INVESTIGATING SAMPLE-LEVEL SHARPNESS AS A CURRICULUM SCORING FUNCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Curriculum learning presents training examples to a model in order of increasing difficulty, with sample-level difficulty estimations typically relying on metrics that are implicitly deduced from the model's loss landscape. Yet, the connection to minima sharpness, a key characteristic of the local loss landscape, which is often investigated in the context of generalisation on a model or dataset level, has barely been made in the literature. In this work, we introduce per-sample sharpness as a difficulty estimation that measures how quickly the loss on a single example rises under random, filter-normalized weight perturbations of a trained model, along with a Gradient-Guided variant that perturbs along the example's own loss gradient. We evaluate both on CIFAR-10, CIFAR-100, and the Minimum Viewing Time dataset, comparing sharpness against five established scoring functions and human difficulty labels.We find that (i) sharper minima on a sample-level minima correlate with higher difficulty according to other difficulty metrics, implying a related but distinct notion of difficulty, which, however, does not correlate with human perception, (ii) random effects w. r. t. to underlying filter-normalized directions can be mitigated through ensembles, which also increase correlation with gradient-guided sharpness, (iii) sharpness-based curriculum learning outperforms a random-order baseline but not standard training (iv) under post-training weight quantization, model accuracy collapses at the same perturbation magnitude at which the Sharpness Score ordering itself becomes unreliable, consistent with a transition from a locally convex to a chaotic region of the loss landscape. We share our code on https://anonymous.4open.science/r/curriculum-learning-7302/.

1 INTRODUCTION

Human educational curricula are explicitly structured to introduce simple material before complex material, and humans teaching an unfamiliar concept or a classification boundary spontaneously order their examples from easy to hard, without being instructed to do so (Khan et al., 2011; Basu & Christensen, 2013). The research line of *curriculum learning* (CL) transfers this idea to connectionist networks under the hypothesis that the order in which samples are presented to an *artificial neural network* (ANN) during training shapes the training dynamics and thus the final behavior of the model (Elman, 1993; Bengio et al., 2009), with a particular emphasis on the concept of difficulty. In other words, each sample within a dataset commonly needs to be assigned a difficulty score by some *scoring function* (SF) to allow for a corresponding difficulty order.

Common definitions of difficulty may either be rooted in human perception, for instance, in the form of sentence length or computation-based, often through some interpretation of the training dynamics observed in the interplay with a model (Kesgin & Amasyali, 2022; Rampp et al., 2024a). The intuition behind CL, according to Bengio et al. (2009), is that a subset of easy samples should have a smoother loss landscape, which may allow the optimisation algorithm to move faster and more robustly to suitable regions before converging on the full data. However, most computational SFs in fact do not take the complexity of the loss landscape into account, but instead rely on rather simple metrics like the sample loss at convergence, the average sample-level performance during training or the average performance on held-out samples (Rampp et al., 2024a). Some of these scoring functions may nevertheless capture the loss geometry indirectly. A sample whose prediction only stabilizes late in training (Toneva et al., 2019), or whose representation only becomes linearly separable in a deep layer (Baldock et al., 2021), plausibly lies in a sharp region of the loss surface.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This is consistent with the finding that the largest eigenvalues of the loss Hessian are determined by the data distribution, such as the number of classes and how separable they are, rather than by the architecture (Sagun et al., 2016; 2017).

A separate line of research investigating the geometry of the loss landscape revolves around flat and sharp minima. Sharp minima are regions where small parameter perturbations in a converged model state cause the loss to increase steeply and have long been suspected to generalize worse than flat minima (Hochreiter & Schmidhuber, 1997) and this link between the generalization gap of large-batch training and sharpness was given influential empirical support by Keskar et al. (2017). Despite some works contesting the universal benefit of flat minima (Andriushchenko et al., 2023; Milling et al., 2024) and some limitations regarding sharpness metrics, such as sensitivity to reparameterization (Dinh et al., 2017), the common trust in the benefits of flat minima has persisted and led to methods explicitly optimizing for flatness, such as SAM (Foret et al., 2021). Notably, this geometric curvature is dictated by the training data – the Hessian of the training loss has a small number of large outlier eigenvalues whose count tracks the number of classes in the data (Sagun et al., 2016; 2017). On the parameter side, only a small subset of a network's weights is actually needed to reach the full network's accuracy (Frankle & Carbin, 2019). Connections between CL, or more specifically *sample difficulty* (SD), and minima sharpness in the literature are very sparse (Stretcu, 2021; Sitawarin et al., 2021) and do not offer a deep understanding of the relationship between these two aspects.

Nevertheless, we argue that SD and minima sharpness should be closely connected. With the loss function for a given training set in a supervised setting $\mathcal{T} = \{X_i, y_i\}$ of size N and a model parameter state θ, the total loss landscape is essentially built as the sum of the loss landscape of each individual sample

$$L(\mathcal{T}, \theta) = \frac{1}{N} \sum_i L(X_i, y_i, \theta). \tag{1}$$

Under the assumption of a better generalisation being correlated with flatter minima, we hypothesise that easy samples should have overall flatter minima in a converged model state, as, for instance, noise should, on the one hand, lead to harder samples and, on the other hand, provide more opportunities for models to overfit, which should result in sharper minima.

Even though a universal benefit of CL for performance improvement is heavily disputed (Wu et al., 2021; Rampp et al., 2024a) – though faster convergence speed may be a more reliable effect – the interpretation of difficulty concepts themselves have proven insightful (Meding et al., 2022; Milling et al., 2026). In this work, we thus aim to investigate per-sample sharpness as an underlying geometric quantity for SD in the context of CL, which, to our knowledge, no prior work has addressed. We approach this setting in two stages: we introduce sharpness-driven scoring function and measure its characteristics in the context of established scoring functions from the literature. We then train models and examine their performance under different optimization strategies, including SAM, which perturbs weights in much the same way our scoring function does and may therefore benefit more from a sharpness-based curriculum than plain SGD.

This question also matters beyond the image benchmarks studied here, as difficulty scores are increasingly used to curate and order the pretraining data of large language models: difficulty-based curricula reduce the steps needed to reach baseline performance by 18–45% (Zhang et al., 2025), per-token scores select which tokens are worth training on (Lin et al., 2024), and pruning with a good difficulty metric can beat power-law data scaling (Sorscher et al., 2022). Understanding what a difficulty score actually measures, as we do here for sharpness, is therefore of direct practical value.

Contributions: Concretely, we (i) introduce a per-sample filter-normalized and a *Gradient-Guided* sharpness as a SF for difficulty estimation; (ii) we provide a systematic analysis of the characteristics of these SFs including stability towards random directions; (iii) acomparison of sharpness-based SFs to other established CL scoring functions and human difficulty judgments (iv) an empirical study of the sharpness-based SF in CL; Our results show that the sharpness-based SF captures a difficulty signal that is largely distinct from loss, established CL scores, and human judgment, and it outperforms a random-difficulty curriculum baseline but not standard training.

2 Related Work

Curriculum learning and scoring functions: Curriculum learning has grown from the hand-designed difficulty signals of Bengio et al. (2009) into a broad literature (Soviany et al., 2022; Wang et al., 2022; Rampp et al., 2024a). Self-paced learning (Kumar et al., 2010) replaced external difficulty labels with the model's own loss, and later scores look at how an example behaves over training: the epoch at which its prediction stops being forgotten (Toneva et al., 2019), its consistency across independently trained models (Jiang et al., 2021), a running estimate from the loss trajectory (Zhou et al., 2020), the depth at which a stable prediction first appears (Baldock et al., 2021), or the confidence of a separately trained teacher (Hacohen & Weinshall, 2019). All of these infer difficulty from training dynamics or learned representations rather than from the loss landscape itself. Whether curricula help at all is contested: Hacohen & Weinshall (2019) report gains in learning speed and final accuracy, and that hard-to-easy ordering hurts, whereas Wu et al. (2021) find these benefits largely vanish once curricula are compared against a random-order baseline with the same growing training pool – the comparison we adopt.

Difficulty scores for data selection: The same per-example signals also decide which examples to keep: Paul et al. (2021) identify half of CIFAR-10 as removable from early-training error or gradient norms, Sorscher et al. (2022) beat power-law scaling of error with dataset size by pruning on a difficulty metric, and Swayamdipta et al. (2020) use confidence and variability over training to find the examples that matter most out-of-distribution. None measures loss-landscape geometry directly.

Sharpness, SAM, and low-dimensional loss geometry: Flat minima have been linked to generalization since Hochreiter & Schmidhuber (1997), and Keskar et al. (2017) observe that large-batch training converges to sharper minima that generalize worse. Dinh et al. (2017), however, show that unnormalized sharpness can be changed arbitrarily by reparameterizing a network without changing the function it computes. We therefore perturb along the filter-normalized directions of Li et al. (2018b), which factor out this scale dependence. Milling et al. (2024) use the resulting ϵ-sharpness to score whole models and find that it predicts generalization to unseen recording devices; we apply the same measure to individual examples. On the optimization side, Foret et al. (2021) derive SAM from a PAC-Bayes generalization bound (McAllester, 1999): it minimizes the worst-case loss in a neighborhood of the current weights. ASAM (Kwon et al., 2021) and GSAM (Zhuang et al., 2022) refine this objective; we use the original formulation. Finally, several results indicate that the curvature that matters around a trained model is low-dimensional: the training-loss Hessian has only a few large outlier eigenvalues (Sagun et al., 2016; 2017), gradient descent quickly moves into the subspace they span (Gur-Ari et al., 2018), and many tasks can be solved by optimizing within a small random subspace of parameter space (Li et al., 2018a). We return to this view in Section 5.1 to explain why ensembling random directions stabilizes our score.

Weight perturbation and quantization: Post-training quantization (Jacob et al., 2018; Han et al., 2016; Gholami et al., 2021) perturbs every weight at once and so offers an independent probe of the same worst-case-perturbation question. Na et al. (2022) report that SAM-trained models survive quantization better than SGD-trained ones; we test this directly and relate it to the reliability of the Sharpness Score.

Difficulty-aware benchmarks: Meding et al. (2022) find that under half of ImageNet validation images discriminate between models at all. Mayo et al. (2023) release human minimum viewing times as difficulty labels for ImageNet and ObjectNet images, and Peterson et al. (2019) collect full human label distributions for CIFAR-10. We use both to test whether our score resembles human difficulty.

3 Per-Sample Sharpness Scores

Our interpretation of sharpness follows Milling et al. (2024) on a sample level, as a combination of filter-normalised sharpness, as introduced by Li et al. (2018b), and ϵ-sharpness of Keskar et al. (2017). Concretely, let $\theta^\star \in \mathbb{R}^P$ be the weights of a trained reference model and let $\mathcal{L}_e(\theta) = \ell\big(f_\theta(x_e), y_e\big)$ be the loss of training example e at weights θ. The ϵ-sharpness measures how much the *training* loss can rise inside a Euclidean ball of radius ϵ around $\theta^\star$. We write this radius as $k\epsilon$, where ϵ is a unit radius and k its multiple: for the random-direction score, ϵ is one filter norm (Section 3.1);

for the gradient-guided score, $\epsilon = 1$ in absolute terms (Section 3.2). Since ϵ is fixed, we report only k. We evaluate the same quantity for each example separately, and we measure it directly by evaluating $\mathcal{L}_e$ at explicit perturbations of $\theta^\star$ rather than through the Hessian, whose top eigenvalue only describes the loss surface in the limit $\epsilon \to 0$ and costs at least one extra backward pass per Hessian-vector product (Pearlmutter, 1994). Both scores below share the same normalization,

$$s_e \;=\; 100 \cdot \frac{\mathcal{L}_e^{\max} - \mathcal{L}_e(\theta^\star)}{1 + \mathcal{L}_e(\theta^\star)}, \tag{2}$$

where $\mathcal{L}_e^{\max}$ is the largest loss found in the neighborhood. Dividing by $1+\mathcal{L}_e(\theta^\star)$ prevents examples that are already poorly fit from scoring high merely because their loss is large. The two scores differ only in how the neighborhood is explored.

3.1 Sharpness Score: random filter-normalized directions

Following Li et al. (2018b), we draw a direction $\mathbf{d} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_P)$ over all weights of the network. We partition the weights of all layers into filters $i = 1, \ldots, F$, one per convolutional filter or row of a linear layer, and write $\mathbf{d}^{(i)}$ and $\theta^{\star(i)}$ for the entries of $\mathbf{d}$ and $\theta^\star$ belonging to filter i. Each filter of the direction is rescaled to the norm of the corresponding trained filter,

$$\hat{\mathbf{d}}^{(i)} \;=\; \mathbf{d}^{(i)} \, \frac{\|\theta^{\star(i)}\|_2}{\|\mathbf{d}^{(i)}\|_2}, \qquad i = 1, \ldots, F, \tag{3}$$

so that a step of $\alpha\hat{\mathbf{d}}$ moves every filter by the same fraction α of its own norm; as in Li et al. (2018b), entries belonging to biases and normalization layers are left unperturbed. The perturbation coefficient α is taken from a symmetric grid $\mathcal{A}_k \subset [-k, k]$ containing 0, where k is the relative perturbation radius. Evaluating $\mathcal{L}_e(\theta^\star + \alpha\hat{\mathbf{d}})$ for $\alpha \in \mathcal{A}_k$ gives the loss profile of example e along one line through $\theta^\star$. A single line covers only one direction of the P-dimensional weight space, so its profile is a noisy estimate of the neighborhood (Section 5.1). We therefore draw D independent directions $\hat{\mathbf{d}}_1, \ldots, \hat{\mathbf{d}}_D$ and average the loss profile over them,

$$\bar{\mathcal{L}}_e(\alpha) \;=\; \frac{1}{D} \sum_{j=1}^{D} \mathcal{L}_e\big(\theta^\star + \alpha\, \hat{\mathbf{d}}_j\big), \qquad \mathcal{L}_e^{\max} \;=\; \max_{\alpha \in \mathcal{A}_k \setminus \{0\}} \bar{\mathcal{L}}_e(\alpha), \tag{4}$$

and insert $\mathcal{L}_e^{\max}$ into equation 2. The directions are shared by all examples, so scoring the whole training set costs $D\,|\mathcal{A}_k|$ forward passes over it and does not require gradient calculation; unless stated otherwise, we use $D = 50$ directions, enough to study how the ordering varies between them. To keep the cost low, we evaluate each direction only at its two endpoints, $\mathcal{A}_k = \{-k, 0, k\}$, assuming that the loss along a direction is usually highest at the largest distance from the minimum. Since the loss can also decrease along a random direction, s_e can be negative.

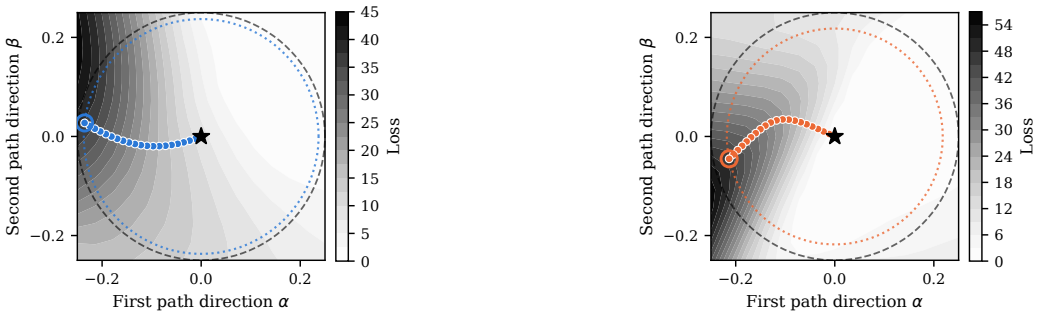
3.2 Gradient-Guided Sharpness Score

The random directions are the same for every example and ignore where that example's loss rises fastest, so their average measures the *typical* loss increase around $\theta^\star$ rather than the *largest* one. To first order, the largest increase lies along the gradient, which is what SAM (Foret et al., 2021) uses: it perturbs the weights by one step of length ρ along the normalized gradient of the batch loss, $\hat{\boldsymbol{\epsilon}} = \rho\, \nabla_\theta \mathcal{L} / \|\nabla_\theta \mathcal{L}\|_2$.

The Gradient-Guided Sharpness Score applies the same idea to each example separately and repeats the step. Starting from $\theta_e^{(0)} = \theta^\star$, example e takes up to T steps of length $\Delta = k/T$ along the normalized gradient of its own loss,

$$\hat{\mathbf{g}}_e(\theta) = \frac{\mathbf{M}\, \nabla_\theta \mathcal{L}_e(\theta)}{\|\mathbf{M}\, \nabla_\theta \mathcal{L}_e(\theta)\|_2}, \qquad \theta_e^{(t)} = \begin{cases} \theta_e^{(t-1)} + \Delta\, \hat{\mathbf{g}}_e\big(\theta_e^{(t-1)}\big) & \text{if this increases } \mathcal{L}_e, \\ \theta_e^{(t-1)} & \text{otherwise (and for all later } t), \end{cases} \tag{5}$$

where $\mathbf{M}$ is a diagonal mask that removes bias and normalization parameters as above. The walk stops at the first step that does not increase the loss. Its endpoint $\theta_e^{\max} = \theta_e^{(T)}$ is at most T steps



(a) Easiest example

(b) Hardest example

Figure 1: Gradient-guided ascent (equation 5, $k = 0.25$, $T = 20$) for the examples that the random-direction score ranks easiest and hardest on CIFAR-100/EfficientNet-B0, projected onto the two leading singular directions of each trajectory ($\geq 99.9\%$ of its variance) and shown over the loss in that plane. Dots are accepted steps, the star is $\theta^\star$ and the ring the endpoint. Both walks accept all 20 steps, so their path length is k (dashed circle), but their distance from $\theta^\star$ (dotted circle) is shorter because the path curves.

of length Δ away, so it lies within distance k of $\theta^\star$, and because the loss only rises along the walk, $\mathcal{L}_e^{\max} = \mathcal{L}_e(\theta_e^{\max})$ gives a non-negative score in equation 2. Figure 1 shows two such trajectories.

Two choices differ from Section 3.1. First, the gradient is normalized as a whole rather than per filter. For a given example, some filters have a gradient close to zero (e.g., filters that a ReLU switches off for that input), and scaling them up to their weight norm with equation 3 would inflate the step by orders of magnitude. As a result, k is an absolute radius here but relative to the filter norms in equation 4, so we compare the two scores only through the rank correlation of their orderings. Second, every example follows its own path and therefore needs its own copy of the weights. We compute all examples of a batch in parallel with per-example gradients (Goodfellow, 2015), so each step costs one batched forward and backward pass, and memory grows with batch size times P.

4 Experimental Setup

Datasets: Our main experiments for training and difficulty scoring are based on the established Tiny Image Datasets CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). With the MVT dataset (Mayo et al., 2023) and CIFAR-10H (Peterson et al., 2019), we further leverage two datasets, which contain difficulty scores originating from human interaction. Further information can be found in appendix A.

Experimental setting: Both scores of Section 3 are computed on a trained the reference models reference models EfficientNet-B0 (Tan & Le, 2019), ResNet-50 (He et al., 2016) as well as WideResNet-28-4 (Zagoruyko & Komodakis, 2016) for curriculum experiments. Unless stated otherwise we use $D = 50$ directions, $\mathcal{A}_k = \{-k, 0, k\}$ and $k = 0.25$; the gradient-guided score uses $k = 0.25$ and $T = 20$ steps. In the curriculum experiments, we further sweep the radius multiple $k \in \{0.001, 0.01, 0.05, 0.1, 0.25, 1\}$. For more concrete information about the individual training settings, please refer to appendix B.

Baseline scoring functions: We pace the identical protocol with five established scoring functions. CELoss and ProbabilityScore rank examples by the cross-entropy loss and by the maximum softmax probability of a single converged checkpoint, in the spirit of self-paced learning (Kumar et al., 2010). CumulativeAccuracy, FirstIteration (Toneva et al., 2019), and PredictionDepth (Baldock et al., 2021) are defined over a training trajectory and are computed from the 200 per-epoch checkpoints of the non-curriculum SAM baseline. All five are therefore derived from the same reference run as the sharpness scores.

Quantization and implementation: Trained checkpoints are quantized post hoc with symmetric, signed, per-output-channel linear quantization at b bits, without fine-tuning or calibration data (Jacob et al., 2018).

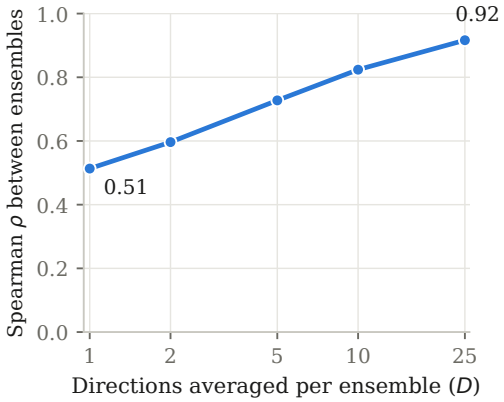


Figure 2: Mean pairwise Spearman correlation between independently formed ensembles of random directions, against ensemble size D, on the CIFAR-100/EfficientNet-B0 checkpoint at $k = 0.25$. The 50 directions are split into non-overlapping groups of D, averaged within a group, and the resulting orderings compared across groups.

Experiments are run with the `autrainer` (Rampp et al., 2024b) and `aucurriculum` (Rampp et al., 2024a) training frameworks and our code is publicly available. Difficulty orderings are compared throughout with the Spearman rank correlation (Spearman, 1904), and curriculum runs are reported by the best validation accuracy.

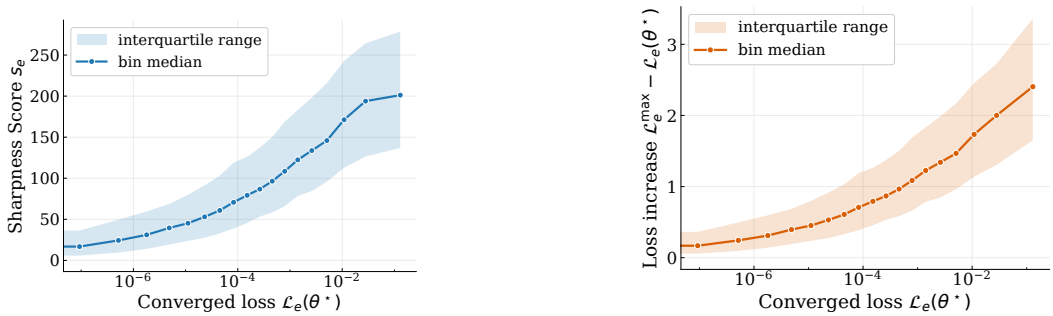
5 Results

5.1 Ensemble of Directions

Training an architecture on a dataset does not produce one deterministic model but a sample from a distribution over models, induced by initialization, data order, and other random effects generally controlled via specifying random seeds in modern frameworks. A difficulty score is meant to describe the data rather than one training run, so it should estimate how hard an example is for a model of this architecture – if across architectures – on average. Variation between two particular models, caused by randomness, is then the variance of that estimate, not an error in it. Two consequences follow: orderings from independently initialized models can disagree and averaging over multiple randomly initialized models of the same architecture should move the difficulty estimate for that specific architecture towards the mean.

Agreement within model: The perturbation direction is one source of randomness, and it is the one we can average over cheaply. On the CIFAR-100 checkpoint at $k = 0.25$, the 50 individual direction orderings have a mean pairwise Spearman correlation of 0.513 over all $1{,}225$ pairs (sd 0.025, range 0.400 to 0.585), so one direction already recovers about half the eventual ordering, but no two directions agree closely. Grouping directions into non-overlapping ensembles and comparing ensembles of equal size raises agreement to up to 0.916 at $D = 25$ (Figure 2).

Agreement between models: Averaging directions removes the variance contributed by the probe, not the variance contributed by the model. Scoring ten independently trained CIFAR-100/EfficientNet-B0 checkpoints with 50 directions each, the resulting orderings agree at $\rho = 0.550$ (sd 0.023) over all 45 pairs. That figure is the model-to-model term, and no ensemble over multiple directions reduces it. Its size can be put in context with the gradient-guided score of Section 3.2, which replaces the random direction with the example's own steepest-ascent direction. On a fixed checkpoint, the two sharpness methods agree at $\rho = 0.625$, which lies considerably above the 0.550 across seeds, while the gradient-guided score on one checkpoint against random directions on the other nine falls to $\rho = 0.344$ (sd 0.009). Which direction is probed, therefore, matters less than which model is probed. That a single informed direction recovers as much of the ensemble ordering as it does is consistent with the relevant curvature lying in a low-dimensional subspace. Gur-Ari et al. (2018) find that gradient descent trajectories collapse into the small subspace spanned by the



(a) Sharpness Score, equation 2

(b) Loss increase without normalization

Figure 3: Sharpness against converged loss on CIFAR-100/EfficientNet-B0 at $k = 0.25$, with examples in 20 equal-size bins of $\mathcal{L}_e(\theta^\star)$; bin medians and interquartile ranges. Training images are scored with data augmentation, so $\mathcal{L}_e(\theta^\star)$ is the average over the same 50 passes used for scoring.

top Hessian eigenvectors, and Li et al. (2018a) that networks reach most of their accuracy while confined to a randomly oriented subspace of a few hundred dimensions.

5.2 Sharpness, Converged Loss and Normalization

How much does the Sharpness Score simply reflect how well the model fits an example? On CIFAR-100/EfficientNet-B0 at $k = 0.25$, it ranks examples much like their converged loss does (Spearman 0.73). This means that examples the model fits worse tend to lie in sharper neighborhoods. This is not caused by the loss term in the score. Dividing by $1 + \mathcal{L}_e(\theta^\star)$ in equation 2 nearly halves the linear dependence on the loss (Pearson $0.17 \rightarrow 0.10$) and leaves the ranking almost unchanged (Spearman $0.74 \rightarrow 0.73$; Figure 3).

The correlation with the loss also grows with the number of directions, from 0.58 for a single direction to 0.73 for $D = 50$. Averaging therefore does not add a new signal; it removes the variance of individual directions.

This mirrors, at the level of single examples, the flat-minima view that models in sharp regions of the loss surface generalize worse (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017): examples with a sharp neighborhood are the ones the model fits worse, and in the sense of Bengio et al. (2009) the harder ones. This only holds for small radii; Section 5.4 shows that it breaks down for $k \geq 1$.

Still, the Sharpness Score is not interchangeable with existing scores. CELoss and ProbabilityScore agree almost perfectly with each other ($\rho = 0.979$), while the Sharpness Score agrees only moderately with the five established scoring functions ($\rho = 0.25$ to 0.52, all computed on the WideResNet-28-4 baseline; Figure 21) and weakly with human difficulty labels on MVT and CIFAR-10H ($\rho \leq 0.23$).

5.3 Curriculum Performance

Table 1 reports the best validation performance of every run. Ordering by the Sharpness Score beats the Random Curriculum at every radius except $k = 1$, where the two are level, and works best at $k = 0.25$. It still falls short of four of the five established scoring functions. The gradient-guided score, although it follows each example's steepest ascent, gives a weaker curriculum than random directions at the same radius.

No curriculum, whichever scoring function it uses, matches standard training of the same architecture, which is in line with prior findings (Wu et al., 2021; Rampp et al., 2024a). Since all curricula were already trained with SAM, we find no sign that a sharpness-based ordering and a sharpness-aware optimizer complement each other. Reversing the ordering hurts – anti-curriculum is the weakest run over the first third of training and still the weakest at convergence, although the gap narrows once the growing pool has admitted most of each class. All runs use a single seed, so differences of a few tenths of a point should not be over-interpreted.

Table 1: Best validation performance on CIFAR-100 (single seed). Curriculum runs use the protocol of Section 4. † uses an adaptive variant of equation 5: when a step fails to increase the loss, the step size is halved and the step retried, instead of the walk stopping.

Curriculum	k	Accuracy (%)	F1 (%)	Loss
Standard training (no curriculum)				
WideResNet-28-4 (SGD)	—	80.78	80.74	0.7207
WideResNet-28-4 (SAM)	—	81.46	81.48	0.6632
WideResNet-28-10 (SGD)	—	84.11	84.12	0.6117
Curriculum learning				
Random Curriculum	—	77.43	77.35	0.9463
Sharpness Score	0.001	77.74	77.63	0.9003
Sharpness Score	0.01	78.11	78.03	0.9009
Sharpness Score	0.05	77.81	77.75	0.9105
Sharpness Score	0.1	77.76	77.73	0.8982
Sharpness Score	0.25	78.40	78.34	0.8876
Sharpness Score	1.0	77.45	77.32	0.9661
Gradient Sharpness Score†	0.25	77.72	77.61	0.9502
CELoss	—	78.36	78.26	0.8661
CumulativeAccuracy	—	78.64	78.63	0.8468
FirstIteration	—	78.80	78.77	0.8542
PredictionDepth	—	**78.89**	**78.86**	0.8591
ProbabilityScore	—	78.83	78.78	**0.8461**

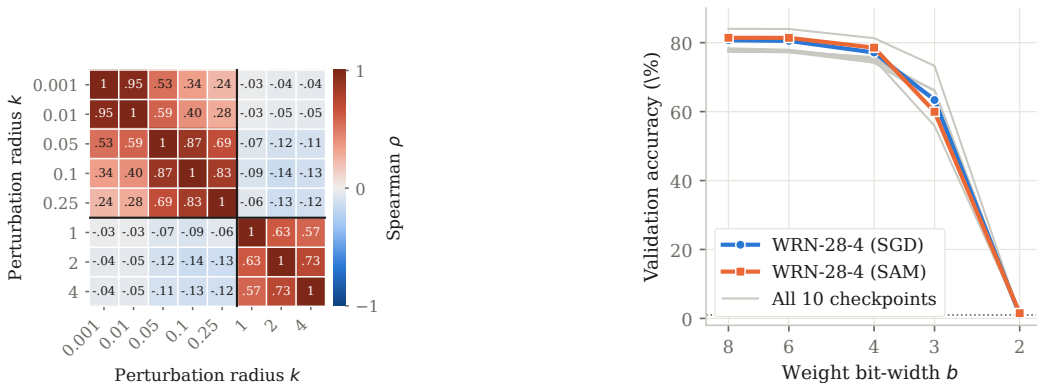
5.4 Perturbation Radius, Quantization and Collapse

Perturbation radius: The radius k sets how far the weights are pushed before the loss increase is measured. To test whether a larger k simply gives a coarser reading of the same property, we scored one checkpoint at eleven radii and correlated every pair of the resulting difficulty orderings (Figure 4a). The orderings split into two groups: small radii ($k \leq 0.25$) agree with each other and with the converged loss, large radii ($k \geq 1$) agree with each other, and the two groups are uncorrelated. We read this as two regimes of the loss surface. Small radii stay inside the locally convex region around the minimum, while large radii reach the chaotic region beyond it, where the loss no longer changes smoothly. The large-radius orderings are not just noise: they agree with one another far more than noise would, even though the raw scores become very large. Within the presumably convex group, $k = 0.25$ is also the radius at which a single direction is most informative (has highest correlation between ensembles of random directions), so we use it throughout.

Quantization as random perturbation: Post-training quantization perturbs every weight randomly and simultaneously, much in the way random directions perturbed model weights. All ten checkpoints behave alike (Figure 4b) – accuracy is unchanged at $b = 8$ and $b = 6$ ($|\Delta| \leq 0.45$pp), falls by 1.97 to 3.63pp at $b = 4$, and comes within one point of the 1% chance level at $b = 2$ – catastrophically degrading the model. Neither the scoring function nor the presence of a curriculum moves that threshold. Between the two directly comparable baselines, the SAM-trained model degrades less than the SGD-trained one initially but the improvement is minimal. SAM and SGD weights differ in ℓ_2 norm by 0.6%, which rules out weight scale as an explanation.

Transition to chaotic domain: Quantization and the Sharpness Score both work by moving the trained weights, so both can be described by how far they move them. For the score that distance is the radius k. For quantization, we measure it as the mean relative weight error $\delta = \|w_q - w\|_2 / \|w\|_2$, which puts $b = 8$ at $\delta = 0.011$, $b = 4$ at 0.19, $b = 3$ at 0.41 and $b = 2$ at 0.85.

Figure 5 places both on that axis against the boundary of the previous section. Every bit-width that keeps its accuracy lies below $\delta = 0.19$, inside the convex region where all the difficulty orderings agree with one another. The bit-width that destroys accuracy, $b = 2$, lies at $\delta = 0.85$, in the chaotic region next to $k = 1$, the first radius whose ordering is unrelated to the convex ones. Between them sits $b = 3$, which is where SAM loses its advantage over SGD. The difficulty ordering and the network's accuracy therefore break down after the same amount of weight movement.



(a) Agreement between radii

(b) Accuracy under quantization

Figure 4: A boundary between $k = 0.25$ and $k = 1$ separates two regimes of the loss surface. *(a)* Spearman correlation between the fully averaged difficulty orderings at each perturbation radius, on one CIFAR-100 checkpoint. Below the rule the perturbation stays inside a locally convex region and every radius ranks the examples the same way. Above it the perturbation reaches the chaotic region further from the minimum, where the rankings agree with each other but not with the convex ones. $k = 8$ and $k = 16$ overflow float32 and are omitted. *(b)* Validation accuracy against weight bit-width for all ten quantized checkpoints, with the two directly comparable standard baselines marked. Accuracy survives to $b = 6$, drops at $b = 4$, and reaches the 1% chance level (dotted) at $b = 2$.

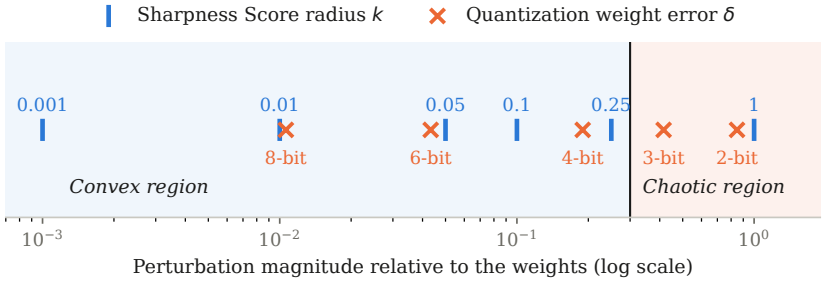


Figure 5: The Sharpness Score radii k and the mean relative weight error δ of each quantization bit-width, on one axis of perturbation magnitude relative to the weights. The rule marks convex/chaotic boundary of Figure 4a.

6 Conclusions

Inspired by common assumptions about loss landscape characteristics, we proposed and investigated sample-level minima sharpness as a scoring function for sample difficulty estimation in the context of curriculum learning. We showed that the sharpness-based difficulty scores correlate to some extent with other established scoring function, indicating that samples with sharper minima for a given converged model state are indeed more difficult than those with flat minima, but that sharpness provides a distinct measure of difficulty. This is also underlined by the fact that curriculum learning with sharpness-based orderings outperform random orderings, even though they do not reach a curriculum-less performance. Effects of randomness in the choice of projection directions mostly disappeared with ensembles of projection directions. Moreover, we observed that different Euclidean ball radii for sharpness estimations led to different regimes of difficulty ordering, presumably shifting from more convex basins to a more chaotic regions loss landscape of the samples. Similar effects could be observed for the quantization of model parameters. Overall, we conclude that sharpness reveals insightful characteristics w. r. t. the loss landscape of individual samples in the context of difficulty estimation and curriculum learning.

AI USE STATEMENT

In this work, we used generative AI tools (Claude, Claude Code and OSS models) to help implement methods: we asked them questions about how third-party frameworks and libraries implement specific functionality, and had them write code that computes the equations of Section 3, speed up experiments through vectorization, parallelization and better use of GPU memory, modify existing code, and generate and modify visualizations. Each of these tasks was specified by the authors, and the tools did not run experiments or make changes on their own. We have not used generative AI tools to propose hypotheses, to design the methodology or experiments, to formulate or prove mathematical claims, or to interpret results; generating synthetic data, translation and dataset cleaning are not applicable to this work. Additionally, we used generative AI tools to polish the writing of the paper and to search for and identify related literature. We have reviewed all AI-assisted work: the authors checked and tested all AI-written code, verified every reported number against the experiment outputs, and checked every suggested reference against the original publication. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All datasets used in this work are publicly available: CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), CIFAR-10H (Peterson et al., 2019) and MVT (Mayo et al., 2023). Appendix A describes how we use them, including the MVT split, and Appendix B lists the training settings of all reference and curriculum models. Our code for scoring, curriculum training and quantization, together with the configuration files of all experiments, is available at `https://anonymous.4open.science/r/curriculum-learning-7302/`. It builds on the open-source `autrainer` (Rampp et al., 2024b) and `aucurriculum` (Rampp et al., 2024a) frameworks. The random directions are drawn with fixed seeds (1 to D), and all curriculum runs use the same fixed seed.

REFERENCES

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Robert J. N. Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10876–10889, 2021.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 9448–9458, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/97af07a14cacba681feacf3012730892-Abstract.html`.

Sumit Basu and Janara Christensen. Teaching classification boundaries to humans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pp. 109–115, 2013.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 41–48, 2009.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848.

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 (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017.

Jeffrey L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993. doi: 10.1016/0010-0277(93)90058-4.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2019.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*, 2021.

Ian Goodfellow. Efficient per-example gradient computations. *arXiv preprint arXiv:1510.01799*, 2015.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2535–2544. PMLR, 2019.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *4th International Conference on Learning Representations (ICLR)*, 2016.

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

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997. doi: 10.1162/neco.1997.9.1.1.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2704–2713. IEEE, 2018. doi: 10.1109/CVPR.2018.00286.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C. Mozer. Characterizing structural regularities of labeled data in overparameterized models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5034–5044. PMLR, 2021.

H. Toprak Kesgin and M. Fatih Amasyali. Development and comparison of scoring functions in curriculum learning. *arXiv preprint arXiv:2202.06823*, 2022.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Faisal Khan, Bilge Mutlu, and Xiaojin Zhu. How do humans teach: On curriculum learning and teaching dimension. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5905–5914. PMLR, 2021.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *6th International Conference on Learning Representations (ICLR)*, 2018a.

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 (NeurIPS) 31*, pp. 6389–6399, 2018b.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*, 2024.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

David Mayo, Jesse Cummings, Xinyu Lin, Dan Gutfreund, Boris Katz, and Andrei Barbu. How hard are computer vision datasets? calibrating dataset difficulty to viewing time. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11008–11036. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2432e9646556f3a98dc78c1f4a10481b-Paper-Datasets_and_Benchmarks.pdf.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT)*, pp. 164–170. ACM, 1999.

Kristof Meding, Luca M. Schulze Buschoff, Robert Geirhos, and Felix A. Wichmann. Trivial or impossible – dichotomous data difficulty masks model differences (on ImageNet and beyond). In *10th International Conference on Learning Representations (ICLR)*, 2022.

Manuel Milling, Andreas Triantafyllopoulos, Iosif Tsangko, Simon David Noel Rampp, and Björn Wolfgang Schuller. Bringing the discussion of minima sharpness to the audio domain: A filter-normalised evaluation for acoustic scene classification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 391–395. IEEE, 2024.

Manuel Milling, Andreas Triantafyllopoulos, Simon Rampp, Alican Akman, and Björn W. Schuller. Leveraging sample difficulty in computer audition analysis. *IEEE Access*, 14:67756–67764, 2026. doi: 10.1109/ACCESS.2026.3688974.

Clara Na, Sanket Vaibhav Mehta, and Emma Strubell. Train flat, then compress: Sharpness-aware minimization learns more compressible models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4909–4936, 2022.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20596–20607, 2021.

Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Computation*, 6(1):147–160, 1994. doi: 10.1162/neco.1994.6.1.147.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9617–9626, 2019.

Simon Rampp, Manuel Milling, Andreas Triantafyllopoulos, and Björn W. Schuller. Does the definition of difficulty matter? Scoring functions and their role for curriculum learning. *arXiv preprint arXiv:2411.00973*, 2024a.

Simon Rampp, Andreas Triantafyllopoulos, Manuel Milling, and Björn W. Schuller. autrainer: A modular and extensible deep learning toolkit for computer audition tasks. *arXiv preprint arXiv:2412.11943*, 2024b.

Levent Sagun, Léon Bottou, and Yann LeCun. Eigenvalues of the Hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

Levent Sagun, Utku Evci, V. Uğur Güney, Yann Dauphin, and Léon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Chawin Sitawarin, Supriyo Chakraborty, and David Wagner. SAT: Improving adversarial training via curriculum-based loss smoothing. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, pp. 25–36, 2021.

Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*, 2017. URL `http://arxiv.org/abs/1708.07120`.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Otilia Stretcu. *Curriculum Learning*. Phd thesis, Carnegie Mellon University, Pittsburgh, PA, 2021.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 1139–1147, 2013.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, 2020.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 2019.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *7th International Conference on Learning Representations (ICLR)*, 2019.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2022. doi: 10.1109/TPAMI.2021.3069908.

Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. When do curricula work? In *International Conference on Learning Representations (ICLR)*, 2021.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016. doi: 10.5244/C.30.87.

Yang Zhang, Amr Mohamed, Hadi Abdine, Guokan Shang, and Michalis Vazirgiannis. Beyond random sampling: Efficient language model pretraining via curriculum learning. *arXiv preprint arXiv:2506.11300*, 2025.

Tianyi Zhou, Shengjie Wang, and Jeffrey Bilmes. Curriculum learning by dynamic instance hardness. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8602–8613, 2020.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C. Dvornek, Sekhar Tatikonda, James S. Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *10th International Conference on Learning Representations (ICLR)*, 2022.

A Extended dataset information

The MVT dataset (Mayo et al., 2023) contains $5{,}000$ images over the 50 classes shared by ImageNet (Deng et al., 2009) and ObjectNet (Barbu et al., 2019), each labeled by how often human subjects misidentified it across six viewing times; we use a stratified $80/10/10$ split ($4{,}000/500/500$ images). CIFAR-10H (Peterson et al., 2019) provides on average 51 human labels for every image of the CIFAR-10 test set, and we take the Shannon entropy of each image's response distribution as its human difficulty.

B Experimental Setting

For CIFAR-10 and CIFAR-100 this is an EfficientNet-B0 (Tan & Le, 2019) trained without a curriculum using Adam (Kingma & Ba, 2015) (learning rate 10^{-3}, batch size 128, 100 epochs). For the MVT dataset it is a ResNet-50 (He et al., 2016) fine-tuned from ImageNet-1K weights with SAM and a one-cycle schedule (Smith & Topin, 2017), reaching 86.4% development accuracy.

All curriculum runs train a WideResNet-28-4 (Zagoruyko & Komodakis, 2016) on CIFAR-100 with SAM ($\rho = 0.05$) over SGD with Nesterov momentum (Sutskever et al., 2013) 0.9 and weight decay 5×10^{-4}, learning rate 0.1 annealed by cosine schedule (Loshchilov & Hutter, 2017), batch size 128, and $78{,}200$ steps. Curriculum pacing (Wu et al., 2021) is linear: the eligible pool grows from 10% to 90% of the training set over the run, and mini-batches are drawn uniformly from the current pool, so examples admitted early are seen more often. The primary baseline is a *Random Curriculum*: identical pacing schedule with a random ordering, which isolates the contribution of the difficulty signal from that of the growing pool of samples (Wu et al., 2021). We also report standard, uniformly-shuffled training of the same WideResNet-28-4 under SGD and SAM, and of a WideResNet-28-10 under SGD, as a reference point outside the curriculum regime.

C Additional Results

This appendix collects further figures from the underlying experiments. Unless stated otherwise, all Sharpness Scores are computed on the CIFAR-100/EfficientNet-B0 reference checkpoint with $D = 50$ directions.

C.1 Random-Direction Sharpness Score

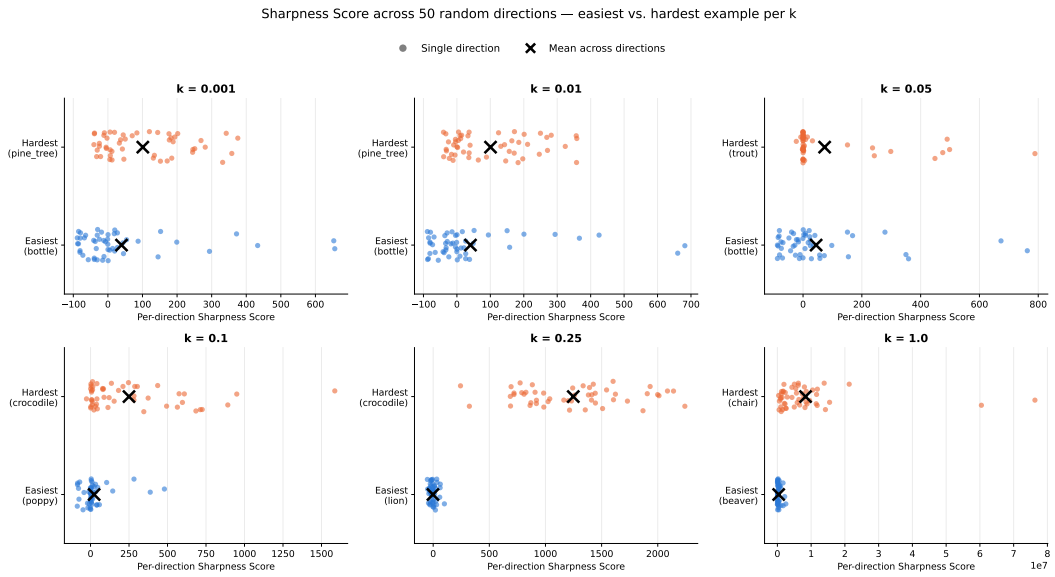


Figure 6: Score of each of the 50 individual directions (dots) and their mean ($\times$) for the example ranked easiest and the example ranked hardest, at each radius k. Single directions scatter widely around the mean, which is why we average over many.

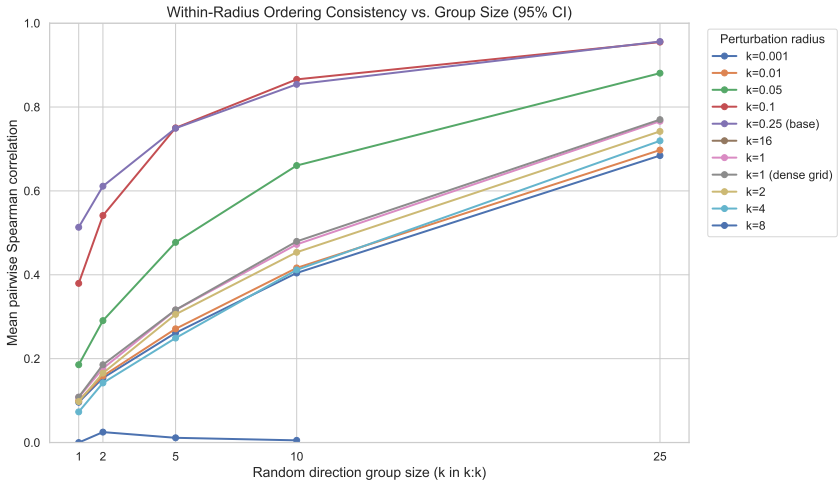


Figure 7: Agreement between independently averaged ensembles of directions against ensemble size, separately for every radius k (Figure 2 shows $k = 0.25$ only). A single direction is most informative at $k = 0.25$; at larger radii even large ensembles agree poorly.

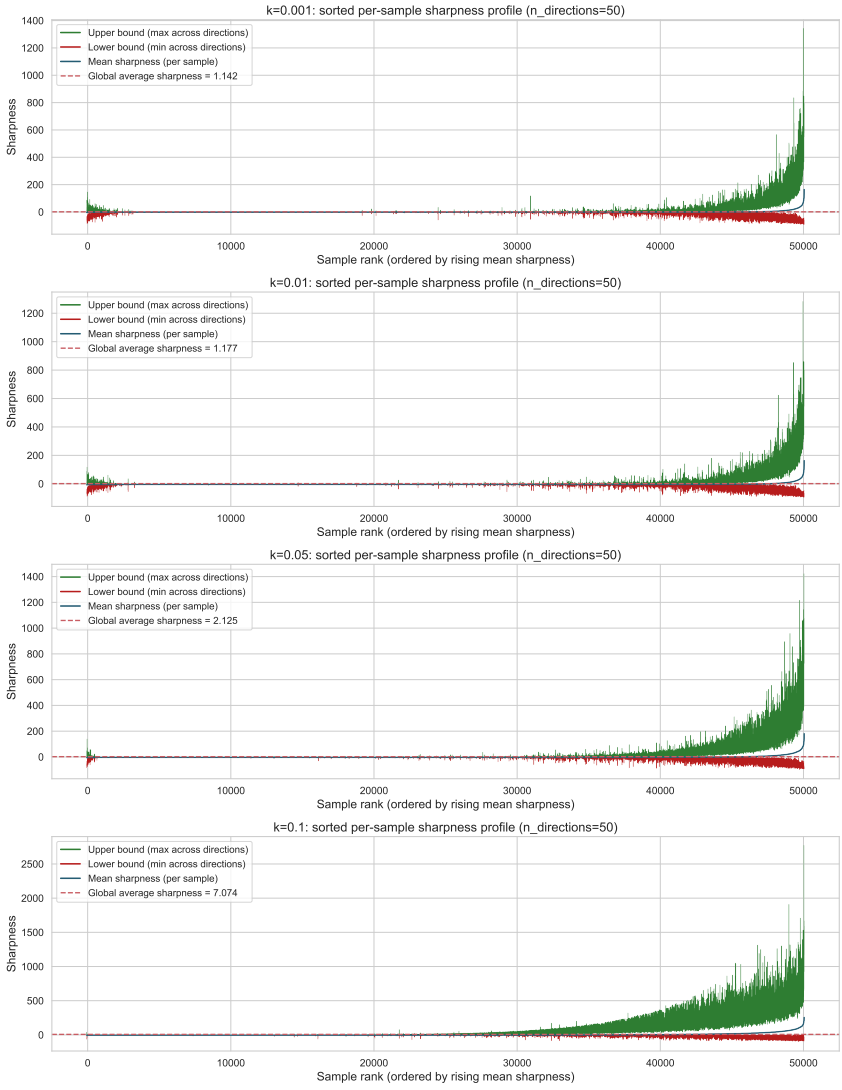


Figure 8: Per-example Sharpness Score at the four smallest radii, examples sorted by their mean score. The band shows the maximum and minimum over directions, the line the mean over directions.

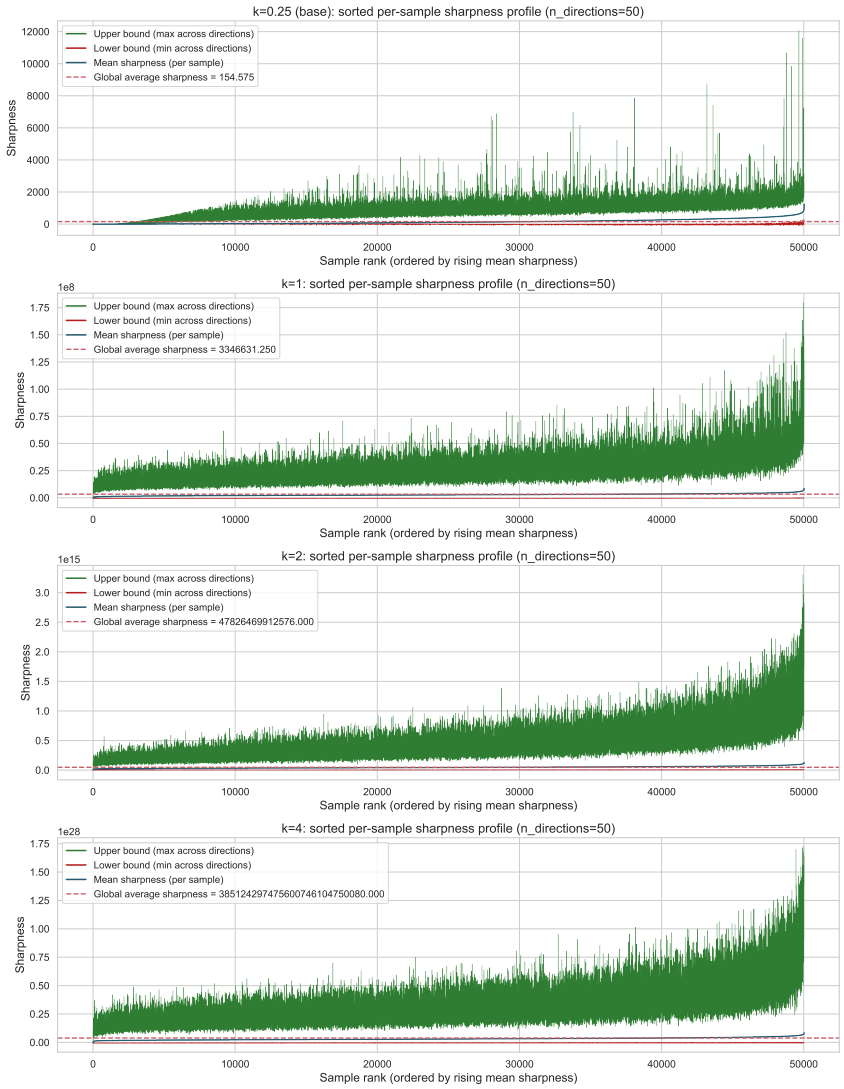


Figure 9: The same profile at $k = 0.25$ and at the larger radii $k \in \{1, 2, 4\}$. The shape of the distribution stays the same, but its scale grows by many orders of magnitude once $k \geq 1$ (Section 5.4).

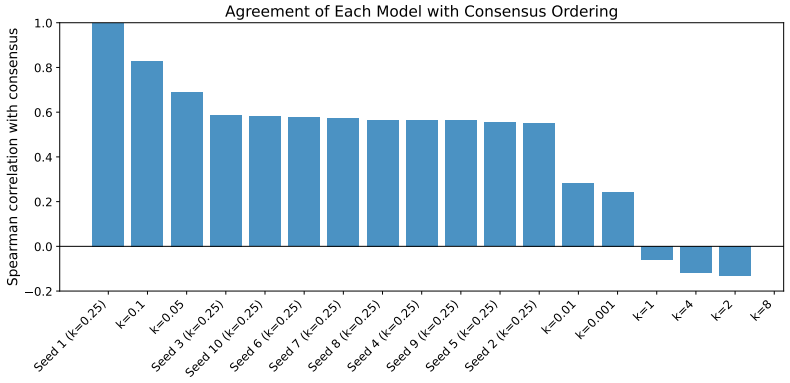


Figure 10: Agreement of other orderings with the reference ordering (seed 1, $k = 0.25$): the same score on nine independently trained seeds, and other radii on the same checkpoint. Other seeds agree moderately, nearby small radii agree strongly, and large radii ($k \geq 1$) not at all.

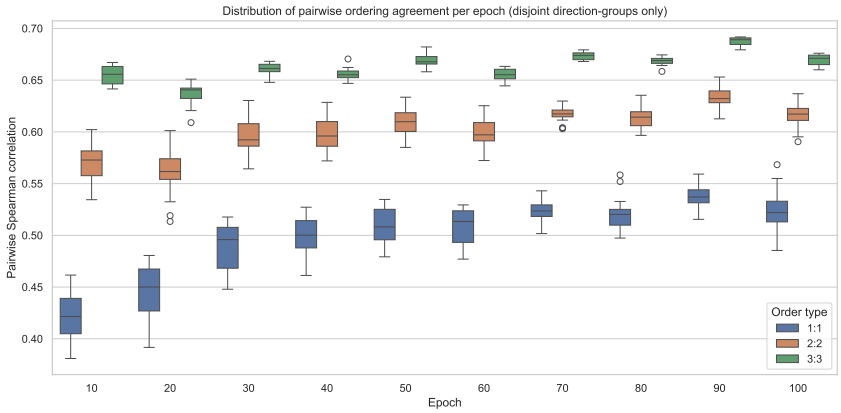


Figure 11: Agreement between disjoint groups of 1, 2 or 3 directions at the same checkpoint, for checkpoints saved every 10 epochs during training of a second EfficientNet-B0 (6 directions per checkpoint). Agreement rises with training and with the number of directions per group.

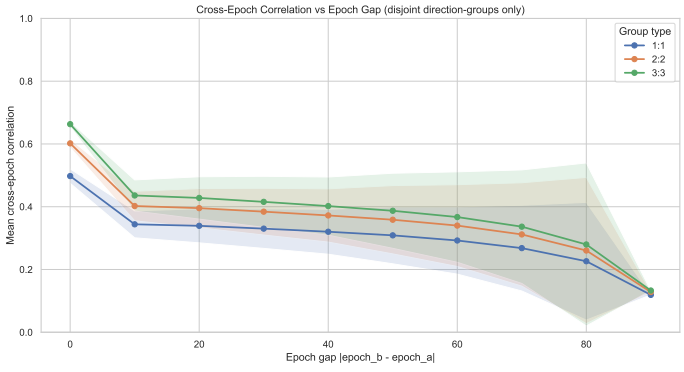


Figure 12: Agreement between orderings computed at two different epochs of the same training run, against the gap between the epochs. The ordering changes considerably during training, so the reference model should be fully trained.

C.2 Example Images and Loss Landscapes

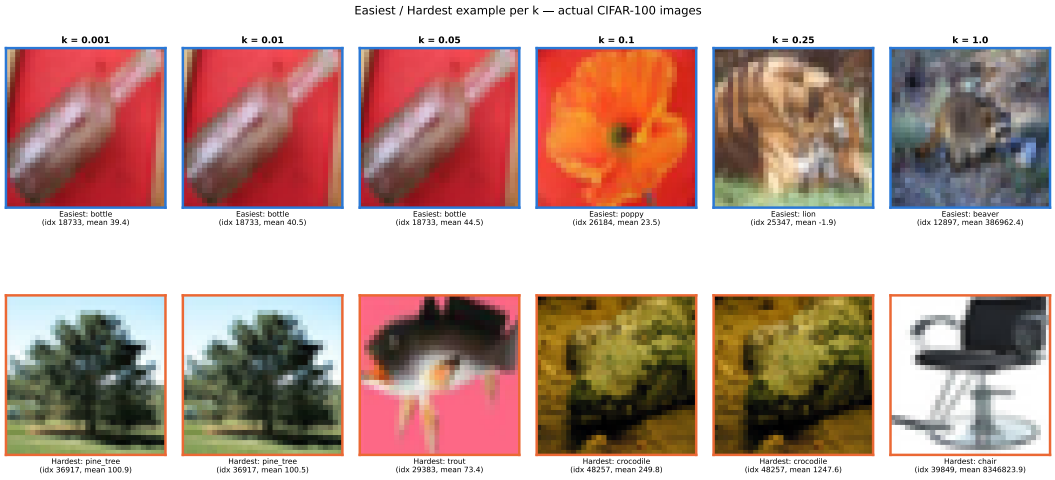


Figure 13: The CIFAR-100 training images ranked easiest (top) and hardest (bottom) by the Sharpness Score at each radius k.

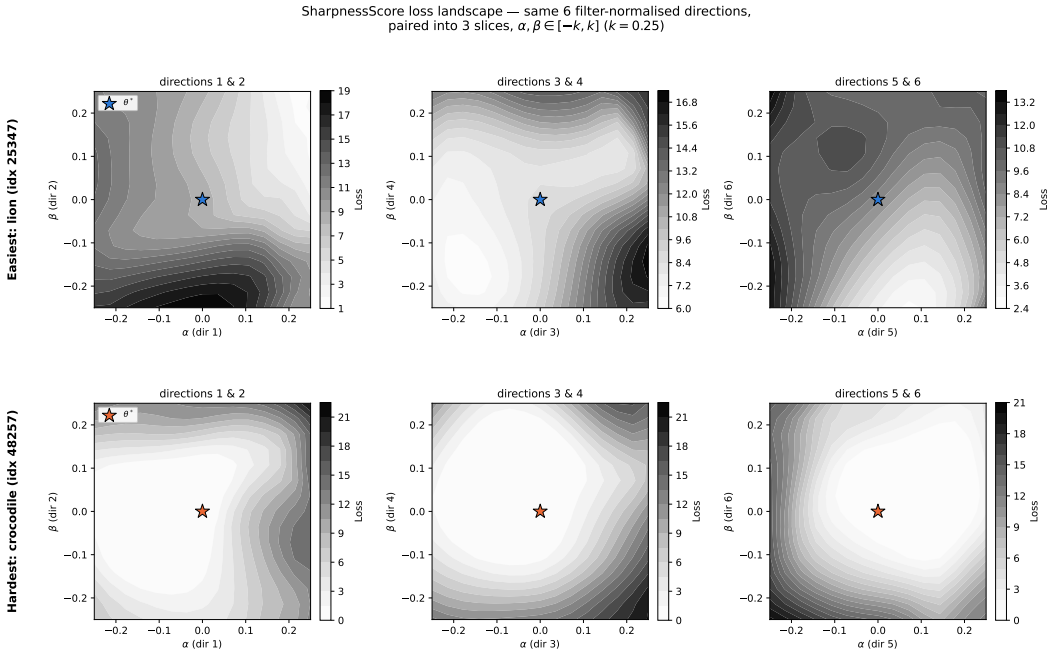


Figure 14: Two-dimensional loss slices around $\theta^\star$ (star) for the easiest (top) and hardest (bottom) example at $k = 0.25$, each spanned by a pair of the random directions used for scoring. Each panel has its own color scale.

C.3 GRADIENT-GUIDED SHARPNESS SCORE

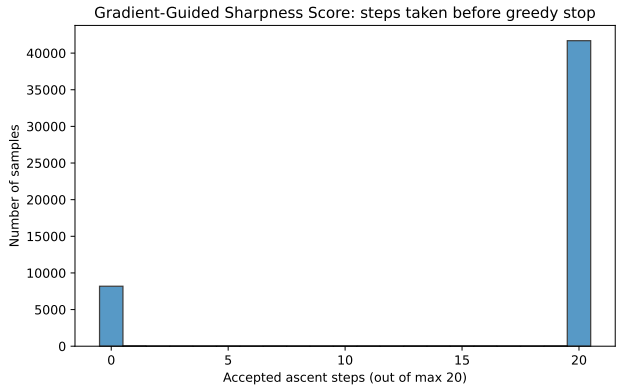


Figure 15: Number of accepted ascent steps per example ($T = 20$). Almost every example either accepts all steps or stops at the first one.

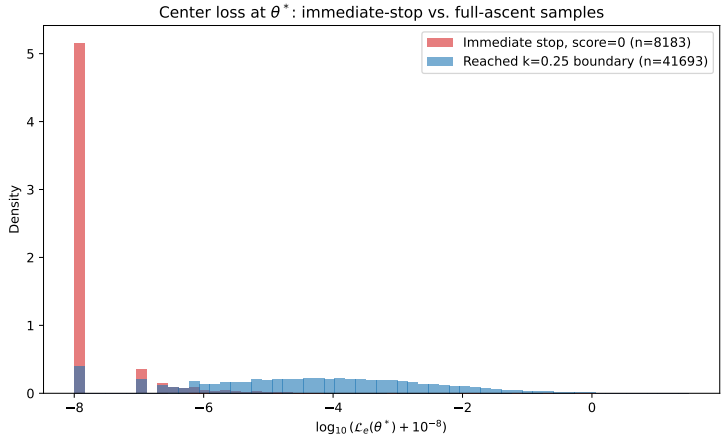


Figure 16: Converged loss $\mathcal{L}_e(\theta^\star)$ of the examples that stop at the first step (score 0) and of those that reach the radius k. Examples that stop immediately are fitted almost perfectly, so their gradient is too small to provide an ascent direction.

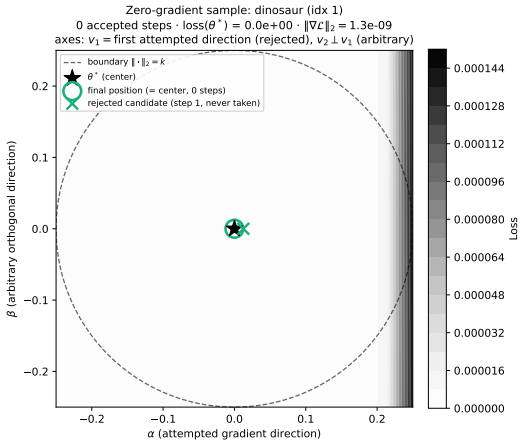


Figure 17: An example whose first step already fails to increase the loss, so the walk stays at $\theta^\star$ and the score is 0. The horizontal axis is the rejected gradient direction, the vertical axis an arbitrary orthogonal direction.

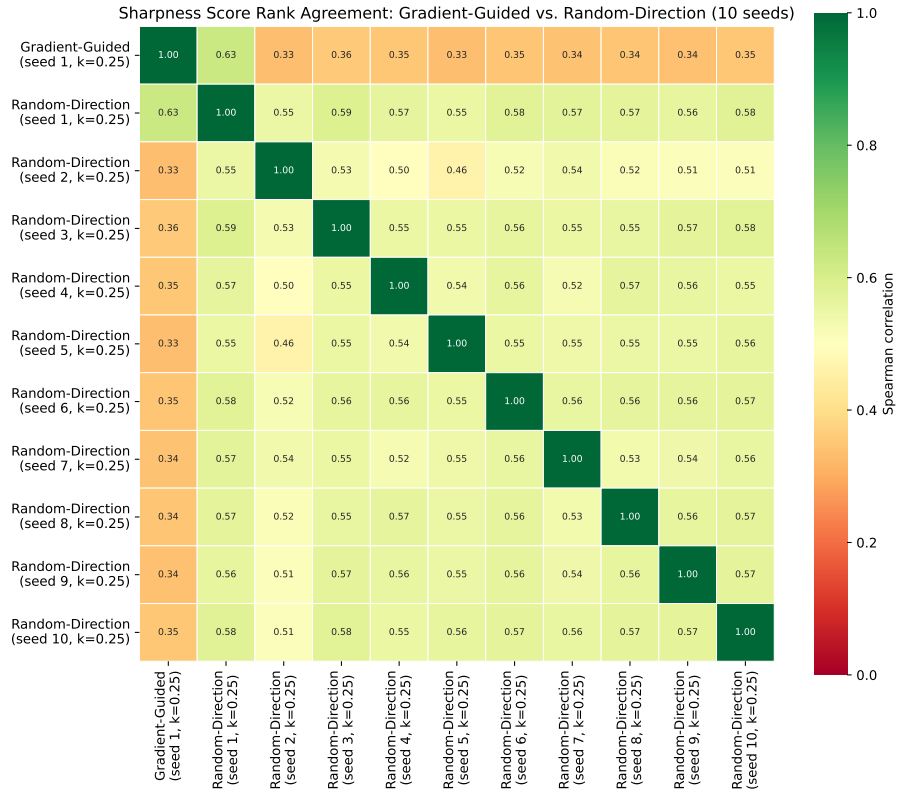


Figure 18: Spearman correlation between the gradient-guided ordering (seed 1) and the random-direction orderings of ten independently trained seeds, all at $k = 0.25$ (Section 5.1).

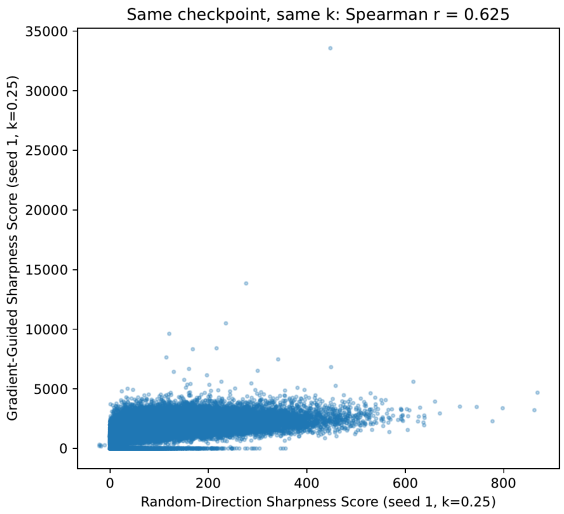


Figure 19: Gradient-guided against random-direction score for every training example on the same checkpoint. The two scores live on different scales, so we compare them only by rank.

C.4 CURRICULUM TRAINING

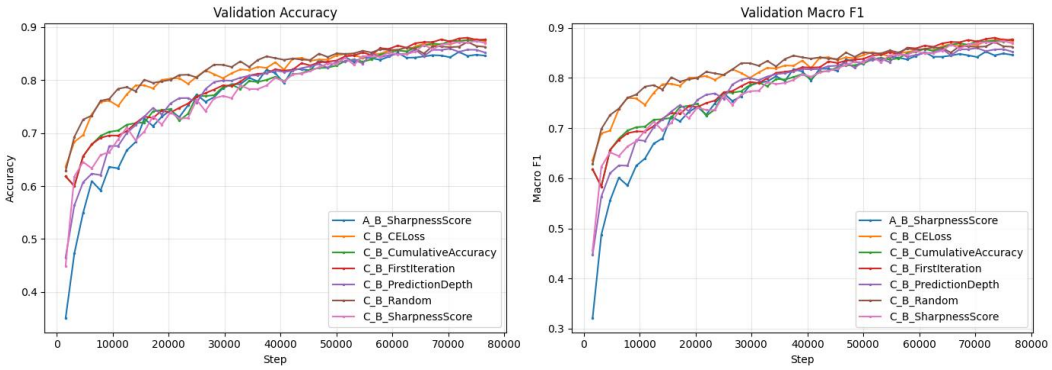


Figure 20: Validation accuracy (left) and macro F1 (right) during training for the four best established scoring functions, the Random Curriculum, and the Sharpness Score as curriculum and anti-curriculum. The anti-curriculum falls behind early and does not fully recover.

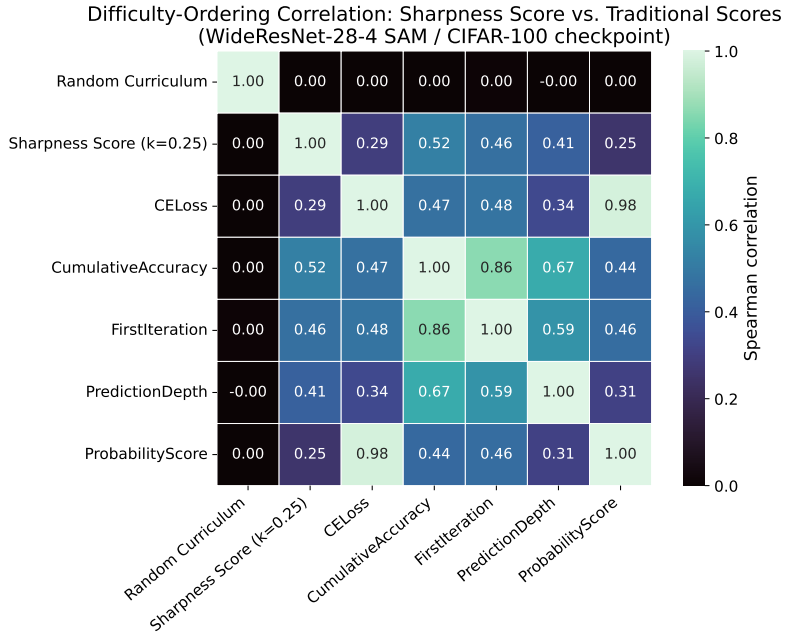


Figure 21: Spearman correlation between the difficulty orderings of the Sharpness Score ($k = 0.25$), the five established scoring functions and the Random Curriculum, all computed on the WideResNet-28-4 (SAM) baseline.

C.5 MVT DATASET

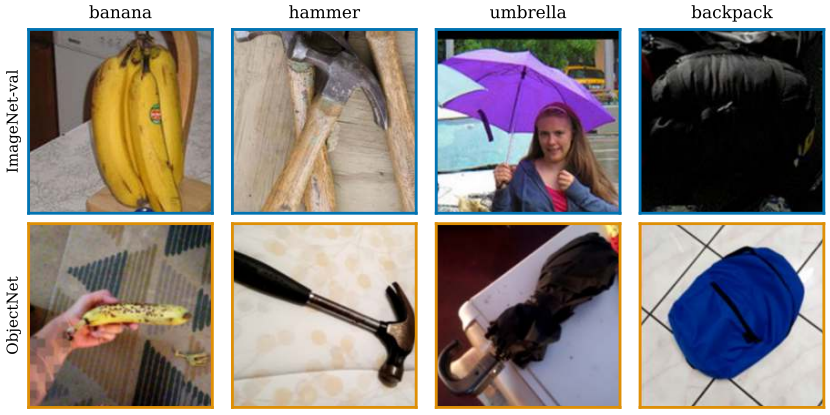


Figure 22: Example images from four of the 50 MVT classes: web-sourced ImageNet images (top) and crowd-sourced ObjectNet images (bottom).

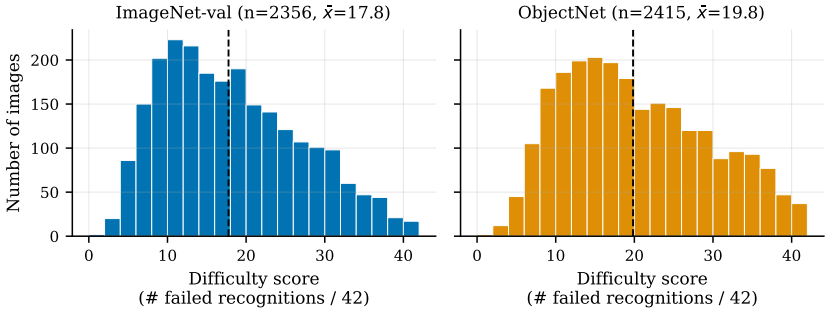


Figure 23: Distribution of the human difficulty score (failed recognitions out of 42 judgments) for the ImageNet and ObjectNet images of MVT; dashed lines mark the means.

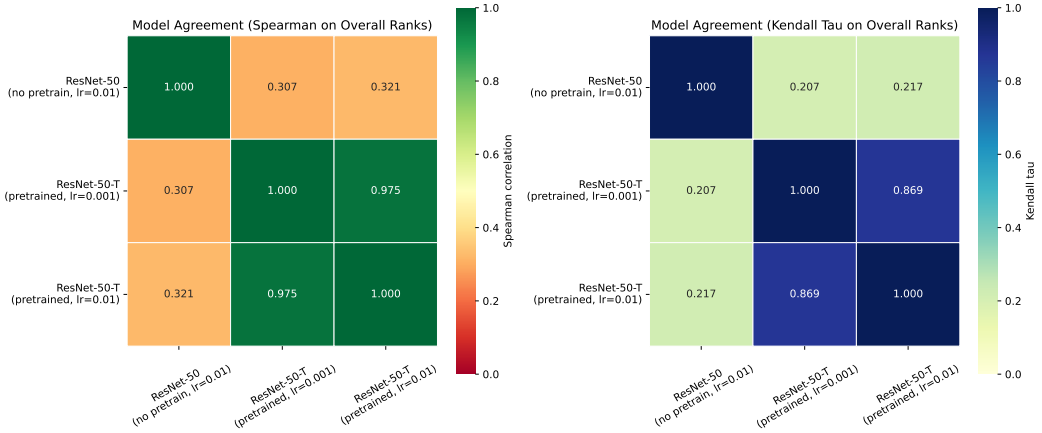


Figure 24: Agreement of Sharpness Score orderings ($k = 0.25$) on MVT between three ResNet-50 checkpoints: two fine-tuned from ImageNet weights and one trained from scratch that failed to generalize. *Left:* Spearman correlation; *right:* Kendall τ. The two fine-tuned checkpoints agree almost perfectly, while the from-scratch checkpoint agrees only weakly with either.