendix B provides the full training and sharpness-evaluation details. Results for VGG and ViT are given in Appendix E and F, where they support the same overall picture. Tables 3, 6, 9 below summarise how each training control affects sharpness and reliability evaluations. Means $\pm$ SEM are reported per metric. Tiny ImageNet results exclude Relative Flatness due to the Relative Flatness calculation exceeding the memory of an 80GB A100 GPU for any sample size of the training set. Additional batch-size (256 and 128) and learning-rate ($1e^{-3}$ and $1e^{-2}$) sweeps for ResNet and VGG are reported in Appendix D.1 and E.1, and support the same trends. In Section 7.2 we further explore how increasing the SAM radius ρ can increase a model's sharpness relative to the baseline while also improving performance. Finally, in Section 7.3, we artificially increase the complexity of the training data and observe that minima become sharper when classes have increasingly disjoint examples. Taken together, these results suggest that function complexity and minima geometry are closely related in high-dimensional optimisation.

7.1 Regularisation and Sharpness

Regularisers Improve Generalisation and Often Increase Sharpness. Across all CIFAR data sets and architectures, we observe a recurrent trend: the Baseline condition yields the flattest minima (lowest values across FR, RF, SAM), yet performs worst on test accuracy and reliability-related metrics: calibration (ECE), robustness (Corruption Accuracy), and functional consistency (Prediction Disagreement) (Tables 3, 6). Conversely, controls with stronger regularisation tend to yield sharper solutions while also achieving stronger

evaluations. For example, in Table 3 for the conditions of augmentation and augmentation + SAM, the sharpness values are two orders of magnitude larger for Fisher-Rao norm and Relative Flatness and several orders of magnitude larger for SAM-Sharpness against the Baseline, while test accuracy increases by more than 15%. This calls into question the conventional view that flatter minima are inherently preferable, and instead supports the function-centric perspective that sharper minima can reflect more complex, well-generalising solutions. Crucially, we also find that sharper minima can coincide empirically with better reliability-related performance. Across the majority of regularisers, the increase in reliability performance against the Baseline condition is statistically supported.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline	28.050 ±0.175	0.720 ±0.002	0.186 ±0.001	58.614 ±0.201	0.282 ±0.001	1.366E-05 $\pm 1.206E-06$	0.032 ±0.001	34.607 ±0.757
Baseline + SAM	20.588 ±0.125	0.794 ±0.001	0.108 ±0.001	66.342 ±0.164	0.168 ±0.000	5.823E-05 $\pm 9.056E-06$	0.107 ±0.006	75.093 ±1.693
Augmentation	10.399 ±0.067	0.886 ±0.001	0.077 ±0.001	68.755 ±0.219	0.121 ±0.001	1.905E-01 $\pm 2.203E-02$	3.940 ±0.207	2903.220 ±89.243
Augmentation + SAM	**6.864** ±0.038	**0.908** ±0.000	**0.014** ±0.001	**71.419** ±0.283	**0.069** ±0.000	1.303E-01 $\pm 1.547E-02$	5.571 ±0.035	4970.972 ±30.139
Weight Decay	27.942 ±0.196	0.721 ±0.002	0.174 ±0.002	58.562 ±0.227	0.281 ±0.001	3.391E-05 $\pm 4.494E-06$	0.065 ±0.004	59.767 ±3.009
Weight Decay + SAM	19.788 ±0.149	0.802 ±0.001	0.096 ±0.001	67.079 ±0.117	0.162 ±0.001	8.733E-05 $\pm 1.430E-05$	0.127 ±0.006	88.807 ±2.336

Table 3: Results for ResNet18 trained on CIFAR10. **Bolded** values indicate the best performance per metric. For sharpness metrics, lower values correspond to flatter models.

We observe in Tables 4, 7 and 10, that regularisers often coincide with a statistically significant increase in sharpness against the baselines. However, there are some instances in the Tiny ImageNet significance Table 10 where the SAM and weight decay controls do not increase sharpness when applied independently, but when applied together, they result in a statistically significant increase in sharpness. As a result, we argue that, for the training conditions explored in this paper, regularisation of neural networks under equal training budgets often leads to models navigating to sharper minima, which may indicate tighter or more constrained learned solutions rather than poorer generalisation.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✗	✗	✓	✗	✗	✓	✓	✓
Weight Decay + SAM	✓	✓	✓	✓	✓	✓	✓	✓

Table 4: Significance results for the control conditions against the baseline for ResNet18 trained on CIFAR10. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

SAM Can Improve Performance While Increasing Sharpness: Contrary to claims that SAM necessarily finds flatter solutions (Foret et al., 2021; Cha et al., 2021), our results show that SAM often increases sharpness across metrics and conditions (Tables 3, 6 and 9, as well as Appendix Sections E and F). Notably, augmentation + SAM achieves the best performance across evaluations while also being among the sharpest model. There are limited exceptions; for example, SAM-Sharpness decreases for augmentation + SAM on CIFAR10 and CIFAR100 (Tables 3, 6), but this behaviour is not consistent across metrics. On more complex data sets (Tiny ImageNet; Table 9), SAM can sometimes lead to flatter solutions, though again not uniformly across measures. Largely, when considering significance, we find that there is statistical support that SAM increases the sharpness of models in Tables 4, 7 and 10, with a few exceptions. This finding is robust across hyperparameter settings and architectures.

Overall, these findings suggest that SAM can support the learning of higher-performing functions that reside in sharper regions of the loss landscape. In Section 7.2, we further show that modifying the SAM

hyperparameter, ρ, which controls the radius of the exploration space, directly alters sharpness of the minima found at the end of training. The results again show that the strongest-performing models can occupy sharper minima than the baseline as the perturbation radius grows, although there is a critical perturbation-radius threshold at which performance begins to deteriorate.

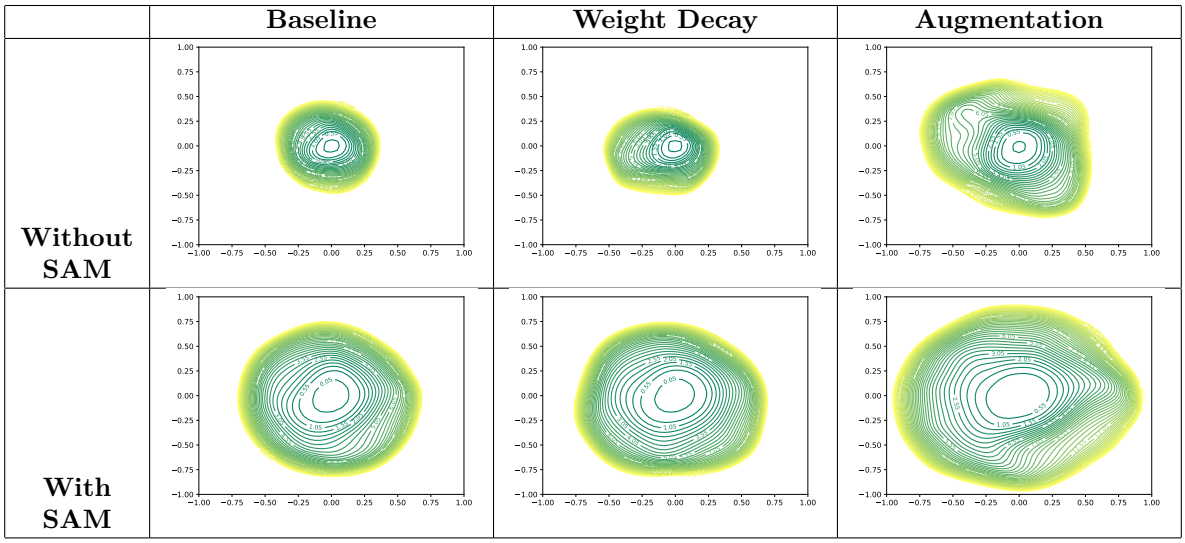


Table 5: Loss landscape visualisation Li et al. (2018) of ResNet18 landscape on CIFAR10, exploring the loss in the domain of the perturbations $[-1, 1]^2$ with 51 steps in both directions.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline	47.010 ±0.166	0.530 ±0.002	0.220 ±0.001	38.760 ±0.085	0.452 ±0.000	2.607E-04 $\pm 3.147E-05$	0.294 ±0.028	32.085 ±0.313
Baseline + SAM	44.421 ±0.168	0.556 ±0.002	0.191 ±0.002	41.888 ±0.098	0.410 ±0.000	4.231E-04 $\pm 4.973E-05$	0.399 ±0.014	123.791 ±4.185
Augmentation	29.642 ±0.133	0.697 ±0.002	0.185 ±0.001	44.613 ±0.169	0.288 ±0.001	1.110E-01 $\pm 9.173E-03$	3.587 ±0.150	2766.925 ±178.669
Augmentation + SAM	**28.999** ±0.092	**0.705** ±0.001	0.145 ±0.001	**45.428** ±0.217	**0.269** ±0.000	1.081E-01 $\pm 1.636E-02$	4.179 ±0.032	4196.832 ±52.606
Weight Decay	47.838 ±0.301	0.521 ±0.003	**0.099** ±0.005	37.868 ±0.265	0.474 ±0.001	5.192E-04 $\pm 8.009E-05$	0.861 ±0.116	136.969 ±7.484
Weight Decay + SAM	45.644 ±0.117	0.543 ±0.001	0.106 ±0.002	40.604 ±0.222	0.444 ±0.001	1.528E-03 $\pm 1.427E-04$	1.788 ±0.069	360.271 ±16.190

Table 6: Results for ResNet18 trained on CIFAR100. **Bolded** values indicate the best performance per metric. For sharpness metrics, lower values correspond to flatter models.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✗	✗	✓	✗	✗	✓	✓	✓
Weight Decay + SAM	✓	✓	✓	✓	✓	✓	✓	✓

Table 7: Significance results for the control conditions against the baseline for ResNet18 trained on CIFAR100. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

Reconciling SAM's Objective with Increased Sharpness. While SAM is commonly understood as a flatness-promoting method (Foret et al., 2021), its objective encourages local robustness rather than global flatness. Specifically, SAM minimises the loss at the worst-case perturbation within a small neighbourhood around

the current weights, thereby promoting low curvature in that vicinity. However, this does not guarantee low values under all global or reparameterisation-invariant sharpness metrics. Our findings – where SAM often increases Fisher-Rao norm, Relative Flatness, and SAM-sharpness – highlight that sharper solutions can still emerge, especially when the model learns more complex (more structured or tightly constrained) functions. This suggests that SAM may improve generalisation and reliability metrics not solely by flattening the landscape, but by guiding the model toward solutions that are robust in important local directions, even if they remain globally sharp under broader measures.

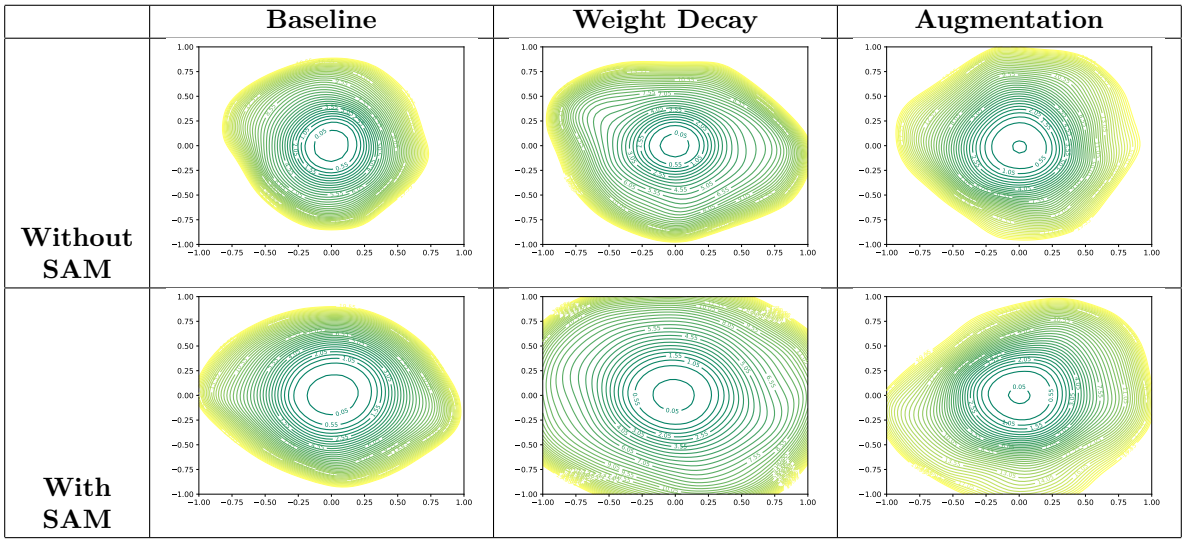


Table 8: Loss landscape visualisation Li et al. (2018) of ResNet18 landscape on CIFAR100, exploring the loss in the domain of the perturbations $[-1, 1]^2$ with 51 steps in both directions.

Limitations of Loss Landscape Visualisations. Loss landscape visualisations (Figures 5, 8), produced using the method of Li et al. (2018), have been used to support the benefits of flatness generally (Foret et al., 2021; Huang et al., 2020) and qualitatively illustrate that regularisation, especially SAM, alters the geometry of the solution. However, these plots often show that applying regularisation results in broader minima in some directions, even when reparametrisation-invariant sharpness metrics increase in sharpness in a statistically supported manner. This apparent mismatch underscores the limitations of low-dimensional loss surface plots, which capture only 2D projections of high-dimensional landscapes. In contrast, sharpness metrics reflect geometric properties beyond local projections. While visualisations can help convey functional changes, we argue that metric-based evaluations provide a more consistent and interpretable picture of sharpness.
Sharper Minima Can Coincide with Stronger Reliability-Related Properties. Across all CIFAR data sets, we consistently observe that the Baseline control yields the flattest solutions, yet performs worst on reliability-related evaluations. In contrast, the controls that achieve the best performance on these metrics are always sharper than the Baseline under most of the sharpness measures considered. The relationship between regularisation, increased sharpness and improved reliability performance is largely statistically supported in Tables Tables 4, 7 and 10. These results suggest that sharper minima can coincide with improved reliability properties, indicating that sharpness may in some settings be an important factor in the learning of more reliable models. Notably, there are instances, although limited in number, where a regularisation technique which sharpness the minima found at the end of training does not improve reliability and generalisation performance against the baseline. This shows that sharpness can be beneficial in some settings but does not by itself necessitate improved performance. However, from the results in this paper a sharper solution has a higher probability of increasing performance against flatter minima and therefore may provide an improved bias for some learning problems.

A possible explanation for the relationship we observe between generalisation, reliability and sharpness in high-dimensional learning problems is suggested by our decision-boundary experiments in Section 5. There

we show that sharper minima correspond not only to memorisation, but also to well-generalising models with tighter decision boundaries, which may be beneficial in certain tasks. This provides a useful lens through which to interpret our findings: improved reliability-related performance does not require flatness, and may in some cases arise alongside sharper solutions.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm
Baseline	49.643 ±0.103	0.503 ±0.001	0.257 ±0.001	28.347 ±0.945	0.385 ±0.000	3.202E-04 $\pm 9.872E-06$	0.479 ±0.002
Baseline + SAM	46.255 ±0.128	0.537 ±0.001	0.223 ±0.001	28.984 ±0.865	0.344 ±0.000	3.080E-04 $\pm 8.424E-06$	0.427 ±0.004
Augmentation	19.993 ±0.091	0.508 ±0.001	0.102 ±0.001	34.457 ±0.875	0.544 ±0.000	1.680E+00 $\pm 8.776E-02$	25.887 ±0.098
Augmentation + SAM	**16.777** ±0.084	0.520 ±0.001	**0.044** ±0.001	**35.324** ±0.865	0.514 ±0.000	1.446E+00 $\pm 6.332E-02$	25.193 ±0.034
Weight Decay	49.689 ±0.092	0.503 ±0.001	0.202 ±0.001	28.997 ±0.959	0.384 ±0.000	2.297E-04 $\pm 9.718E-06$	0.998 ±0.002
Weight Decay + SAM	46.061 ±0.111	**0.539** ±0.001	0.177 ±0.001	34.549 ±0.057	**0.339** ±0.000	3.784E-04 $\pm 9.996E-06$	0.736 ±0.004

Table 9: Results for ResNet18 (Pre-Trained) on Tiny ImageNet. **Bolded** values indicate the best performance per metric. For sharpness metrics, lower values correspond to flatter models.

We Find No Geometric Goldilocks Zone for Sharpness. Although we report statistical support for sharper solutions performing better across generalisation and reliability-related metrics on the CIFAR data sets, the sharpest model is not always the best overall. Still, the top-performing model is typically sharper than the Baseline, suggesting that a learning task may require a level of sharpness beyond what is induced by the architecture's implicit bias alone. This supports the view that neither extreme flatness nor extreme sharpness is universally optimal. This insight offers further refinement of the perspective developed from Section 5 where we saw that extreme sharpness could indicate memorisation while intermediate sharpness showed a strong relationship with tightening decision boundaries. Instead, the "right" level of sharpness appears to be task- and architecture-dependent. More broadly, this highlights the risk of misleading conclusions when aggregating sharpness trends across heterogeneous architectures: we observe that apparent general trends can invert under such aggregation, consistent with Simpson's Paradox (Simpson, 1951). Careful control over architecture-specific inductive biases is therefore essential when studying geometry–function relationships.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm
Baseline + SAM	✓	✓	✓	✗	✓	✗	✗
Augmentation	✓	✓	✓	✓	✗	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✗	✓	✓
Weight Decay	✗	✗	✓	✗	✓	✗	✓
Weight Decay + SAM	✓	✓	✓	✓	✓	✓	✓

Table 10: Significance results for the control conditions against the baseline for ResNet18 trained on Tiny ImageNet. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

7.2 Understanding SAM's Increased Sharpness Through the ρ Hyperparameter

We show that our finding that sharpness can increase under SAM is robust to perturbations of the ρ hyperparameter. We vary ρ across the values $0.5, 0.25, 0.05, 0.025, 0.005, 0.0025$, training ResNet-18 with a **batch size 256 and a learning rate $\mathbf{1e^{-3}}$**. As shown in Figure 9, increasing ρ increases the sharpness of the minima found at the end of training, which often coincides with a reduced generalisation gap. We find statistical support for the majority of ρ values that they coincide with increased reliability performance and sharpness in Table 12. Table 11 shows that when $\rho = 0.25$, we obtain the best test accuracy, calibration, corruption robustness, and functional consistency. Notably, this condition is substantially sharper than the lower-ρ values, again suggesting a relationship between increased sharpness and improved generalisation

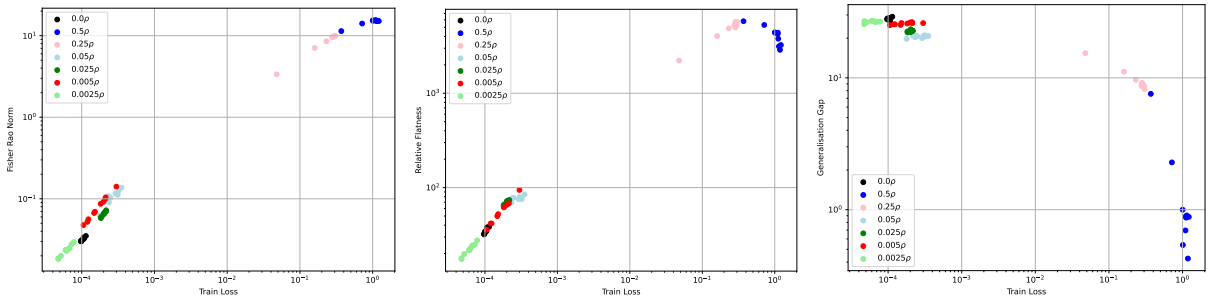
properties. At the same time, it is important to note that the sharpest condition overall, found at $\rho = 0.50$, is not the best-performing model. This reinforces our broader interpretation that the sharpness required to fit a function and associated with strong performance is highly dependent on the learning problem itself, and that there appears to be no universal Goldilocks zone of sharpness that is sufficient across tasks in our settings. We also record an instance in this setting at a very low value of $\rho = 0.0025$ where the resulting model is statistically flatter than the baseline and outperforms on generalisation and reliability metrics, but is generally not the best model compared to sharper models round under an increased ρ radius.

ρ Radius	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
0.000	28.050 ±0.175	0.720 ±0.002	0.186 ±0.001	58.614 ±0.201	0.282 ±0.001	1.366E-05 $\pm 1.206E-06$	0.032 ±0.001	34.607 ±0.757
0.5000	**1.605** ±0.646	0.629 ±0.024	0.079 ±0.007	52.847 ±1.667	0.221 ±0.009	6.814E-02 $\pm 7.326E-03$	14.767 ±0.375	4156.344 ±279.557
0.2500	9.751 ±0.640	**0.835** ±0.002	**0.026** ±0.003	**68.479** ±0.302	**0.089** ±0.002	3.884E-02 $\pm 4.981E-03$	8.712 ±0.623	4876.348 ±314.164
0.0500	20.588 ±0.125	0.794 ±0.001	0.108 ±0.001	66.342 ±0.164	0.168 ±0.000	5.823E-05 $\pm 9.056E-06$	0.107 ±0.006	75.093 ±1.693
0.0250	22.602 ±0.109	0.774 ±0.001	0.124 ±0.001	64.224 ±0.154	0.195 ±0.000	2.587E-05 $\pm 1.987E-06$	0.065 ±0.001	70.223 ±0.941
0.0050	25.793 ±0.137	0.742 ±0.001	0.167 ±0.001	60.985 ±0.280	0.250 ±0.000	4.861E-05 $\pm 7.166E-06$	0.082 ±0.009	57.886 ±5.223
0.0025	26.654 ±0.130	0.733 ±0.001	0.176 ±0.001	60.107 ±0.226	0.262 ±0.001	8.624E-06 $\pm 7.512E-07$	0.023 ±0.001	22.262 ±0.969

Table 11: Results for ResNet-18 trained on CIFAR10 with **batch size 256 and a learning rate of 1e^{-3}**, while varying the SAM ρ hyperperameter (0.5, 0.25 ,0.05 ,0.025, 0.005, 0.0025). Numbers in bold indicate best scores per metric. For sharpness metrics lower values represent flatter models.

ρ Radius	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
0.5000	✓	✗	✓	✗	✓	✓	✓	✓
0.2500	✓	✓	✓	✓	✓	✓	✓	✓
0.0500	✓	✓	✓	✓	✓	✓	✓	✓
0.0250	✓	✓	✓	✓	✓	✓	✓	✓
0.0050	✓	✓	✓	✓	✓	✓	✓	✓
0.0025	✓	✓	✓	✓	✓	✗	✗	✗

Table 12: Significance Table for ResNet-18 trained on CIFAR10 with **batch size 256 and a learning rate of 1e^{-3}**, while varying the SAM ρ hyperperameter (0.5, 0.25 ,0.05 ,0.025, 0.005, 0.0025). ✓ indicates significant difference compared to the baseline ($\rho = 0.000$); ✗ indicates no significance.



(a) Fisher-Rao norm and Loss (b) Relative Flatness and Loss (c) Generalisation Gap and Loss

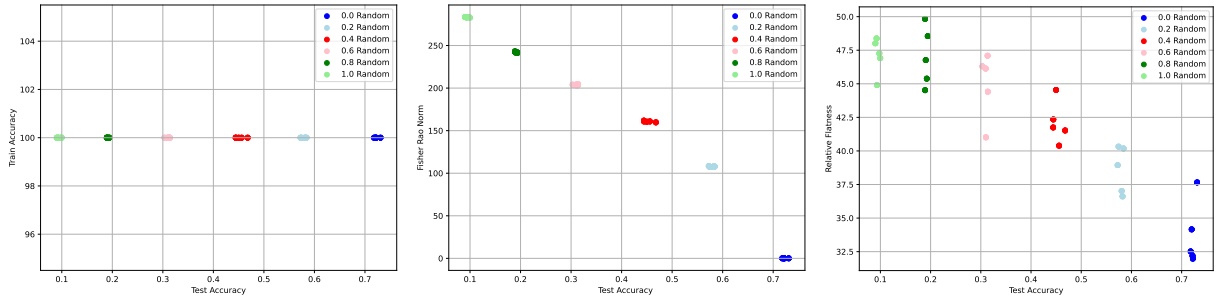
Figure 9: Scatter plots of 60 converged minima for ResNet-18 on CIFAR-10 varying the SAM (ρ) hyperparameter $(0.5, 0.25, 0.05, 0.025, 0.005, 0.0025)$, using batch size 256 and learning rate 1^{-3}: (a) Fisher–Rao norm vs. train loss, (b) Relative Flatness vs. train loss, and (c) generalisation gap vs. train loss (log scale).

7.3 Explicitly Increasing Function Complexity on High-Dimensional Data

To further demonstrate the relationship between sharp minima and increased function complexity, suggested already by the toy-setting results in Section 4 and 5, we conduct an additional experiment in the classification setting on high-dimensional data. Here, we artificially increase the difficulty of the learning task and record

the sharpness of the minima at the end of training. **Importantly, we calculate the sharpness of all models at the end of training on the same unaltered training data set**. To do this, we randomise the labels in the CIFAR10 training data set in 20% intervals and show that the resulting model trained to minimise loss on this data set reaches a sharper minimum than a model trained on the standard training data (0% randomised data). The training setup matches that of the ResNet18 baseline models, with a batch size 256 and a learning rate 0.001, and results are averaged over 5 seeds (0–4) for each condition.

The increased complexity of the learning task is also reflected in the training loss, which is higher for the randomised-data conditions despite all models being trained under the same setup, as shown in Table 13. It is important to note that all models nevertheless achieve 100% training accuracy.



(a) Train Accuracy (b) Fisher Roa norm (c) Relative Flatness

Figure 10: Increasing the portion of random training data on CIFAR10 with Resnet18 from 0.0 to 1.0: plotting train accuracy and sharpness metrics against test accuracy.

Percentage of Random Data	Train Accuracy	Train Loss	SAM Sharpness	Fisher-Rao norm	Relative Flatness
0% (Baseline)	1.000 ±0.000	9.997E-05 $\pm 1.381E-06$	1.344E-05 $\pm 1.968E-06$	0.031 ±0.001	33.700 ±0.949
20%	1.000 ±0.000	1.039E-04 $\pm 1.714E-06$	2.594E-01 $\pm 2.349E-03$	107.830 ±0.183	38.615 ±0.693
40%	1.000 ±0.000	1.071E-04 $\pm 1.759E-06$	4.180E-01 $\pm 2.405E-02$	160.742 ±0.292	42.111 ±0.613
60%	1.000 ±0.000	1.085E-04 $\pm 2.294E-06$	8.927E-01 $\pm 1.848E-02$	203.816 ±0.212	44.990 ±0.972
80%	1.000 ±0.000	1.107E-04 $\pm 2.535E-06$	4.984E-01 $\pm 3.913E-02$	242.238 ±0.260	47.011 ±0.878
100%	1.000 ±0.000	1.094E-04 $\pm 1.490E-06$	3.450E-01 $\pm 3.319E-02$	283.181 ±0.129	47.095 ±0.543

Table 13: Results for ResNet-18 trained on CIFAR10 with **increasingly randomised data**. For sharpness metrics lower values represent flatter models.

The results in Table 13 and Figure 10 show that as the percentage of randomised examples increases, and the class samples are increasingly disjoint through increased randomisation, there is a strong increase in the sharpness of the resulting minima (under Fisher-Rao norm and Relative Flatness) found at the end of training. In this setting, the experiment can be interpreted as simulating a function of increasing complexity, insofar as the class structure becomes progressively more disjoint and the effective decision-boundary geometry becomes more complex (irregular and/or tightly constrained). While the learned function in this case may not be useful from the standpoint of generalisation, the results suggest that more complex relationships in the training data can lead to sharper minima at the end of training, directly associating function complexity and minima geometry.

While we increase complexity in this experiment by increasing the number of disjoint examples the model must fit for each class, the regularisation experiments suggest there exist other sources of increased complexity in learning. These include, for example, a larger number of classes, or greater variability within each class, as introduced by augmentation. Across these scenarios, we observe that the model can navigate to sharper regions of the loss landscape in order to fit more complex data structure, but, crucially these tighter decision boundaries are harder for the model to fit in the same training budget.

8 Conclusion

This work revisits the relationship between geometry and generalisation in deep learning, extending it to include reliability-related evaluations such as calibration, robustness to corruptions, and functional consistency, and the interaction of regularisation techniques. Rather than focusing solely on accuracy, we study how sharpness relates to broader (reliability) properties of model behaviour and optimisation settings. Across diverse architectures and data sets, we find that standard training controls such as weight decay, data augmentation, and SAM often lead to sharper solutions that also achieve stronger performance on reliability-related metrics. These results call into question the conventional assumption that flatter minima are inherently preferable, and instead support a function-centric view in which sharper minima can correspond to more constrained, well-generalising functions. We further reconcile SAM's behaviour by that it promotes local robustness rather than global flatness, helping to explain why improved generalisation can coincide with increased sharpness. More broadly, our findings suggest that sharpness is not universally harmful; rather its meaning depends on the learned function, the inductive bias of the model, and the training control used to obtain the solution. We posit that the geometry of learned solutions is shaped by task-specific demands, including the need to represent functions of greater complexity or tighter decision-boundary structure, which we demonstrate in toy-settings and on high-dimensional tasks. Overall, this work calls for a re-evaluation of geometric intuitions in deep learning, and underscores the importance of connecting training controls, solution geometry, and functional reliability.

9 Limitations

We survey a large range of architectures, single-objective optimisation problems, binary classification problems and high-dimensional learning tasks using a range of hyperparameters, which shows a strong relationship between regularisation, increased sharpness and improved generalisation and reliability performance. However, we do not explore other data modalities in high-dimensional learning problems, such as audio and language tasks; as a result, it would be a natural extension of this work to explore whether our findings hold in these other modalities. Also, we do not explore how specific optimisers impact learning beyond the settings considered here. As a result, it would be interesting to explore how our results may change under optimisers such as Adam (Kingma & Ba, 2017), AdamW (Loshchilov & Hutter, 2019) and RAdam (Liu et al., 2020). Another extension of this work would be to develop a more formal operationalisation of function complexity. Finally, the conclusion supported by our results is that sharpness is not a universal indicator of poor generalisation under the settings and measures considered here. The causal relationship between training control, learned function, and minima geometry remains unresolved.

References

Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389, 2005. URL `https://psycnet.apa.org/buy/2005-16136-002`.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021. URL `https://proceedings.neurips.cc/paper/2021/file/bcb41ccdc4363c6848a1d760f26c28a0-Paper.pdf`.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017. URL `https://proceedings.mlr.press/v70/dinh17b.html`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective, 2020. URL https://arxiv.org/abs/1912.02757.

IJ Good. Explicativity: a mathematical theory of explanation with statistical applications. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 354(1678):303–330, 1977. URL https://royalsocietypublishing.org/rspa/article-abstract/354/1678/303/14412/Explicativity-a-mathematical-theory-of-explanation?redirectedFrom=fulltext.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Ting Han, Linara Adilova, Henning Petzka, Jens Kleesiek, and Michael Kamp. Flatness is necessary, neural collapse is not: Rethinking generalization via grokking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=lbtOctHDQ3.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in Neural Information Processing Systems*, volume 7, 1994. URL https://proceedings.neurips.cc/paper/1994/hash/01882513d5fa7c329e940dda99b12147-Abstract.html.

Sepp Hochreiter and Jurgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997. URL https://direct.mit.edu/neco/article-abstract/9/1/1/6027/Flat-Minima?redirectedFrom=fulltext.

W Ronny Huang, Zeyad Ali Sami Emam, Micah Goldblum, Liam H Fowl, Justin K Terry, Furong Huang, and Tom Goldstein. Understanding generalization through visualizations. In *"I Can't Believe It's Not Better!" NeurIPS 2020 workshop*, 2020. URL https://openreview.net/forum?id=pxqYT_7gToV.

Cheongjae Jang, Sungyoon Lee, Frank C. Park, and Yung-Kyun Noh. A reparametrization-invariant sharpness measure based on information geometry. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=AVh_HTC76u.

Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt Kusner. When do flat minima optimizers work? In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=vDeh2yxTvuh.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. URL http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf.

Jan Kukačka, Vladimir Golkov, and Daniel Cremers. Regularization for deep learning: A taxonomy, 2017. URL https://arxiv.org/pdf/1710.10686.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/f8c0c968632845cd133308b1a494967f-Abstract.html.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

Hyun Kyu Lee and Sung Whan Yoon. Flat reward in policy parameter space implies robust reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4OaO3GjP7k.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 888–896. PMLR, 2019. URL https://proceedings.mlr.press/v89/liang19a.html.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22188–22214. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23ao.html.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgz2aEKDr.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Haiquan Lu, Xiaotian Liu, Yefan Zhou, Qunli Li, Kurt Keutzer, Michael W. Mahoney, Yujun Yan, Huanrui Yang, and Yaoqing Yang. Sharpness-diversity tradeoff: improving flat ensembles with sharpbalance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=wJaCsnT9UE.

Israel Mason-Williams. Neural Network Compression: The Functional Perspective. In *5th Workshop on practical ML for limited and low resource settings*, 2024. URL https://openreview.net/forum?id=Q7GXKjmCSB.

Israel Mason-Williams and Gabryel Mason-Williams. Reproducibility: The new frontier in AI governance. In *ICML Workshop on Technical AI Governance (TAIG)*, 2025. URL https://openreview.net/forum?id=TiR9nkxdrH.

Israel Mason-Williams, Gabryel Mason-Williams, and Mark Sandler. Knowledge distillation: The functional perspective. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=Cgo73ZnAQc.

Reza Moradi, Reza Berangi, and Behrouz Minaei. A survey of regularization strategies for deep models. *Artif. Intell. Rev.*, 53(6):3947–3986, aug 2020. ISSN 0269-2821. doi: 10.1007/s10462-019-09784-7. URL https://doi.org/10.1007/s10462-019-09784-7.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18420–18432, 2021. URL https://openreview.net/forum?id=sygvo7ctb_.

Claudio Filipi Gonçalves Dos Santos and João Paulo Papa. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Comput. Surv.*, 54(10s), sep 2022. ISSN 0360-0300. doi: 10.1145/3510413. URL https://doi.org/10.1145/3510413.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951. URL https://www.jstor.org/stable/2984065?seq=1.

David Stutz, Matthias Hein, and Bernt Schiele. Relating adversarially robust generalization to flat minima. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7807–7817, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Stutz_Relating_Adversarially_Robust_Generalization_to_Flat_Minima_ICCV_2021_paper.pdf.

Yingjie Tian and Yuqi Zhang. A comprehensive survey on regularization strategies in machine learning. *Information Fusion*, 80:146–166, 2022. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2021.11.005. URL https://www.sciencedirect.com/science/article/pii/S156625352100230X.

Yipei Wang, Jeffrey Mark Siskind, and Xiaoqian Wang. Great minds think alike: The universal convergence trend of input salience. In *Advances in Neural Information Processing Systems*, volume 37, pp. 71672–71704, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/83e77607638c4fb17fba4a9b7844800c-Abstract-Conference.html.

Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1024–1035, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0354767c6386386be17cabe4fc59711b-Abstract-Conference.html.

Robert F Woolson. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780471462422.eoct979.

Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1-2):171–196, 1999. URL https://www.sciencedirect.com/science/article/abs/pii/S0378375899000415.

Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *International conference on machine learning*, pp. 26982–26992. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhao22i/zhao22i.pdf.

A Single-Objective Optimisation Functions

The Sphere, Rosenbrock, Rastrigin, Beale, Booth, Three Hump Camel and the Himmelblaus functions are defined in equations 8 to 14 respectively.

$$f(\boldsymbol{x}, \boldsymbol{y}) = (x^2 + y^2) \tag{8}$$

$$f(\boldsymbol{x}, \boldsymbol{y}) = (a - x)^2 + b(y - x^2)^2, \text{ where } a = 1 \text{ and } b = 100 \tag{9}$$

$$f(\boldsymbol{x}, \boldsymbol{y}) = 2a + x^2 - a\cos(2x\pi) + y^2 - a\cos(2y\pi), \text{ where } a = 10 \tag{10}$$

$$f(\boldsymbol{x}, \boldsymbol{y}) = (1.5 - x + xy)^2 + (2.25 - x + xy^2)^2 + (2.625 - x + xy^3)^2 \tag{11}$$

$$f(\boldsymbol{x}, \boldsymbol{y}) = (x + 2y - 7)^2 + (2x + y - 5)^2 \tag{12}$$

$$f(\boldsymbol{x}, \boldsymbol{y}) = 2x^2 - 1.05x^4 + \frac{x^6}{6} + xy + y^2 \tag{13}$$

$$f(\boldsymbol{x}, \boldsymbol{y}) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2 \tag{14}$$

A.1 Training Details

We trained a 3 layer ReLU multi-layered perception with an input width of two, a hidden width of 64 and output width of 1 with the Adam Optimizer and a learning rate of 1e-3. The train and test data set consisted of 10,000 input pairs (X, Y) generated by independently sampling X and Y from a uniform distribution $\mathcal{U}(-3.5, 3.5)$. For each of the seven functions (Sphere, Rosenbrock, Rastrigin, Beale, Booth, Three Hump Camel and the Himmelblaus), every input pair was evaluated using that specific function, yielding a target output T for each function such that $F(X, Y) = T$. This procedure resulted in seven distinct data sets with identical input distributions but unique output transformations determined by their respective functions allowing for a clear assessment and comparison of the model's capacity to learn each target function under controlled input conditions.

For Figure 3 the model was trained 10 times with the same initialisation with ten different data sets for the respective function for 10^6 epochs, where the mean sam sharpness based on the training data and train and test loss where recorded for initialisation and epochs $10^0, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6$.

For Figures 11-15, the model was trained 10 times with the same initialisation with ten different data sets for the respective function until the mode the model reached the specified training target loss of $300, 150, 100, 10,$ and 1. For the Beale function, the model was unable to achieve a train loss of than 150 and lower within 10^6 epochs, and for the Rosenbrock function the model was unable to achieve a train loss of 100 and lower within 10^6 epochs.

A.2 Training to Equivalent Loss

Because the model achieves different final losses after training for $[10^0, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6]$ epochs, we control for training duration by fixing a target train loss. We then investigate how reaching an approximate target train loss influences model sharpness and the generalisation gap.

To illustrate these findings, we compare the mean sam sharpness achieved at a train loss of 300 (Figure 11) and observe clear patterns. The model trained on the Rosenbrock and Beale tasks has sharpness values between 20 and 50. In contrast, when trained on Rastrigin, Booth, and Himmelblau tasks, sharpness values range between 5 and 10, while the Sphere and Three-Hump Camel tasks produce the flattest results. At this fixed loss, Figure 11(centre) highlights that the model trained on different tasks shows varied generalisation gaps. Notably, Figure 11(right) reveals several tasks (Sphere, Rastrigin, Booth, Three-Hump Camel, Himmelblaus) yielding similar generalisation gaps but differing sharpness values. Interestingly, the Rosenbrock task produces significantly higher sharpness while overlapping in generalisation gap with the Three-Hump Camel task. These observations collectively underscore that sharpness reflects the learning task rather than model generalisation.

Recognising that measuring at a train loss of 300 is arbitrary, we also examine target losses of 150, 100, 10, and 1 (Figures 12-15). Across these settings, we again find that the model can achieve similar train losses but differ in sharpness depending on the learned function, which further supports the initial claim.

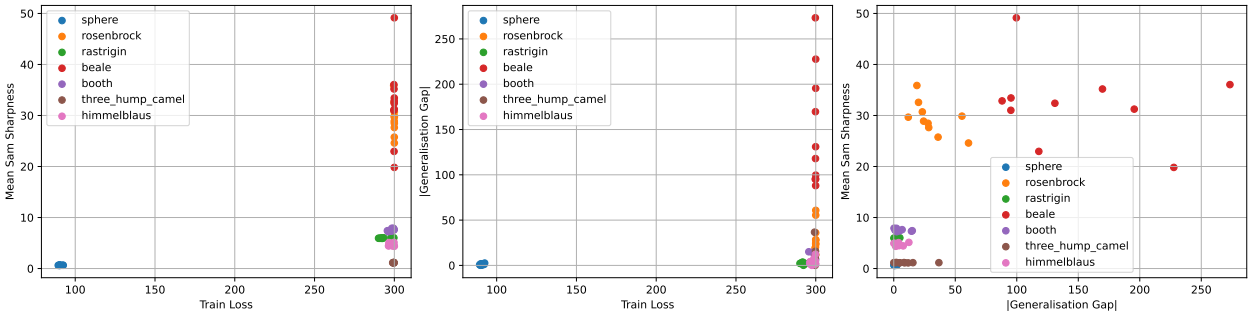


Figure 11: Scatter plots an MLP trained on the sphere, rosenbrock, rastrigin, beale, booth, three-hump camel, and himmelblaus functions for 10 different data sets till reaching a target train loss of 300: (left) mean sam sharpness vs. train loss, (centre) | generalisation gap | vs. train loss, and (right) |generalisation gap| vs. mean sam sharpness.

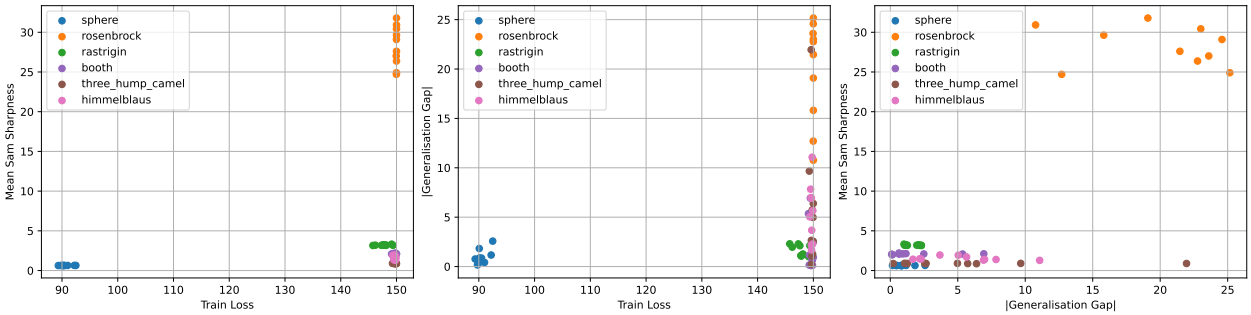


Figure 12: Scatter plots an MLP trained on the sphere, rosenbrock, rastrigin, booth, three-hump camel, and himmelblaus functions for 10 different data sets till reaching a target train loss of 150: (left) mean sam sharpness vs. train loss, (centre) | generalisation gap | vs. train loss, and (right) |generalisation gap| vs. mean sam sharpness.

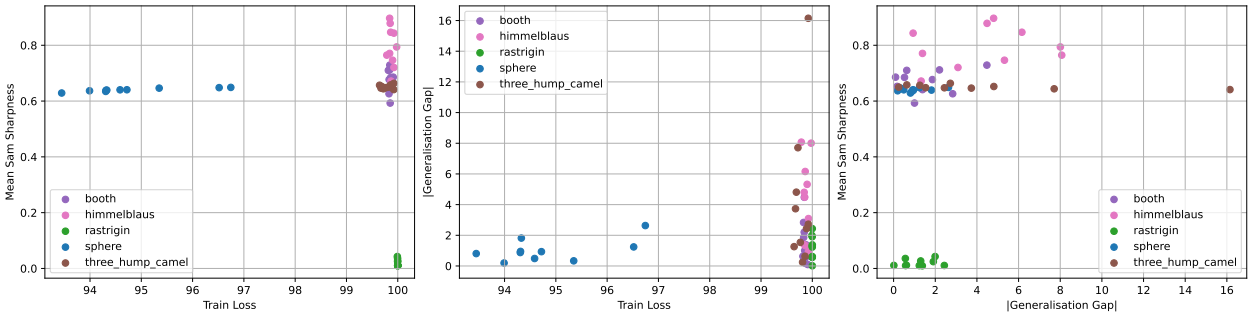


Figure 13: Scatter plots an MLP trained on the sphere, rastrigin, booth, three-hump camel, and himmelblaus functions for 10 different data sets till reaching a target train loss of 100: (left) mean sam sharpness vs. train loss, (centre) | generalisation gap | vs. train loss, and (right) |generalisation gap| vs. mean sam sharpness.

Some functions drop off as we reach particular target losses. This happens because functions with more complicated landscapes, such as Beale and Rosenbrock, cannot exceed a train loss of 150 MSE. Less complex

functions, such as the Sphere, can surpass this threshold. This supports our understanding that function complexity and solution geometry impact how easily a function can be fit. Less complex functions are more easily fit and tend to record lower sharpness values than complex functions, even when they achieve the same relative loss and generalisation gaps. Therefore, to improve optimisation outcomes for complex functions, it may be necessary to design or adopt inductive biases specifically tailored to their landscapes, rather than relying solely on traditional initialisation strategies.

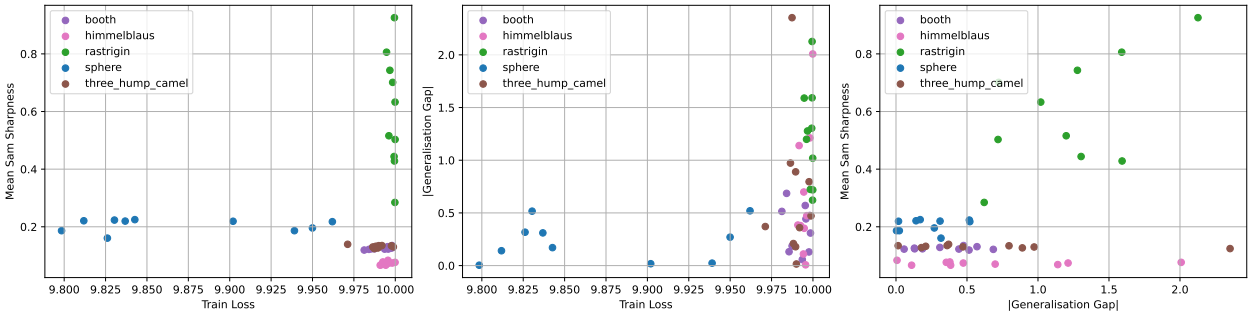


Figure 14: Scatter plots an MLP trained on the sphere, booth, three-hump camel, and himmelblaus functions for 10 different data sets till reaching a target train loss of 10: (left) mean sam sharpness vs. train loss, (centre) | generalisation gap | vs. train loss, and (right) |generalisation gap| vs. mean sam sharpness.

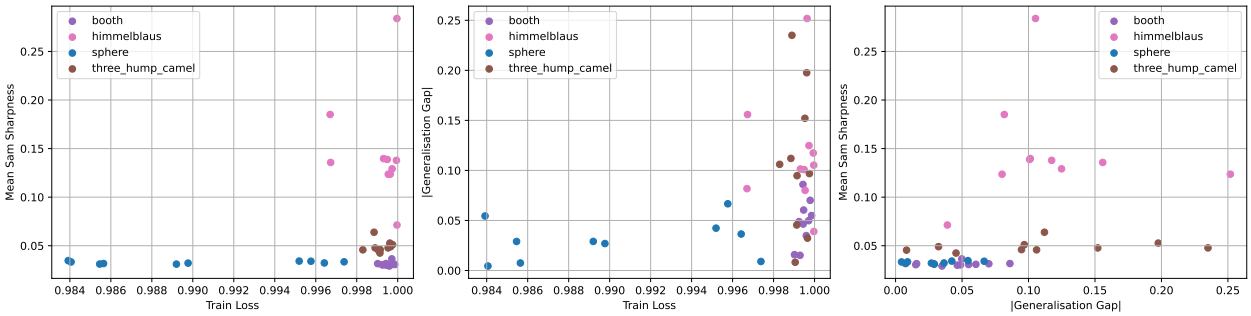


Figure 15: Scatter plots an MLP trained on the sphere, booth, three-hump camel, and himmelblaus functions for 10 different data sets till reaching a target train loss of 1: (left) mean sam sharpness vs. train loss, (centre) | generalisation gap | vs. train loss, and (right) |generalisation gap| vs. mean sam sharpness.

B Experimental Settings

All models are trained using NVIDIA A100 GPU's and each sharpness metric is calculated using the same GPU setup - as models output layer becomes larger for transitions between CIFAR10, CIFAR100 and Tiny ImageNet the computational cost of the calculation of sharpness metrics increases (by an order of magnitude between CIFAR10 and CIFAR100). It should be noted that while Fisher-Rao norm is computationally inexpensive to calculate, SAM sharpness takes a factor of time longer and Relative Flatness is the most computationally expensive measure from a time and memory perspective. All models are trained such that they converge on the training data set or approximately converge in the case of augmentation conditions - it is important to note that all models are given **100 epochs to reduce loss on the training** set to make comparisons fair. As a result, the test error is appropriate for assessing the generalisation gap as a high test accuracy is indicative of a small generalisation gap.

CIFAR10 Training: To train the **baseline** architectures on the CIFAR10 data set we use the following settings: We use SGD with the momentum hyperparameter at 0.9 to minimize cross entropy loss for 100 epochs, using a batch size of 256 a learning rate of 0.001. For all architectures in the **SAM condition** we use the same settings as above but with SAM an extra optimization step occurs. We use SAM with the

hyperparameter ρ at the standard value of 0.05. For the **augmentation condition** we use the Baseline conditions with the augmentations Random Crop with a padding of 4 and a fill of 128 alongside a Random Horizontal Flip with a probability of 0.5. Finally for the **weight decay condition** we use the same setup as the Baseline condition but with the addition of the weight decay value set at $5e^{-4}$.

CIFAR10 Sharpness: For all sharpness metrics on CIFAR10 we used the entire training data set to calculate sharpness across Fisher-Rao norm, SAM Sharpness and Relative Flatness. For the augmentation condition, the training data set is the augmentations data used to train the model. We show in Sections C.1 and C.2 that calculating sharpness on the augmented training data set for the models in the augmentation condition is approximately equivalent to calculating with the original training data set without augmentation, thus preserving the same trends of increased sharpness for models trained with augmentation.

CIFAR100 Training: To train the **baseline** architectures on the CIFAR100 data set we use the following settings: We use SGD with the momentum hyperparameter at 0.9 to minimize cross entropy loss for 100 epochs, using a batch size of 256 a learning rate of $1e^{-2}$, we also use a Pytorch's (Paszke et al., 2019) Cosine Annealing learning rate scheduler with a Maximum number of iterations of 100. For all architectures in the **SAM condition** we use the same settings as above but with SAM as an extra optimization step occurs and for this we use SAM with the hyperparameter ρ at the standard value of 0.05. For the **augmentation condition** we use the Baseline conditions with the augmentations Random Crop with a padding of 4 and a fill of 128 alongside a Random Horizontal Flip with a probability of 0.5. Finally for the **weight decay condition** we use the same setup as the Baseline condition but with the addition of the weight decay value set at $5e^{-4}$.

CIFAR100 Sharpness: For both the Fisher-Rao norm and SAM Sharpness metrics on CIFAR100 we used the entire training data set to calculate sharpness. However, due to the computational burden of calculating Relative Flatness, we only employ 20% of the training data set to calculate sharpness for this metrics. Once again, for the augmentation condition, the training data set is the augmentations data used to train the model.

Tiny ImageNet Training: On the Tiny ImageNet data set we use use pre-trained weights provided for the ResNet18[4]and VGG19BN[5] by Pytorch - we modify these architectures by removing the existing final layer and replacing it with a final layer with a 200 output classification layer.

To train the **baseline** condition on these architectures using the following settings: We use SGD with the momentum hyperparameter at 0.9 to minimize cross entropy loss for 100 epochs, using a batch size of 256 a learning rate of 0.001. For all architectures in the **SAM condition** we use the same settings as above but with SAM as an extra optimization step occurs and for this we use SAM with the hyperparameter ρ at the standard value of 0.05. For the **augmentation condition** we use the Baseline conditions with the augmentations Random Resized Crop to the size of 64 and a Random Horizontal Flip with a probability of 0.5. Finally for the **weight decay condition** we use the same setup as the Baseline condition but with the addition of the weight decay value set at $5e^{-4}$. For the Tiny ImageNet-C data set[6], only consists of 182/200 test classes, as a result we created a Tiny ImageNet-C data set that contains all 200 classes using the methodology from Hendrycks & Dietterich (2019) to ensure the corruption accuracy results presented in the paper accurately estimate the variability of the models under corruption for the Tiny ImageNet data set.

Tiny ImageNet Sharpness: For the Fisher-Rao norm sharpness metric on Tiny ImageNet we used the entire training data set to calculate sharpness. However, due to the computational burden of calculating SAM Sharpness, we only employ 20% of the training data set to calculate sharpness for this metrics. Due to memory constraints on the A100 GPU's we were unable to calculate Relative Flatness for any size of the

[4]Pytorch ResNet18 ImageNet1K Pretrained Model: `https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html`

[5]Pytorch VGG19BN ImageNet1K Pretrained Model: `https://docs.pytorch.org/vision/main/models/generated/torchvision.models.vgg19_bn.html`

[6]Official download link for the Tiny ImageNet-C data set:`https://www.google.com/search?client=safari&rls=en&q=TinyImageNet-C&ie=UTF-8&oe=UTF-8`

training data set on this architecture. Once again, for the augmentation condition, the training data set is the augmentations data used to train the model.

C Augmented or Standard Training Data Sharpness Calculation?

We argue that the standard data set is a subset of the augmented training data set, which leads to similar sharpness trends for both data sets. Our results indicate that calculating sharpness using augmented data produces nearly identical outcomes to those from the standard data set for Fisher-Rao norm, Sam Sharpness, Relative Flatness, and loss landscape visualisations. Specifically, we present these findings for the ResNet18 and VGG19 architectures in Subsections C.1 and C.2.

C.1 ResNet18

Sharpness Metrics When calculating the sharpness metrics, it can be seen that the difference between using augmented training data, in Table 15, or standard training data, in Table 16, for each of the metrics provides no difference for the trends of results observed.

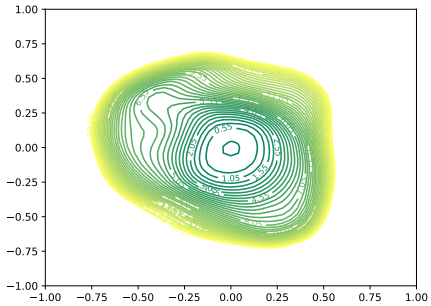
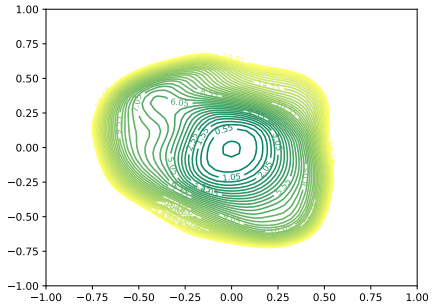
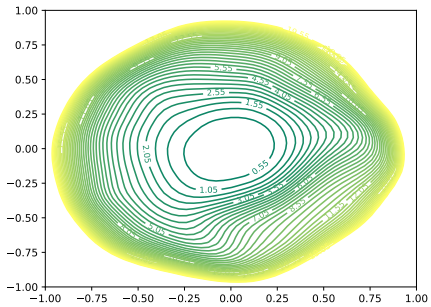
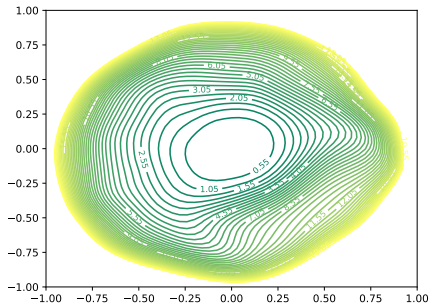
	Augmentation Augmented Training Data	**Augmentation Standard Training Data**
Without SAM		
With SAM		

Table 14: Loss landscape visualisation (Li et al., 2018)of ResNet18 landscape on CIFAR10 exploring the loss in the domain of the perturbations $[-1, 1]^2$ with 51 steps in both directions on models trained with augmentation visualising landscape with standard training data and augmented training data.

Control Condition	**SAM Sharpness**	**Fisher-Rao norm**	**Relative Flatness**
Augmentation	1.905E-01 $\pm 2.203E-02$	3.940 ± 0.207	2903.220 ± 89.243
Augmentation + SAM	1.303E-01 $\pm 1.547E-02$	5.571 ± 0.035	4970.972 ± 30.139

Table 15: Sharpness Calculation for ResNet18 landscape on CIFAR10 trained with batch size of 256 and learning rate of 0.001 using augmented training data for sharpness calculations.

Control Condition	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Augmentation	1.591E-02 $\pm 1.609E-03$	3.962 ± 0.292	2972.554 ± 137.079
Augmentation + SAM	2.035E-02 $\pm 1.203E-03$	5.084 ± 0.032	5105.327 ± 43.058

Table 16: Sharpness Calculation for ResNet18 landscape on CIFAR10 trained with batch size of 256 and learning rate of 0.001 using standard training data for sharpness calculations.

Loss Landscape Visualisations Table 14 shows that using augmented or standard training data has little effect on the loss landscape visualisation. This confirms that calculating sharpness for models trained with augmented data is valid. The data set used does not significantly affect landscape sharpness or sharpness values. The standard data set is a subset of the augmented data. Sharpness calculations depend more on model weights than on data and should represent any data set given identical weight permutations.

C.2 VGG

Sharpness Metrics When calculating the sharpness metrics, comparing the results from augmented training data (Table 18) and standard training data (Table 19) shows that both present consistent trends. In both cases, training with augmentation or augmentation plus SAM leads to a minima that is substantially sharper than the baseline model, as evidenced by the values in the respective tables.

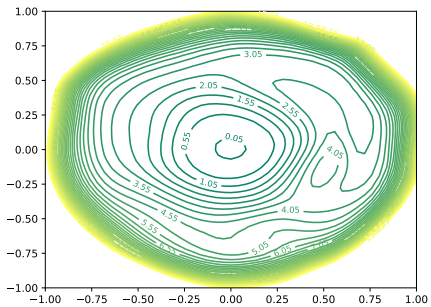
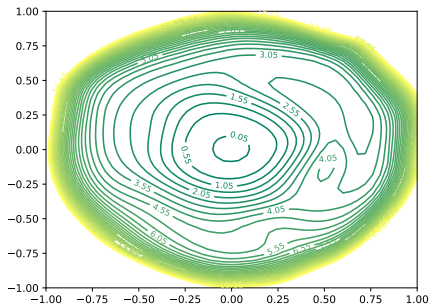
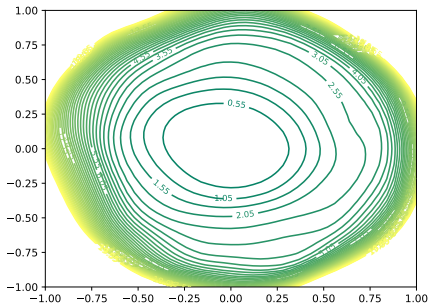
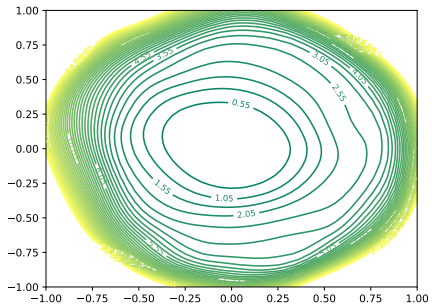
	Augmentation Augmented Training Data	**Augmentation Standard Training Data**
Without SAM		
With SAM		

Table 17: Loss landscape visualisation (Li et al., 2018) of VGG19 landscape on CIFAR10 exploring the loss in the domain of the perturbations $[-1, 1]^2$ with 51 steps in both directions on models trained with augmentation visualising landscape with standard training data and augmented training data.

Control Condition	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Augmentation	1.967E-01 ±2.298E-02	3.505 ±0.155	688.897 ±26.348
Augmentation + SAM	9.777E-02 ±1.126E-02	4.278 ±0.027	1609.212 ±22.719

Table 18: Sharpness Calculation for VGG19 on CIFAR10 trained with batch size of 256 and learning rate of 0.001 using augmented training data for sharpness calculations.

Loss Landscape Visualisations: In Table 17, we show that the use of augmented or standard training data has little impact on the resulting loss landscape visualisation. This reaffirms that it is valid to calculate sharpness for models using augmented training data. The data set used does not significantly impact the sharpness of the landscape or the resulting sharpness values. The standard data set is simply a subset of the augmented data. Furthermore, sharpness calculation depends more on the model weights than on the data, and should be a representative value for any data set given the same weight permutations.

Control Condition	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Augmentation	9.589E-03 ±1.438E-03	2.322 ±0.125	481.312 ±23.358
Augmentation + SAM	1.497E-02 ±7.363E-04	3.756 ±0.030	1413.712 ±22.212

Table 19: Sharpness Calculation for VGG19 on CIFAR10 trained with batch size of 256 and learning rate of 0.001 using standard training data for sharpness calculations.

D ResNet-18 Further Results

D.1 ResNet-18 Batch Size and Learning Rate Hyperparameter Sweep

Here we observe how two core hyperparameters, batch size and learning rate, impact the general finding that models under the use of training regularisation navigate to sharper points and thus tighter decision boundaries than base models without their application. In line with the findings in the main paper, we observe that the best-performing models in each condition are those that are sharper than the baseline models for each respective experiment.

We observe that baseline sharpness values are impacted by a change in learning rate and batch size. For example, a learning rate of $1e^{-2}$ results in far flatter baseline models for both batch sizes. The flatness observed in these cases does correspond to better-performing baseline models. However, we still observe that in all cases when regularisation is applied, the models become sharper than the respective baseline, and it still remains that these regularised models that are sharper outperform the baseline models across all hyperparameter settings.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline	28.050 ±0.175	0.720 ±0.002	0.186 ±0.001	58.614 ±0.201	0.282 ±0.001	1.366E-05 ±$1.206E-06$	0.032 ±0.001	34.607 ±0.757
Baseline + SAM	20.588 ±0.125	0.794 ±0.001	0.108 ±0.001	66.342 ±0.164	0.168 ±0.000	5.823E-05 ±$9.056E-06$	0.107 ±0.006	75.093 ±1.693
Augmentation	10.399 ±0.067	0.886 ±0.001	0.077 ±0.001	68.755 ±0.219	0.121 ±0.001	1.905E-01 ±$2.203E-02$	3.940 ±0.207	2903.220 ±89.243
Augmentation + SAM	**6.864** ±0.038	**0.908** ±0.000	**0.014** ±0.001	**71.419** ±0.283	**0.069** ±0.000	1.303E-01 ±$1.547E-02$	5.571 ±0.035	4970.972 ±30.139
Weight Decay	27.942 ±0.196	0.721 ±0.002	0.174 ±0.002	58.562 ±0.227	0.281 ±0.001	3.391E-05 ±$4.494E-06$	0.065 ±0.004	59.767 ±3.009
Weight Decay + SAM	19.788 ±0.149	0.802 ±0.001	0.096 ±0.001	67.079 ±0.117	0.162 ±0.001	8.733E-05 ±$1.430E-05$	0.127 ±0.006	88.807 ±2.336

Table 20: Results for ResNet-18 trained on CIFAR10 with **batch size 256 and a learning rate of $1e^{-3}$**. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline	16.203 ±0.266	0.838 ±0.003	0.109 ±0.004	70.814 ±0.390	0.138 ±0.001	7.648E-06 $\pm 3.794E-06$	0.015 ±0.006	16.641 ±5.960
Baseline + SAM	14.549 ±0.059	0.855 ±0.001	0.084 ±0.001	72.618 ±0.161	0.110 ±0.000	2.019E-05 $\pm 2.683E-06$	0.042 ±0.002	49.022 ±1.901
Augmentation	7.593 ±0.092	0.921 ±0.001	0.056 ±0.003	72.923 ±0.223	0.078 ±0.001	1.091E-01 $\pm 1.958E-02$	2.390 ±0.268	1604.778 ±103.972
Augmentation + SAM	**6.920** ±0.056	**0.931** ±0.001	**0.037** ±0.001	**73.483** ±0.212	**0.058** ±0.000	2.267E-02 $\pm 2.070E-03$	1.165 ±0.014	1173.090 ±15.607
Weight Decay	16.791 ±0.122	0.832 ±0.001	0.071 ±0.001	68.538 ±0.159	0.157 ±0.000	3.673E-05 $\pm 4.662E-06$	0.097 ±0.001	99.041 ±0.736
Weight Decay + SAM	14.022 ±0.089	0.860 ±0.001	0.050 ±0.001	73.100 ±0.137	0.116 ±0.000	3.238E-04 $\pm 6.384E-05$	0.446 ±0.013	178.169 ±4.103

Table 21: Results for ResNet-18 trained on CIFAR10 with **batch size 256 and a learning rate of $1e^{-2}$**. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✗	✗	✓	✗	✗	✓	✓	✓
Weight Decay + SAM	✓	✓	✓	✓	✓	✓	✓	✓

Table 22: Significance results for the control conditions against the baseline for ResNet18 trained on CIFAR10 with **batch size 256 and a learning rate of $1e^{-2}$**. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline	23.325 ±0.140	0.767 ±0.001	0.154 ±0.001	63.035 ±0.204	0.227 ±0.000	5.181E-06 $\pm 5.856E-07$	0.013 ±0.000	27.916 ±0.340
Baseline + SAM	16.714 ±0.125	0.833 ±0.001	0.083 ±0.001	69.769 ±0.108	0.126 ±0.000	2.640E-05 $\pm 4.345E-06$	0.072 ±0.006	139.589 ±2.679
Augmentation	9.110 ±0.079	0.905 ±0.001	0.065 ±0.001	71.516 ±0.308	0.099 ±0.000	9.266E-02 $\pm 5.276E-03$	2.465 ±0.105	3735.018 ±173.247
Augmentation + SAM	**6.869** ±0.022	**0.921** ±0.000	**0.013** ±0.000	**72.870** ±0.207	**0.058** ±0.000	8.913E-02 $\pm 8.054E-03$	4.070 ±0.027	7532.582 ±69.191
Weight Decay	23.504 ±0.136	0.765 ±0.001	0.137 ±0.001	62.879 ±0.214	0.231 ±0.000	2.241E-05 $\pm 3.426E-06$	0.047 ±0.000	80.599 ±0.548
Weight Decay + SAM	16.433 ±0.096	0.836 ±0.001	0.072 ±0.001	70.226 ±0.158	0.124 ±0.000	4.797E-05 $\pm 7.285E-06$	0.110 ±0.004	194.034 ±2.840

Table 23: Results for ResNet-18 trained on CIFAR10 with **batch size 128 and a learning rate of $1e^{-3}$**. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✗	✗	✓	✗	✗	✓	✓	✓
Weight Decay + SAM	✓	✓	✓	✓	✓	✓	✓	✓

Table 24: Significance results for the control conditions against the baseline for ResNet18 trained on CIFAR10 with **batch size 128 and a learning rate of $1e^{-3}$**. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline	15.027 ±0.069	0.850 ±0.001	0.109 ±0.001	71.960 ±0.158	0.125 ±0.000	1.094E-06 $\pm 9.618E-08$	0.003 ±0.000	8.785 ±0.142
Baseline + SAM	13.231 ±0.065	0.868 ±0.001	0.081 ±0.001	73.053 ±0.164	0.099 ±0.000	1.021E-05 $\pm 5.519E-07$	0.024 ±0.001	70.694 ±2.273
Augmentation	7.455 ±0.062	0.923 ±0.001	0.057 ±0.001	72.594 ±0.152	0.076 ±0.000	8.274E-02 $\pm 6.784E-03$	2.086 ±0.140	2864.657 ±151.088
Augmentation + SAM	**6.678** ±0.060	**0.933** ±0.001	**0.036** ±0.001	**73.565** ±0.245	**0.056** ±0.000	2.173E-02 $\pm 2.042E-03$	1.012 ±0.014	2354.005 ±38.058
Weight Decay	12.695 ±0.072	0.873 ±0.001	0.057 ±0.001	70.979 ±0.124	0.103 ±0.000	1.265E-04 $\pm 3.140E-06$	0.159 ±0.003	355.345 ±12.866
Weight Decay + SAM	12.606 ±0.069	0.874 ±0.001	0.036 ±0.001	72.795 ±0.162	0.107 ±0.000	5.880E-04 $\pm 6.127E-05$	0.745 ±0.017	439.467 ±8.439

Table 25: Results for ResNet-18 trained on CIFAR10 with **batch size 128 and a learning rate of $1e^{-2}$**. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

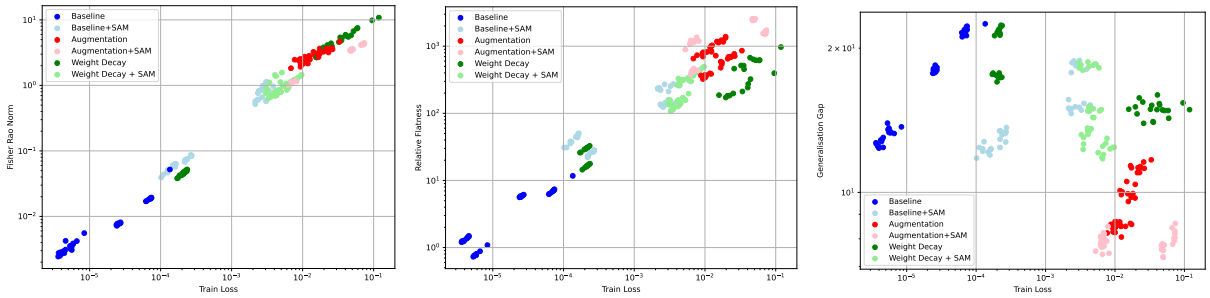
Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✓	✓	✓	✗	✓	✓	✓	✓
Weight Decay + SAM	✓	✓	✓	✓	✓	✓	✓	✓

Table 26: Significance results for the control conditions against the baseline for ResNet18 trained on CIFAR10 with **batch size 128 and a learning rate of $1e^{-2}$**. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

We observe three core insights across statistically supported results in Tables 22, 24, and 26. 1: improved generalisation performance can be found at sharper minima, 2: improved reliability mostly coincides with sharper minima, and 3: sharper minima often represent the best performing models. However, it should be noted that there are instances where weight decay increases the sharpness of minima against the baseline but does not provide improved reliability or generalisation performance.

E VGG-19

E.1 VGG19 Batch Size and Learning Rate Hyperparameter Sweep



(a) Fisher-Rao norm and Loss (b) Relative Flatness and Loss (c) Generalisation Gap and Loss

Figure 16: Plot of 240 minima using reparametrisation invariant sharpness metrics against train loss and generalisation gap against train loss using log-scale for the VGG19 with different training hyperparameters (batch size of 256, 128 and learning rate of 0.001 and $1e^{-2}$) trained on CIFAR10.

Here we observe how two core hyperparameters, batch size and learning rate, impact the general finding that models using training regularisation navigate to sharper points and thus tighter decision boundaries

than base models without regularisation. As found in the main paper, the best-performing models in each condition are those that are sharper than the baseline models for each experiment. However, changing the learning rate and batch size does influence sharpness values in each condition. A larger learning rate typically makes the minima flatter compared to a smaller learning rate. Still, these augmented models navigate to sharper landscapes than the baseline and achieve the best performance across generalisation and reliability evaluations.

Control Condition	**Genralisation Gap**	**Test Accuracy**	**Test ECE**	**Corruption Accuracy**	**Prediction Disagreement**	**SAM Sharpness**	**Fisher-Rao norm**	**Relative Flatness**
Baseline	21.805 $\pm$0.128	0.782 $\pm$0.001	0.160 $\pm$0.001	64.316 $\pm$0.193	0.204 $\pm$0.000	7.649E-06 $\pm 2.207E-06$	0.022 $\pm$0.003	7.374 $\pm$0.470
Baseline + SAM	18.444 $\pm$0.097	0.815 $\pm$0.001	0.108 $\pm$0.001	66.655 $\pm$0.296	0.150 $\pm$0.000	1.495E-03 $\pm 1.703E-04$	0.938 $\pm$0.036	140.164 $\pm$3.149
Augmentation	11.289 $\pm$0.066	0.879 $\pm$0.001	0.084 $\pm$0.001	68.497 $\pm$0.199	0.121 $\pm$0.000	1.967E-01 $\pm 2.298E-02$	3.505 $\pm$0.155	688.897 $\pm$26.348
Augmentation + SAM	**8.139** $\pm$0.074	**0.903** $\pm$0.001	**0.019** $\pm$0.001	**71.268** $\pm$0.196	**0.075** $\pm$0.000	9.777E-02 $\pm 1.126E-02$	4.278 $\pm$0.027	1609.212 $\pm$22.719
Weight Decay	21.801 $\pm$0.121	0.782 $\pm$0.001	0.151 $\pm$0.001	64.405 $\pm$0.217	0.202 $\pm$0.000	1.315E-05 $\pm 1.143E-06$	0.048 $\pm$0.001	16.494 $\pm$0.292
Weight Decay + SAM	18.394 $\pm$0.067	0.816 $\pm$0.001	0.104 $\pm$0.001	66.827 $\pm$0.286	0.151 $\pm$0.000	3.210E-03 $\pm 6.174E-04$	1.121 $\pm$0.080	157.592 $\pm$5.360

Table 27: Results for VGG19 trained on CIFAR10 with **batch size 256 and a learning rate of $\mathbf{1e^{-3}}$**. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control Condition	**Generalisation Gap**	**Test Accuracy**	**Test ECE**	**Corruption Accuracy**	**Prediction Disagreement**	**SAM Sharpness**	**Fisher-Rao norm**	**Relative Flatness**
Baseline + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✗	✗	✓	✗	✓	✗	✓	✓
Weight Decay + SAM	✓	✓	✓	✓	✓	✓	✓	✓

Table 28: Significance results for the control conditions against the baseline for VGG19 trained on CIFAR10 with **batch size 256 and a learning rate of $\mathbf{1e^{-3}}$**. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

Control Condition	**Generalisation Gap**	**Test Accuracy**	**Test ECE**	**Corruption Accuracy**	**Prediction Disagreement**	**SAM Sharpness**	**Fisher-Rao norm**	**Relative Flatness**
Baseline	13.507 $\pm$0.063	0.865 $\pm$0.001	0.105 $\pm$0.001	71.476 $\pm$0.125	0.119 $\pm$0.000	1.604E-06 $\pm 1.637E-07$	0.004 $\pm$0.000	0.807 $\pm$0.032
Baseline + SAM	13.183 $\pm$0.115	0.868 $\pm$0.001	0.081 $\pm$0.001	71.908 $\pm$0.121	0.103 $\pm$0.000	4.290E-05 $\pm 5.280E-06$	0.077 $\pm$0.002	25.287 $\pm$0.720
Augmentation	8.565 $\pm$0.030	0.910 $\pm$0.000	0.065 $\pm$0.000	71.491 $\pm$0.442	0.092 $\pm$0.000	1.146E-01 $\pm 9.917E-03$	2.555 $\pm$0.154	396.136 $\pm$22.479
Augmentation + SAM	**7.969** $\pm$0.080	**0.920** $\pm$0.001	**0.040** $\pm$0.001	**73.037** $\pm$0.160	**0.071** $\pm$0.000	3.122E-02 $\pm 3.347E-03$	1.087 $\pm$0.014	429.679 $\pm$6.763
Weight Decay	15.241 $\pm$0.124	0.836 $\pm$0.003	0.118 $\pm$0.002	68.294 $\pm$0.390	0.184 $\pm$0.002	2.302E-02 $\pm 4.987E-03$	5.120 $\pm$0.586	237.081 $\pm$22.113
Weight Decay + SAM	13.188 $\pm$0.130	0.868 $\pm$0.001	0.063 $\pm$0.001	71.618 $\pm$0.284	0.117 $\pm$0.000	5.764E-04 $\pm 5.989E-05$	0.742 $\pm$0.022	124.537 $\pm$3.152

Table 29: Results for VGG19 trained on CIFAR10 with **batch size 256 and a learning rate of $\mathbf{1e^{-2}}$**. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation	✓	✓	✓	✗	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✗	✗	✗	✗	✗	✓	✓	✓
Weight Decay + SAM	✗	✗	✓	✗	✓	✓	✓	✓

Table 30: Significance results for the control conditions against the baseline for VGG19 trained on CIFAR10 with **batch size 256 and a learning rate of $\mathbf{1e^{-2}}$**. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline	18.026 ±0.066	0.820 ±0.001	0.140 ±0.001	68.325 ±0.110	0.167 ±0.000	8.460E-07 $\pm 5.355E-08$	0.008 ±0.000	5.824 ±0.058
Baseline + SAM	15.059 ±0.085	0.849 ±0.001	0.089 ±0.001	69.791 ±0.174	0.115 ±0.000	5.689E-04 $\pm 1.248E-04$	0.649 ±0.032	242.483 ±6.334
Augmentation	9.988 ±0.088	0.895 ±0.001	0.074 ±0.001	70.671 ±0.288	0.107 ±0.000	1.708E-01 $\pm 1.967E-02$	2.851 ±0.121	1158.004 ±40.307
Augmentation + SAM	**7.594** ±0.050	**0.916** ±0.000	**0.017** ±0.000	**72.194** ±0.175	**0.066** ±0.000	7.664E-02 $\pm 1.149E-02$	3.469 ±0.023	2487.050 ±14.772
Weight Decay	17.485 ±0.066	0.825 ±0.001	0.127 ±0.001	68.655 ±0.156	0.158 ±0.000	4.918E-06 $\pm 6.710E-07$	0.044 ±0.001	29.566 ±0.647
Weight Decay + SAM	14.803 ±0.082	0.851 ±0.001	0.082 ±0.001	69.699 ±0.181	0.114 ±0.000	1.089E-03 $\pm 3.140E-04$	0.802 ±0.022	298.386 ±7.049

Table 31: Results for VGG19 trained on CIFAR10 with **batch size 128 and a learning rate of $\mathbf{1e^{-3}}$**. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✓	✓	✓	✗	✓	✓	✓	✓
Weight Decay + SAM	✓	✓	✓	✓	✓	✓	✓	✓

Table 32: Significance results for the control conditions against the baseline for VGG19 trained on CIFAR10 with **batch size 128 and a learning rate of $\mathbf{1e^{-3}}$**. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline	12.661 ±0.066	0.873 ±0.001	0.100 ±0.001	71.689 ±0.113	0.111 ±0.000	3.227E-07 $\pm 3.198E-08$	0.003 ±0.000	1.331 ±0.031
Baseline + SAM	12.248 ±0.078	0.878 ±0.001	0.077 ±0.001	71.838 ±0.256	0.100 ±0.000	8.424E-06 $\pm 9.410E-07$	0.054 ±0.003	40.853 ±2.177
Augmentation	8.333 ±0.060	0.913 ±0.001	0.063 ±0.001	72.114 ±0.259	0.091 ±0.000	1.605E-01 $\pm 2.042E-02$	2.510 ±0.122	772.205 ±19.954
Augmentation + SAM	**7.791** ±0.072	**0.922** ±0.001	**0.038** ±0.001	**72.470** ±0.190	**0.072** ±0.000	3.497E-02 $\pm 4.347E-03$	1.076 ±0.019	1155.243 ±35.334
Weight Decay	14.589 ±0.143	0.839 ±0.003	0.110 ±0.002	66.692 ±0.603	0.187 ±0.001	7.018E-03 $\pm 8.372E-04$	6.253 ±0.596	608.657 ±44.645
Weight Decay + SAM	12.295 ±0.112	0.877 ±0.001	0.052 ±0.001	71.167 ±0.301	0.116 ±0.001	5.407E-04 $\pm 1.109E-04$	1.193 ±0.043	406.163 ±15.541

Table 33: Results for VGG19 trained on CIFAR10 with **batch size 128 and a learning rate of $\mathbf{1e^{-2}}$**. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline + SAM	✓	✓	✓	✗	✓	✓	✓	✓
Augmentation	✓	✓	✓	✗	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✗	✗	✗	✗	✗	✓	✓	✓
Weight Decay + SAM	✓	✓	✓	✗	✗	✓	✓	✓

Table 34: Significance results for the control conditions against the baseline for VGG19 trained on CIFAR10 with **batch size 128 and a learning rate of 1e^{-2}**. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

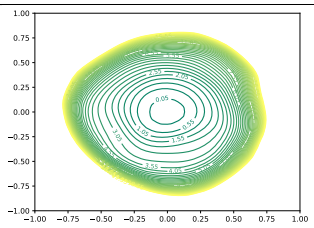
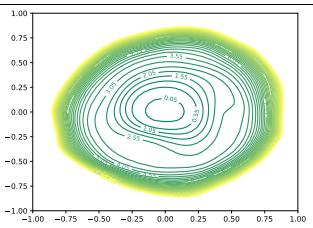
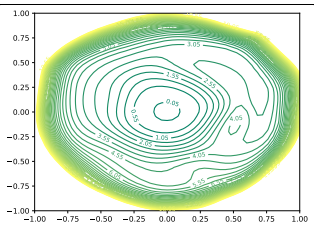
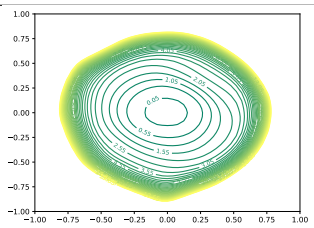
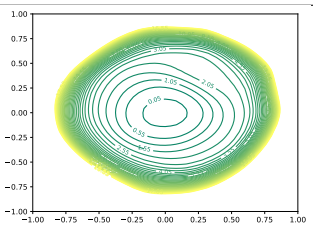
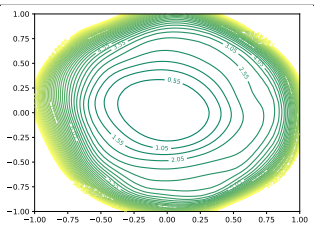
	Baseline	**Weight Decay**	**Augmentation**
Without SAM			
With SAM			

Table 35: Loss landscape visualisation (Li et al., 2018) of VGG19 landscape on CIFAR10 exploring the loss in the domain of the perturbations $[-1, 1]^2$ with 51 steps in both directions

.

In line with the hyperparameter sweeps of the ResNet, we find that the three main core findings hold for the VGG architecture. We observe across hyperparameter settings in Tables 32, 34, 23 and 34 that the three core insights remain supported. **1**: improved generalisation performance can be found at sharper minima, **2**: improved reliability mostly coincides with sharper minima and **3**: sharper minima often represent the best performing models. Once again, it should be noted that there are instances where weight decay increases the sharpness of minima against the baseline but does not provide improved reliability or generalisation performance.

CIFAR10 Landscape Visualisation: Here we observe that the loss landscapes show that the use of regularisation does change the function learned by the model and that this can often increase in complexity. For example, in Table 35 we can see that the use of weight decay, augmentation and SAM all change the minima that is reached at the end of training, with weight decay and augmentation showing a big increase in complexity compared to the baseline landscape.

E.2 CIFAR100:

The augmentation and SAM condition perform the best for test accuracy, Corruption Accuracy and Prediction Disagreement. However, for ECE, we see that weight decay is the best condition. Augmentation and SAM are the second sharpest models for Fisher-Rao norm and SAM sharpness, and have the highest

value for Relative Flatness. It is important to note that for weight decay, with the lowest ECE, it has higher sharpness values than the Baseline condition.

When considering the significance of these results, we see that the overall findings hold. In Table 38, it can be observed that regularisers' weight decay, augmentation and augmentation +SAM result in models that have statistically supported improvements across generalisation and reliability metrics as well as reaching statistically supported increases in minima sharpness. As a result, the findings in the main paper hold in this case. We also do observe that there are instances where models reach sharper minima but are not as performant in the case of Baseline+SAM and weight decay + SAM, which shows that SAM regularisation may be too strong and that the sharpness that is reached through optimisation may not be universally preferential for this learning task.

Control Condition	**Generalisation Gap**	**Test Accuracy**	**Test ECE**	**Corruption Accuracy**	**Prediction Disagreement**	**SAM Sharpness**	**Fisher-Rao norm**	**Relative Flatness**
Baseline	42.454 $\pm$0.092	0.575 $\pm$0.001	0.253 $\pm$0.000	40.749 $\pm$0.124	0.396 $\pm$0.000	2.123E-04 $\pm 2.649E-05$	0.158 $\pm$0.017	8.384 $\pm$0.151
Baseline + SAM	43.815 $\pm$0.224	0.561 $\pm$0.002	0.232 $\pm$0.002	39.690 $\pm$0.196	0.399 $\pm$0.001	7.520E-04 $\pm 5.791E-05$	0.529 $\pm$0.017	67.485 $\pm$1.802
Augmentation	32.519 $\pm$0.156	0.646 $\pm$0.002	0.222 $\pm$0.002	40.832 $\pm$0.321	0.358 $\pm$0.001	2.835E-01 $\pm 1.439E-02$	7.156 $\pm$0.270	1430.826 $\pm$53.977
Augmentation + SAM	**32.008** $\pm$0.099	**0.656** $\pm$0.001	**0.157** $\pm$0.001	**41.276** $\pm$0.089	**0.326** $\pm$0.001	1.971E-01 $\pm 1.170E-02$	5.653 $\pm$0.073	2085.080 $\pm$31.648
Weight Decay	41.579 $\pm$0.107	0.584 $\pm$0.001	0.138 $\pm$0.000	41.266 $\pm$0.112	0.384 $\pm$0.000	3.302E-04 $\pm 3.256E-05$	0.678 $\pm$0.008	45.728 $\pm$0.073
Weight Decay + SAM	44.631 $\pm$0.228	0.553 $\pm$0.002	0.189 $\pm$0.002	38.961 $\pm$0.191	0.429 $\pm$0.001	2.630E-03 $\pm 2.111E-04$	2.138 $\pm$0.084	153.194 $\pm$6.495

Table 36: Results for VGG-19 trained on CIFAR100, the Mean and $\pm$ 1 SEM are recorded over 10 models. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

CIFAR100 Landscape Visualisation: Once again, we confirm through the loss landscape visualisation in Table 37, that the application of regularisers does indeed change the properties of the minima that a network reaches at the end of training. This, corroborates our findings that state that regularisation can change the function complexity of a network and thus impact the geometric properties of the minima found at the end of training.

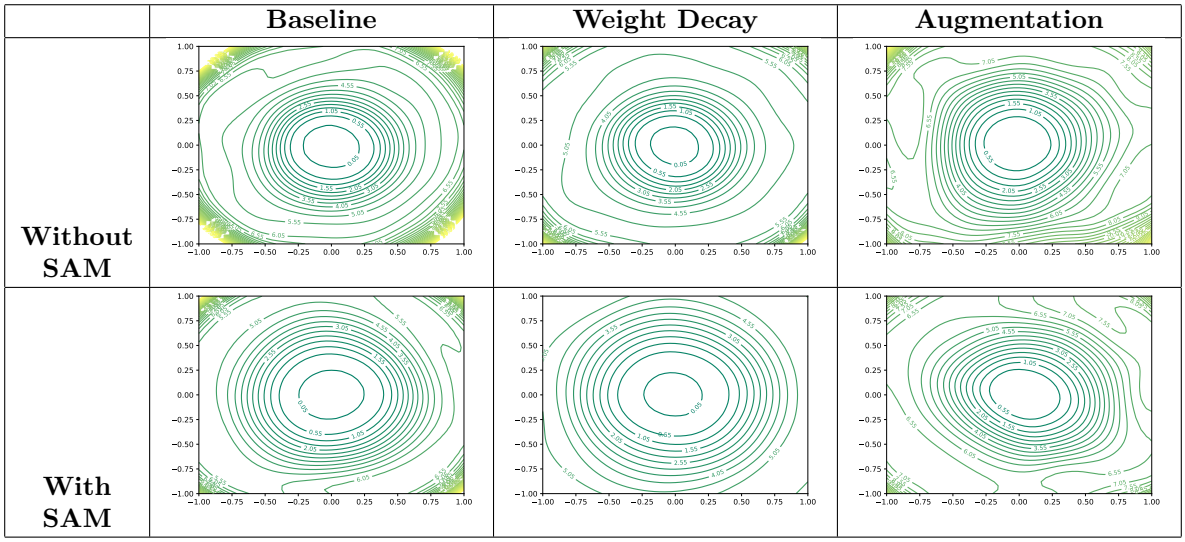


Table 37: Loss landscape visualisation (Li et al., 2018) VGG19 landscape on CIFAR100 exploring the loss in the domain of the perturbations $[-1, 1]^2$ with 51 steps in both directions

.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline + SAM	✗	✗	✓	✗	✗	✓	✓	✓
Augmentation	✓	✓	✓	✗	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay + SAM	✗	✗	✓	✗	✗	✓	✓	✓

Table 38: Significance results for the control conditions against the baseline for VGG19 trained on CIFAR100. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

E.3 Tiny ImageNet:

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao Norm
Baseline	39.588 ±0.063	0.604 ±0.001	0.303 ±0.001	39.082 ±0.042	0.238 ±0.000	2.642E-04 $\pm 1.986E-05$	0.337 ±0.118
Baseline + SAM	36.131 ±0.048	0.638 ±0.000	0.199 ±0.001	41.815 ±0.038	0.186 ±0.000	3.364E-04 $\pm 2.684E-05$	0.419 ±0.097
Augmentation	20.952 ±0.080	0.578 ±0.001	0.119 ±0.001	42.359 ±0.123	0.473 ±0.000	1.893E+00 $\pm 7.702E-02$	20.033 ±0.076
Augmentation + SAM	17.927 ±0.048	0.594 ±0.000	0.056 ±0.002	44.449 ±0.125	0.440 ±0.000	1.665E+00 $\pm 5.137E-02$	19.230 ±0.035
Weight Decay	39.622 ±0.069	0.604 ±0.001	0.265 ±0.000	39.148 ±0.026	0.222 ±0.000	2.679E-04 $\pm 1.084E-05$	0.207 ±0.026
Weight Decay + SAM	35.922 ±0.050	0.641 ±0.001	0.180 ±0.001	42.106 ±0.048	0.185 ±0.000	3.072E-04 $\pm 4.621E-06$	0.342 ±0.015

Table 39: Results for VGG19-BN (Pre-Trained) on Tiny ImageNet. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

The weight decay and SAM condition performs best for test accuracy and Prediction Disagreement. For weight decay and SAM condition we see no real difference in the sharpness values. For ECE we see that augmentation + SAM is the best condition. augmentation and SAM is the second sharpest model for Fisher-Rao norm and SAM sharpness. In this instance, the results indicate that the models trained in the augmentation and augmentation + SAM conditions are under optimised as their generalisation gaps are small while the generalisation gap remain low indicating that the models have not fully converge on the training set for this setting.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm
Baseline + SAM	✓	✓	✓	✓	✓	✗	✗
Augmentation	✓	✗	✓	✓	✗	✓	✓
Augmentation + SAM	✓	✗	✓	✓	✗	✓	✓
Weight Decay	✗	✗	✓	✗	✓	✗	✗
Weight Decay + SAM	✓	✓	✓	✓	✓	✗	✗

Table 40: Significance results for the control conditions against the baseline for VGG19 trained on Tiny ImageNet. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

These results show that both SAM and weight decay + SAM produce flatter models than the Baseline, with improved reliability and generalisation, as shown in Table 40. In contrast, the Augmented and Augmented + SAM conditions may be under-optimised and perform worse; they are also sharper. This case demonstrates a preference for flatness and serves as a reminder to apply caution when making broad statements about geometric property preference.

F Vision Transformer

F.1 CIFAR10

For the Vision Transformer (ViT), we see that our findings are consistent with the findings observed with ResNet and VGG architectures. Often regularisers increase the sharpness of the resulting model; more frequently than not, this increased sharpness coincides with improves generalisation and reliability.

For this architecture, the baseline sharpness exceeds that of ResNet and VGG, which are trained with the same batch size and learning rate. This may indicate that ViT's learning biases the models towards narrower decision boundaries in general; however, further investigation is needed to confirm this. In Table 41, augmentation and augmentation + SAM conditions achieve the best performance and exhibit the highest sharpness values across metrics, which is consistent with our previous findings.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline	39.040 $\pm$0.177	0.610 $\pm$0.002	0.308 $\pm$0.002	54.805 $\pm$0.147	0.408 $\pm$0.001	8.769E-05 $\pm 4.974E-06$	0.221 $\pm$0.003	347.198 $\pm$6.425
Baseline + SAM	39.935 $\pm$0.144	0.600 $\pm$0.001	0.276 $\pm$0.001	54.792 $\pm$0.113	0.421 $\pm$0.001	1.458E-03 $\pm 8.995E-05$	1.576 $\pm$0.083	1459.292 $\pm$82.220
Augmentation	1.305 $\pm$0.076	**0.724** $\pm$0.001	**0.019** $\pm$0.001	**64.092** $\pm$0.152	0.217 $\pm$0.001	4.741E-01 $\pm 3.822E-02$	22.809 $\pm$0.117	38465.647 $\pm$139.905
Augmentation + SAM	**-1.199** $\pm$0.097	0.668 $\pm$0.002	0.030 $\pm$0.001	60.535 $\pm$0.179	**0.201** $\pm$0.001	4.352E-01 $\pm 2.420E-02$	22.372 $\pm$0.042	18412.664 $\pm$617.822
Weight Decay	38.746 $\pm$0.196	0.613 $\pm$0.002	0.301 $\pm$0.002	55.077 $\pm$0.159	0.402 $\pm$0.001	1.359E-04 $\pm 1.030E-05$	0.328 $\pm$0.003	422.966 $\pm$6.897
Weight Decay + SAM	39.881 $\pm$0.162	0.600 $\pm$0.002	0.268 $\pm$0.001	54.797 $\pm$0.125	0.419 $\pm$0.001	2.890E-03 $\pm 3.102E-04$	2.250 $\pm$0.099	1908.688 $\pm$97.800

Table 41: Results for ViT trained on CIFAR10, the Mean and $\pm$ 1 SEM are recorded over 10 models. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline + SAM	✗	✗	✓	✗	✗	✓	✓	✓
Augmentation	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay + SAM	✗	✗	✓	✗	✗	✓	✓	✓

Table 42: Significance results for the control conditions against the baseline for ViT trained on CIFAR10. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.

With respect to significance in Table 42, we can see that the application of regularisers has statistically supported sharpness for each condition. Further aligning that regularisation can induce performance. However, we do find that only for augmentation, augmentation + SAM and weight decay this does result in statistically supported improved reliability and generalisation performance.

F.2 CIFAR100

In line with the CIFAR10 results for this architecture, we find that regularisers increase the sharpness of the resulting model and, again, that more frequently than not, this increased sharpness coincides with improves generalisation and reliability. The ViT's baseline model sharpness also exceeds that of ResNet and VGG, which are trained with the same batch size and learning rate. As a result, this adds strength to the idea that ViT's learning biases the models towards narrower decision boundaries in general. Table 43, augmentation and augmentation + SAM conditions achieve the best performance and exhibit the highest sharpness values across metrics, which is consistent with our previous findings.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline	69.048 ±0.164	0.309 ±0.002	0.402 ±0.002	25.936 ±0.088	0.723 ±0.000	3.428E-04 $\pm 4.954E-05$	0.646 ±0.061	112.185 ±4.246
Baseline + SAM	67.376 ±0.126	0.326 ±0.001	0.386 ±0.001	27.628 ±0.097	0.697 ±0.000	4.539E-04 $\pm 6.066E-05$	0.821 ±0.070	124.472 ±30.314
Augmentation	37.472 ±0.249	0.508 ±0.001	0.227 ±0.001	38.680 ±0.091	0.483 ±0.001	5.995E-01 $\pm 8.815E-02$	17.321 ±0.192	17401.462 ±143.009
Augmentation + SAM	**32.136** ±0.262	**0.523** ±0.001	**0.146** ±0.002	**40.275** ±0.097	0.446 ±0.000	4.649E-01 $\pm 2.505E-02$	19.664 ±0.127	17812.985 ±55.523
Weight Decay	67.524 ±0.148	0.325 ±0.001	0.324 ±0.001	27.364 ±0.103	0.700 ±0.000	8.440E-04 $\pm 1.251E-04$	1.563 ±0.073	251.148 ±15.330
Weight Decay + SAM	67.227 ±0.077	0.327 ±0.001	0.284 ±0.001	27.739 ±0.069	0.695 ±0.001	4.323E-03 $\pm 3.837E-04$	5.181 ±0.260	1554.595 ±91.649

Table 43: Results for ViT trained on CIFAR100, the Mean and ± 1 SEM are recorded over 10 models. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

It can be observed in Table 44 that the application of regularisation results in statistically supported increased sharpness. However, only for the conditions of weight decay, augmentation and augmentation + SAM do we observe that this increased sharpness coincides with improved generalisation and reliability. We hypothesise that the nuanced relationship with the application of SAM for the ViT could be related to its initial sharp inductive bias, however, more experiments would be needed to confirm this.

Control Condition	Generalisation Gap	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	SAM Sharpness	Fisher-Rao norm	Relative Flatness
Baseline + SAM	✗	✗	✓	✗	✗	✓	✓	✓
Augmentation	✓	✓	✓	✓	✓	✓	✓	✓
Augmentation + SAM	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay	✓	✓	✓	✓	✓	✓	✓	✓
Weight Decay + SAM	✗	✗	✓	✗	✗	✓	✓	✓

Table 44: Significance results for the control conditions against the baseline for ViT trained on CIFAR100. ✓ indicates significant difference compared to the baseline; ✗ indicates no significance.