A Modern Look at the Relationship between Sharpness and Generalization

Maksym Andriushchenko [1] **Francesco Croce** [2 3] **Maximilian Müller** [2 3] **Matthias Hein** [2 3] **Nicolas Flammarion** [1]

Abstract

Sharpness of minima is a promising quantity that can positively correlate with test error for deep networks and, when optimized during training, can improve generalization. However, standard sharpness is not invariant under reparametrizations of neural networks, and, to fix this, reparametrization-invariant sharpness definitions have been proposed, most prominently *adaptive sharpness* (Kwon et al., 2021). But does it really capture generalization in modern practical settings? We comprehensively explore this question in a detailed study of various definitions of adaptive sharpness in settings ranging from training from scratch on ImageNet and CIFAR-10 to fine-tuning CLIP on ImageNet and BERT on MNLI. We focus mostly on *transformers* for which little is known in terms of sharpness despite their widespread usage. Overall, we observe that sharpness *does not correlate well* with generalization but rather with some training parameters like the learning rate that can be positively or negatively correlated with generalization depending on the setup. Interestingly, in multiple cases, we observe a consistent *negative* correlation of sharpness with out-of-distribution error implying that *sharper* minima can generalize *better*. Finally, we illustrate on a simple model that the right sharpness measure is highly data-dependent, and that we do not understand well this aspect for realistic data distributions. Our code is available at `https://github.com/tml-epfl/sharpness-vs-generalization`.

1. Introduction

Considering the sharpness of the training objective at a minimum has intuitive appeal: if the loss surface is slightly perturbed due to a train vs. test or out-of-distribution (OOD) discrepancy, flat minima of deep networks should still have low loss (Hochreiter & Schmidhuber, 1995; Keskar et al., 2016). On the theoretical side, sharpness appears in generalization bounds (Neyshabur et al., 2017; Dziugaite & Roy, 2018; Foret et al., 2021) but this fact alone is not necessarily informative for practical settings. For example, quantities like the VC-dimension typically correlate *negatively* with generalization contrary to what the generalization bound might suggest (Jiang et al., 2020). Importantly, it has been shown empirically that sharpness can also correlate well with generalization in common deep learning setups (Keskar et al., 2016; Jiang et al., 2020) which makes it a promising generalization measure that can potentially distinguish well-generalizing solutions. Additionally, empirical success of training methods that minimize sharpness such as sharpness-aware minimization (SAM) (Zheng et al., 2021; Wu et al., 2020; Foret et al., 2021) further suggests that sharpness can be an important quantity for generalization.

Motivation: why revisiting sharpness? Many works imply or conjecture that flatter minima should generalize better (Xing et al., 2018; Zhou et al., 2020; Cha et al., 2021; Park & Kim, 2022; Lyu et al., 2022) for standard or OOD data. However, standard sharpness definitions do not correlate well with generalization (Jiang et al., 2020; Kaur et al., 2022) which can be partially due to their lack of invariance under reparametrizations that leave the model unchanged (Dinh et al., 2017; Granziol, 2020; Zhang et al., 2021). Adaptive sharpness appears to be more promising since it fixes the reparametrization issue and is shown to empirically correlate better with generalization (Kwon et al., 2021). However, the empirical evidence in Kwon et al. (2021) and other works that discuss sharpness (Keskar et al., 2016; Jiang et al., 2020; Dziugaite et al., 2020; Bisla et al., 2022) is restricted to small datasets like CIFAR-10 or SVHN. In addition, SAM appears to be particularly useful for new architectures like vision transformers (Chen et al., 2022) for which there has been no systematic studies of sharpness vs. generalization. Moreover, transfer learning is becoming the default option for vision and language tasks but not much is known about sharpness there. Finally, the relationship between sharpness and OOD generalization is also underexplored. These new developments motivate us to revisit the role of sharpness in these new settings.

[1]EPFL [2]Tübingen AI Center [3]University of Tübingen. Correspondence to: Maksym Andriushchenko <maksym.andriushchenko@epfl.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Contributions. We aim to provide a comprehensive study focusing specifically on adaptive sharpness in order to answer the following fundamental question:

Can reparametrization-invariant sharpness capture generalization in modern practical settings?

Towards this goal, we make the following contributions:

- We provide extensive evaluations of multiple reparametrization-invariant sharpness measures for (1) training from scratch on ImageNet and CIFAR-10 using transformers and ConvNets, and (2) fine-tuning CLIP and BERT transformers on ImageNet and MNLI.
- We observe that sharpness *does not correlate well* with generalization but rather with some training parameters like the learning rate which can be positively or negatively correlated with generalization depending on the setup.
- Interestingly, in multiple cases, we observe a consistent *negative* correlation of sharpness with OOD generalization implying that *sharper* minima can generalize *better*.
- Finally, we provide an analysis on a simple model where we know the measure responsible for generalization. Our analysis suggests that (1) different sharpness definitions can capture totally different trends, and (2) the right sharpness measure is highly *data-dependent*.

2. Related work

Here we discuss the most related papers to our work.

Systematic studies on sharpness vs. generalization. The seminal work of Keskar et al. (2016) shows that the performance degradation of large-batch SGD (LeCun et al., 2012) is correlated with sharpness of minima. Neyshabur et al. (2017) explore different generalization measure that may explain generalization for deep networks suggesting that sharpness can be a promising measure. Jiang et al. (2020) perform a systematic study that shows a strong correlation between sharpness and generalization on a large set of CIFAR-10/SVHN models trained with many different hyperparameters. Their experimental protocol is, however, criticized in Dziugaite et al. (2020) since it can obscure failures of generalization measures and instead should be evaluated within the framework of distributional robustness. Vedantam et al. (2021) discuss OOD generalization on small datasets and evaluate a definition of sharpness which, however, does not correlate well with OOD generalization. Stutz et al. (2021) study the relationship between sharpness and generalization under ℓ_p-bounded adversarial perturbations. Andriushchenko & Flammarion (2022) study reasons behind the success of SAM and highlight the importance of using sharpness computed on a small subset of training points. Kaur et al. (2022) discuss that the maximum eigenvalue of the Hessian is not always predictive to generalization even for models obtained via standard training methods.

Reparametrization-invariant sharpness definitions. The magnitude-aware sharpness of Keskar et al. (2016) mitigates but does not completely resolve reparametrization invariance. Liang et al. (2019) consider the Fisher-Rao metric related to sharpness and invariant to network reparametrization. Petzka et al. (2021) propose a sharpness measure based on the trace of the Hessian and show correlation for a small ConvNet on CIFAR-10. Tsuzuku et al. (2020) suggest to use a specifically rescaled sharpness inspired by the PAC-Bayes theory and report high correlation with generalization for ResNets on CIFAR-10. Most importantly for our work, Kwon et al. (2021) introduce adaptive sharpness which is reparametrization invariant, correlates well with generalization, and generalizes multiple existing sharpness definitions.

Explicit and implicit sharpness minimization. The idea that flat minima can be beneficial for generalization dates back to Hochreiter & Schmidhuber (1995) and inspires multiple methods that optimize for more robust minima. These methods optimize different criteria ranging from random perturbations such as dropout (Srivastava et al., 2014) and Entropy-SGD (Chaudhari et al., 2016) to worst-case perturbations such as SAM (Foret et al., 2021) and its variations (Kwon et al., 2021; Zhuang et al., 2022; Du et al., 2022). Notably, Chen et al. (2022) suggest that SAM is particularly helpful for vision transformers on ImageNet scale and that standard transformers by default converge to very sharp minima. Concurrently, works on the implicit bias of SGD suggest *implicit* minimization of some hidden complexity measures related to flatness of minima (Keskar et al., 2016; Smith & Le, 2018; Xing et al., 2018). Izmailov et al. (2018) propose to average weights during SGD to improve generalization and motivate it by sharpness reduction. Smith et al. (2021) derive an implicit regularization term of SGD based on the gradient norm. Sharpness-related quantities based on the Hessian have been a focus of many recent works. E.g., Cohen et al. (2021); Arora et al. (2022); Damian et al. (2023) empirically and theoretically characterize the regime of full-batch gradient descent where the maximum eigenvalue of the Hessian becomes inversely proportional to the learning rate used for training. Blanc et al. (2020); Li et al. (2021); Damian et al. (2021) discover implicit minimization of the trace of the Hessian for label-noise SGD used as a proxy of standard SGD. The common theme behind these works is a focus on sharpness-related metrics as a tool to better understand generalization for deep networks.

3. Adaptive Sharpness, its Invariances, and Computation

In this section, we first provide background on adaptive sharpness, then discuss its invariance properties for modern

architectures, and propose a way to compute worst-case sharpness efficiently.

3.1. Background on Sharpness

Sharpness definitions. We denote the loss on a set of *training* points $\mathcal{S}$ as $L_{\mathcal{S}}(\boldsymbol{w}) = \frac{1}{|S|}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{S}} \ell_{\boldsymbol{xy}}(\boldsymbol{w})$, where $\ell_{\boldsymbol{xy}}(\boldsymbol{w}) \in \mathbb{R}_+$ represents some loss function (e.g., cross-entropy) on the training pair $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{S}$ computed with the network weights $\boldsymbol{w}$. For arbitrary $\boldsymbol{w} \in \mathbb{R}^p$ (i.e., not necessarily a minimum), we define the *adaptive average-case* and *adaptive worst-case m-sharpness* with radius ρ and with respect to a vector $\boldsymbol{c} \in \mathbb{R}^p$ as:

$$S^{\rho}_{avg}(\boldsymbol{w}, \boldsymbol{c}) \triangleq \mathbb{E}_{\substack{\mathcal{S}\sim P_m \\ \boldsymbol{\delta}\sim\mathcal{N}(0,\rho^2 diag(\boldsymbol{c}^2))}} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w}), \quad (1)$$

$$S^{\rho}_{max}(\boldsymbol{w}, \boldsymbol{c}) \triangleq \mathbb{E}_{\mathcal{S}\sim P_m} \max_{\|\boldsymbol{\delta}\odot\boldsymbol{c}^{-1}\|_p\leq\rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w}),$$

where $\odot/^{-1}$ denotes elementwise multiplication/inversion and P_m is the data distribution that returns m training pairs $(\boldsymbol{x}, \boldsymbol{y})$. Both average-case and worst-case sharpness have often been considered in the literature, and worst-case sharpness is mostly determined to correlate better with generalization (Jiang et al., 2020; Dziugaite et al., 2020; Kwon et al., 2021), especially with a small m (i.e., $|\mathcal{S}|$) in worst-case sharpness (Foret et al., 2021). Using $\boldsymbol{c} = |\boldsymbol{w}|$ leads to *elementwise* adaptive sharpness (Kwon et al., 2021) and makes the sharpness invariant under multiplicative reparametrizations that preserve the network, i.e., for any $\boldsymbol{c} \in \mathbb{R}^p$ such that $f(\boldsymbol{w} \odot \boldsymbol{c}) = f(\boldsymbol{w})$ we have:

$$\begin{aligned}
&S^{\rho}_{max}(\boldsymbol{w}\odot\boldsymbol{c}, |\boldsymbol{w}\odot\boldsymbol{c}|) = \\
&\mathbb{E}_{\mathcal{S}} \max_{\|\boldsymbol{\delta}\odot(|\boldsymbol{w}|\odot\boldsymbol{c})^{-1}\|_p\leq\rho} L_{\mathcal{S}}(\boldsymbol{w}\odot\boldsymbol{c}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w}\odot\boldsymbol{c}) = \\
&\mathbb{E}_{\mathcal{S}} \max_{\|\boldsymbol{\delta}'\odot|\boldsymbol{w}|^{-1}\|_p\leq\rho} L_{\mathcal{S}}((\boldsymbol{w}+\boldsymbol{\delta}')\odot\boldsymbol{c}) - L_{\mathcal{S}}(\boldsymbol{w}\odot\boldsymbol{c}) = \\
&\mathbb{E}_{\mathcal{S}} \max_{\|\boldsymbol{\delta}'\odot|\boldsymbol{w}|^{-1}\|_p\leq\rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}') - L_{\mathcal{S}}(\boldsymbol{w}) = S^{\rho}_{max}(\boldsymbol{w}, |\boldsymbol{w}|),
\end{aligned}$$

where we used the substitution $\boldsymbol{\delta}' := \boldsymbol{\delta} \odot \boldsymbol{c}^{-1}$. Similarly, one can show that $S^{\rho}_{avg}(\boldsymbol{w}\odot\boldsymbol{c}, |\boldsymbol{w}\odot\boldsymbol{c}|) = S^{\rho}_{avg}(\boldsymbol{w}, |\boldsymbol{w}|)$. Thus, this illustrates that *the criticism of sharpness stated in Dinh et al. (2017) does not apply to adaptive sharpness*, and there is no need to "balance" the network in a pre-processing step like, e.g., done in Bisla et al. (2022).

Connections between different sharpness definitions. Here we generalize the analytical expressions of standard sharpness for radius $\rho \to 0$ that depend on the first- or second-order terms which are frequently used in the literature (Blanc et al., 2020; Tsuzuku et al., 2020; Li et al., 2021; Damian et al., 2021). For a thrice differentiable loss $L(\boldsymbol{w})$, the average-case elementwise adaptive sharpness can be computed as (see App. A.1 for proofs):

$$\begin{aligned}
S^{\rho}_{avg}(\boldsymbol{w}, |\boldsymbol{w}|) = &\mathbb{E}_{\mathcal{S}\sim P_m} \frac{\rho^2}{2} \mathrm{tr}(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w}) \odot |\boldsymbol{w}||\boldsymbol{w}|^{\top}) \\
&+ O(\rho^3). \quad (2)
\end{aligned}$$

We note that the first-order term cancels out completely and plays no role. This is not the case for worst-case adaptive sharpness where we get for $p = 2$ the following expression for every critical point that is not a local maximum:

$$\begin{aligned}
S^{\rho}_{max}(\boldsymbol{w}, |\boldsymbol{w}|) = &\mathbb{E}_{\mathcal{S}\sim P_m} \frac{\rho^2}{2} \lambda_{\max}(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w}) \odot |\boldsymbol{w}||\boldsymbol{w}|^{\top}) \\
&+ O(\rho^3), \quad (3)
\end{aligned}$$

otherwise the first-order term dominates and we get $\rho\,\mathbb{E}_{\mathcal{S}\sim P_m} \|\nabla L(\boldsymbol{w}) \odot |\boldsymbol{w}|\|_2$, which resembles the implicit gradient regularization of Smith et al. (2021). Thus, worst-case sharpness with a small radius captures different properties of the loss surface depending on whether $\boldsymbol{w}$ is close to a minimum or not. We make use of these quantities in the last section to discuss insights from simple models. For the experiments, however, we evaluate a range of ρ where the smallest ρ well-approximates the above quantities.

What do we expect sharpness to capture? We are looking for a sharpness measure that can be *predictive for generalization* meaning that it satisfies either of these two hypotheses:

- **Strong hypothesis**: sharpness is highly correlated with generalization suggesting a *possibility* of a causal relation.
- **Weak hypothesis**: models with the lowest sharpness generalize well suggesting that sharpness might be *sufficient but not necessary* for generalization.

To detect correlation, we follow the previous works by Jiang et al. (2020); Dziugaite et al. (2020); Kwon et al. (2021) and use the Kendall rank correlation coefficient:

$$\tau(\boldsymbol{t}, \boldsymbol{s}) = \frac{2}{M(M-1)} \sum_{i<j} \mathrm{sign}(t_i - t_j)\,\mathrm{sign}(s_i - s_j) \quad (4)$$

where $\boldsymbol{t}, \boldsymbol{s} \in \mathbb{R}^M$ are vectors of test error and sharpness values for M different models. We adopt a less demanding setting than in the previous works of Neyshabur et al. (2017); Jiang et al. (2020); Dziugaite et al. (2020), and only compare models *within the same loss surface* motivated by the geometric motivation behind sharpness. This restriction rules out comparing models with different architectures (including different width and depth) or measuring sharpness on a different set of points since both changes would change the loss surface. According to the same reason, we also do not consider the ability of sharpness to capture robustness to different amounts of noisy labels (unlike, e.g., Neyshabur et al. (2017)). We always evaluate sharpness on the *same*

training points taken without any data augmentations. Moreover, we always compare models trained with exactly the same training sets but, at the same time, we allow the usage of algorithmic techniques such as data augmentation or mixup for training.

3.2. Which Invariances Do We Need Sharpness to Capture for Modern Architectures?

Throughout the paper, we focus on *elementwise* adaptive sharpness which, as we show, satisfies the main reparametrization invariances for ResNets and ViTs. Let us denote $f_{\boldsymbol{w}} : \mathbb{R}^d \to \mathbb{R}^K$ a network with parameters $\boldsymbol{w}$, which returns the logits $f_{\boldsymbol{w}}(\boldsymbol{x}) \in \mathbb{R}^K$ for an input $\boldsymbol{x} \in \mathbb{R}^d$. By a reparametrization invariance we mean a function $T : \mathbb{R}^p \to \mathbb{R}^p$ such that for every $\boldsymbol{w} \in \mathbb{R}^p$ and $\boldsymbol{x} \in \mathbb{R}^d$ it holds $f_{\boldsymbol{w}}(\boldsymbol{x}) = f_{T(\boldsymbol{w})}(\boldsymbol{x})$. We briefly discuss here that adaptive sharpness also stays invariant for *modern* architectures like ResNets and ViTs involving normalization layers and self-attention. Finally, we discuss how to treat the scale-sensitivity of classification losses.

Adaptive sharpness for ResNets. A typical block of a pre-activation ResNet between skip connections includes the following sequence of operations: `BN`→`ReLU`→`conv`→`BN`→`ReLU`→`conv` where `BN` denotes BatchNorm. So we need to make sure that the sharpness definition we use is invariant to transformations that leave the network unchanged: (1) multiplication of the affine BatchNorm parameters by $\alpha \in \mathbb{R}_+$ and division of the subsequent convolutional parameters by the same α (since ReLU is positive one-homogeneous and $\text{ReLU}(\alpha z)/\alpha = \text{ReLU}(z)$), and (2) multiplying the convolutional layer by any $\alpha \in \mathbb{R}_+$ due to scale-invariance of the subsequent BatchNorm layer. Both multiplicative invariances are satisfied by elementwise adaptive sharpness since $S^{\rho}_{max}(\boldsymbol{w} \odot \boldsymbol{c}, |\boldsymbol{w} \odot \boldsymbol{c}|) = S^{\rho}_{max}(\boldsymbol{w}, |\boldsymbol{w}|)$ as shown above.

Adaptive sharpness for ViTs. A typical MLP block of ViTs contains the following operations: `LN`→`Linear`→`GELU`→`Linear` where `LN` denotes LayerNorm, and pre-softmax self-attention weights are computed as $ZW_QW_K^\top Z^\top$ where $Z \in \mathbb{R}^{P \times D}$ is the matrix of P D-dimensional tokens. The network thus has the following invariances to multiplication/division by α: (1) between `LN` and `Linear` in MLP, (2) between W_Q in W_K in self-attention, (3) between two `Linear` layers that have GELU in-between for which $\text{GELU}(\alpha z)/\alpha \approx \text{GELU}(z)$. Moreover, at the beginning of the network there is a part of the network which is invariant to the scale of the `Linear` layer (`Linear`→`LN`). Similarly to ResNets, all these invariances are multiplicative, so the argument about the invariance of elementwise adaptive sharpness is the same.

Scale-sensitivity for classification losses. However, adaptive sharpness remains sensitive to the *scale* of the classifier,

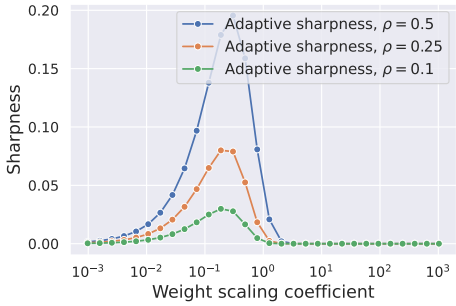


Figure 1: Sensitivity of adaptive sharpness to weight scaling for a linear model that achieves zero training error.

meaning that the sharpness together with the cross-entropy loss keep decreasing to zero after reaching zero training error. This can be seen even for linear models for which scaling the weight vector by a constant changes the adaptive sharpness as shown in Fig. 1. To fix this issue, Tsuzuku et al. (2020) propose to use normalization of the logits $f_{\boldsymbol{w}}$, i.e.:

$$\tilde{f}_{\boldsymbol{w}}(\boldsymbol{x}) \triangleq \frac{f_{\boldsymbol{w}}(\boldsymbol{x})}{\sqrt{\frac{1}{K}\sum_{i=1}^{K}(f_{\boldsymbol{w}}(\boldsymbol{x})_i - f_{avg}(\boldsymbol{x}))^2}}, \qquad (5)$$

where $f_{avg}(\boldsymbol{x}) = \frac{1}{K}\sum_{j=1}^{K} f_{\boldsymbol{w}}(\boldsymbol{x})_j$. This provably fixes the scaling issue meaning that scaling the output layer by $\alpha \in \mathbb{R}_+$ does not affect the logits. Moreover, this change can make models having different training loss more comparable to each other.

3.3. How to Compute Worst-Case Sharpness Efficiently?

Estimation of worst-case sharpness involves solving a constrained maximization problem typically using projected gradient ascent which can be sensitive to its hyperparameters, primarily the step size. To avoid doing extensive grid searches over the hyperparameters of gradient ascent for each model, we choose to use *Auto-PGD* (Croce & Hein, 2020) (see Algorithm 1 in Appendix for the precise formulation). Auto-PGD is a *hyperparameter-free* method designed to accurately estimate adversarial robustness by solving a similar optimization problem to worst-case sharpness but over the input space instead of the parameter space. As in ℓ_∞ and ℓ_2 versions of Auto-PGD, for each gradient step, we use gradient-sign and plain-gradient updates, respectively, but we make them proportional to $|\boldsymbol{w}|$, to better take into account the geometry induced by elementwise adaptive sharpness. We show in Sec. H.2 in Appendix that as few as 20 steps are typically sufficient to converge with Auto-PGD.

4. Sharpness vs. Generalization: Modern Setup

The current understanding of the relationship between sharpness and generalization is based on experiments on

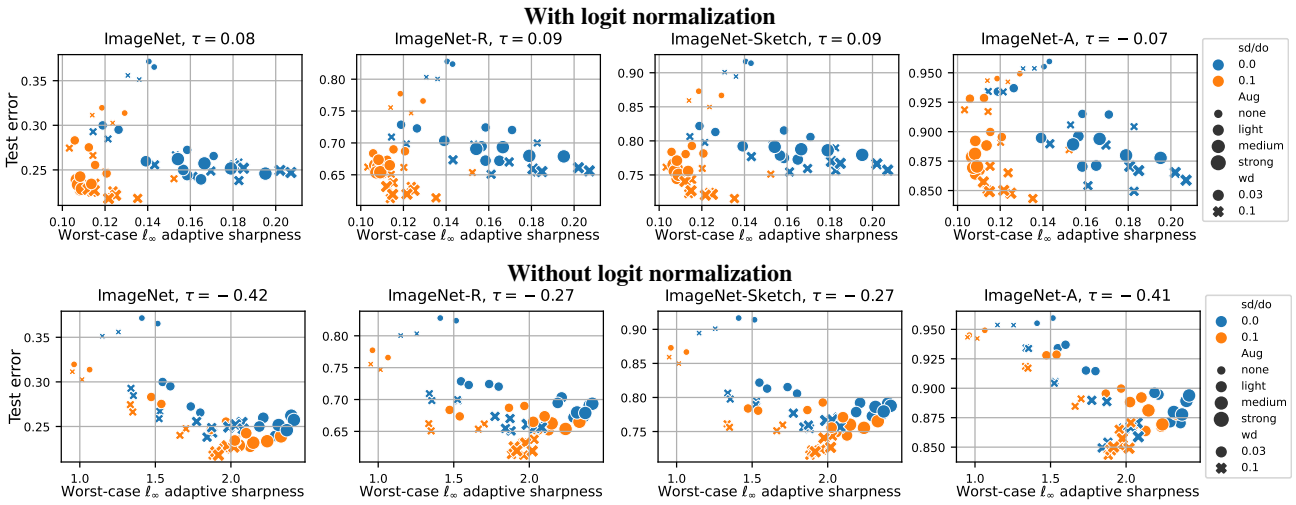


Figure 2: ViT-B/16 trained from scratch on ImageNet-1k. We show for 56 models from Steiner et al. (2021) the test error on ImageNet and its OOD variants vs. worst-case ℓ_∞ sharpness with (top) or without (bottom) normalization at $\rho = 0.002$. The color indicates models trained with stochastic depth (sd) and dropout (do), markers and their size indicate the strength of weight decay (wd) and augmentations (aug), and τ indicates the rank correlation coefficient from Eq. (4). Overall, the correlation of sharpness with test error is either close to zero or even negative.

non-residual convolution networks and small datasets like CIFAR-10 and SVHN (Jiang et al., 2020). We revisit here this relationship for state-of-the-art transformers trained from scratch on ImageNet-1k and CLIP / BERT fine-tuned on ImageNet-1k / MNLI. We explore both in-distribution (ID) and out-of-distribution (OOD) generalization due to the common intuition that flatter models are expected to be more robust (Cha et al., 2021). We focus on worst-case ℓ_∞ adaptive sharpness with low m (256) since it appears to be one of the most promising sharpness definitions (Kwon et al., 2021). We compute sharpness with and without logit normalization, and provide *average-case* sharpness for different radii ρ in Appendix. We focus primarily on the relationship between sharpness and *test error* but we also discuss sharpness vs. *generalization gap* in Sec. B in Appendix.

Training on ImageNet-1k from scratch. To investigate the relationship between sharpness and generalization for large-scale settings, we evaluate ViT models from Steiner et al. (2021), using ViT-B/16-224 weights. Those were trained from scratch on ImageNet-1k for 300 epochs with different hyperparameter settings, and subsequently fine-tuned on the same dataset for 20.000 steps with 2 different learning rates. The different hyperparameters include augmentations, weight decay, and stochastic depth / dropout, leading to a rich pool of 56 models with test errors ranging from 21.8% to 37.2%. As shown in Figure 2 (first column), neither the sharpness measure computed with nor without logit normalization can effectively distinguish model performance. Logit-normalized sharpness effectively separates models with stochastic depth / dropout (sd/do from now on) from those without by grouping them into two distinct clusters (blue and orange). However, these clusters do not correspond to a separation by test error. For the OOD tasks (ImageNet-R, ImageNet-Sketch, ImageNet-A), within each cluster, the models trained with higher weight decay yield lower test error fairly consistently. However, this ranking is not captured by sharpness, which only disentangles the sd/do clusters. For sharpness without logit normalization, the sd/do clusters are not well-separated. Surprisingly, there is a consistent *negative* correlation between sharpness and test error, both on ID and OOD data, i.e. the flattest models tend to have the largest test error. Evaluation for other radii, average-case sharpness measures (App. C) and for ViTs pretrained on IN-21k and fine-tuned on IN-1k (App. D) similarly suggest that sharpness does not consistently capture generalization properties. When considering IN-1k and IN-21k pre-trained models together (App. E) we even find similar or *higher* sharpness for significantly better-generalizing models. Then, for none of the settings studied, we can confirm either the strong or weak hypotheses.

Fine-tuning on ImageNet-1k from CLIP. We investigate fine-tuning from CLIP (Radford et al., 2021), which is a crucial approach due to the popularity of CLIP features (Ramesh et al., 2022), its fast training time, and its ability to achieve higher accuracy. We study the pool of classifiers obtained by Wortsman et al. (2022a) who fine-tuned a CLIP ViT-B/32 model on ImageNet multiple times by randomly selecting training hyperparameters such as learning rate, number of epochs, weight decay, label smoothing and augmentations. This set of 71 fine-tuned models, along with the base model, allows us to study how well generalization and training hyperparameters are captured by sharpness.

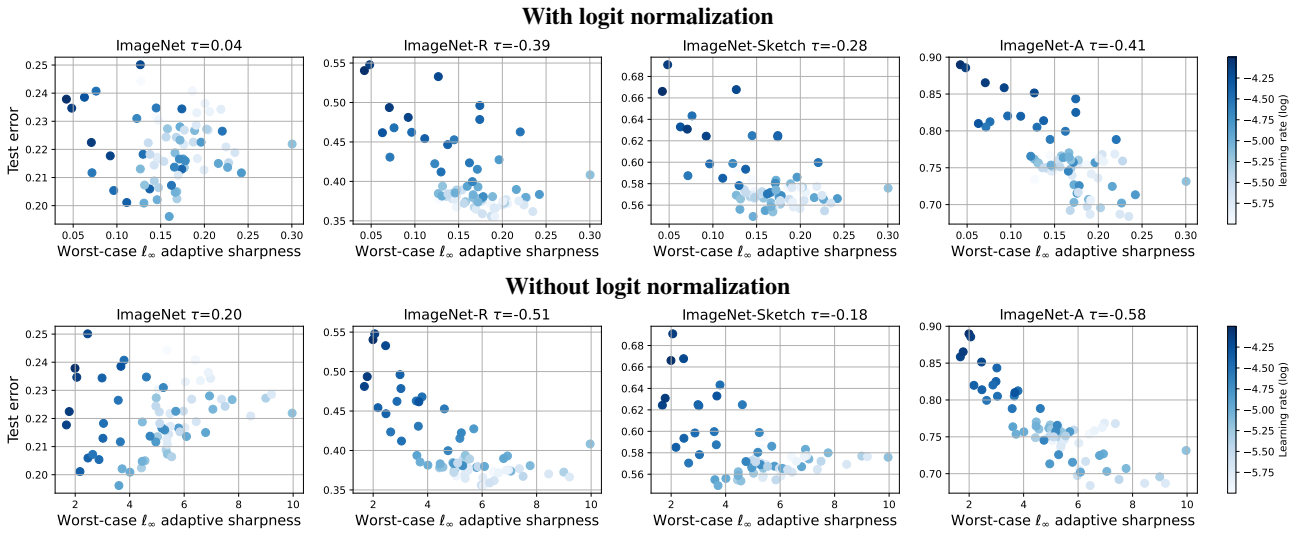


Figure 3: Fine-tuning CLIP ViT-B/32 on ImageNet-1k. We show for 72 models from Wortsman et al. (2022a) the test error on ImageNet or its variants (distribution shifts) vs worst-case ℓ_∞ sharpness with (top) or without (bottom) normalization at $\rho = 0.002$. Darker color indicates larger learning rate used for fine-tuning.

The leftmost column of Fig. 3 illustrates that worst-case ℓ_∞ adaptive sharpness does not effectively predict which classifiers have the lowest test error on ImageNet. Furthermore, there is a consistent negative correlation between sharpness and test error when evaluating classifiers on the distribution shifts ImageNet-R (Hendrycks et al., 2021a), ImageNet-Sketch (Wang et al., 2019) and ImageNet-A (Hendrycks et al., 2021b) (second to fourth columns). We further notice that, in contrast with ImageNet, higher test errors on these datasets go in parallel with higher learning rates used for fine-tuning (darker color in the plots). Indeed, smaller learning rates lead to smaller changes in the features of the base CLIP model which are more robust to distribution shifts since they were obtained from a much larger dataset than ImageNet. Finally, similar observations hold for the other sharpness definition and radii (App. F).

Fine-tuning on MNLI from BERT. We explore fine-tuning from BERT (Devlin et al., 2019), to expand our analysis beyond vision tasks. To study the linguistic generalization of multiple classifiers trained on the same dataset, McCoy et al. (2020) have fine-tuned BERT 100 times on the Multi-genre Natural Language Inference (MNLI) dataset (Williams et al., 2018) varying exclusively the random seed across runs. These random seeds affect the initialization of the classifier and the scanning order of the training data for SGD. All these classifiers achieve very similar in-distribution generalization, i.e. on MNLI test points, but behave differently on the out-of-distribution tasks represented by the HANS dataset (McCoy et al., 2019). For example, in one of HANS sub-domains the accuracy of the models ranges from 5% to 55%. We randomly choose 50 of the 100 available classifiers, and compute the different measures of sharpness for various radii. Fig. 4 shows how the worst-case ℓ_∞ adaptive sharpness, with and without logit normalization, correlates with test error on MNLI and three HANS tasks. We observe that the correlation is weak and does not exceed 0.04, even for datasets like HANS lexical (second column) where test errors vary significantly (between 45% and 95%). Moreover, in some cases the correlation is weakly negative suggesting that on average sharper models tend to generalize slightly better. Results for other radii can be found in App. G.

Summary of the findings. To conclude, *none* of the settings studied above support either the strong or weak hypotheses about the role of sharpness. Contrary to our expectations, CLIP models fine-tuned on ImageNet suggest that flatter solutions consistently generalize *worse* on OOD data. Finally, sharpness is not useful to distinguish different solutions found by fine-tuning BERT on MNLI. All this evidence suggests that the intuitive ideas about the generalization benefits of flat minima are *not supported in the modern settings*.

5. Why Doesn't Sharpness Correlate Well with Generalization?

The goal of this section is to clarify the disconnect between sharpness and generalization in the modern setup. We first revisit sharpness in a controlled environment on CIFAR-10, then explore the different sharpness definitions for a simple model where generalization is well understood.

5.1. The Role of Sharpness in a Controlled Setup

Motivation. We consider three potential explanations for why sharpness does not correlate well with generalization in

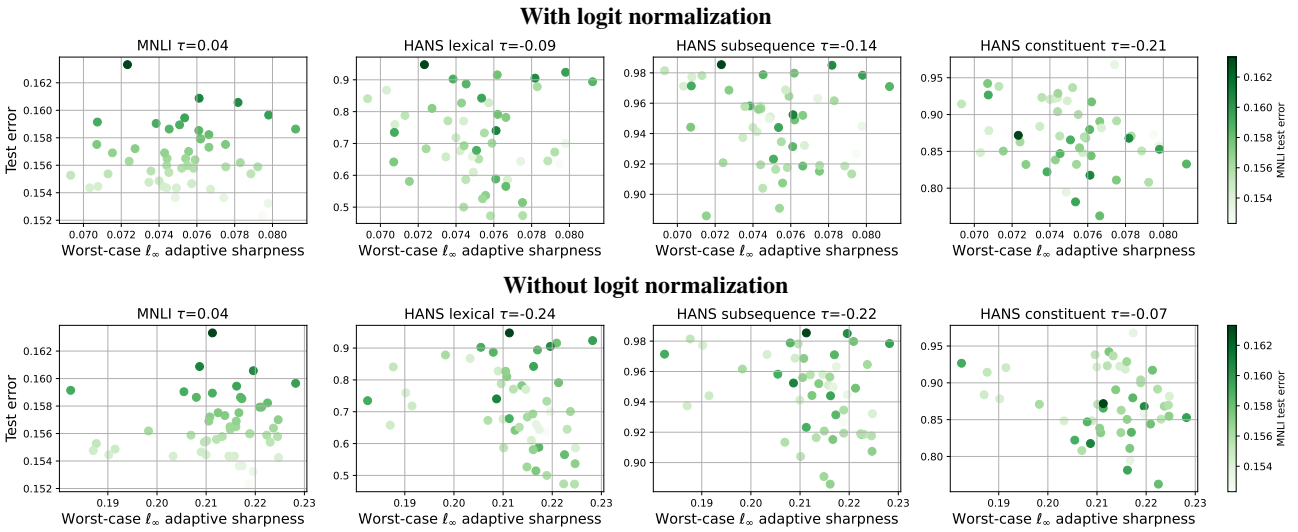


Figure 4: Fine-tuning BERT on MNLI. We show for 50 models the error on MNLI or out-of-distribution domains (HANS subsets) vs worst-case ℓ_∞ sharpness with (top) or without (bottom) normalization at $\rho = 0.0005$. Darker color indicates higher test error on MNLI.

the previous section: (1) the use of transformers instead of typical convolutional networks, (2) the use of much larger datasets (ImageNet vs. CIFAR-10), (3) the need to measure sharpness closer to a global minimum. We thus train 200 ResNets-18 and 200 ViTs on CIFAR-10 in a setting similar to Jiang et al. (2020) and Kwon et al. (2021), and evaluate sharpness only for models that reach *at most* 1% *training error*. This is in contrast to the ImageNet models from the previous section that are not necessarily trained to $\approx 0\%$ training error as it is usually not necessary in practice. Being closer to a global minimum ensures that the worst-case sharpness captures more the curvature by preventing first-order terms from dominating in Eq. 3.

Setup. We train models for 200 epochs using SGD with momentum and linearly decreasing learning rates after a linear warm-up for the first 40% iterations. We use the SimpleViT architecture from the `vit-pytorch` library which is a modification of the standard ViT (Dosovitskiy et al., 2021) with a fixed positional embedding and global average pooling instead of the CLS embedding. We vary the learning rate, $\rho \in \{0, 0.05, 0.1\}$ of SAM (Foret et al., 2021), mixup ($\alpha = 0.5$) (Zhang et al., 2018), and standard augmentations combined with RandAugment (Cubuk et al., 2020). We only show models that have $\leq 1\%$ training error.

Observations. We benchmark 12 different sharpness definitions: ℓ_2 vs. ℓ_∞, average- vs. worst-case, standard vs. adaptive, with vs. without logit normalization, and consider different perturbation radii ρ. We report most of these results in App. H and here highlight only ℓ_∞ adaptive sharpness in Fig. 5. We observe that for ResNets, there is a strong correlation between sharpness and test error but *only within each subgroup of training parameters* such as augmentations and mixup. Importantly, sharpness does not correctly capture generalization between different subgroups leading to low positive or negative correlation (0.30 and -0.36). For ViTs, we do not observe strong positive correlation even within each subgroup (in fact, without logit normalization the correlation is noticeably negative -0.68), and many models with an order of magnitude difference in sharpness can have the same test error. Moreover, we do not consistently observe that models with the lowest sharpness generalize best. For OOD generalization on common image corruptions (Hendrycks & Dietterich, 2019), the trend is even less clear and the subgroups are mixed. We note that similar conclusions hold for other sharpness radii ρ and definitions which we show in App. H.4. Moreover, in App. H we also analyze the role of data points used to evaluate sharpness (with and without augmentations), number of iterations of Auto-PGD for worst-case sharpness, and different m in worst-case m-sharpness (Foret et al., 2021). In conclusion, even in this controlled small-scale setup that includes more established architectures like ResNets, we find no empirical support to either the strong or weak hypothesis.

Sharpness captures the learning rate even when it is not helpful to predict generalization. Prior works have shown a robust link between the learning rate of first-order methods and standard sharpness definitions such as $\lambda_{\max}(\nabla^2 L(\boldsymbol{w}))$ and $\mathrm{tr}(\nabla^2 L(\boldsymbol{w}))$ (Cohen et al., 2021; Wu et al., 2022). However, the connection between the learning rate and *adaptive* sharpness remains elusive, so we investigate it empirically in Fig. 6. For both ResNets and ViTs, we observe a significant negative correlation, especially within each subgroup defined by the same values of `augment` $\times$ `mixup`. This is

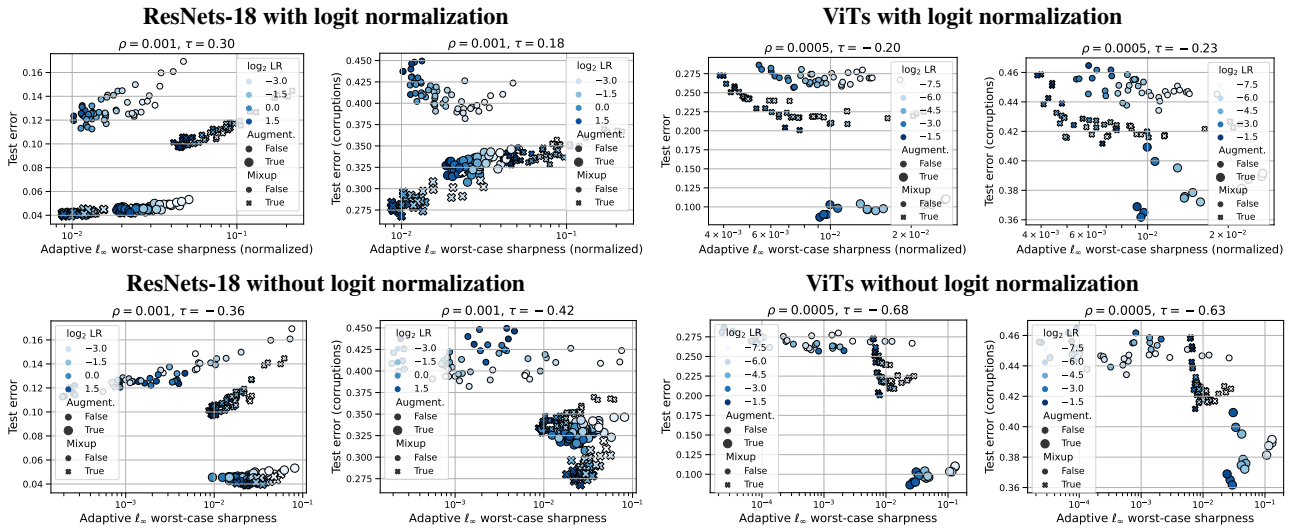


Figure 5: Training from scratch on CIFAR-10. Normalized and unnormalized ℓ_∞ adaptive sharpness vs. standard and OOD test error on common corruptions for ResNets-18 and ViTs. For other sharpness definitions (ℓ_2/ℓ_∞, average-/worst-case, etc) and multiple sharpness radii ρ, see App. H.4.

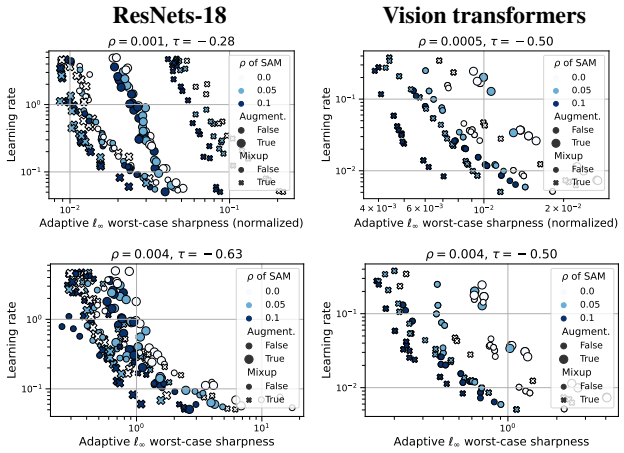


Figure 6: Training from scratch on CIFAR-10. Sharpness negatively correlates with the *learning rate*, especially within each subgroup defined by the same values of `augment` × `mixup`.

however *not* always a desirable property for predicting generalization. On the one hand, monotonically capturing the learning rates can be useful in setting like training ResNets from scratch (Li et al., 2019). On the other hand, large learning rates do not preserve the original features and can significantly harm OOD generalization for fine-tuning (Wortsman et al., 2022b). We also see a negative correlation between sharpness and learning rate for CLIP models fine-tuned on ImageNet in Fig. 20, shown in App. F. However, for these models, we do not have subgroups as clearly defined as for the CIFAR-10 models so we cannot see a more fine-grained trend. Finally, we note that whenever learning rates have a beneficial regularization effect, it is closely tied to the amount of stochastic noise in SGD (Jastrzebski et al., 2017; Andriushchenko et al., 2023). This amount is equally determined by other hyperparameters like batch size, momentum coefficient, or weight decay for normalized networks (see Li et al. (2020) for a discussion on the intrinsic learning rate). These parameters are commonly varied in studies on sharpness vs. generalization (Jiang et al., 2020; Kwon et al., 2021; Bisla et al., 2022) but all reflect essentially the same underlying trend.

5.2. Is Sharpness the Right Quantity in the First Place? Insights from Simple Models

Here, we study the link between sharpness and generalization for sparse regression with *diagonal linear networks* for which the ℓ_1 norm of the solution is predictive of generalization. This simple model suggests that sharpness measures which are universally correlated with better generalization across all possible data distributions simply do not exist.

Diagonal linear networks are defined as predictors $\langle \boldsymbol{x}, \boldsymbol{\beta} \rangle$ with parameterization $\boldsymbol{\beta} = \boldsymbol{u} \odot \boldsymbol{v}$ for weights $\boldsymbol{w} = \left[\begin{smallmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{smallmatrix}\right] \in \mathbb{R}^{2d}$. They have been widely studied as the simplest non-trivial neural network (Woodworth et al., 2020; Pesme et al., 2021). We consider an overparametrized sparse regression problem for a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and label vector $\boldsymbol{y}$:

$$L(\boldsymbol{w}) := \|\boldsymbol{X}(\boldsymbol{u} \odot \boldsymbol{v}) - \boldsymbol{y}\|_2^2, \tag{6}$$

for which the ground truth $\boldsymbol{\beta}^*$ is a sparse vector (i.e., most coordinates are zeros) and there exist many solutions $\boldsymbol{w}$ such that $L(\boldsymbol{w}) = 0$. Assuming whitened data $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}$ and that $\boldsymbol{w}$ is a global minimum, the Hessian of the loss L

simplifies to

$$\nabla^2 L(\boldsymbol{w}) = \begin{bmatrix} \operatorname{diag}(\boldsymbol{v} \odot \boldsymbol{v}) & \operatorname{diag}(\boldsymbol{u} \odot \boldsymbol{v}) \\ \operatorname{diag}(\boldsymbol{u} \odot \boldsymbol{v}) & \operatorname{diag}(\boldsymbol{u} \odot \boldsymbol{u}) \end{bmatrix}.$$

We first consider standard definitions of *local* (i.e., $\rho \to 0$) sharpness for which we have a closed-form expression. The average-case local sharpness is equal to $\operatorname{tr}(\nabla^2 L(\boldsymbol{w})) = \sum_{i=1}^{d} u_i^2 + v_i^2$ while the worst-case local sharpness at a minimum is $\lambda_{\max}(\nabla^2 L(\boldsymbol{w})) = \max_{1 \le i \le d} v_i^2 + u_i^2$ (see Sec. A.2 for details). Importantly, both average- and worst-case local sharpness are not invariant under α-reparametrization $(\alpha \boldsymbol{u}, \boldsymbol{v}/\alpha)$ while the predictor $\boldsymbol{\beta} = \boldsymbol{u} \odot \boldsymbol{v}$ is. This fact emphasizes the need for a measure of the sharpness that adjusts to the changing scale of the parameters as the adaptive sharpness. Indeed, with the carefully selected elementwise scaling $c_i = \sqrt{|v_i|/|u_i|}$ for $1 \le i \le d$ and $c_i = \sqrt{|u_i|/|v_i|}$ for $d < i \le 2d$, we obtain for the average-case and worst-case adaptive local sharpness

$$S^{\rho}_{avg}(\boldsymbol{w}, \boldsymbol{c}) = \frac{1}{2}\sum_{i=1}^{d} u_i^2 |v_i|/|u_i| + \frac{1}{2}\sum_{i=1}^{d} v_i^2 |u_i|/|v_i| = \|\boldsymbol{\beta}\|_1,$$
$$S^{\rho}_{max}(\boldsymbol{w}, \boldsymbol{c}) = \max_{1 \le i \le d} |u_i||v_i| = \|\boldsymbol{\beta}\|_\infty.$$

We first note that both definitions of adaptive sharpness are invariant under α-reparametrization as they only depend on the predictor $\boldsymbol{\beta}$. However, average and worst-case sharpness do not capture the same properties of $\boldsymbol{\beta}$. In particular, $\|\boldsymbol{\beta}\|_1$ is a generalization measure that correctly captures the sparsity of the linear predictor which is a good indicator of generalization for a *sparse* $\boldsymbol{\beta}^*$. In contrast, $\|\boldsymbol{\beta}\|_\infty$ is a generalization measure that is more suitable to capture how uniform the weights of $\boldsymbol{\beta}$ are which is a good predictor of generalization for a *dense* $\boldsymbol{\beta}^*$. Finally, we note that using $\boldsymbol{c} = \boldsymbol{w}$ in adaptive sharpness would instead lead to $\|\boldsymbol{\beta}\|_2^2$ and $\|\boldsymbol{\beta}\|_\infty^2$ that would have a different interpretation. This simple model highlights that the sharpness definition that correlates well with generalization is data-dependent and in general S_{avg} and S_{max} capture very different trends.

To further illustrate this point, we train 200 diagonal linear networks to 10^{-5} training loss on a sparse regression task ($d = 200$ with 90% sparsity) with different learning rates and random initializations. We show the results in Fig. 7 which illustrate that (1) $\|\boldsymbol{u} \odot \boldsymbol{v}\|_1$ is approximated well by $\frac{1}{2}\operatorname{tr}(\tilde{\nabla}^2 L(\boldsymbol{w}))$, (2) $\operatorname{tr}(\tilde{\nabla}^2 L(\boldsymbol{w}))$ correlates better than $\operatorname{tr}(\nabla^2 L(\boldsymbol{w}))$ so the adaptive part is important, (3) the relationship between $\operatorname{tr}(\tilde{\nabla}^2 L(\boldsymbol{w}))$ and $\lambda_{\max}(\tilde{\nabla}^2 L(\boldsymbol{w}))$ can be even reverse showing that different sharpness definitions capture totally different trends. We also note that even with the right definition of sharpness, the correlation is not perfect (around $\tau = 0.8$) and there is always some non-negligible gap in predicting the test loss. Overall, we conclude that finding a sharpness definition that correlates well with generalization requires understanding both the role of the data distribution and its interaction with the architecture. It is possible in very simple cases but appears extremely challenging for complex architectures like vision transformers on complex real-world datasets like ImageNet.

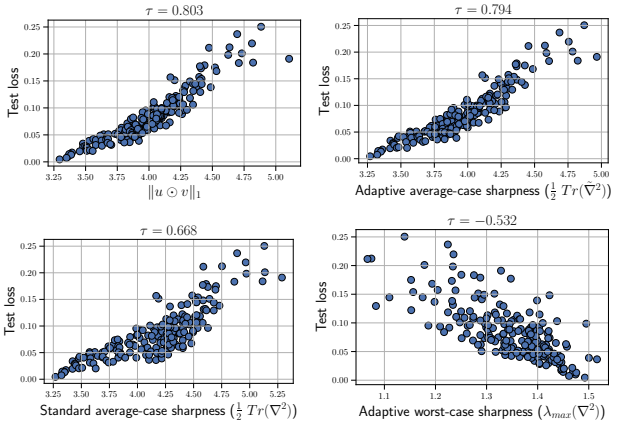


Figure 7: Different generalization measures for diagonal linear networks. $\tilde{\nabla}^2$ denotes the rescaled Hessian corresponding to adaptive sharpness.

6. Conclusions

Our results suggest that even reparametrization-invariant sharpness is *not* a good indicator of generalization in the modern setting. While there definitely exist restricted settings where correlation between sharpness and generalization is significantly positive (e.g., for ResNets on CIFAR-10 with a specific combination of augmentations and mixup), it is not true anymore when we compare all models *jointly*. Moreover, the correlation, even within subgroups of models defined by augmentations, is much lower for vision transformers. Thus, we believe it is important to rethink the intuitive understanding of sharpness based on the geometric intuition about the shift of the loss surface. Moreover, our findings suggest that one should avoid blanket statements like *"flatter minima generalize better"* since even when they are only intended to imply *correlation*, their correctness still depends on a number of factors such as data distribution, model family, or initialization schemes (i.e., random vs. from pretrained weights).

Acknowledgements

M.A. was supported by the Google Fellowship and Open Phil AI Fellowship. M.M. and M.H. were supported by the Carl Zeiss Foundation in the project "Certification and Foundations of Safe Machine Learning Systems in Healthcare". We thank David Stutz for very fruitful discussions at the initial stage of the project, Jana Vuckovic for experiments on sharpness that helped us to shape the project and Aditya Varre for discussions on sharpness for diagonal networks.

References

Andriushchenko, M. and Flammarion, N. Towards understanding sharpness-aware minimization. In ICML, 2022.

Andriushchenko, M., Varre, A., Pillaud-Vivien, L., and Flammarion, N. SGD with large step sizes learns sparse features. In ICML, 2023.

Arora, S., Li, Z., and Panigrahi, A. Understanding gradient descent on edge of stability in deep learning. In ICML, 2022.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In NeurIPS, 2019.

Bisla, D., Wang, J., and Choromanska, A. Low-pass filtering sgd for recovering flat optima in the deep learning optimization landscape. AISTATS, 2022.

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In COLT, 2020.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., and Park, S. Swad: Domain generalization by seeking flat minima. NeurIPS, 34:22405–22418, 2021.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. Journal of Statistical Mechanics: Theory and Experiment, 2019(12):124018, 2016.

Chen, X., Hsieh, C.-J., and Gong, B. When vision transformers outperform resnets without pre-training or strong data augmentations? ICLR, 2022.

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. ICLR, 2021.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. ICML, 2020.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. NeurIPS, 2020.

Damian, A., Ma, T., and Lee, J. D. Label noise sgd provably prefers flat global minimizers. NeurIPS, 34:27449–27461, 2021.

Damian, A., Nichani, E., and Lee, J. D. Self-stabilization: The implicit bias of gradient descent at the edge of stability. ICLR, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In ICML, pp. 1019–1028. PMLR, 2017.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR, 2021.

Du, J., Daquan, Z., Feng, J., Tan, V., and Zhou, J. T. Sharpness-aware training for free. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), Advances in Neural Information Processing Systems, 2022. URL https://openreview.net/forum?id=xK6wRfL2mv7.

Dziugaite, G. K. and Roy, D. Entropy-sgd optimizes the prior of a pac-bayes bound: Generalization properties of entropy-sgd and data-dependent priors. In ICML, pp. 1377–1386. PMLR, 2018.

Dziugaite, G. K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I., and Roy, D. M. In search of robust measures of generalization. NeurIPS, 33: 11723–11733, 2020.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In ICLR, 2021.

Fort, S., Brock, A., Pascanu, R., De, S., and Smith, S. L. Drawing multiple augmentation samples per image during training efficiently decreases test error. arXiv preprint arXiv:2105.13343, 2021.

Granziol, D. Flatness is a false friend. arXiv preprint arXiv:2006.09091, 2020.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In ICLR, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. ICCV, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. CVPR, 2021b.

Hochreiter, S. and Schmidhuber, J. Simplifying neural nets by discovering flat minima. In NeurIPS, pp. 529–536, 1995.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. UAI, 2018.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in sgd. arXiv preprint arXiv:1711.04623, 2017.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. ICLR, 2020.

Kaur, S., Cohen, J., and Lipton, Z. C. On the maximum hessian eigenvalue and generalization. arXiv preprint arXiv:2206.10654, 2022.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. ICLR, 2016.

Kwon, J., Kim, J., Park, H., and Choi, I. K. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. ICML, 2021.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. Efficient backprop. In Neural networks: Tricks of the trade, pp. 9–48. Springer, 2012.

Li, Y., Wei, C., and Ma, T. Towards explaining the regularization effect of initial large learning rate in training neural networks. In NeurIPS, 2019.

Li, Z., Lyu, K., and Arora, S. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. NeurIPS, 33:14544–14555, 2020.

Li, Z., Wang, T., and Arora, S. What happens after sgd reaches zero loss?–a mathematical framework. arXiv preprint arXiv:2110.06914, 2021.

Liang, T., Poggio, T., Rakhlin, A., and Stokes, J. Fisher-rao metric, geometry, and complexity of neural networks. In AISTATS. PMLR, 2019.

Lyu, K., Li, Z., and Arora, S. Understanding the generalization benefit of normalization layers: Sharpness reduction. NeurIPS, 2022.

McCoy, R. T., Min, J., and Linzen, T. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, November 2020.

McCoy, T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In ACL, 2019.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In NeurIPS, pp. 5947–5956, 2017.

Park, N. and Kim, S. How do vision transformers work? ICLR, 2022.

Pesme, S., Pillaud-Vivien, L., and Flammarion, N. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. In NeurIPS, 2021.

Petzka, H., Kamp, M., Adilova, L., Sminchisescu, C., and Boley, M. Relative flatness and generalization. NeurIPS, 34:18420–18432, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In ICML, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125, 2022.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In ICML, pp. 5389–5400. PMLR, 2019.

Smith, S. L. and Le, Q. V. A Bayesian perspective on generalization and stochastic gradient descent. In ICLR, 2018.

Smith, S. L., Dherin, B., Barrett, D. G., and De, S. On the origin of implicit regularization in stochastic gradient descent. ICLR, 2021.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. JMLR, 15(1), 2014.

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. How to train your vit? data, augmentation, and regularization in vision transformers. TMLR, 2021.

Stutz, D., Hein, M., and Schiele, B. Relating adversarially robust generalization to flat minima. ICCV, 2021.

Tsuzuku, Y., Sato, I., and Sugiyama, M. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis. In ICML, pp. 9636–9647. PMLR, 2020.

Vedantam, S. R., Lopez-Paz, D., and Schwab, D. J. An empirical investigation of domain generalization with empirical risk minimizers. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), NeurIPS, 2021.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In NeurIPS, pp. 10506–10518, 2019.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In NAACL, 2018.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In COLT. PMLR, 2020.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In ICML, pp. 23965–23998. PMLR, 2022a.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. Robust fine-tuning of zero-shot models. In CVPR, pp. 7959–7971, 2022b.

Wu, D., Xia, S.-t., and Wang, Y. Adversarial weight perturbation helps robust generalization. NeurIPS, 2020.

Wu, L., Wang, M., and Su, W. When does sgd favor flat minima? a quantitative characterization via linear stability. NeurIPS, 2022.

Xing, C., Arpit, D., Tsirigotis, C., and Bengio, Y. A walk with sgd. arXiv preprint arXiv:1802.08770, 2018.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. ICLR, 2018.

Zhang, S., Reid, I., Pérez, G. V., and Louis, A. Why flatness does and does not correlate with generalization for deep neural networks. arXiv preprint arXiv:2103.06219, 2021.

Zheng, Y., Zhang, R., and Mao, Y. Regularizing neural networks via adversarial model perturbation. CVPR, 2021.

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H., et al. Towards theoretically understanding why SGD generalizes better than Adam in deep learning. NeurIPS, 2020.

Zhuang, J., Gong, B., Yuan, L., Cui, Y., Adam, H., Dvornek, N. C., sekhar tatikonda, s Duncan, J., and Liu, T. Surrogate gap minimization improves sharpness-aware training. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=edONMAnhLu-.

Appendix

The appendix is organized as follows:

- Sec. A: omitted derivations for sharpness when $\rho \to 0$, first for the general case and then specifically for diagonal linear networks.
- Sec. B: figures with correlation between sharpness and *generalization gap*. We observe a similar trend between sharpness and *generalization gap* as between sharpness and *test error* which is reported in the main part.
- Sec. C: additional figures about ViTs from Steiner et al. (2021) trained with different hyperparameter settings on ImageNet-1k. We observe that different sharpness variants are not predictive of the performance on ImageNet and the OOD datasets, typically only separating models by stochastic depth / dropout, but not ranking them according to generalization, and often even yielding a negative correlation with OOD test error.
- Sec. D: figures about ViTs from Steiner et al. (2021) pre-trained on ImageNet-21k and then fine-tuned on ImageNet-1k. The observations are very similar to those for training on ImageNet-1k from scratch: sharpness variants are not predictive of the performance on ImageNet, and they often lead to a negative correlation with OOD test error.
- Sec. E: figures for combined analysis of ViTs from Steiner et al. (2021) both with and without ImageNet-21k pre-training. We find the better-generalizing models pretrained on ImageNet-21k to have significantly higher worst-case sharpness and roughly equal or higher logit-normalized average-case adaptive sharpness, underlining that the models' generalization properties resulting from different pretraining datasets are not captured.
- Sec. F: additional details and figures for CLIP models fine-tuned on ImageNet. We observe that sharpness variants are not predictive of the performance on ImageNet and ImageNet-V2. Moreover, there is in most cases a negative correlation with test error in presence of distribution shifts which is likely to be related to the influence that the learning rate has on sharpness.
- Sec. G: additional details and figures for BERT models fine-tuned on MNLI. We find that all sharpness variants we consider are not predictive of the generalization performance of the model, and in some cases there is rather a weak negative correlation between sharpness and test error on out-of-distribution tasks from HANS.
- Sec. H: additional details and ablation studies for CIFAR-10 models. We analyze the role of data used to evaluate sharpness, the role of the number of iterations in Auto-PGD, the role of m in m-sharpness, and the influence of different sharpness definitions and radii on correlation with generalization. Overall, we conclude that none of the considered sharpness definitions or radii correlates positively with generalization nor that low sharpness implies good performance of the model.

Also, for the sake of convenience, we provide in Table 1, Table 2, Table 3, and Table 4 a summary of correlation coefficients τ between sharpness and generalization for all our experiments (except ablation studies).

ImageNet-1k models trained from scratch

			Rank correlation coefficient τ					
Sharpness	**LogitNorm**	ρ	**IN**	**IN-v2**	**IN-R**	**IN-Sketch**	**IN-A**	**ObjectNet**
Worst-case ℓ_∞	Yes	0.001	0.09	0.08	0.10	0.10	-0.06	0.04
Worst-case ℓ_∞	Yes	0.002	0.08	0.08	0.09	0.09	-0.07	0.03
Worst-case ℓ_∞	Yes	0.004	-0.11	-0.11	-0.06	-0.06	-0.23	-0.16
Worst-case ℓ_∞	No	0.001	-0.42	-0.43	-0.27	-0.28	-0.45	-0.45
Worst-case ℓ_∞	No	0.002	-0.42	-0.42	-0.27	-0.27	-0.41	-0.45
Worst-case ℓ_∞	No	0.004	-0.34	-0.34	-0.20	-0.20	-0.36	-0.36
Avg-case ℓ_∞	Yes	0.05	0.46	0.44	0.38	0.42	0.31	0.39
Avg-case ℓ_∞	Yes	0.1	0.44	0.43	0.39	0.43	0.29	0.39
Avg-case ℓ_∞	Yes	0.2	0.42	0.42	0.39	0.42	0.29	0.38
Avg-case ℓ_∞	No	0.05	**-0.55**	**-0.56**	-0.40	-0.42	**-0.57**	**-0.60**
Avg-case ℓ_∞	No	0.1	-0.44	-0.43	-0.28	-0.32	-0.47	-0.47
Avg-case ℓ_∞	No	0.2	0.13	0.15	0.26	0.23	0.05	0.11

ImageNet-1k models fine-tuned from IN-21k

			Rank correlation coefficient τ					
Sharpness	**LogitNorm**	ρ	**IN**	**IN-v2**	**IN-R**	**IN-Sketch**	**IN-A**	**ObjectNet**
Worst-case ℓ_∞	Yes	0.001	-0.49	-0.49	-0.44	-0.33	**-0.53**	-0.46
Worst-case ℓ_∞	Yes	0.002	-0.48	-0.48	-0.46	-0.33	**-0.51**	-0.44
Worst-case ℓ_∞	Yes	0.004	-0.45	-0.43	-0.41	-0.33	-0.45	-0.42
Worst-case ℓ_∞	No	0.001	-0.13	-0.09	-0.05	0.05	-0.13	-0.09
Worst-case ℓ_∞	No	0.002	-0.10	-0.03	-0.01	0.11	-0.07	-0.02
Worst-case ℓ_∞	No	0.004	-0.10	-0.01	-0.01	0.11	-0.06	0.00
Avg-case ℓ_∞	Yes	0.05	-0.11	-0.08	-0.11	-0.07	-0.06	-0.06
Avg-case ℓ_∞	Yes	0.1	-0.12	-0.11	-0.14	-0.10	-0.09	-0.08
Avg-case ℓ_∞	Yes	0.2	-0.25	-0.24	-0.25	-0.23	-0.25	-0.24
Avg-case ℓ_∞	No	0.05	-0.02	-0.04	-0.03	-0.02	-0.05	-0.06
Avg-case ℓ_∞	No	0.1	-0.07	-0.10	-0.08	-0.08	-0.11	-0.10
Avg-case ℓ_∞	No	0.2	-0.11	-0.11	-0.10	-0.11	-0.12	-0.13

ImageNet-1k models fine-tuned from CLIP

			Rank correlation coefficient τ					
Sharpness	**LogitNorm**	ρ	**IN**	**IN-v2**	**IN-R**	**IN-Sketch**	**IN-A**	**ObjectNet**
Worst-case ℓ_∞	Yes	0.001	-0.04	-0.16	-0.23	-0.26	-0.25	-0.36
Worst-case ℓ_∞	Yes	0.002	0.04	-0.10	-0.39	-0.28	-0.41	-0.47
Worst-case ℓ_∞	Yes	0.004	-0.08	-0.19	-0.12	-0.16	-0.17	-0.27
Worst-case ℓ_∞	No	0.001	0.19	0.09	-0.37	-0.06	**-0.57**	-0.48
Worst-case ℓ_∞	No	0.002	0.20	0.08	**-0.51**	-0.18	**-0.58**	**-0.51**
Worst-case ℓ_∞	No	0.004	0.02	-0.05	**-0.51**	-0.27	-0.45	-0.33
Avg-case ℓ_∞	Yes	0.001	-0.03	-0.18	-0.36	-0.34	-0.33	-0.46
Avg-case ℓ_∞	Yes	0.002	-0.21	-0.32	-0.02	-0.27	-0.06	-0.21
Avg-case ℓ_∞	Yes	0.004	-0.19	-0.21	0.26	-0.03	0.23	0.06
Avg-case ℓ_∞	No	0.001	0.13	-0.01	**-0.62**	-0.26	**-0.67**	**-0.60**
Avg-case ℓ_∞	No	0.002	0.06	0.03	-0.34	-0.12	**-0.50**	-0.37
Avg-case ℓ_∞	No	0.004	0.19	0.21	-0.12	0.09	-0.21	-0.08

Table 1: A summary of correlation between sharpness and generalization for all experiments on **ImageNet**. We boldface entries with $|\tau| > 0.5$ suggesting a reasonably strong correlation. LogitNorm stands for *logit normalization* and IN stands for *ImageNet*.

MNLI models fine-tuned from BERT

			Rank correlation coefficient τ			
Sharpness	**LogitNorm**	ρ	**MNLI**	**HANS-L**	**HANS-S**	**HANS-C**
Worst-case ℓ_∞	Yes	0.0005	0.04	-0.09	-0.14	-0.21
Worst-case ℓ_∞	Yes	0.001	-0.09	-0.09	-0.13	-0.18
Worst-case ℓ_∞	Yes	0.002	0.05	-0.09	-0.14	-0.17
Worst-case ℓ_∞	No	0.0005	0.04	-0.24	-0.22	-0.07
Worst-case ℓ_∞	No	0.001	0.04	-0.13	-0.15	-0.15
Worst-case ℓ_∞	No	0.002	-0.11	-0.15	-0.12	-0.13
Avg-case ℓ_∞	Yes	0.1	-0.35	-0.46	-0.28	0.17
Avg-case ℓ_∞	Yes	0.2	-0.37	-0.48	-0.28	0.24
Avg-case ℓ_∞	Yes	0.4	0.01	-0.29	-0.27	0.05
Avg-case ℓ_∞	No	0.1	-0.34	-0.31	-0.23	0.13
Avg-case ℓ_∞	No	0.2	-0.34	**-0.58**	-0.39	0.16
Avg-case ℓ_∞	No	0.4	0.04	-0.16	-0.09	0.05

Table 2: A summary of correlation between sharpness and generalization for all experiments on **MNLI** for models fine-tuned from BERT. We boldface entries with $|\tau| > 0.5$ suggesting a reasonably strong correlation. LogitNorm stands for *logit normalization*.

ResNets-18 trained from scratch on CIFAR-10

			Rank correlation coefficient τ	
Sharpness	**LogitNorm**	ρ	**CIFAR-10**	**CIFAR-10-C**
Standard avg-case ℓ_2	No	0.05	0.14	0.04
Standard avg-case ℓ_2	No	0.1	0.26	0.19
Standard avg-case ℓ_2	No	0.2	0.28	0.21
Standard avg-case ℓ_2	No	0.4	0.28	0.20
Standard worst-case ℓ_2	No	0.25	0.17	0.10
Standard worst-case ℓ_2	No	0.5	0.24	0.16
Standard worst-case ℓ_2	No	1.0	0.25	0.18
Standard worst-case ℓ_2	No	2.0	0.22	0.14
Adaptive avg-case ℓ_2	No	0.05	-0.37	-0.46
Adaptive avg-case ℓ_2	No	0.1	**-0.50**	**-0.53**
Adaptive avg-case ℓ_2	No	0.2	-0.42	-0.41
Adaptive avg-case ℓ_2	No	0.4	-0.31	-0.31
Adaptive worst-case ℓ_2	No	0.25	-0.36	-0.39
Adaptive worst-case ℓ_2	No	0.5	-0.42	-0.36
Adaptive worst-case ℓ_2	No	1.0	-0.27	-0.17
Adaptive worst-case ℓ_2	No	2.0	-0.17	-0.07
Adaptive avg-case ℓ_2	Yes	0.05	0.18	0.07
Adaptive avg-case ℓ_2	Yes	0.1	0.07	-0.04
Adaptive avg-case ℓ_2	Yes	0.2	-0.14	-0.26
Adaptive avg-case ℓ_2	Yes	0.4	-0.43	-0.58
Adaptive worst-case ℓ_2	Yes	0.25	0.19	0.14
Adaptive worst-case ℓ_2	Yes	0.5	0.07	0.00
Adaptive worst-case ℓ_2	Yes	1.0	-0.13	-0.22
Adaptive worst-case ℓ_2	Yes	2.0	**-0.52**	**-0.58**
Standard avg-case ℓ_∞	No	0.1	0.16	0.08
Standard avg-case ℓ_∞	No	0.2	0.28	0.21
Standard avg-case ℓ_∞	No	0.4	0.28	0.20
Standard avg-case ℓ_∞	No	0.8	0.28	0.20
Standard worst-case ℓ_∞	No	0.0005	0.29	0.23
Standard worst-case ℓ_∞	No	0.001	0.30	0.24
Standard worst-case ℓ_∞	No	0.002	0.30	0.24
Standard worst-case ℓ_∞	No	0.004	0.29	0.23
Adaptive avg-case ℓ_∞	No	0.1	-0.36	-0.47
Adaptive avg-case ℓ_∞	No	0.2	**-0.53**	**-0.56**
Adaptive avg-case ℓ_∞	No	0.4	-0.41	-0.41
Adaptive avg-case ℓ_∞	No	0.8	-0.20	-0.18
Adaptive worst-case ℓ_∞	No	0.001	-0.36	-0.42
Adaptive worst-case ℓ_∞	No	0.002	-0.05	-0.10
Adaptive worst-case ℓ_∞	No	0.004	0.25	0.20
Adaptive worst-case ℓ_∞	No	0.008	0.26	0.24
Adaptive avg-case ℓ_∞	Yes	0.1	0.18	0.07
Adaptive avg-case ℓ_∞	Yes	0.2	0.05	-0.06
Adaptive avg-case ℓ_∞	Yes	0.4	-0.23	-0.37
Adaptive avg-case ℓ_∞	Yes	0.8	-0.46	**-0.62**
Adaptive worst-case ℓ_∞	Yes	0.001	0.30	0.18
Adaptive worst-case ℓ_∞	Yes	0.002	0.29	0.16
Adaptive worst-case ℓ_∞	Yes	0.004	0.21	0.07
Adaptive worst-case ℓ_∞	Yes	0.008	-0.04	-0.19

Table 3: A summary of correlation between sharpness and generalization for all experiments on **CIFAR-10** for ResNets-18 trained from scratch. We boldface entries with $|\tau| > 0.5$ suggesting a reasonably strong correlation. LogitNorm stands for *logit normalization*.

Vision transformers trained from scratch on CIFAR-10

			Rank correlation coefficient τ	
Sharpness	**LogitNorm**	ρ	**CIFAR-10**	**CIFAR-10-C**
Standard avg-case ℓ_2	No	0.005	-0.45	**-0.54**
Standard avg-case ℓ_2	No	0.01	-0.39	-0.49
Standard avg-case ℓ_2	No	0.02	-0.20	-0.31
Standard avg-case ℓ_2	No	0.04	-0.08	-0.20
Standard worst-case ℓ_2	No	0.025	**-0.59**	**-0.62**
Standard worst-case ℓ_2	No	0.05	-0.37	-0.43
Standard worst-case ℓ_2	No	0.1	-0.16	-0.24
Standard worst-case ℓ_2	No	0.2	-0.12	-0.20
Adaptive avg-case ℓ_2	No	0.1	-0.45	**-0.50**
Adaptive avg-case ℓ_2	No	0.2	-0.45	-0.45
Adaptive avg-case ℓ_2	No	0.4	-0.42	-0.47
Adaptive avg-case ℓ_2	No	0.8	-0.10	0.08
Adaptive worst-case ℓ_2	No	0.5	**-0.64**	**-0.53**
Adaptive worst-case ℓ_2	No	1.0	-0.32	-0.19
Adaptive worst-case ℓ_2	No	2.0	-0.11	-0.01
Adaptive worst-case ℓ_2	No	4.0	-0.07	-0.03
Adaptive avg-case ℓ_2	Yes	0.1	-0.18	-0.31
Adaptive avg-case ℓ_2	Yes	0.2	-0.28	-0.40
Adaptive avg-case ℓ_2	Yes	0.4	-0.39	-0.46
Adaptive avg-case ℓ_2	Yes	0.8	-0.44	**-0.52**
Adaptive worst-case ℓ_2	Yes	0.25	-0.21	-0.12
Adaptive worst-case ℓ_2	Yes	0.5	-0.24	-0.17
Adaptive worst-case ℓ_2	Yes	1.0	-0.22	-0.19
Adaptive worst-case ℓ_2	Yes	2.0	-0.14	-0.11
Standard avg-case ℓ_∞	No	0.01	-0.44	**-0.54**
Standard avg-case ℓ_∞	No	0.02	-0.35	-0.45
Standard avg-case ℓ_∞	No	0.04	-0.17	-0.28
Standard avg-case ℓ_∞	No	0.08	-0.04	-0.14
Standard worst-case ℓ_∞	No	0.00001	**-0.61**	**-0.63**
Standard worst-case ℓ_∞	No	0.00002	-0.46	**-0.51**
Standard worst-case ℓ_∞	No	0.00004	-0.25	-0.31
Standard worst-case ℓ_∞	No	0.00008	-0.16	-0.22
Adaptive avg-case ℓ_∞	No	0.1	-0.45	**-0.53**
Adaptive avg-case ℓ_∞	No	0.2	-0.46	**-0.50**
Adaptive avg-case ℓ_∞	No	0.4	-0.45	-0.44
Adaptive avg-case ℓ_∞	No	0.8	-0.41	-0.47
Adaptive worst-case ℓ_∞	No	0.0005	**-0.68**	**-0.63**
Adaptive worst-case ℓ_∞	No	0.001	-0.43	-0.40
Adaptive worst-case ℓ_∞	No	0.002	-0.26	-0.23
Adaptive worst-case ℓ_∞	No	0.004	-0.18	-0.18
Adaptive avg-case ℓ_∞	Yes	0.1	-0.11	-0.23
Adaptive avg-case ℓ_∞	Yes	0.2	-0.16	-0.29
Adaptive avg-case ℓ_∞	Yes	0.4	-0.31	-0.42
Adaptive avg-case ℓ_∞	Yes	0.8	-0.40	-0.47
Adaptive worst-case ℓ_∞	Yes	0.0005	-0.20	-0.23
Adaptive worst-case ℓ_∞	Yes	0.001	-0.22	-0.26
Adaptive worst-case ℓ_∞	Yes	0.002	-0.29	-0.34
Adaptive worst-case ℓ_∞	Yes	0.004	-0.39	-0.44

Table 4: A summary of correlation between sharpness and generalization for all experiments on **CIFAR-10** for ViTs trained from scratch. We boldface entries with $|\tau| > 0.5$ suggesting a reasonably strong correlation. LogitNorm stands for *logit normalization*.

A. Omitted Proofs

A.1. Asymptotic Analysis of Adaptive Sharpness Measures

For the convenience of the reader we repeat here quickly the definitions of adaptive sharpness measures. Let $L_{\mathcal{S}}(\boldsymbol{w}) = \frac{1}{|S|}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{S}} \ell_{\boldsymbol{xy}}(\boldsymbol{w})$ be the loss on a set of *training* points $\mathcal{S}$. For arbitrary weights $\boldsymbol{w}$ (i.e., not necessarily a minimum), then the *average-case* and *worst-case* m*-sharpness* is defined as:

$$S^{\rho}_{avg,p}(\boldsymbol{w},\boldsymbol{c}) \triangleq \mathbb{E}_{\substack{\mathcal{S}\sim P_m \\ \boldsymbol{\delta}\sim\mathcal{N}(0,\rho^2 diag(\boldsymbol{c}^2))}} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w}) \qquad S^{\rho}_{max,p}(\boldsymbol{w},\boldsymbol{c}) \triangleq \mathbb{E}_{\mathcal{S}\sim P_m} \max_{\|\boldsymbol{\delta}\odot\boldsymbol{c}^{-1}\|_p\leq\rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w}),$$

where $\odot/^{-1}$ denotes elementwise multiplication/inversion and P_m is the data distribution that returns m training pairs $(\boldsymbol{x},\boldsymbol{y})$.

If $\boldsymbol{c} = |\boldsymbol{w}|$ then the perturbation set is $\left\|\delta \odot |\boldsymbol{w}|^{-1}\right\|_p \leq \rho$. We first introduce a new variable $\boldsymbol{\gamma} = \boldsymbol{\delta}\odot|\boldsymbol{w}|^{-1}$ and do a Taylor expansion around w:

$$L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) = L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\gamma}\odot|\boldsymbol{w}|) = L_{\mathcal{S}}(\boldsymbol{w}) + \langle\nabla L_{\mathcal{S}}(\boldsymbol{w}), |\boldsymbol{w}|\odot\boldsymbol{\gamma}\rangle + \frac{1}{2}\left\langle\boldsymbol{\gamma}\odot|\boldsymbol{w}|, \nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\boldsymbol{\gamma}\odot|\boldsymbol{w}|\right\rangle + O(\|\boldsymbol{\gamma}\|_p^3),$$

where $\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})$ denotes the Hessian of $L_{\mathcal{S}}$ at $\boldsymbol{w}$.

Proposition 1. *Let $L_{\mathcal{S}} \in C^3(\mathbb{R}^s)$, S be a finite sample of training points $(x_i,y_i)_{i=1}^n$ and let P_m denote the uniform distribution over subsamples of size $m \leq n$ from S. Then we define for $p \geq 1$, $q \in \mathbb{R}$ such that $\frac{1}{p}+\frac{1}{q}=1$, then it holds*

$$\lim_{\rho\to 0} S^{\rho}_{max,p}(\boldsymbol{w},|\boldsymbol{w}|)$$

$$= \mathbb{E}_{\mathcal{S}\sim P_m}\begin{cases} \|\nabla L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}|\|_q\,\rho + O(\rho^2) & \text{if } \nabla L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}| \neq 0, \\ \frac{\rho^2}{2}\max_{\boldsymbol{\gamma}\neq 0}\frac{\left\langle\boldsymbol{\gamma},\left(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot(|\boldsymbol{w}||\boldsymbol{w}|^T)\right)\boldsymbol{\gamma}\right\rangle}{\|\boldsymbol{\gamma}\|_p^2} + O(\rho^3) & \text{if } \nabla L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}| = 0 \text{ and} \\ & \quad \nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot(|\boldsymbol{w}||\boldsymbol{w}|^T) \text{ not negative definite} \\ O(\rho^3) & \text{if } \nabla L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}| = 0 \text{ and} \\ & \quad \nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot(|\boldsymbol{w}||\boldsymbol{w}|^T) \text{ is negative definite} \end{cases}$$

Proof. We get

$$\begin{aligned}\max_{\|\boldsymbol{\gamma}\|_p\leq\rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\gamma}\odot|\boldsymbol{w}|) - L_{\mathcal{S}}(\boldsymbol{w}) &= \max_{\|\boldsymbol{\gamma}\|_p\leq\rho}\langle\nabla L_{\mathcal{S}}(\boldsymbol{w}),|\boldsymbol{w}|\odot\boldsymbol{\gamma}\rangle + \frac{1}{2}\left\langle\boldsymbol{\gamma}\odot|\boldsymbol{w}|,\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\boldsymbol{\gamma}\odot|\boldsymbol{w}|\right\rangle + O(\|\boldsymbol{\gamma}\|_p^3) \\ &= \max_{\|\boldsymbol{\gamma}\|_p\leq\rho}\langle\nabla L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}|,\boldsymbol{\gamma}\rangle + \frac{1}{2}\left\langle\boldsymbol{\gamma},\left(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot(|\boldsymbol{w}||\boldsymbol{w}|^T)\right)\boldsymbol{\gamma}\right\rangle + O(\|\boldsymbol{\gamma}\|_p^3)\end{aligned}$$

If $\nabla L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}| \neq 0$, then the first order term dominates for ρ sufficiently small and we get

$$\max_{\|\boldsymbol{\gamma}\|_p\leq\rho}\langle\nabla L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}|,\boldsymbol{\gamma}\rangle = \max_{\|\boldsymbol{\gamma}\|_p\leq\rho}\|\nabla L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}|\|_q\,\|\boldsymbol{\gamma}\|_p = \rho\,\|\nabla L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}|\|_q\,.$$

Otherwise we have to consider

$$\max_{\|\boldsymbol{\gamma}\|_p\leq\rho}\frac{1}{2}\left\langle\boldsymbol{\gamma},\left(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot(|\boldsymbol{w}||\boldsymbol{w}|^T)\right)\boldsymbol{\gamma}\right\rangle.$$

If $\nabla^2 L(\boldsymbol{w})\odot(|\boldsymbol{w}||\boldsymbol{w}|^T)$ is negative definite, then the maximum is zero attained at $\boldsymbol{\gamma}=0$. In the other case, we get

$$\max_{\|\boldsymbol{\gamma}\|_p\leq\rho}\frac{1}{2}\left\langle\boldsymbol{\gamma},\left(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot(|\boldsymbol{w}||\boldsymbol{w}|^T)\right)\boldsymbol{\gamma}\right\rangle = \frac{\rho^2}{2}\max_{\boldsymbol{\gamma}\neq 0}\frac{\left\langle\boldsymbol{\gamma},\left(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot(|\boldsymbol{w}||\boldsymbol{w}|^T)\right)\boldsymbol{\gamma}\right\rangle}{\|\boldsymbol{\gamma}\|_p^2}.$$

This almost finishes the proof. Finally, it holds

$$\lim_{\rho\to 0} S^{\rho}_{max,p}(\boldsymbol{w}, |\boldsymbol{w}|) = \lim_{\rho\to 0} \mathbb{E}_{\mathcal{S}\sim P_m}\left[\max_{\|\boldsymbol{\gamma}\|_p\leq\rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\gamma}\odot|\boldsymbol{w}|) - L_{\mathcal{S}}(\boldsymbol{w})\right],$$
$$= \mathbb{E}_{\mathcal{S}\sim P_m}\left[\lim_{\rho\to 0}\max_{\|\boldsymbol{\gamma}\|_p\leq\rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\gamma}\odot|\boldsymbol{w}|) - L_{\mathcal{S}}(\boldsymbol{w})\right]$$

where for the last step we have used that $\mathbb{E}_{\mathcal{S}\sim P_m}$ is the expectation over all possible subsamples of size m and thus boils down to a finite sum for which we can drag the limit inside. □

We note that for $p=2$ it holds $q=2$ and

$$\max_{\boldsymbol{\gamma}\neq 0}\frac{\left\langle\boldsymbol{\gamma}, \left(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot(|\boldsymbol{w}||\boldsymbol{w}|^T)\right)\boldsymbol{\gamma}\right\rangle}{\|\boldsymbol{\gamma}\|_2^2} = \lambda_{\max}\Big(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot(|\boldsymbol{w}||\boldsymbol{w}|^T)\Big),$$

which is the result used in the main paper.

Proposition 2. *Let $L_{\mathcal{S}}\in C^3(\mathbb{R}^s)$, S be a finite sample of training points $(x_i,y_i)_{i=1}^n$ and let P_m denote the uniform distribution over subsamples of size $m\leq n$ from S. Then*

$$\lim_{\rho\to 0}\frac{2}{\rho^2}S^{\rho}_{avg}(\boldsymbol{w},|\boldsymbol{w}|) = \mathbb{E}_{\mathcal{S}\sim P_m}\left[tr(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}||\boldsymbol{w}|^\top)\right] + O(\rho)$$

Proof. Let us consider the loss without the subcript for clarity. Then we consider

$$\mathbb{E}_{\boldsymbol{\delta}\sim\mathcal{N}(0,\rho^2 diag(\boldsymbol{c}^2))} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w})$$

When plugging in the Taylor expansion of the loss, we see that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\delta}\sim\mathcal{N}(0,\rho^2 diag(\boldsymbol{c}^2))} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w})\\
&\quad= \mathbb{E}_{\boldsymbol{\gamma}\in\mathcal{N}(0,\rho^2\boldsymbol{I})}\left[\langle\nabla L_{\mathcal{S}}(\boldsymbol{w}),|\boldsymbol{w}|\odot\boldsymbol{\gamma}\rangle + \frac{1}{2}\langle\boldsymbol{\gamma}\odot|\boldsymbol{w}|,\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\boldsymbol{\gamma}\odot|\boldsymbol{w}|\rangle + O(\|\boldsymbol{\gamma}\|_2^3)\right]\\
&\quad= \frac{1}{2}\mathbb{E}_{\boldsymbol{\gamma}\in\mathcal{N}(0,\rho^2\boldsymbol{I})}\left[\langle\boldsymbol{\gamma}\odot|\boldsymbol{w}|,\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\boldsymbol{\gamma}\odot|\boldsymbol{w}|\rangle\right] + O(\rho^3)\\
&\quad= \frac{1}{2}\mathbb{E}_{\boldsymbol{\gamma}\in\mathcal{N}(0,\rho^2\boldsymbol{I})}\left[\langle\boldsymbol{\gamma},\left(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}||\boldsymbol{w}|^T\right)\boldsymbol{\gamma}\rangle\right] + O(\rho^3)\\
&\quad= \frac{\rho^2}{2}\mathrm{tr}(\nabla^2 L_{\mathcal{S}}(\boldsymbol{w})\odot|\boldsymbol{w}||\boldsymbol{w}|^\top) + O(\rho^3)
\end{aligned}$$

where we use that the components of $\boldsymbol{\gamma}$ are independent and have zero mean and thus the first order term vanishes and for the second order term only the diagonal entries remain which are equal to the variance ρ^2. Finally, we take the expectation with respect to P_m. As in the proof of Proposition 1 we can drag the limit inside as the expectation with respect to P_m corresponds to a finite sum. □

A.2. Derivations for Diagonal Linear Networks

Hessian for diagonal linear networks. Denote $\boldsymbol{r}=\boldsymbol{X}(\boldsymbol{u}\odot\boldsymbol{v})-\boldsymbol{y}$, $\boldsymbol{V}=\mathrm{diag}(\boldsymbol{v})$, $\boldsymbol{U}=\mathrm{diag}(\boldsymbol{u})$, then the Hessian of the loss $\nabla^2 L(\boldsymbol{w})$ for diagonal linear networks is given by:

$$L(\boldsymbol{w}) = \begin{bmatrix}\boldsymbol{V}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{V} & \boldsymbol{V}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{U}+\mathrm{diag}(\boldsymbol{X}^\top\boldsymbol{r})\\ \boldsymbol{V}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{U}+\mathrm{diag}(\boldsymbol{X}^\top\boldsymbol{r}) & \boldsymbol{U}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{U}\end{bmatrix}. \tag{7}$$

It is easy to verify that the data-dependent terms disappear due to the assumption of whitened data $\boldsymbol{X}^\top\boldsymbol{X}=\boldsymbol{I}$ and zero residuals $\boldsymbol{r}$ at a minimum. Thus, we arrive at a much simpler expression for the Hessian:

$$L(\boldsymbol{w}) = \begin{bmatrix}\mathrm{diag}(\boldsymbol{v}\odot\boldsymbol{v}) & \mathrm{diag}(\boldsymbol{v}\odot\boldsymbol{u})\\ \mathrm{diag}(\boldsymbol{v}\odot\boldsymbol{u}) & \mathrm{diag}(\boldsymbol{u}\odot\boldsymbol{u})\end{bmatrix}, \tag{8}$$

Maximum eigenvalue for diagonal linear networks. Since the Hessian has a simple block structure, we can rearrange the rows and columns coherently and get a block-diagonal structure as follows

$$\begin{bmatrix} v_1^2 & v_1u_1 & 0 & \dots & 0 \\ v_1u_1 & u_1^2 & 0 & \dots & 0 \\ 0 & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & v_d^2 & v_du_d \\ 0 & \dots & 0 & v_du_d & u_d^2 \end{bmatrix} \tag{9}$$

where eigenvalues of each 2×2 submatrix are $u_i^2 + v_i^2$ and 0. Thus, $\lambda_{\max} = \max_{1 \le i \le d} v_i^2 + u_i^2$ by using the property of block-diagonal matrices.

B. Correlation Between Sharpness and Generalization Gap

Throughout the paper we focused on correlation between sharpness and *test error*, but it is natural to ask if the picture differs if we consider correlation between sharpness and *generalization gap*, i.e., the difference between the test error and training error. We note that in the experiments on CIFAR-10 in Section 5.1, since we consider only models with $\leq 1\%$ training error and since the test error is significantly larger than 1%, the behavior of generalization gap vs. sharpness has to be almost identical to that of test error vs. sharpness. For other datasets, however, the training error is not necessarily close to 0, thus in Figure 8 and Figure 9, we additionally plot the *generalization gap* vs. sharpness (and side-by-side the test error vs. sharpness for the sake of convenience) for the ImageNet experiments. We observe only small differences in the correlation values which do not alter the conclusions about the relationship of sharpness and generalization.

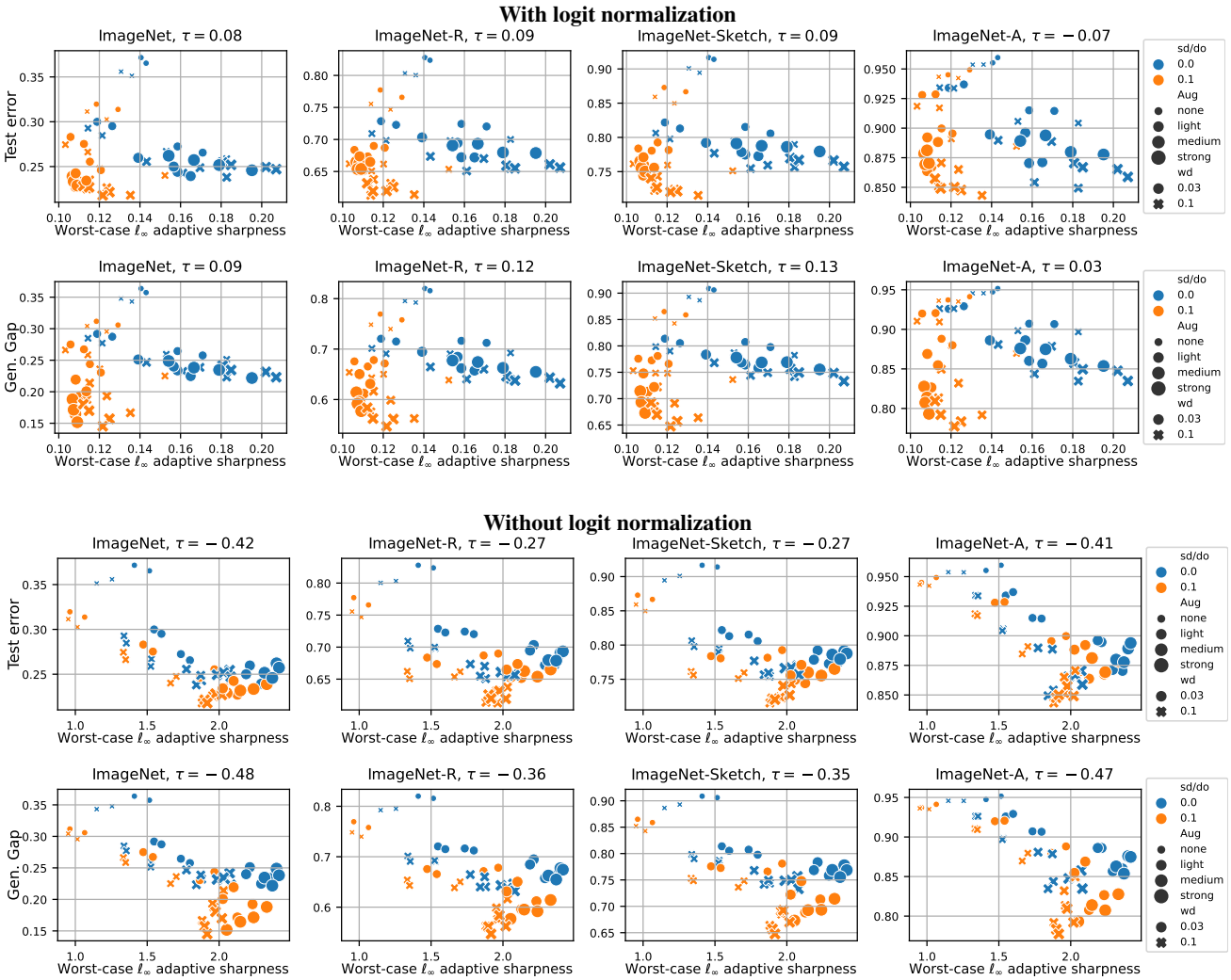


Figure 8: ViT-B/16 trained from scratch on ImageNet-1k. We show side-by-side the test error and **generalization gap** (**Gen. Gap**) for 56 models from Steiner et al. (2021) on ImageNet and its OOD variants vs. worst-case ℓ_∞ sharpness with (top) or without (bottom) normalization at $\rho = 0.002$. The color indicates models trained with stochastic depth (sd) and dropout (do), markers and their size indicate the strength of weight decay (wd) and augmentations (aug), and τ indicates the rank correlation coefficient.

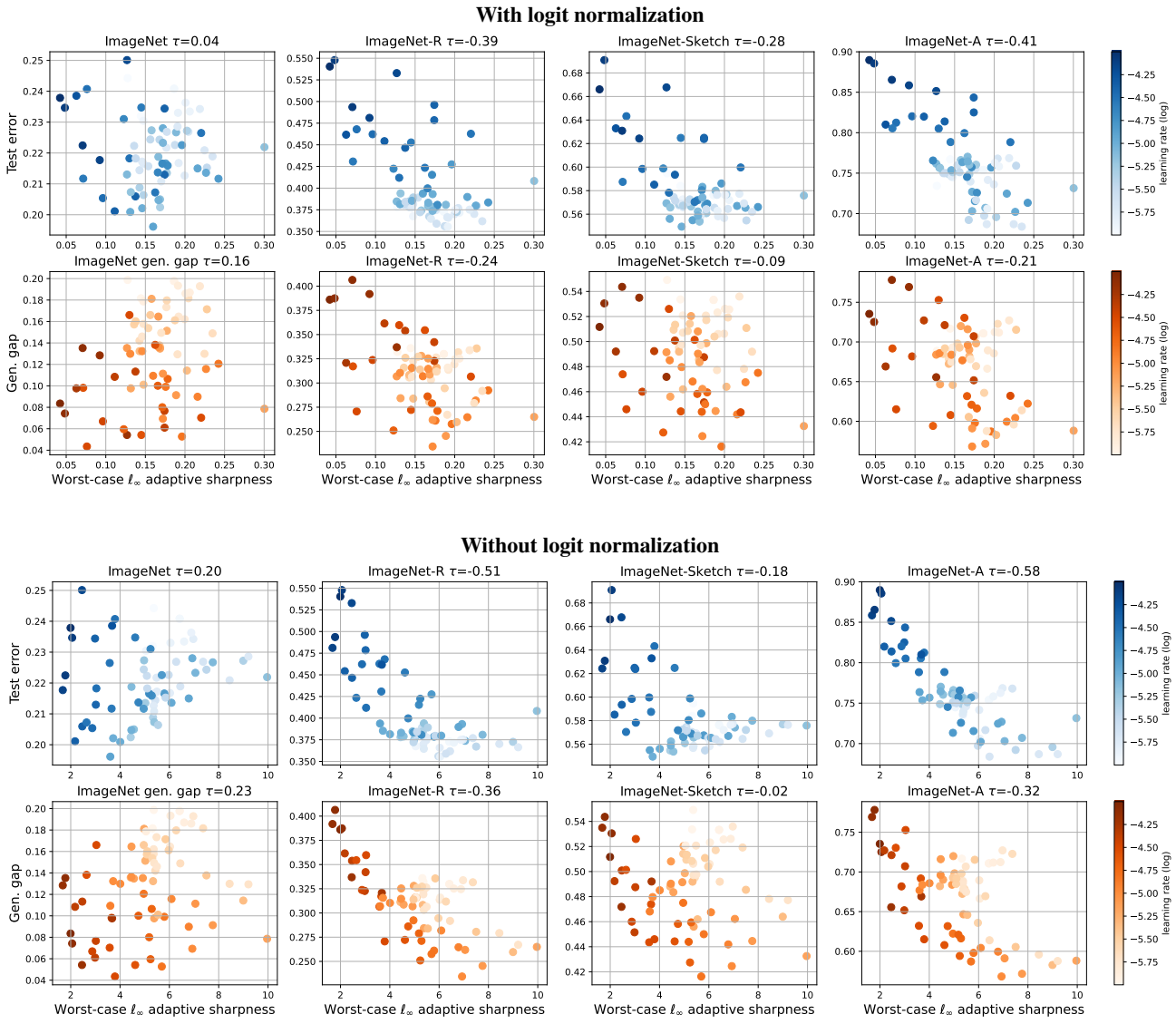


Figure 9: Fine-tuning CLIP ViT-B/32 on ImageNet-1k. We show side-by-side the test error and **generalization gap** (**Gen. gap**) for 72 models from Wortsman et al. (2022) on ImageNet and its OOD variants vs. worst-case ℓ_∞ sharpness with (top) or without (bottom) normalization at $\rho = 0.002$. Darker color indicates larger learning rate used for fine-tuning.

C. ImageNet-1k Models Trained from Scratch from (Steiner et al., 2021): Extra Details and Figures

Experimental details. As explained in the main paper, the ViT-B/16-224 weights were trained on ImageNet-1k for 300 epochs with different hyperparameter settings, and subsequently fine-tuned on the same dataset for 20.000 steps with 2 different learning rates (0.01 and 0.03). The pretraining hyperparameters include 7 augmentation types (*none, light0, light1, medium0, medium1, strong0, strong1*), which we group into (*none, light, medium, strong*) in the plots. Weight decay was either 0.1 or 0.03, and dropout and stochastic depth were either both set to 0 or both set to 0.1. We evaluated the resulting 56 configurations. The model weights can be obtained from `https://github.com/google-research/vision_transformer`.

Sharpness evaluation. For sharpness evaluation we use 2048 data points from the training set split in 8 batches: we compute sharpness on each of them and report the average. For worst-case sharpness we use Auto-PGD for 20 steps (for each batch) with random uniform initialization in the feasible set, while for average-case sharpness we sample 100 different weights perturbations for every batch. We use the same sharpness evaluation for all ImageNet-1k and MNLI models. For convenience we restate the algorithm of Auto-PGD in Algorithm 1: it follows the original version presented in Croce & Hein (2020) while using the network weights $\boldsymbol{w}$ as optimization variables instead of the input image components. In Alg. 1 we denote f the target objective function (cross-entropy loss on the batch of images in our experiments), S the feasible set of perturbations and P_S the projection onto it. Also, η and W are fixed hyperparameters (we keep the original values), and the two conditions in Line 13 can be found in Croce & Hein (2020).

Algorithm 1 Auto-PGD

Input: objective function f, perturbation set S, $\boldsymbol{w}^{(0)}$, η, N_{iter}, $W = \{w_0, \dots, w_n\}$
Output: $\boldsymbol{w}_{\text{max}}$, f_{max}
$\boldsymbol{w}^{(1)} \leftarrow P_{\mathcal{S}}\left(\boldsymbol{w}^{(0)} + \eta \nabla f(\boldsymbol{w}^{(0)})\right)$
$f_{\text{max}} \leftarrow \max\{f(\boldsymbol{w}^{(0)}), f(\boldsymbol{w}^{(1)})\}$
$\boldsymbol{w}_{\text{max}} \leftarrow \boldsymbol{w}^{(0)}$ **if** $f_{\text{max}} \equiv f(\boldsymbol{w}^{(0)})$ **else** $\boldsymbol{w}_{\text{max}} \leftarrow \boldsymbol{w}^{(1)}$
for $k = 1$ **to** $N_{\text{iter}} - 1$ **do**
  $\boldsymbol{z}^{(k+1)} \leftarrow P_{\mathcal{S}}\left(\boldsymbol{w}^{(k)} + \eta \nabla f(\boldsymbol{w}^{(k)})\right)$
  $\boldsymbol{w}^{(k+1)} \leftarrow P_{\mathcal{S}}\left(\boldsymbol{w}^{(k)} + \alpha(\boldsymbol{z}^{(k+1)} - \boldsymbol{w}^{(k)}) + (1-\alpha)(\boldsymbol{w}^{(k)} - \boldsymbol{w}^{(k-1)})\right)$
  if $f(\boldsymbol{w}^{(k+1)}) > f_{\text{max}}$ **then**
    $\boldsymbol{w}_{\text{max}} \leftarrow \boldsymbol{w}^{(k+1)}$ and $f_{\text{max}} \leftarrow f(\boldsymbol{w}^{(k+1)})$
  end if
  if $k \in W$ **then**
    if Condition 1 **or** Condition 2 **then**
      $\eta \leftarrow \eta/2$ and $\boldsymbol{w}^{(k+1)} \leftarrow \boldsymbol{w}_{\text{max}}$
    end if
  end if
end for

Extra figures. For each sharpness definition we show for three values of ρ the correlation between test error on ImageNet (in-distribution) and on the various distribution shifts. In particular, we use worst-case ℓ_∞ adaptive sharpness with (Fig. 10) and without (Fig. 11) logit normalization, and average-case adaptive sharpness with (Fig. 12) and without (Fig. 13) logit normalization. For all figures the color shows stochastic depth / dropout, the marker size corresponds to augmentation strength, and the marker type to weight decay. In addition to the OOD-datasets from the main paper, we here report the results for ImageNet-V2 (Recht et al., 2019) and ObjectNet (Barbu et al., 2019). ImageNet-V2 consists in a new test set for ImageNet models and is sampled from the same image distribution as the existing validation set: then, the performance of the classifiers on it are highly correlated to that on ImageNet validation set, and ImageNet-V2 cannot be considered a distribution shift in the same sense as the other datasets. In general, we observe that sharpness variants are not predictive of the performance on ImageNet and the OOD datasets, typically only separating models by stochastic depth / dropout, but not ranking them according to generalization properties, and often even yielding a negative correlation with OOD test error. The only case where low sharpness indicates low test-error is for logit-normalized average-case adaptive sharpness on ImageNet and ImageNet-v2. For the remaining OOD datasets, however, there are always models with low sharpness and larger test error.

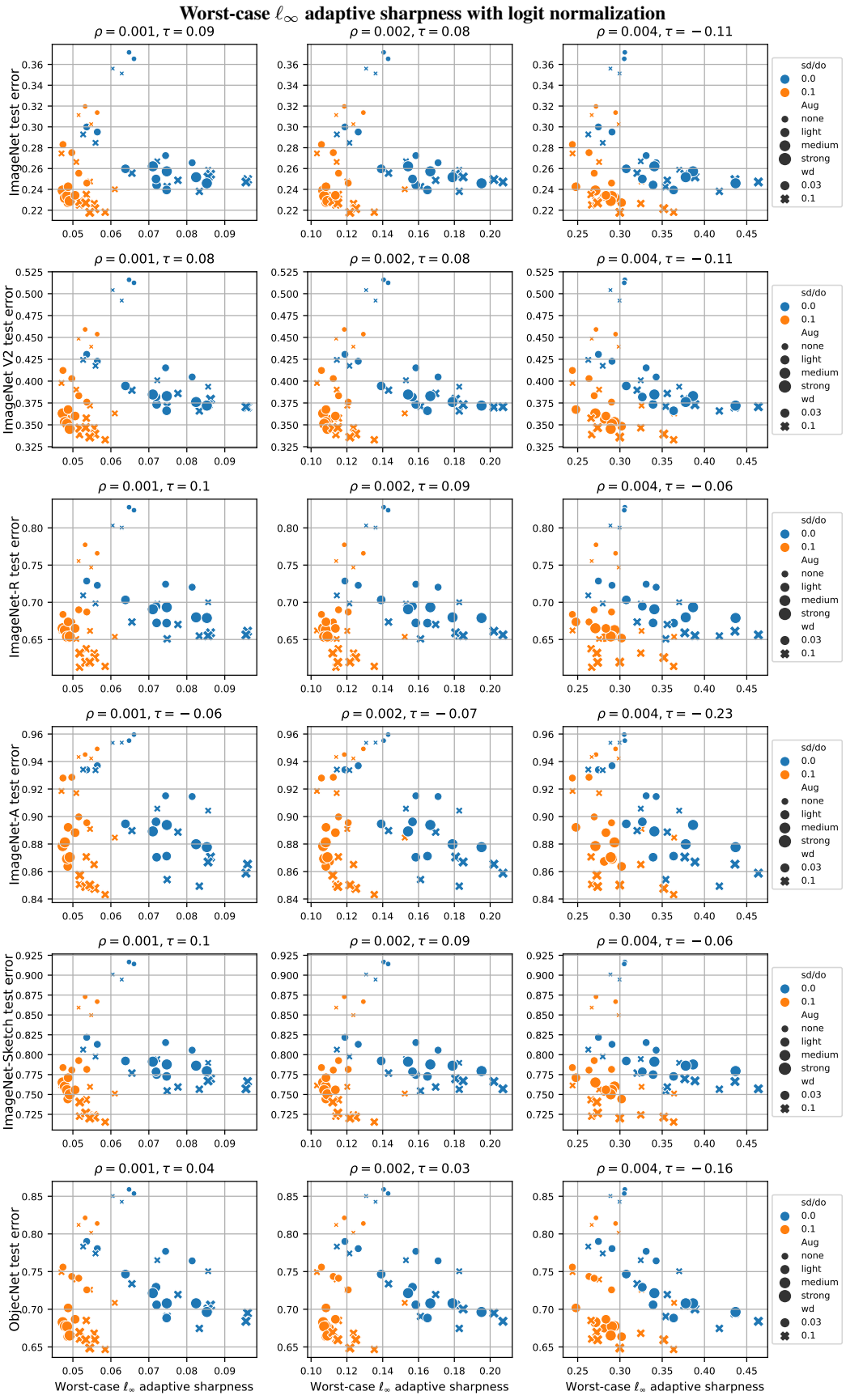


Figure 10: Correlation of sharpness with generalization on ImageNet for different ρ and for different distribution shifts.

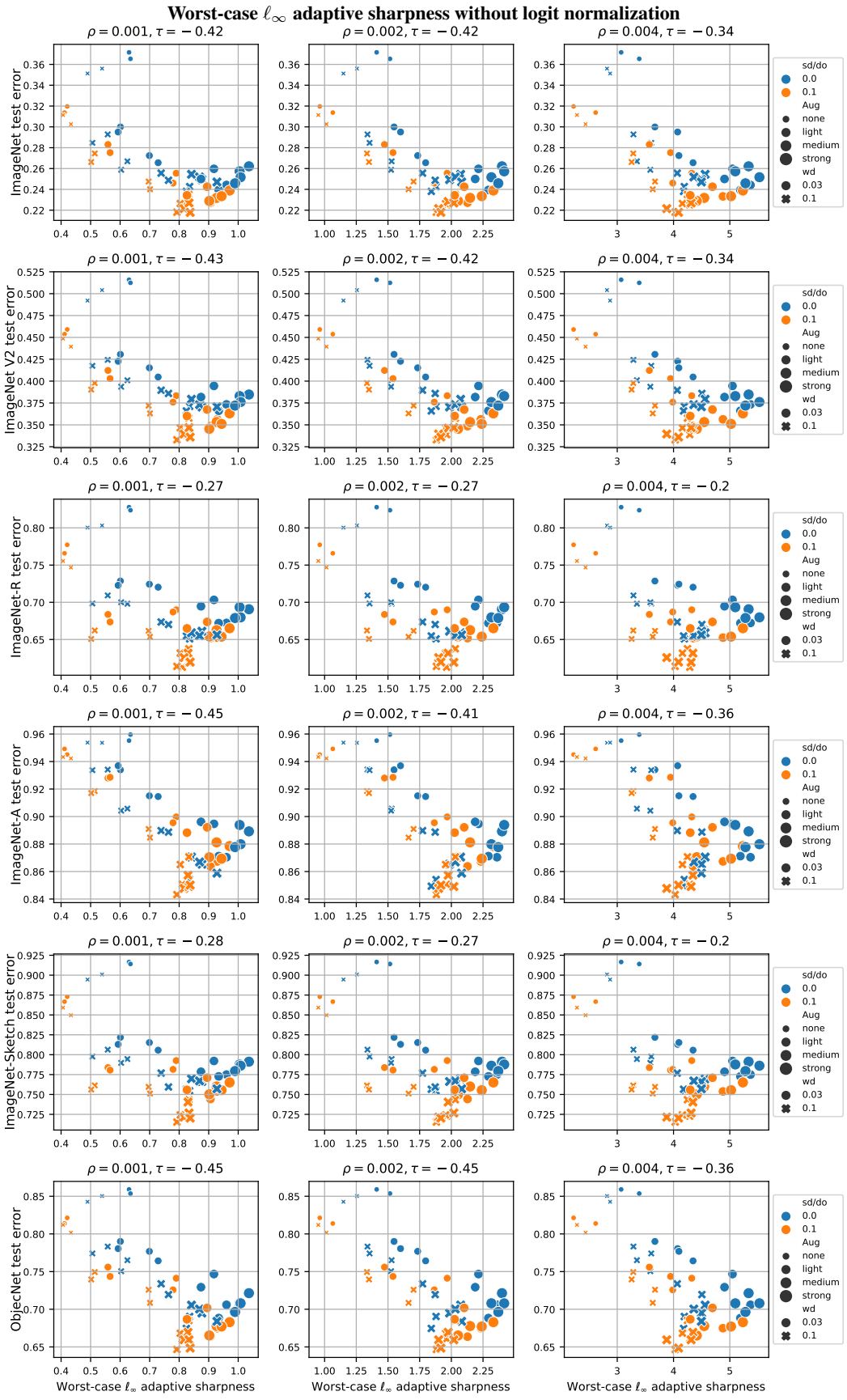


Figure 11: Correlation of sharpness with generalization on ImageNet for different ρ and for different distribution shifts.

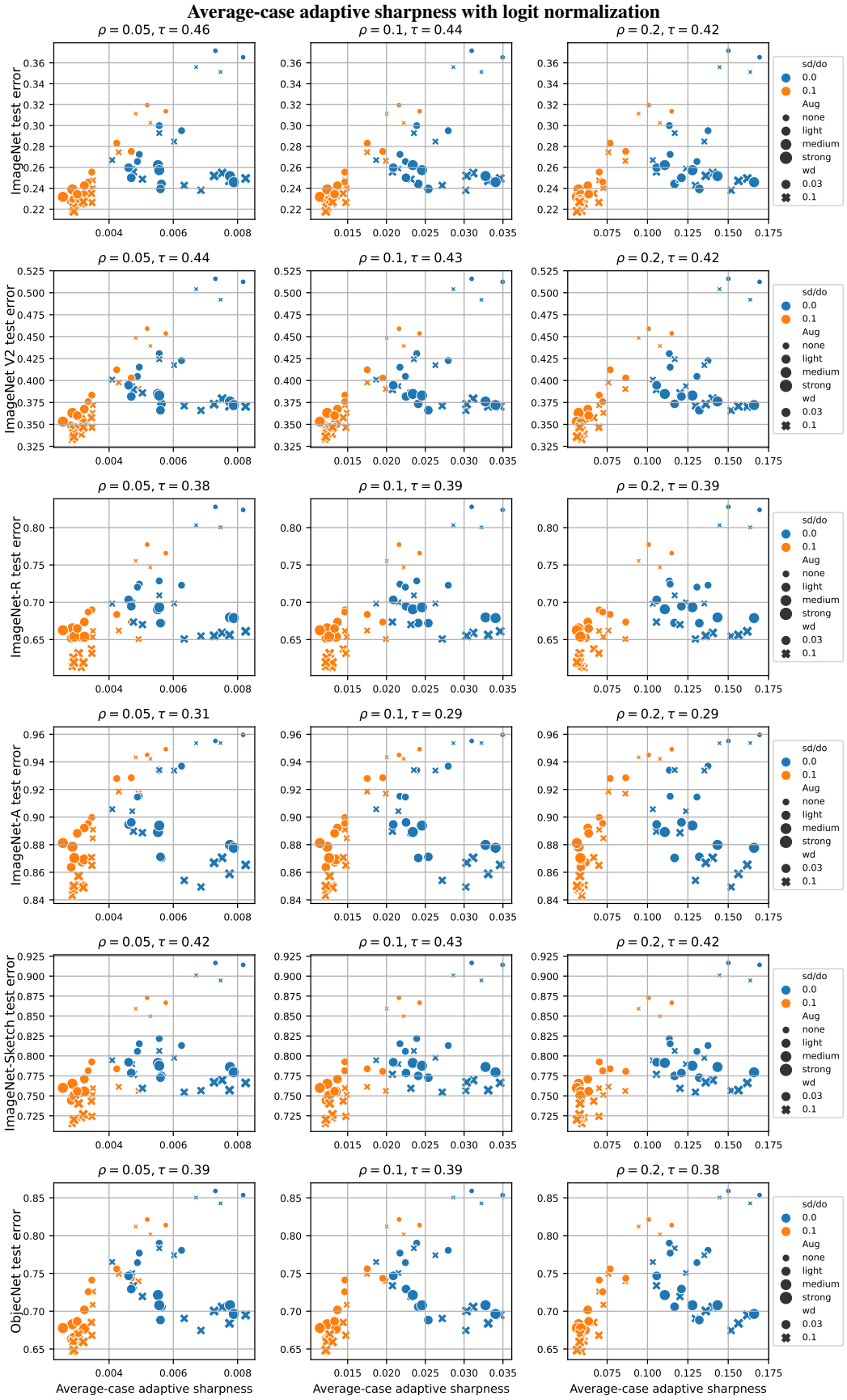


Figure 12: Correlation of sharpness with generalization on ImageNet for different ρ and for different distribution shifts.

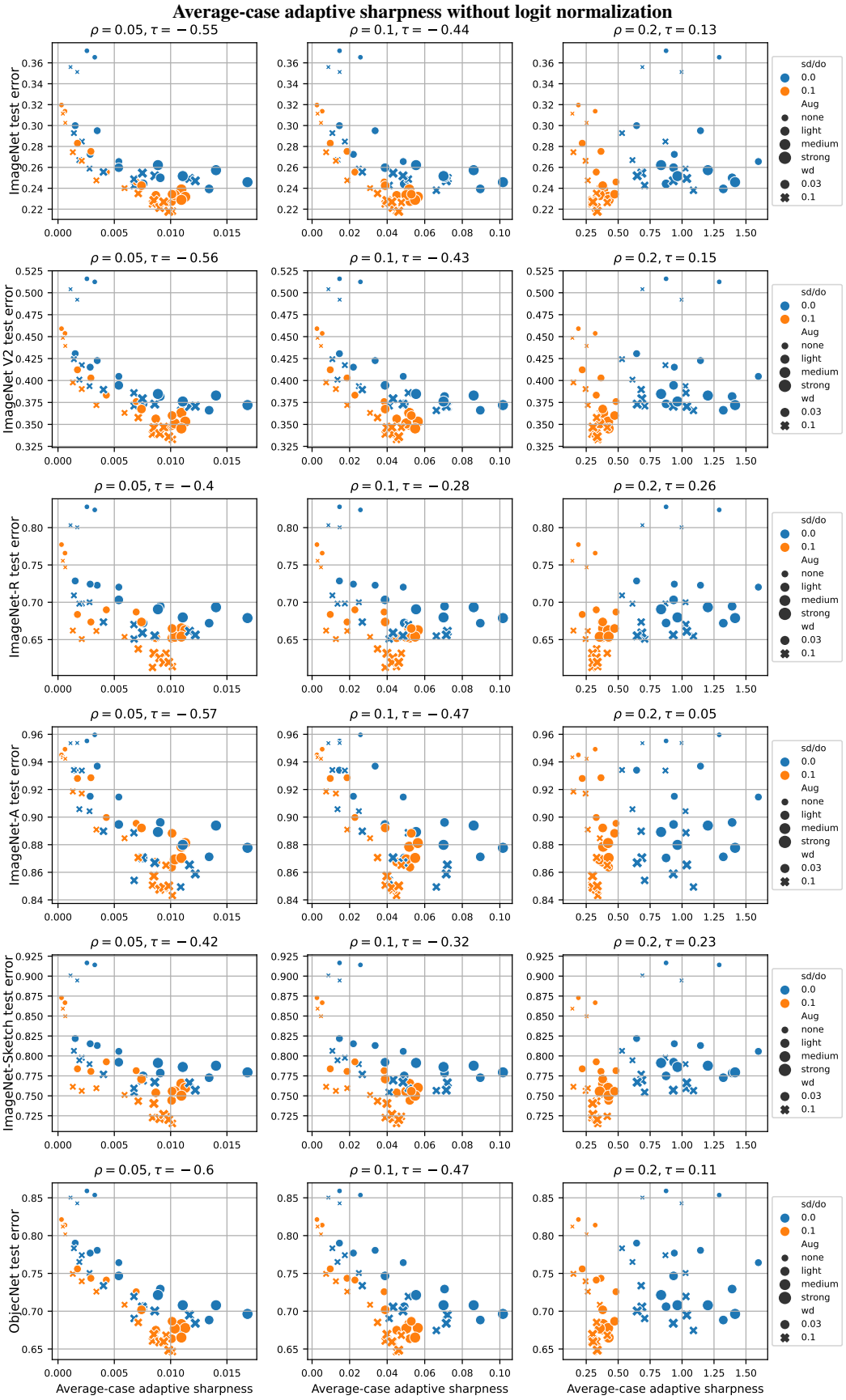


Figure 13: Correlation of sharpness with generalization on ImageNet for different ρ and for different distribution shifts.

D. Fine-tuning of ImageNet-1k Models Pretrained on ImageNet-21k from Steiner et al. (2021): Extra Figures and Details

Experimental details. All hyperparameter settings are identical to those explained in Appendix C, only the pretraining dataset is ImageNet-21k instead of ImageNet-1k. Since two of the models showed close to 100% test error, we did not evaluate them, resulting in 54 instead of 56 models.

Extra figures. Like in Appendix C we show each sharpness definition for three values of ρ and its the correlation to test error on ImageNet (in-distribution) and on the various distribution shifts. The observations are very similar to those on ImageNet-1k pretraining: sharpness variants are not predictive of the performance on ImageNet and the distribution shift datasets, typically only separating models by stochastic depth / dropout, and often even yielding a negative correlation with OOD test error.

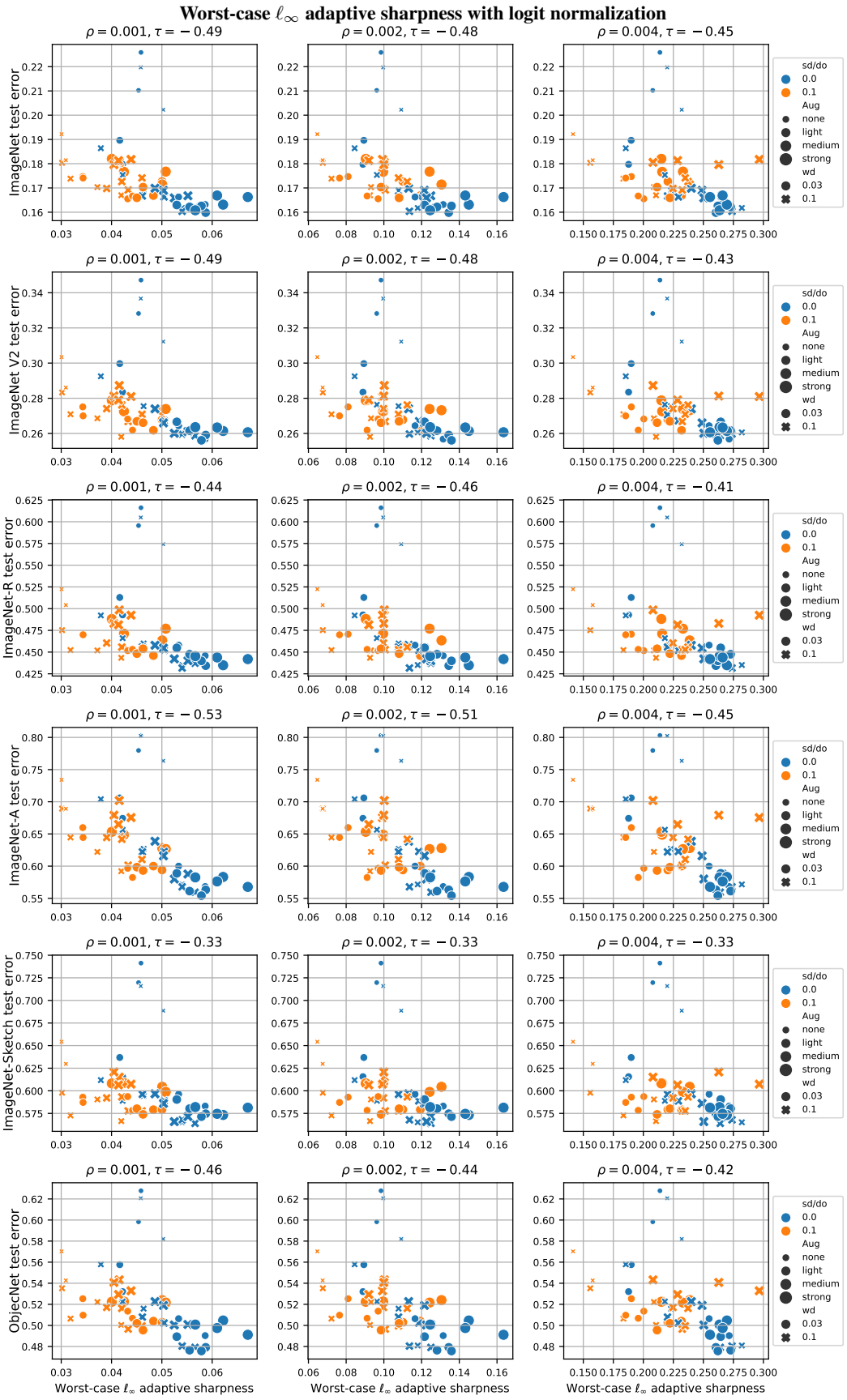


Figure 14: Correlation of sharpness with generalization on ImageNet for different ρ and for different distribution shifts.

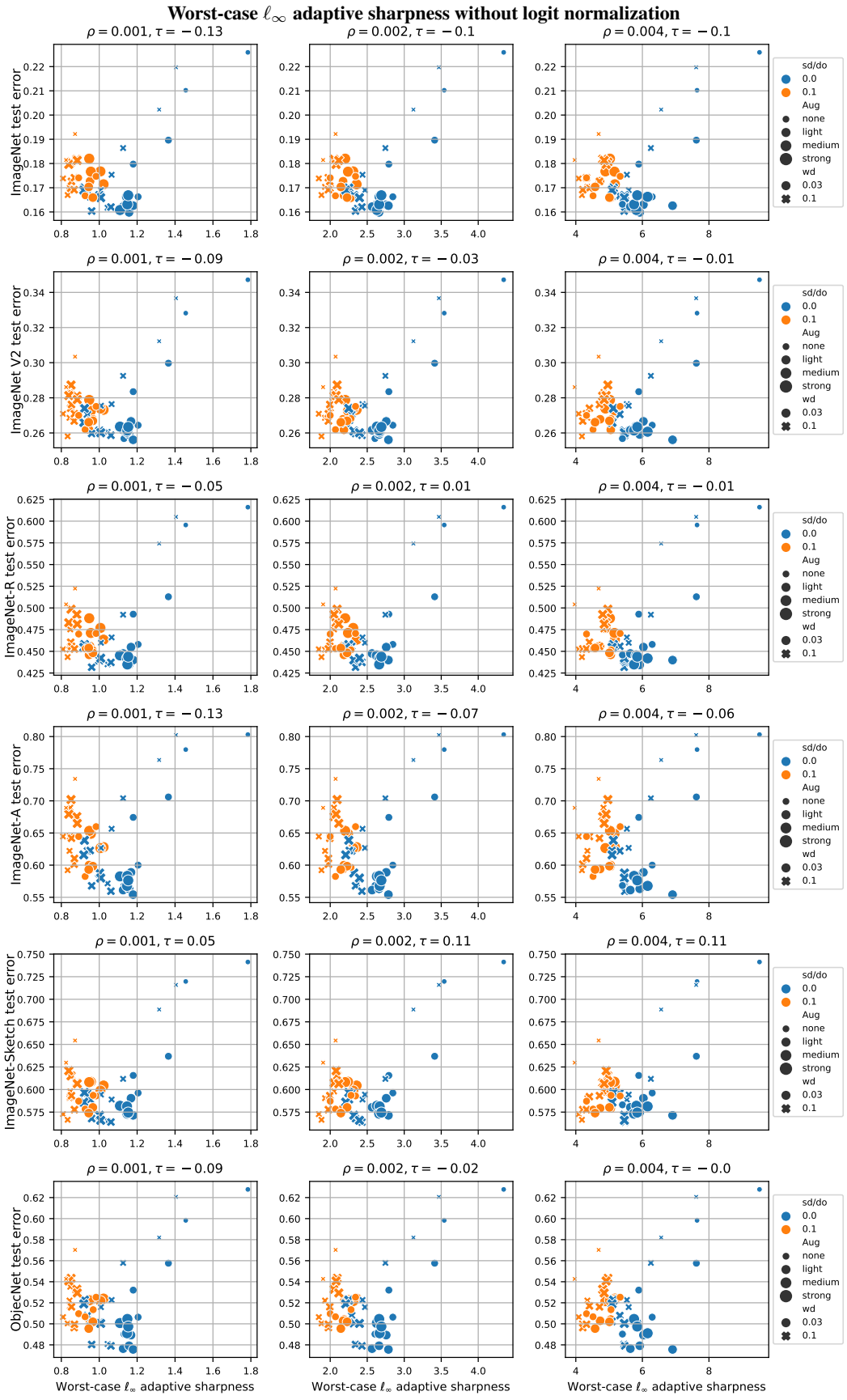


Figure 15: Correlation of sharpness with generalization on ImageNet for different ρ and for different distribution shifts.

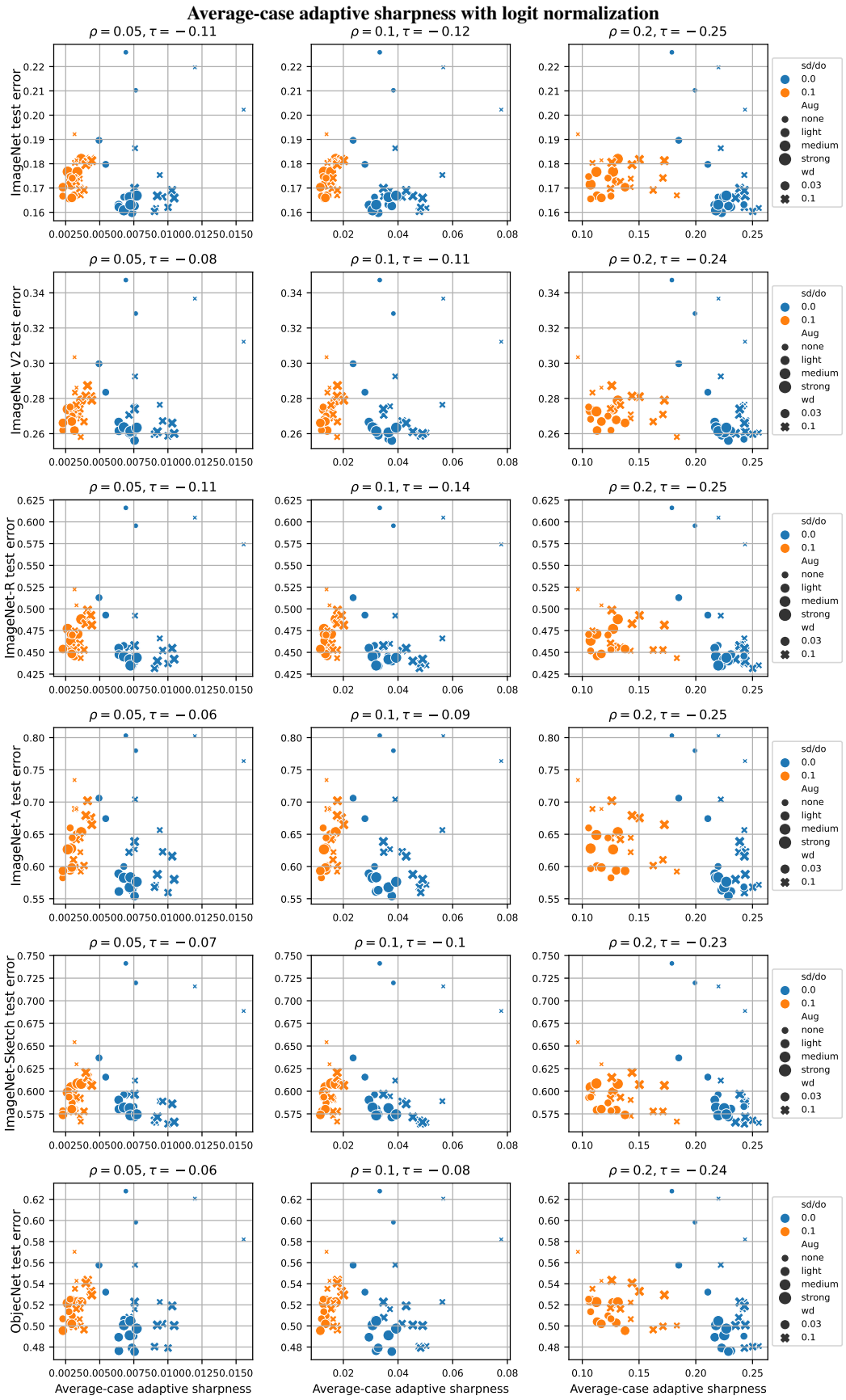


Figure 16: Correlation of sharpness with generalization on ImageNet for different ρ and for different distribution shifts.

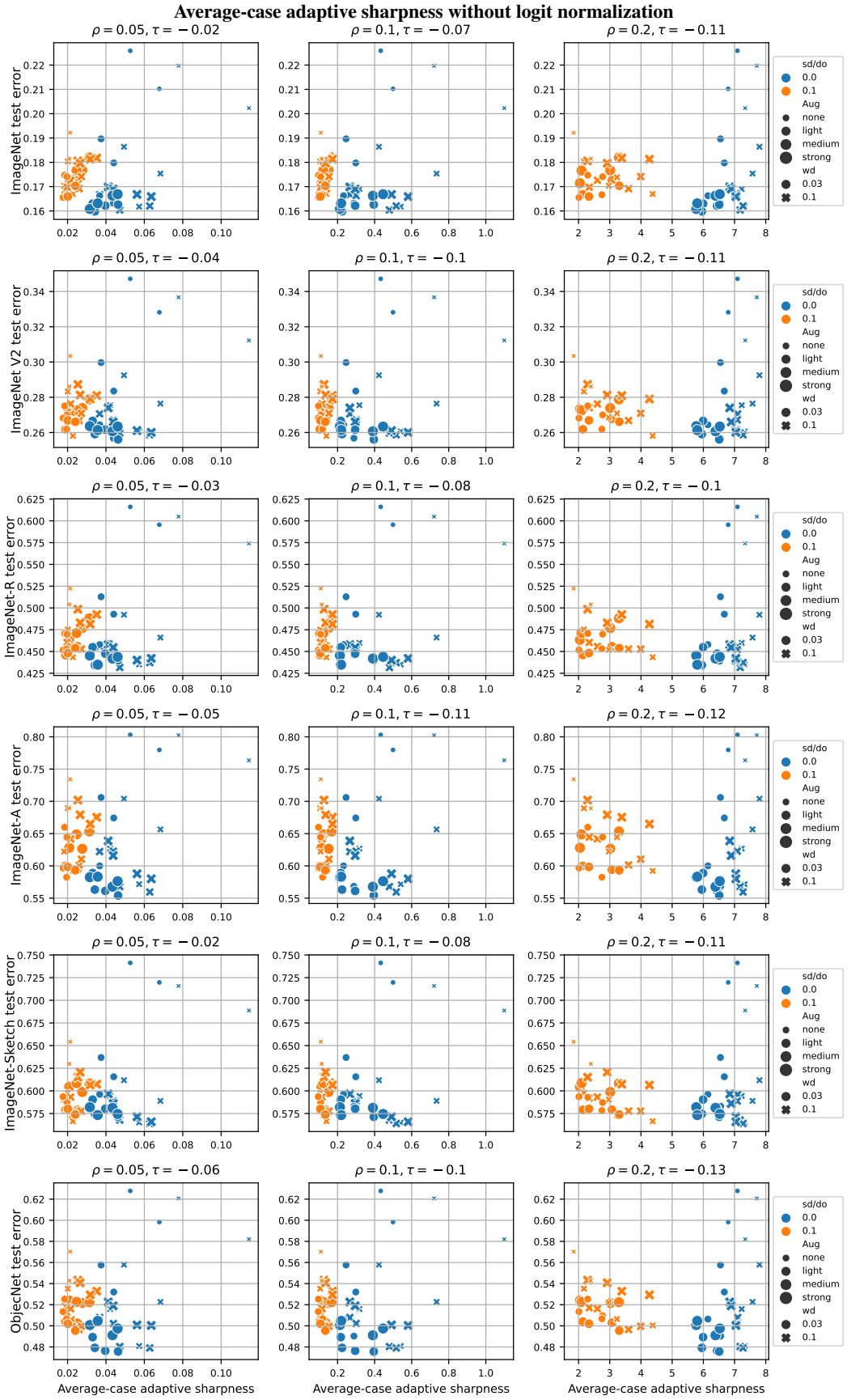


Figure 17: Correlation of sharpness with generalization on ImageNet for different ρ and for different distribution shifts.

E. ImageNet Models both Pretrained on ImageNet-1k and ImageNet-21k from Steiner et al. (2021)

For completeness, we here show for two sharpness definitions the models pretrained on ImageNet-21k and ImageNet-1k together. We find the better-generalizing models pretrained on ImageNet-21k to have significantly higher worst-case sharpness, and roughly equal or higher logit-normalized average-case adaptive sharpness, underlining that the models generalization properties resulting from different pretraining datasets are not captured.

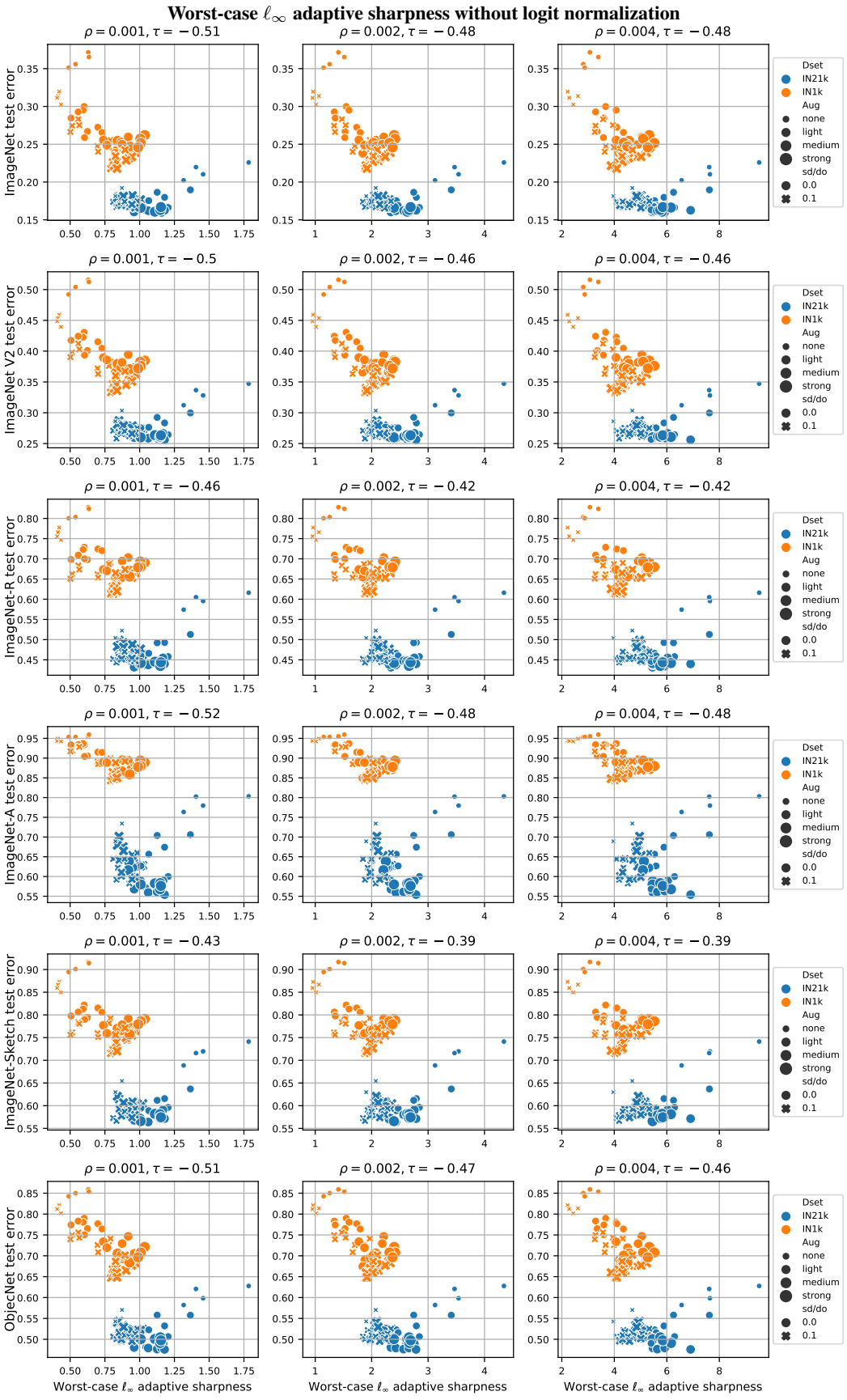


Figure 18: Correlation of sharpness with generalization on ImageNet-1k for different ρ and for different distribution shifts.

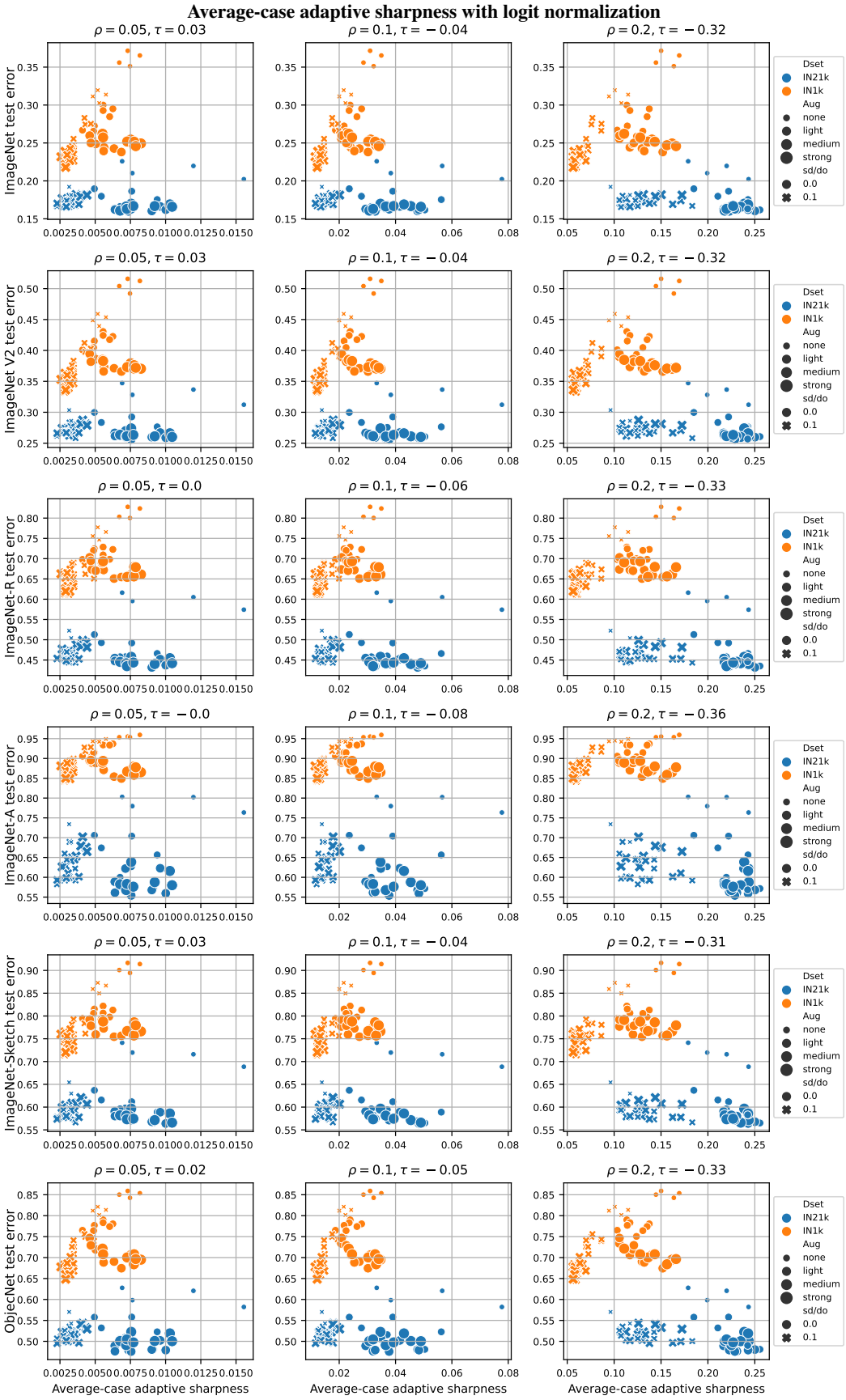


Figure 19: Correlation of sharpness with generalization on ImageNet-1k for different ρ and for different distribution shifts.

F. Fine-tuning CLIP Models on ImageNet: Extra Details and Figures

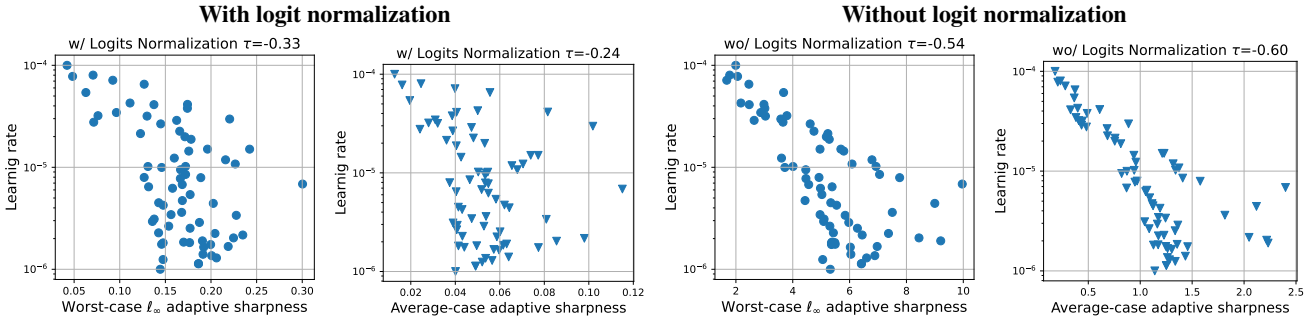


Figure 20: Fine-tuning CLIP ViT-B/32 on ImageNet-1k. Sharpness negatively correlates with the size of learning rate for fine-tuning, both with (left) and without (right) using logit normalization. For worst-case sharpness $\rho = 0.002$ is used, for average-case sharpness $\rho = 0.1$.

Experimental details. We take advantage of the models fine-tuned by Wortsman et al. (2022a) from a pre-trained CLIP ViT-B/32, with randomly sampled training hyperparameters (see *random search* setup in Wortsman et al. (2022a)), for which the evaluation of ImageNet validation set and distribution shifts are provided.

Extra figures. For each sharpness definition we show for three values of ρ the correlation between test error on ImageNet (in-distribution) and on the various distribution shifts. In particular, we use worst-case ℓ_∞ adaptive sharpness with (Fig. 21) and without (Fig. 22) logit normalization, and average-case adaptive sharpness with (Fig. 23) and without (Fig. 24) logit normalization. For all figures we represent with colors represent the size of the learning rate used for fine-tuning (darker color means larger learning rate). In addition to the datasets shown in Sec. 4, we here report the results for ImageNet-V2 (Recht et al., 2019) and ObjectNet (Barbu et al., 2019). ImageNet-V2 consists in a new test set for ImageNet models and is sampled from the same image distribution as the existing validation set: then, the performance of the classifiers on it are highly correlated to that on ImageNet validation set, and ImageNet-V2 cannot be considered a distribution shift in the same sense as the other datasets. In general, we observe that sharpness variants are not predictive of the performance on ImageNet and ImageNet-V2. Moreover, there is in most cases a negative correlation with test error in presence of distribution shifts. We hypothesize that this is related to the influence that the learning rate has on sharpness (see Fig. 20), i.e. lower values lead to sharper models.

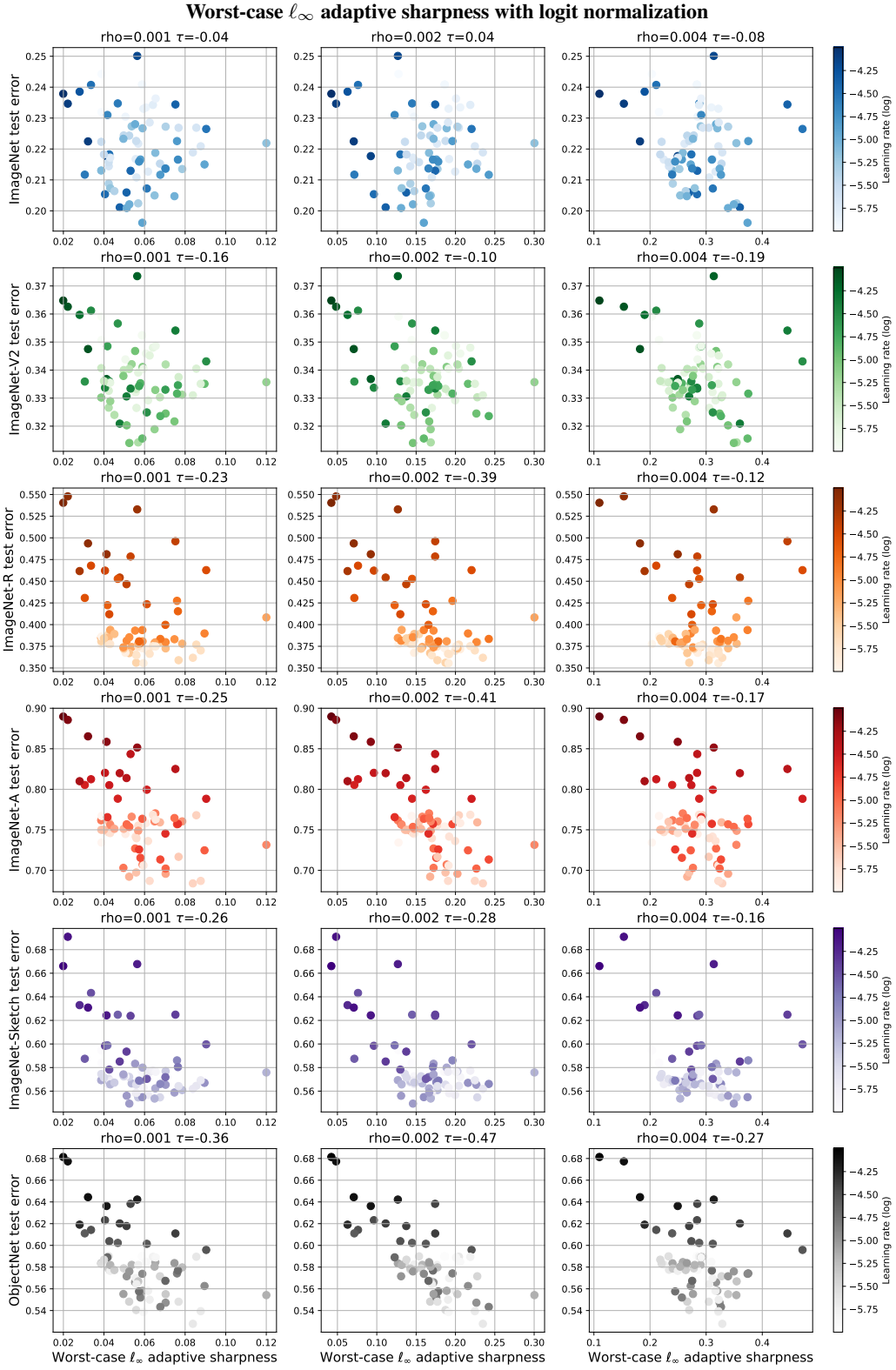


Figure 21: Correlation of sharpness with varying ρ with generalization on ImageNet for different distribution shifts.

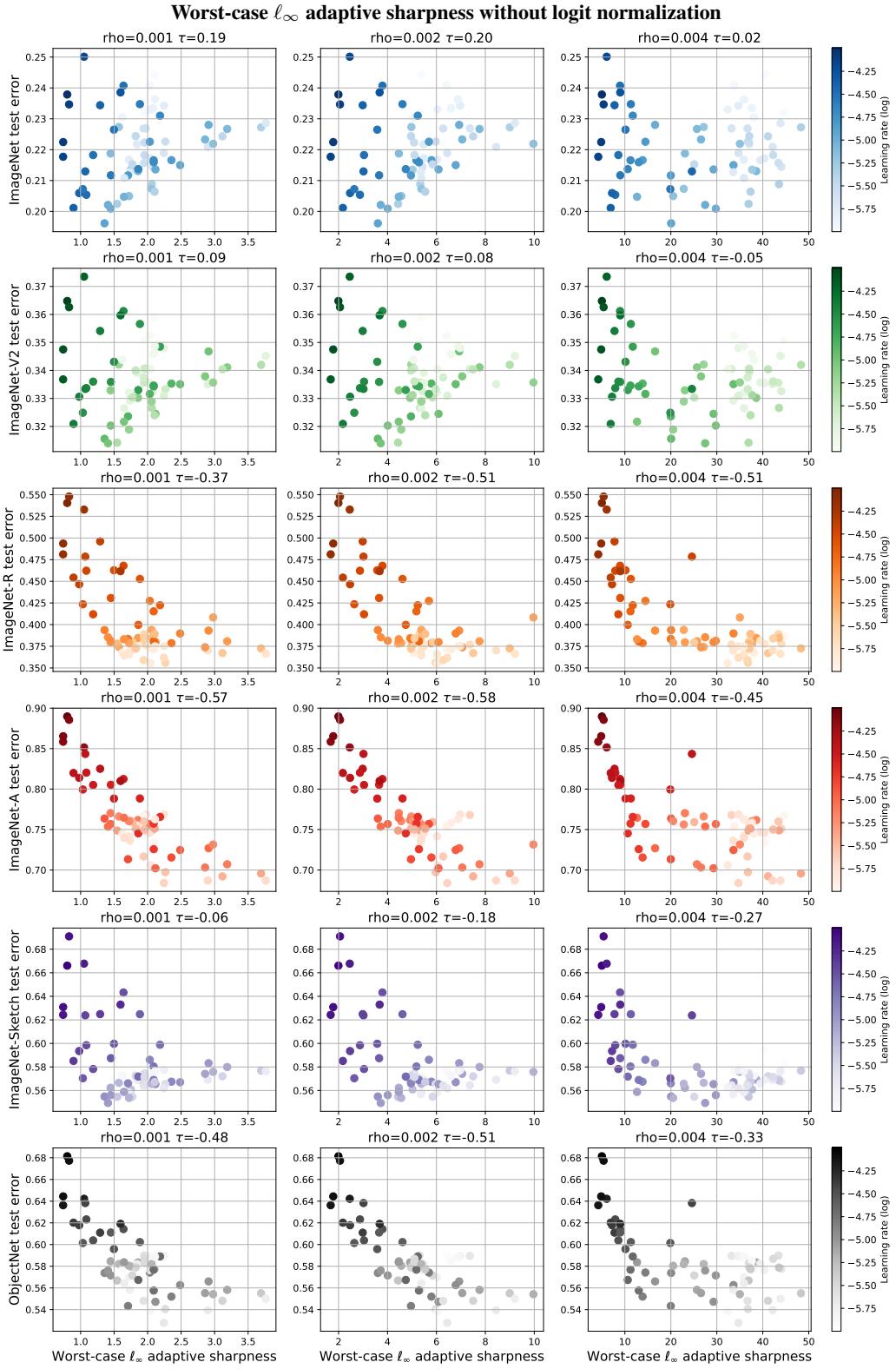


Figure 22: Correlation of sharpness with varying ρ with generalization on ImageNet for different distribution shifts.

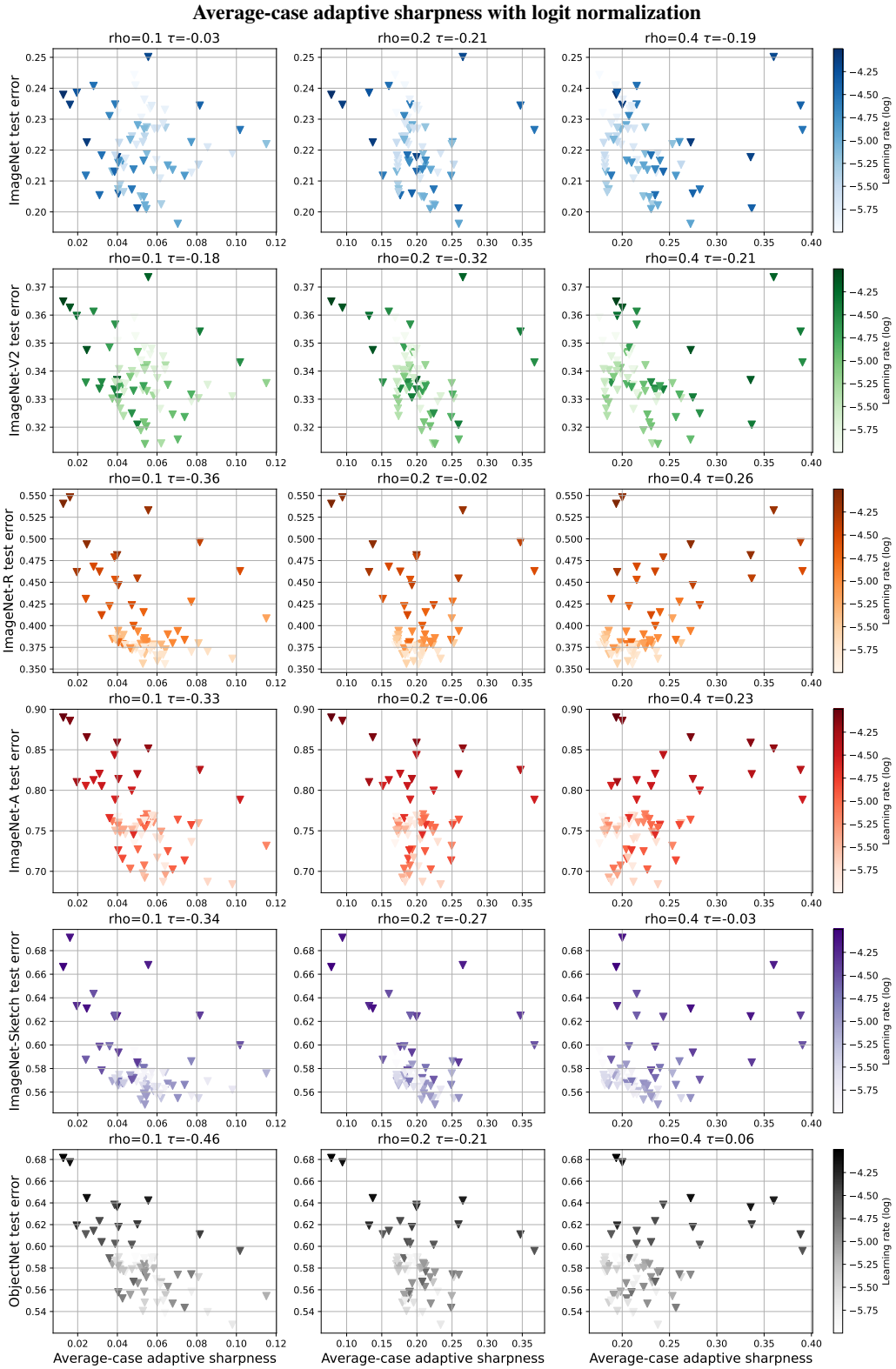


Figure 23: Correlation of sharpness with varying ρ with generalization on ImageNet for different distribution shifts.

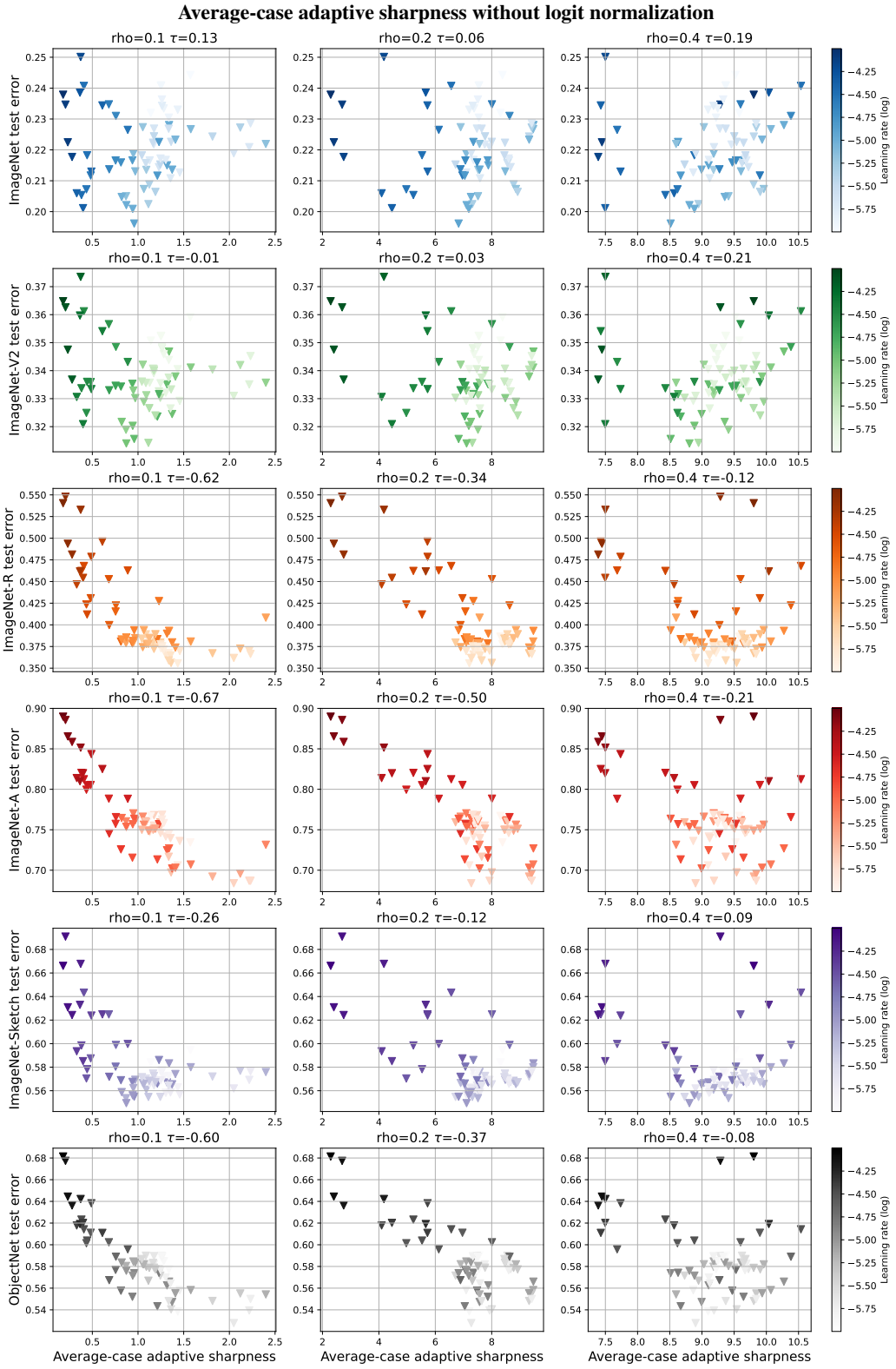


Figure 24: Correlation of sharpness with varying ρ with generalization on ImageNet for different distribution shifts.

G. Fine-tuning on MNLI: Extra Details and Figures

Experimental details. The models from McCoy et al. (2020) we use are BERT fine-tuned with initialization weights of `bert-case-uncased`. The in-distribution test error is computed on the MNLI `matched` development set, that is a classification task with three classes. As out-of-distribution datasets we use three categories of HANS considered "Inconsistent with heuristic" (see McCoy et al. (2020): *Lexical overlap*, on which the classifiers show the largest variance in test error, *Subsequence* and *Constituent*. In this case, there are only two possible classes.

Extra figures. For each sharpness definition we show for three values of ρ the correlation between test error on MNLI (in-distribution) and on various HANS subsets (out-of-distribution). In particular, we use worst-case ℓ_∞ adaptive sharpness with (Fig. 25) and without (Fig. 26) logit normalization, and average-case adaptive sharpness with (Fig. 27) and without (Fig. 28) logit normalization. For all figures we represent with darker colors the models with higher test error on MNLI. In general, all sharpness variants we consider are not predictive of the generalization performance of the model, and in some cases (e.g. Fig. 28) there is rather a weak negative correlation between sharpness and test error on out-of-distribution tasks.

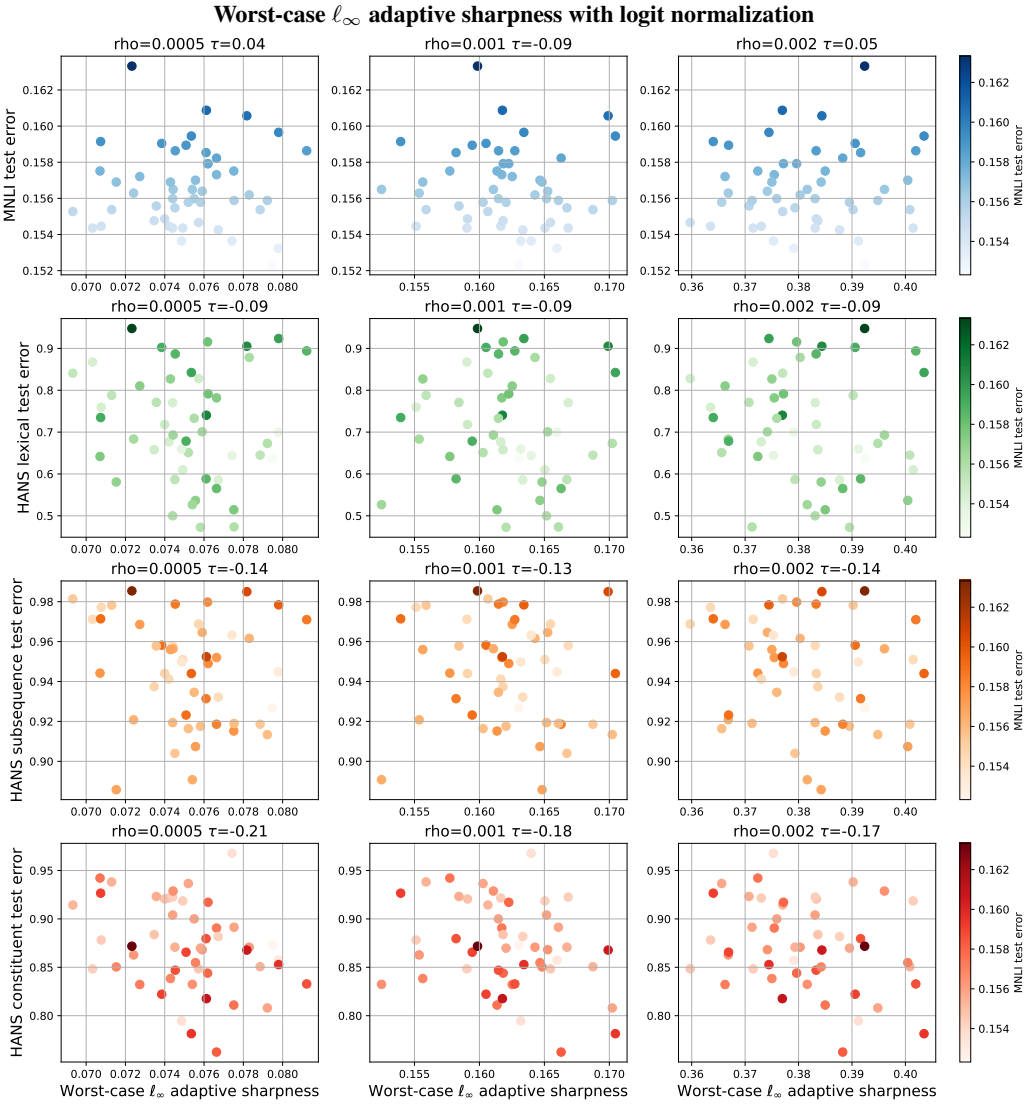


Figure 25: Correlation of sharpness with varying ρ with generalization on MNLI for different distribution shifts.

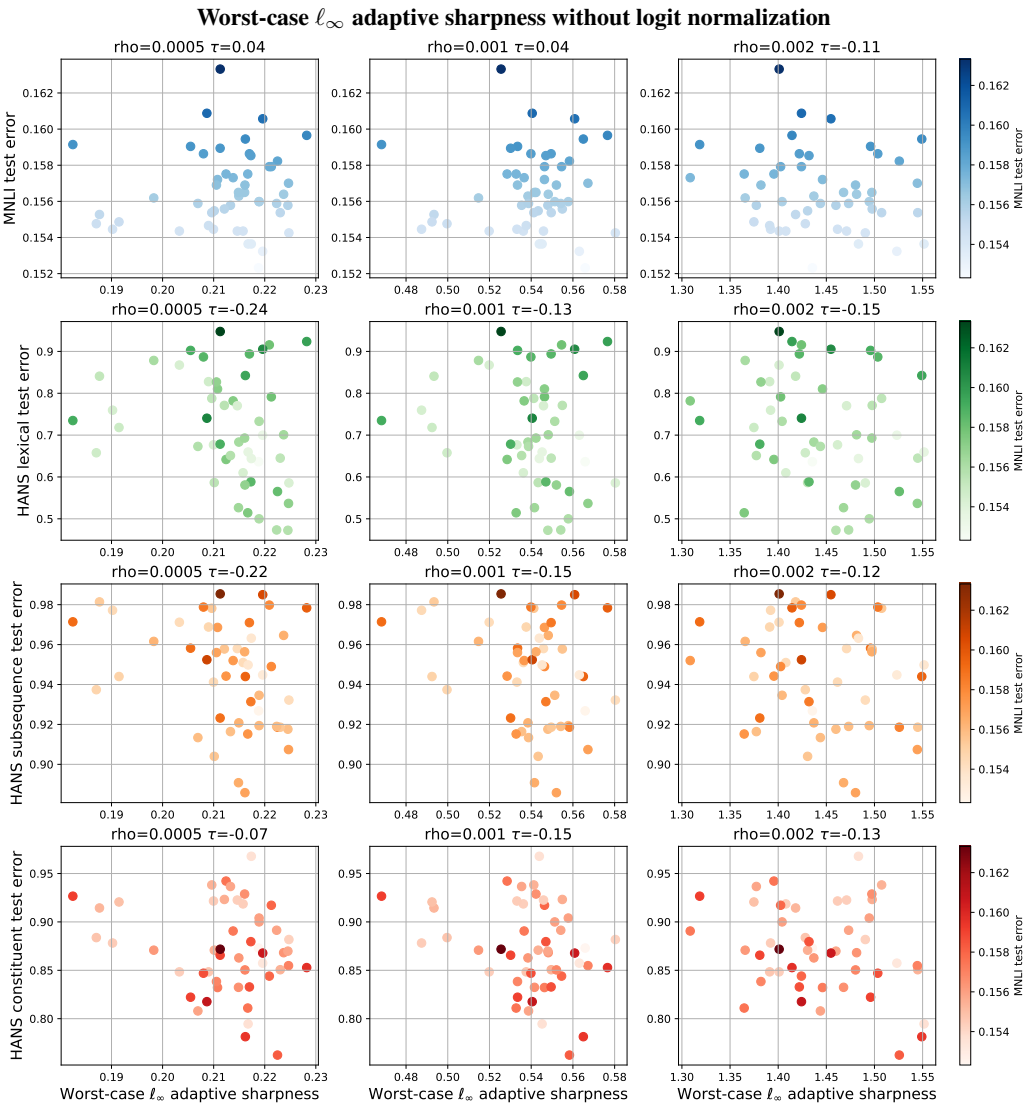


Figure 26: Correlation of sharpness with varying ρ with generalization on MNLI for different distribution shifts.

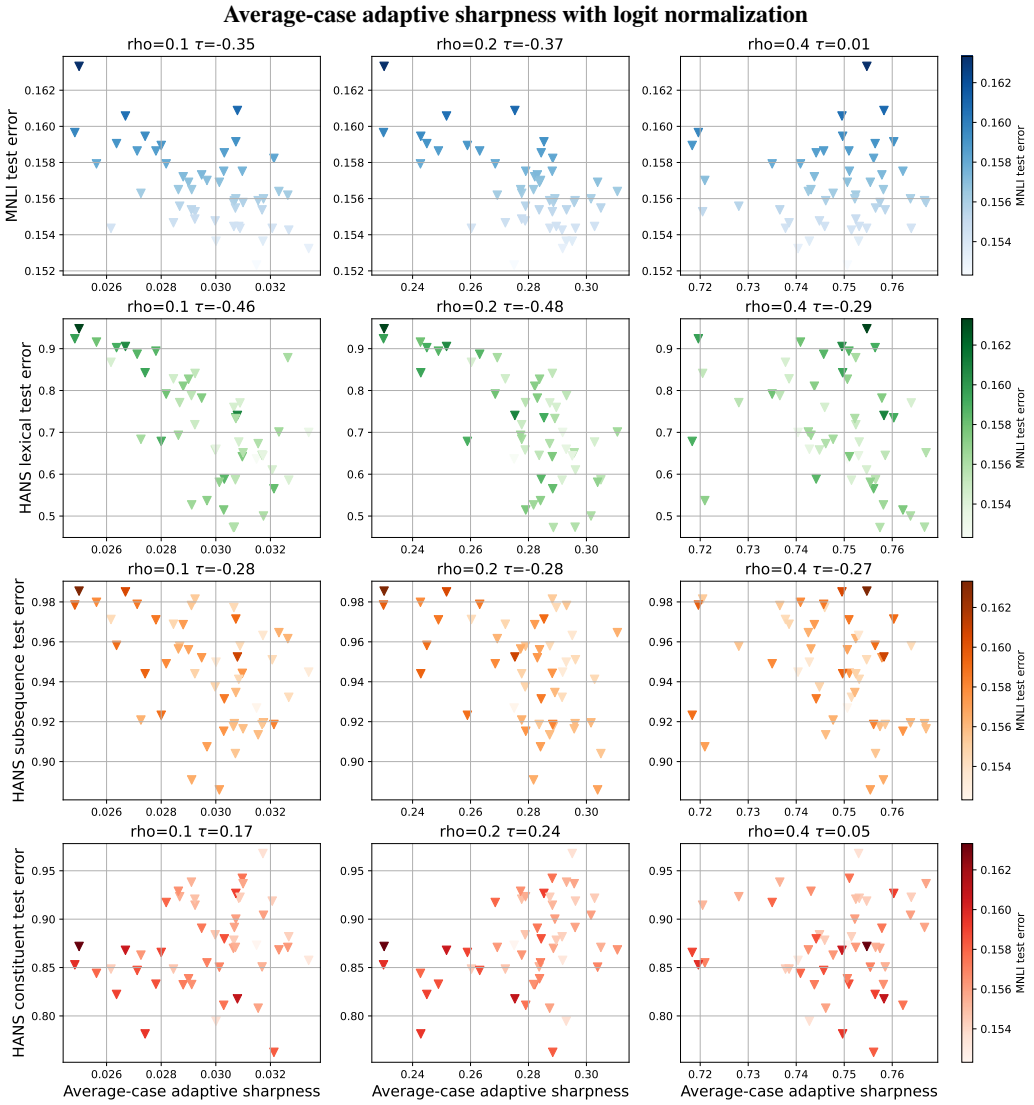


Figure 27: Correlation of sharpness with varying ρ with generalization on MNLI for different distribution shifts.

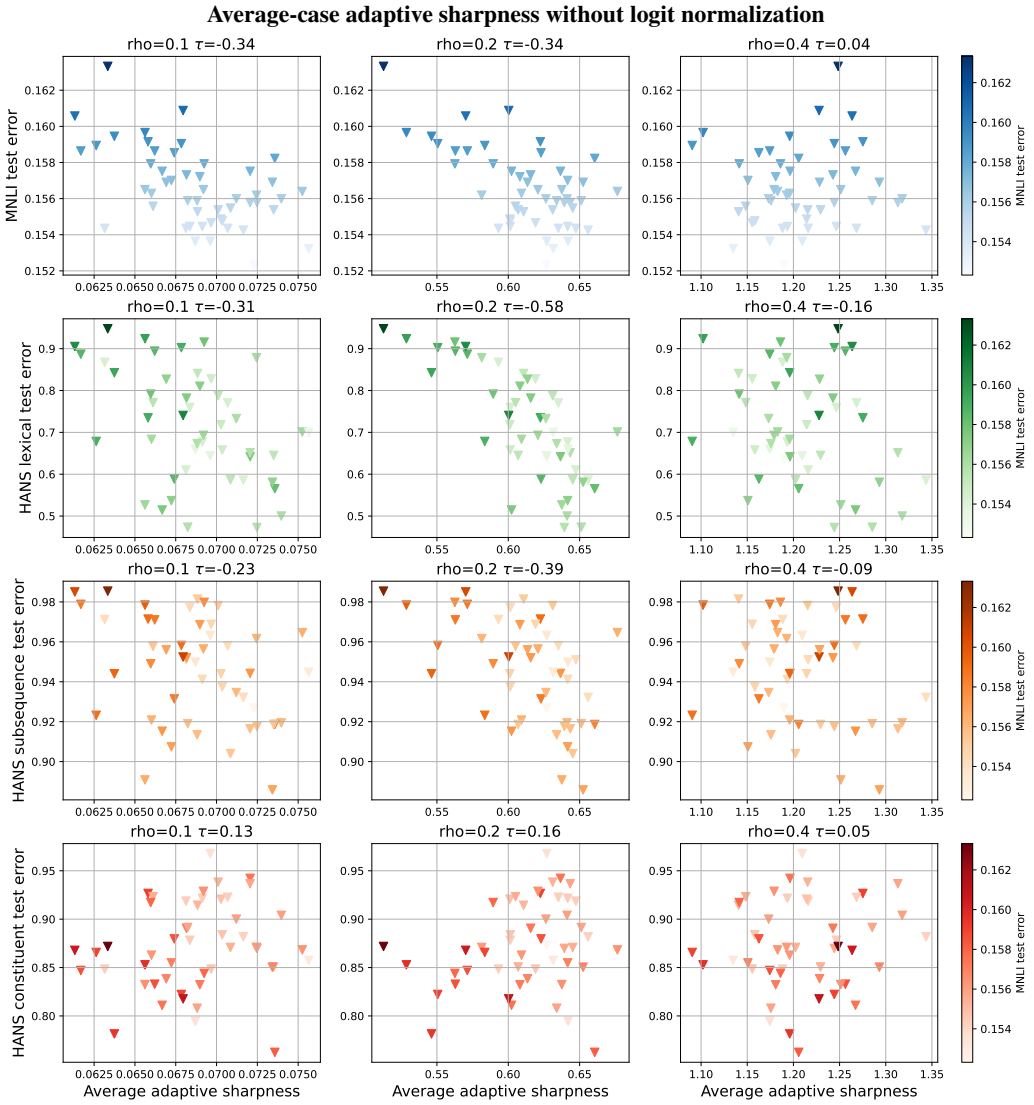


Figure 28: Correlation of sharpness with varying ρ with generalization on MNLI for different distribution shifts.

H. Training from Scratch on CIFAR-10: Extra Details and Figures

Extra details. We train 200 ResNet-18 and 200 ViT models for 200 epochs using SGD with momentum and linearly decreasing learning rates after a linear warm-up for the first 40% iterations. We found that adding such warm-up to SGD allows us to bridge the gap between SGD and Adam training for ViTs. We use the SimpleViT architecture from the `vit-pytorch` library which is a modification of the standard ViT (Dosovitskiy et al., 2021) with a fixed positional embedding and global average pooling instead of the CLS embedding. We use a ViT model with 4×4 patches, depth of 6 blocks, with 16 heads, embedding size 512, and MLP dimension of 1024. We sample the learning rate from the log-uniform distribution in the range $[0.005, 0.5]$ for ViTs and $[0.05, 5.0]$ for ResNets. We sample uniformly $\rho \in \{0, 0.05, 0.1\}$ of SAM (Foret et al., 2021), with probability 50% mixup ($\alpha = 0.5$) (Zhang et al., 2018), and with probability 50% standard augmentations combined with RandAugment (with parameters $N = 2$, $M = 14$) (Cubuk et al., 2020). We use $2\times$ repeated augmentations to reduce the augmentation variance from RandAugment (Fort et al., 2021). For CIFAR-10 models, we only show sharpness for well-trained models that have $\leq 1\%$ training error. We note that this selection criterion leaves more ResNets than ViTs on the figures below.

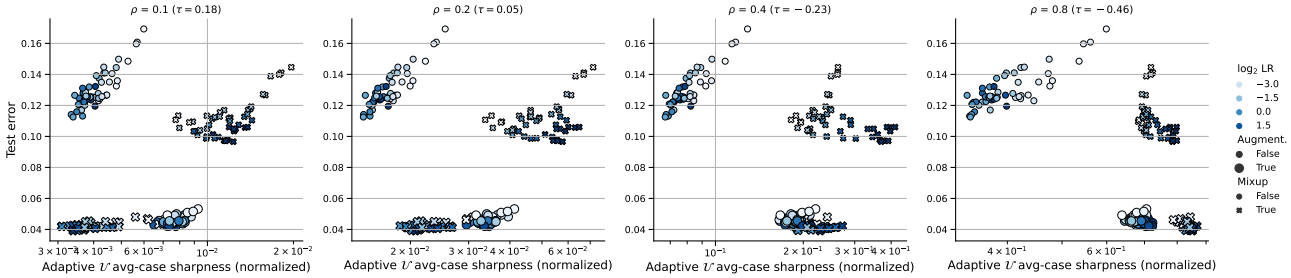
Sharpness evaluation. For sharpness evaluation we use 1024 data points from the training set split in 8 batches: we compute sharpness on each of them and report the average. For worst-case sharpness we use Auto-PGD for 20 steps (for each batch) with random uniform / Gaussian initialization in the feasible set depending on the ℓ_∞ vs. ℓ_2 norm of sharpness. For average-case sharpness, we sample 100 different weights perturbations for every batch.

Extra figures. We present additional figures in Sec. H.1 on the role of data used to evaluate sharpness, in Sec. H.2 on the role of the number of iterations in Auto-PGD to estimate sharpness, in Sec. H.3 on the role of m in m-sharpness, and in Sec. H.4 on the influence of different sharpness definitions and radii on correlation with generalization.

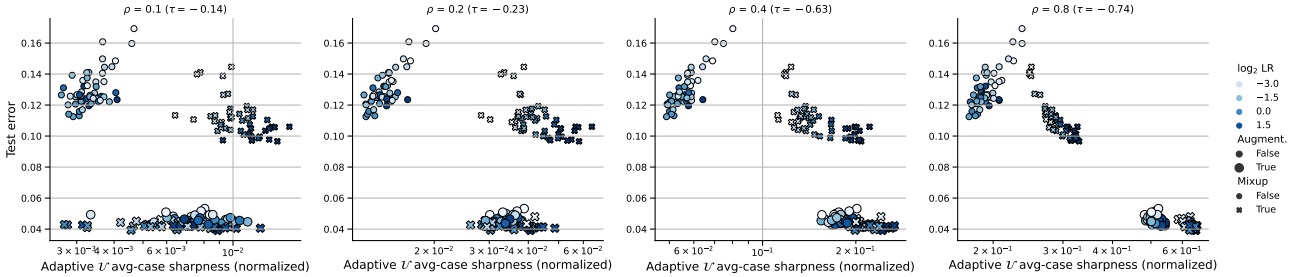
H.1. The Role of Data Used for Sharpness Evaluation

We emphasize that for all experiments, we evaluate sharpness on the original training set (CIFAR-10, ImageNet or MNLI) *without augmentations*. However, one may wonder how sensitive this choice is compared to evaluation on the *augmented* training set, particularly in presence of strong data augmentations such as RandAugment (Cubuk et al., 2020) used for training of some models. To test this, in Fig. 29, we compare adaptive average-case sharpness computed on the original training set and on augmented training set of CIFAR-10 for ResNets-18. We find that the overall trend is nearly the same for small ρ and differs more strongly for larger ρ where the overall correlation with generalization becomes significantly negative (-0.74 for the largest ρ) on augmented data. In addition, a side-by-side comparison of sharpness on standard vs. augmented training shows that the relationship between them does not deviate too much from a linear trend, especially when considering separately models trained with and without augmentations.

Adaptive average-case ℓ_∞ (uniform perturbations) sharpness (normalized) for ResNets-18 on *original* training data



Adaptive average-case ℓ_∞ (uniform perturbations) sharpness (normalized) for ResNets-18 on *augmented* training data



A side-by-side comparison of sharpness on original vs. augmented training data

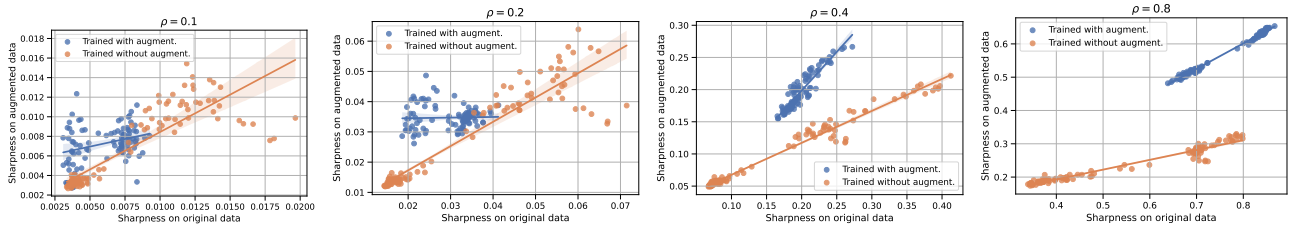


Figure 29: Adaptive average-case ℓ_∞ (uniform perturbations) sharpness (normalized) for ResNets-18 on *original* vs. *augmented* CIFAR-10 training data for ResNets-18 for different radii ρ.

H.2. The Role of the Number of Iterations in Auto-PGD

Here we aim to justify the choice of 20 iterations of Auto-PGD in our experiments. In Fig. 30, we present results for adaptive worst-case ℓ_∞ sharpness (normalized) for ResNets-18 on CIFAR-10 for 20, 50, 100, and 200 iterations. We can see that the sharpness values are not visibly affected by increasing the number of iterations and the overall trend stays exactly the same.

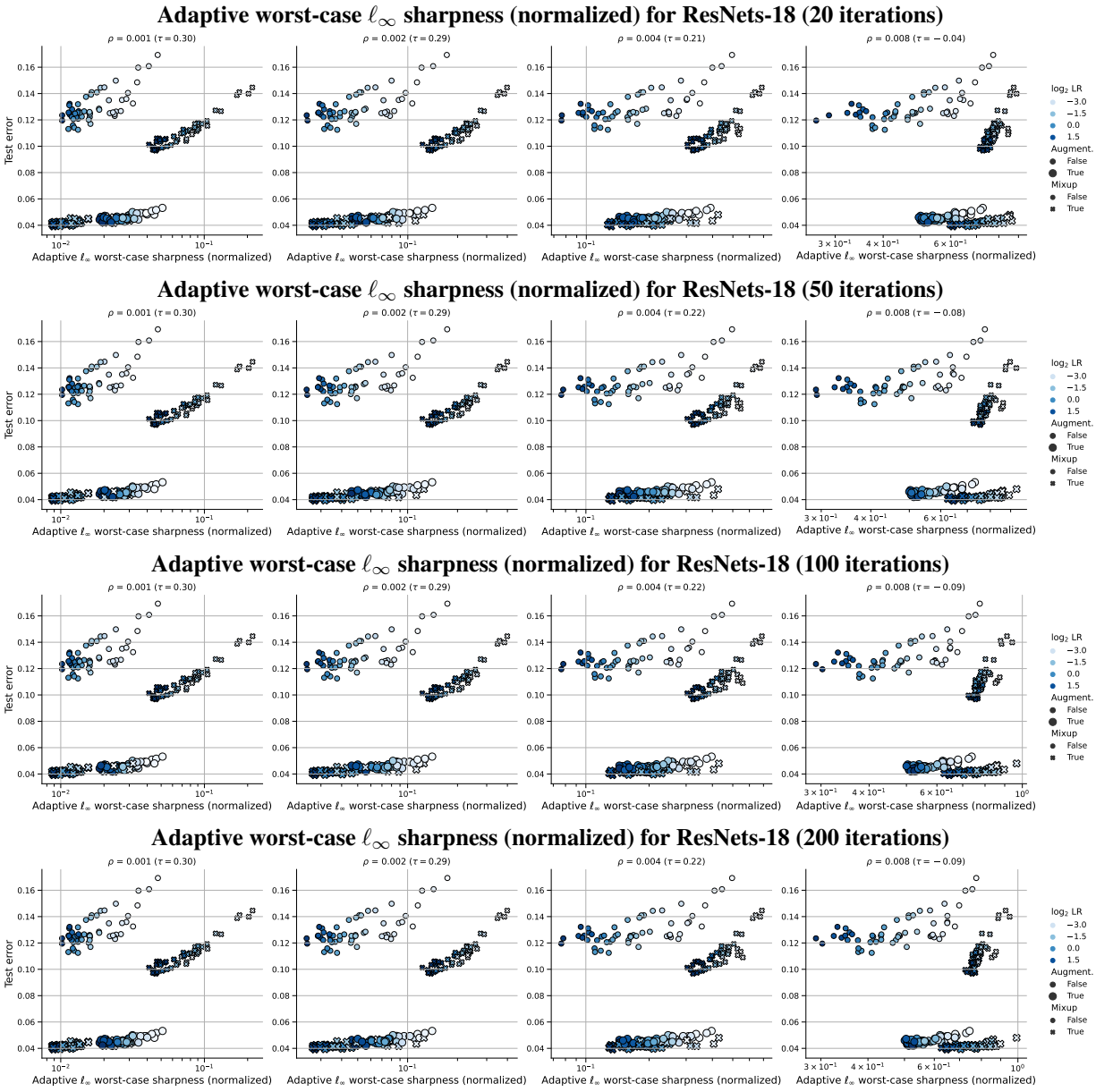


Figure 30: Adaptive worst-case ℓ_∞ sharpness (normalized) **for different number of iterations in Auto-PGD** vs. test error on CIFAR-10 for ResNets-18 for different radii ρ.

H.3. The Role of m in m-Sharpness

Foret et al. (2021) suggested that a lower m in m-sharpness, i.e., the batch size used for maximizing sharpness, can lead to a higher correlation with generalization in some settings. We note that we have already used a small m for all our experiments ($m = 128$ on CIFAR-10 and $m = 256$ on ImageNet and MNLI), but here we check additionally whether even smaller m change the trend. Fig. 31 shows the results sharpness for adaptive worst-case ℓ_∞ sharpness (normalized) for ResNets-18 and ViTs on CIFAR-10 for $m \in \{16, 32, 64, 128\}$. We can see that different m only slightly affects the sharpness values and the overall trend stays unaffected.

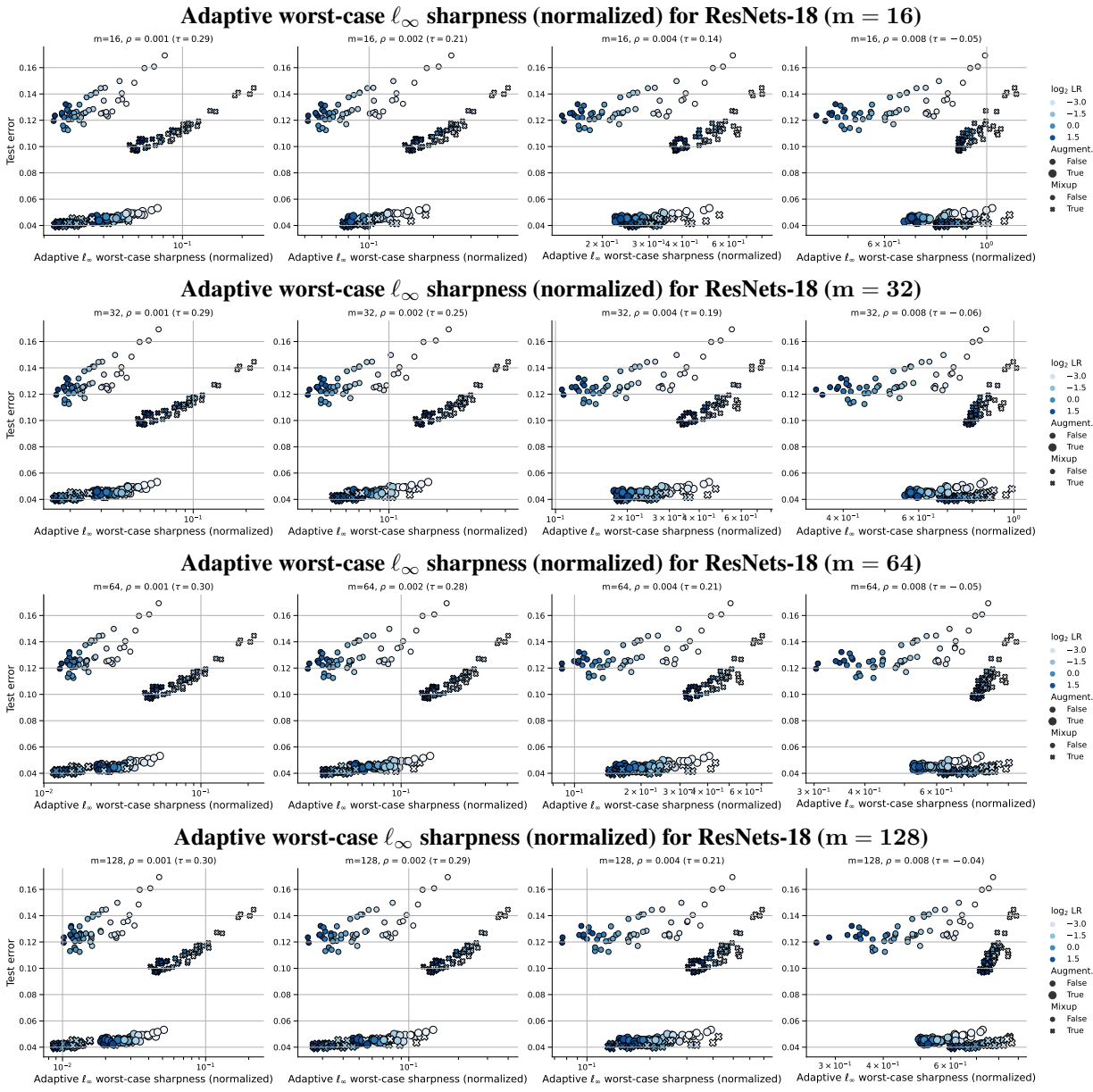


Figure 31: Adaptive worst-case ℓ_∞ sharpness (normalized) **for different m** in m-sharpness vs. test error on CIFAR-10 for ResNets-18 for different radii ρ.

Figure 32: Adaptive worst-case ℓ_∞ sharpness (normalized) **for different m** in m-sharpness vs. test error on CIFAR-10 for ViTs for different radii ρ.

H.4. The Role of Different Sharpness Definitions and Radii

Here we present results for 12 different sharpness definitions:

- standard average-case ℓ_2 (Gaussian perturbations) sharpness without logit normalization,
- standard worst-case ℓ_2 sharpness without logit normalization,
- adaptive average-case ℓ_2 (Gaussian perturbations) sharpness without logit normalization,
- adaptive worst-case ℓ_2 sharpness without logit normalization,
- adaptive average-case ℓ_2 (Gaussian perturbations) sharpness with logit normalization,
- adaptive worst-case ℓ_2 sharpness with logit normalization,
- standard average-case ℓ_∞ (uniform perturbations) sharpness without logit normalization,
- standard worst-case ℓ_∞ sharpness without logit normalization,
- adaptive average-case ℓ_∞ (uniform perturbations) sharpness without logit normalization,
- adaptive worst-case ℓ_∞ sharpness without logit normalization (*shown in the main part for a single* ρ),
- adaptive average-case ℓ_∞ (uniform perturbations) sharpness with logit normalization,
- adaptive worst-case ℓ_∞ sharpness with logit normalization (*shown in the main part for a single* ρ).

We evaluate a wide range of radii for each sharpness definition to make sure that we do not miss the right scale of sharpness. We present results first for ResNets and then for ViTs.

Observations for ResNets. For ResNets, we observe that many sharpness definitions can successfully capture correlation with standard generalization within each subgroup defined by the values of `augment` $\times$ `mixup`. In particular, on average, *adaptive* sharpness shows a better correlation with generalization within each subgroup, and the best correlation within each subgroup is achieved by ℓ_∞ adaptive worst-case sharpness with logit normalization for a small ρ. In many cases, the correlation of sharpness with OOD generalization on CIFAR-10-C is noticeably lower compared to the correlation of sharpness with standard generalization. Overall, we see that there is no coherent global trend of correlation with generalization that would apply to all models at once. We also observe that for some sharpness definitions, the flattest models generalize best (for adaptive worst-case ℓ_2 sharpness with normalization for the smallest ρ and for adaptive worst-case ℓ_∞ sharpness without normalization for the largest ρ) but this appears to be unsystematic and there exist nearly equally flat solutions that generalize much worse.

Observations for ViTs. For ViTs, in contrast to ResNets, we do not observe a consistent correlation with generalization even within subgroups. The only exception is the subgroup of points with augmentations where multiple definitions of sharpness tend to correlate with generalization and capture the effect of larger learning rate. We think it is likely due to the fact that with heavy augmentations optimizing the training objective to smaller values is helpful for generalization, while without augmentations all runs have converged within 200 epochs and the learning rate plays no visible role for generalization there. Globally, when taken over all models, the correlation with standard generalization is close to 0 and tends to slightly decrease when we measure OOD generalization on CIFAR-10-C. Finally, we note that there are no cases where the flattest ViT models achieve the best generalization. Thus, even our weak hypothesis about the role of sharpness is not confirmed here.

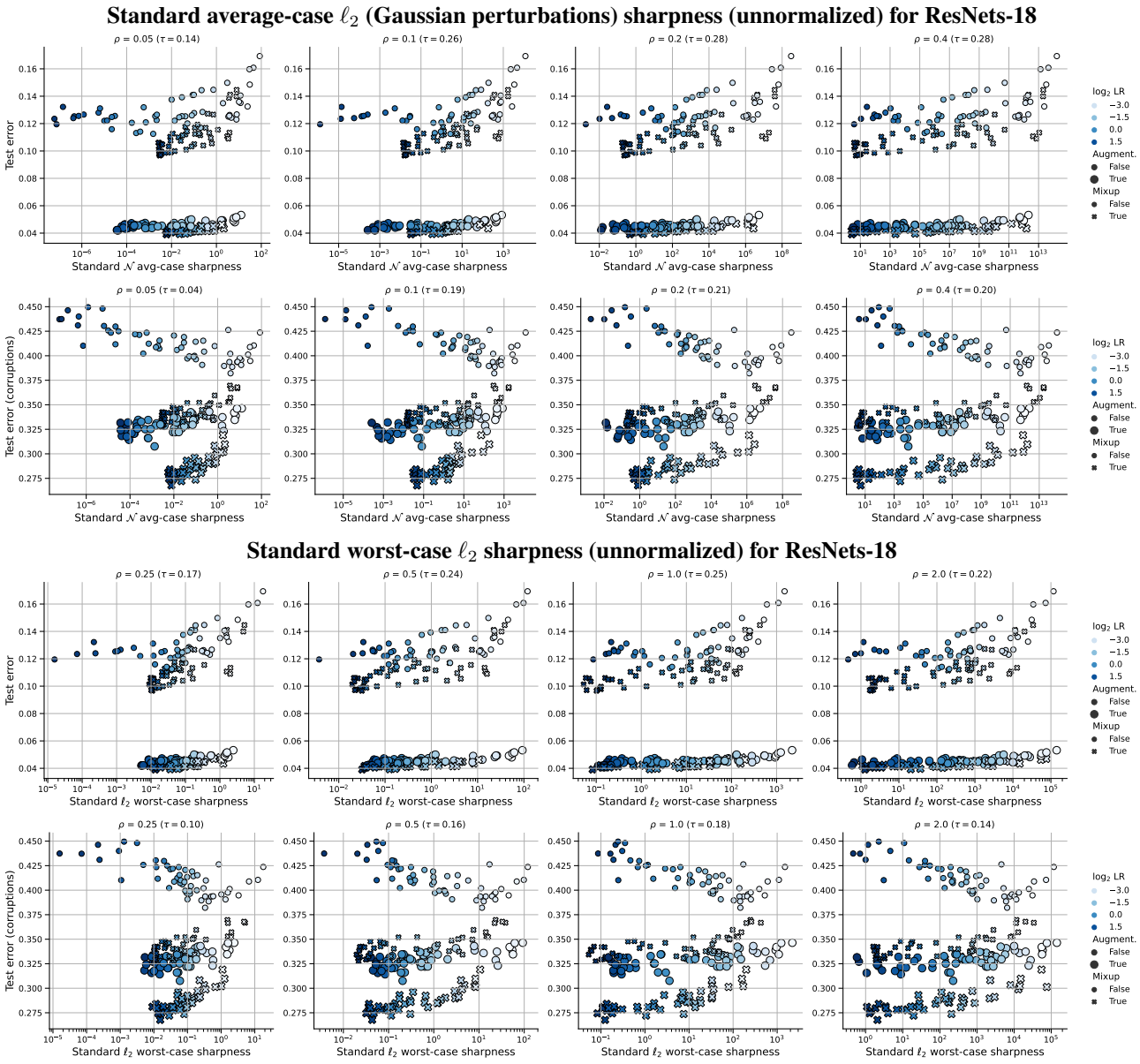


Figure 33: Average and worst-case ℓ_2 standard sharpness definitions (unnormalized) vs. test error and OOD test error (common corruptions) on CIFAR-10 for ResNets-18 for different radii ρ.

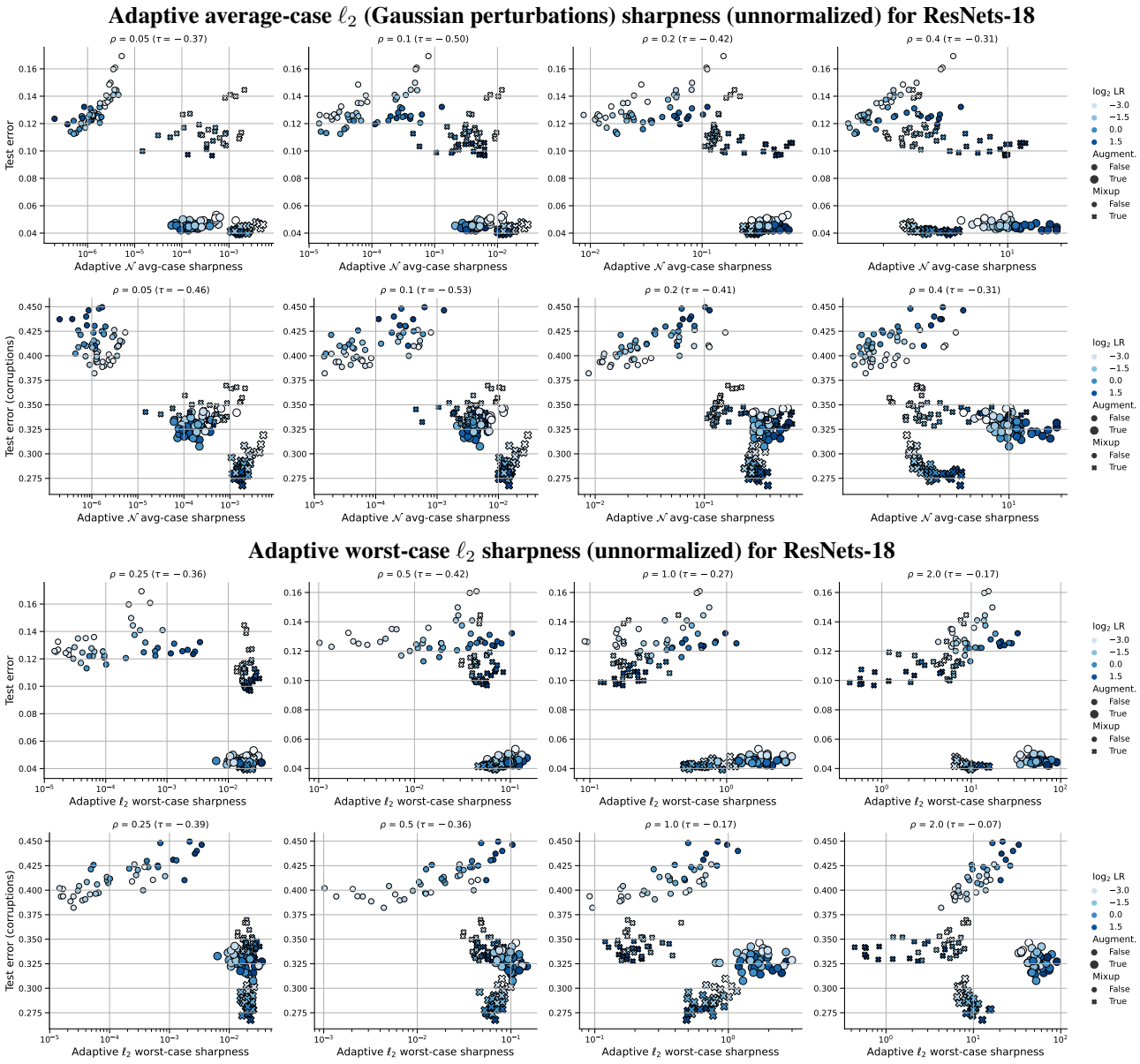


Figure 34: Average and worst-case ℓ_2 adaptive sharpness definitions (unnormalized) vs. test error and OOD test error (common corruptions) on CIFAR-10 for ResNets-18 for different radii ρ.

Figure 35: Average and worst-case ℓ_2 adaptive sharpness definitions (normalized) vs. test error and OOD test error (common corruptions) on CIFAR-10 for ResNets-18 for different radii ρ.

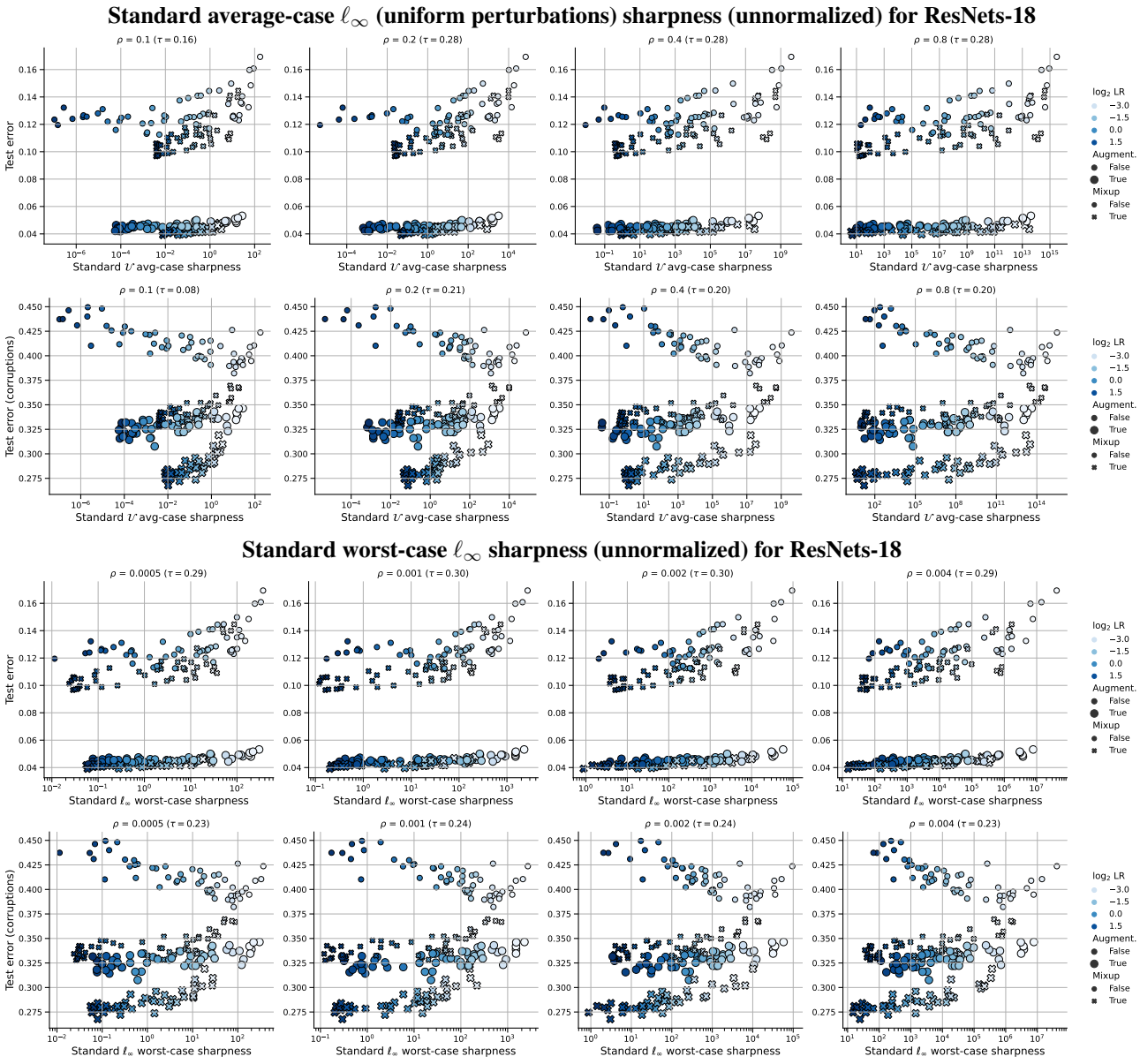


Figure 36: Average and worst-case ℓ_∞ standard sharpness definitions (unnormalized) vs. test error and OOD test error (common corruptions) on CIFAR-10 for ResNets-18 for different radii ρ.

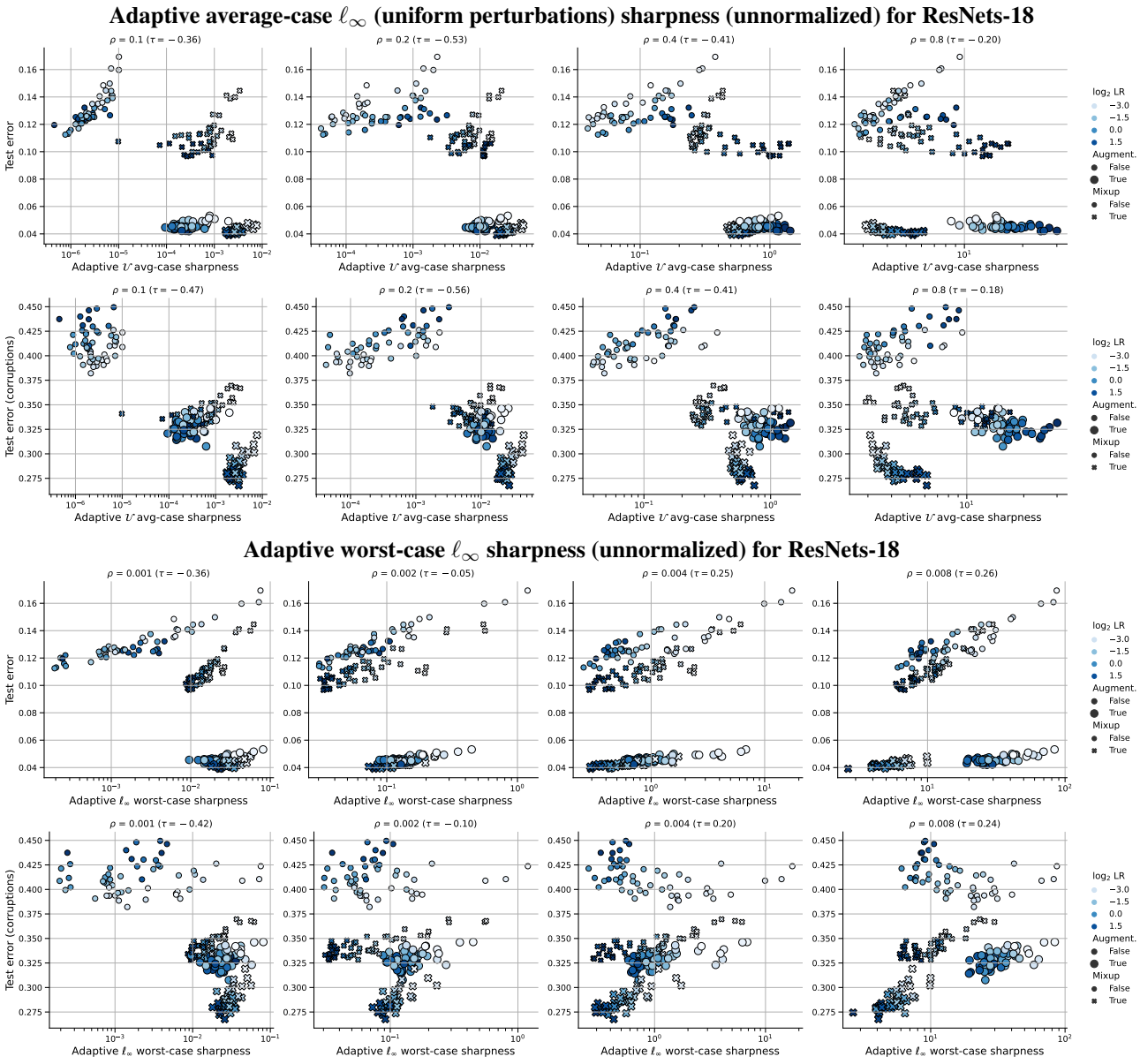


Figure 37: Average and worst-case ℓ_∞ adaptive sharpness definitions (unnormalized) vs. test error and OOD test error (common corruptions) on CIFAR-10 for ResNets-18 for different radii ρ.

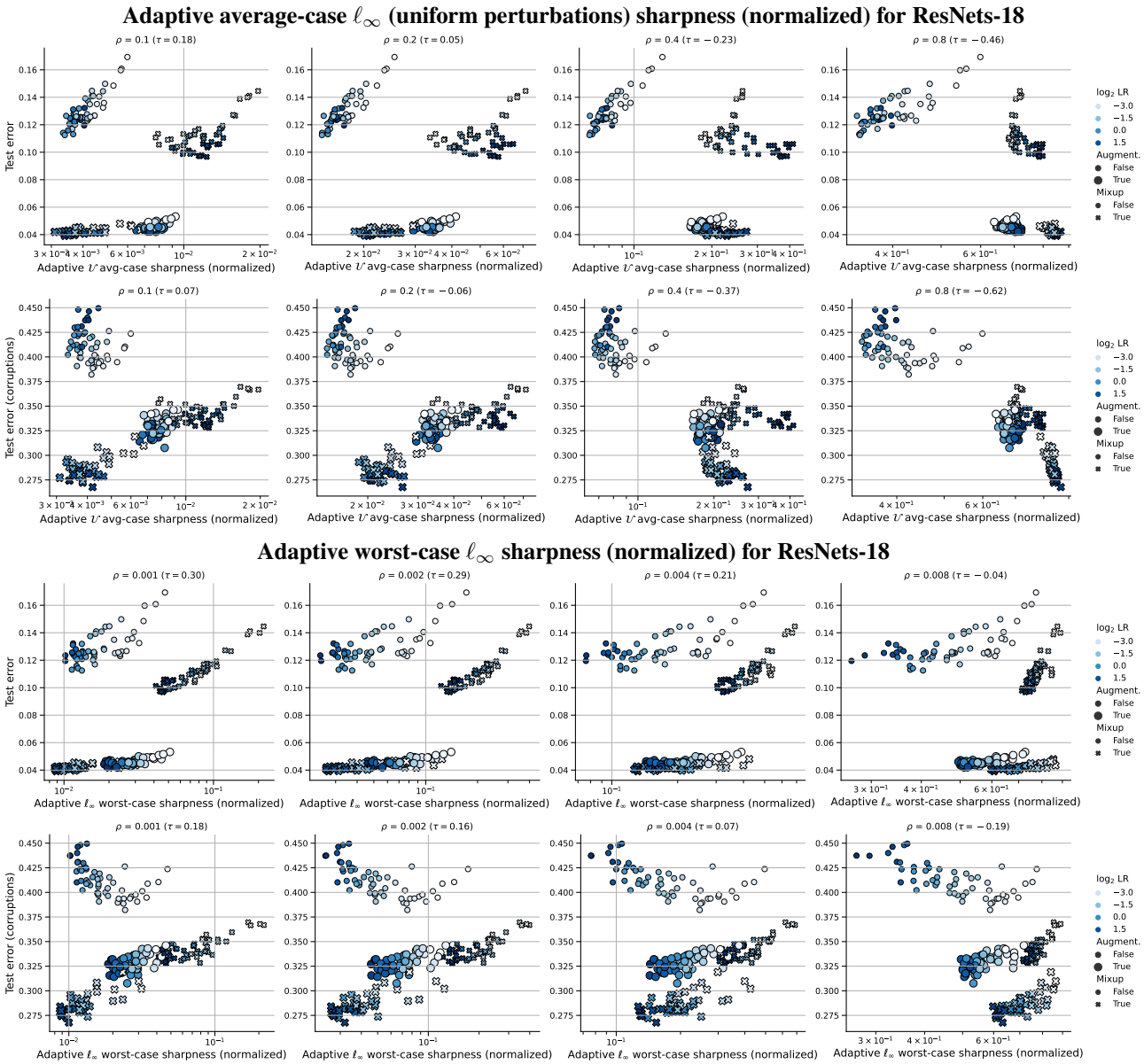


Figure 38: Average and worst-case ℓ_∞ adaptive sharpness definitions (normalized) vs. test error and OOD test error (common corruptions) on CIFAR-10 for ResNets-18 for different radii ρ.

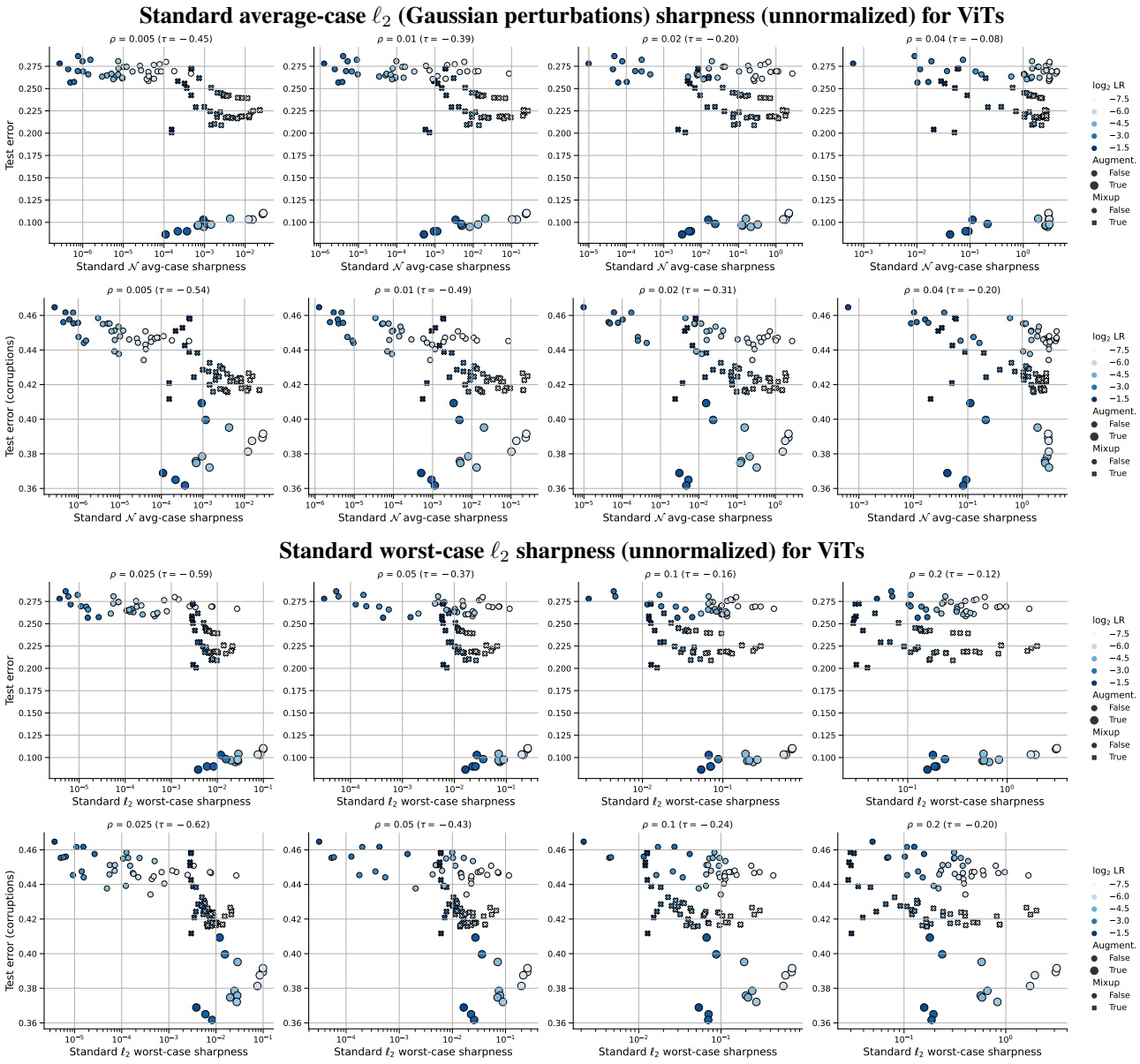


Figure 39: Average and worst-case ℓ_2 standard sharpness definitions (unnormalized) vs. test error and OOD test error (common corruptions) on CIFAR-10 for ViTs for different radii ρ.

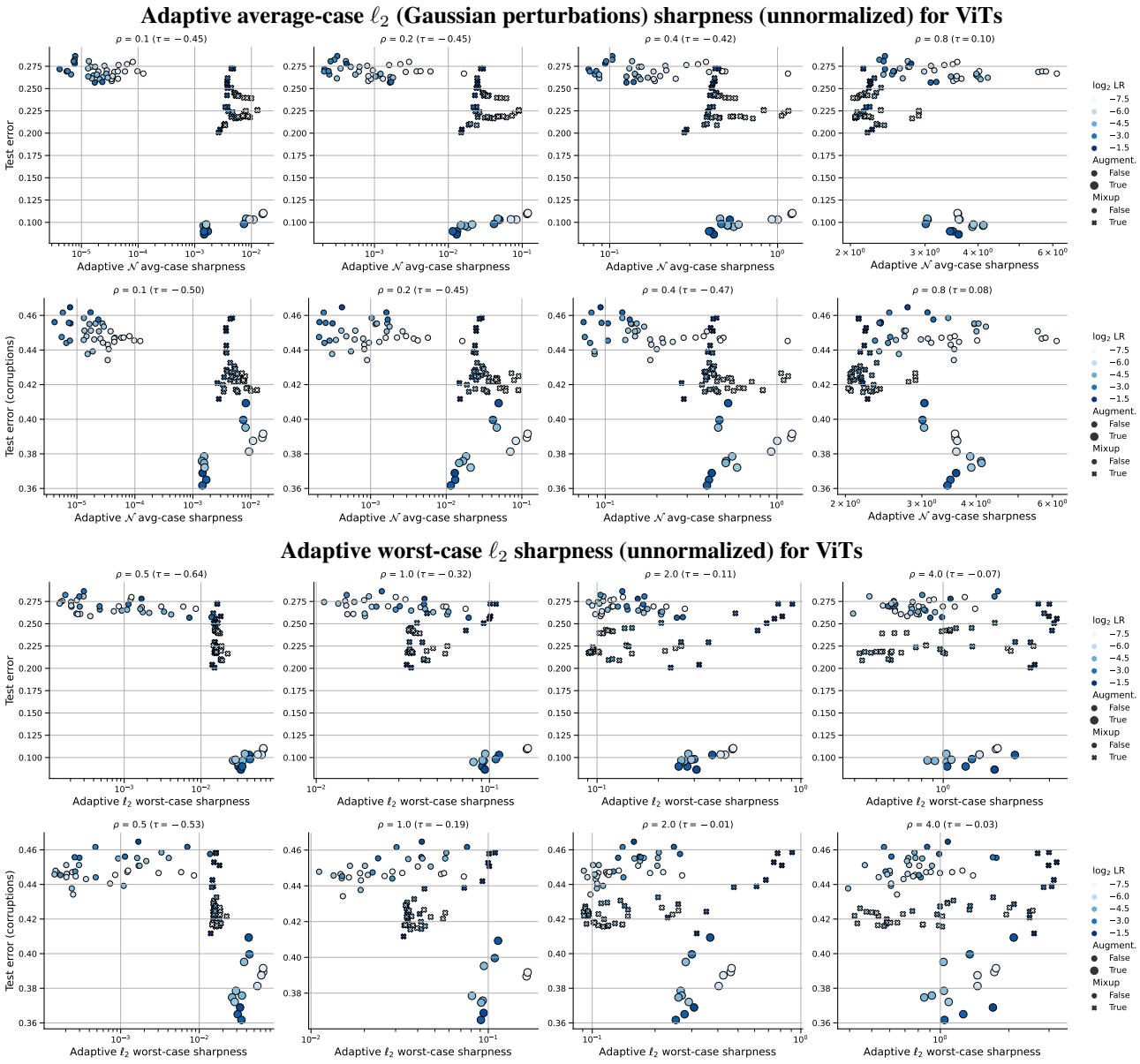


Figure 40: Average and worst-case ℓ_2 adaptive sharpness definitions (unnormalized) vs. test error and OOD test error (common corruptions) on CIFAR-10 for ViTs for different radii ρ.

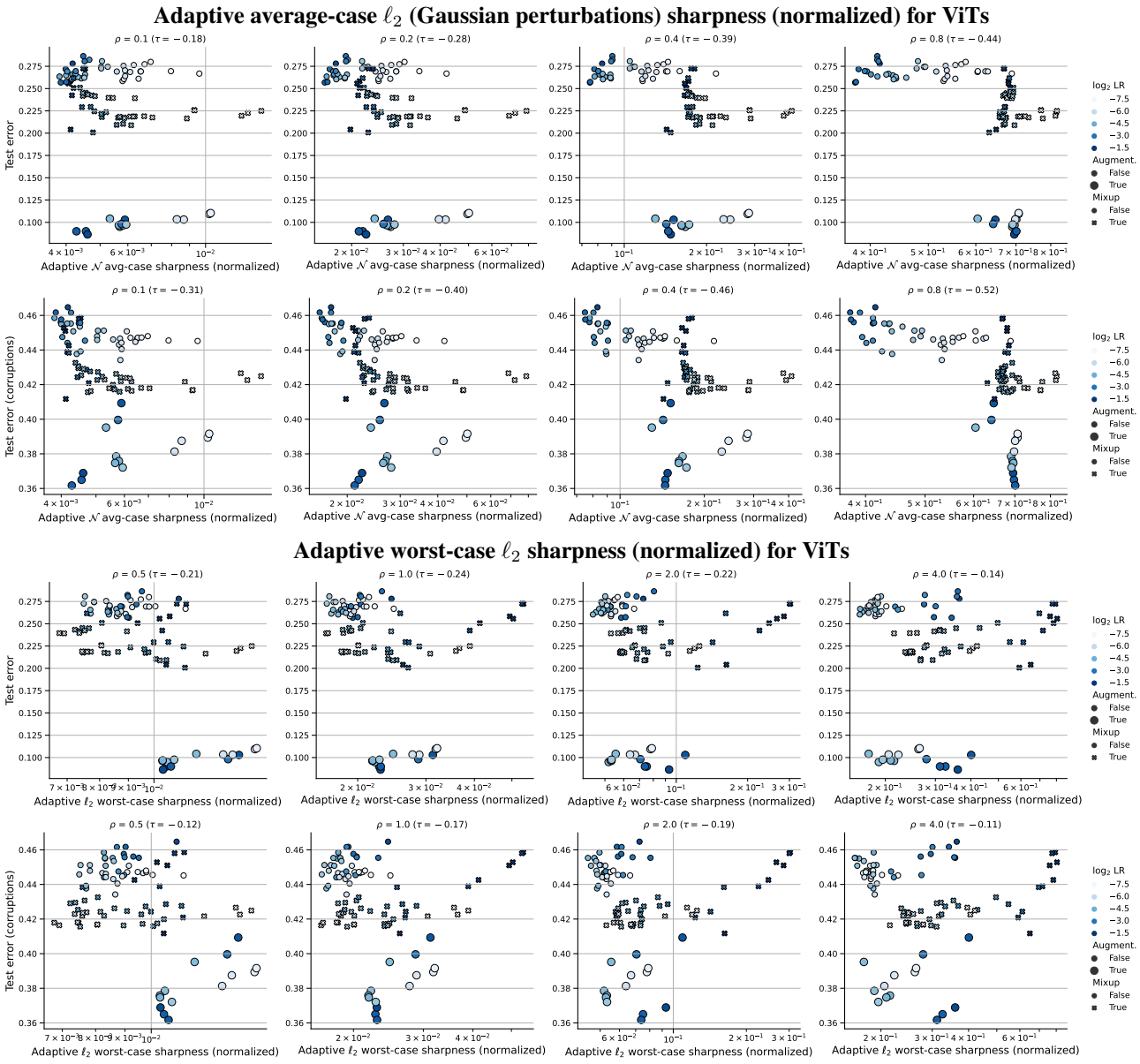


Figure 41: Average and worst-case ℓ_2 adaptive sharpness definitions (normalized) vs. test error and OOD test error (common corruptions) on CIFAR-10 for ViTs for different radii ρ.

Figure 42: Average and worst-case ℓ_∞ standard sharpness definitions (unnormalized) vs. test error and OOD test error (common corruptions) on CIFAR-10 for ViTs for different radii ρ.

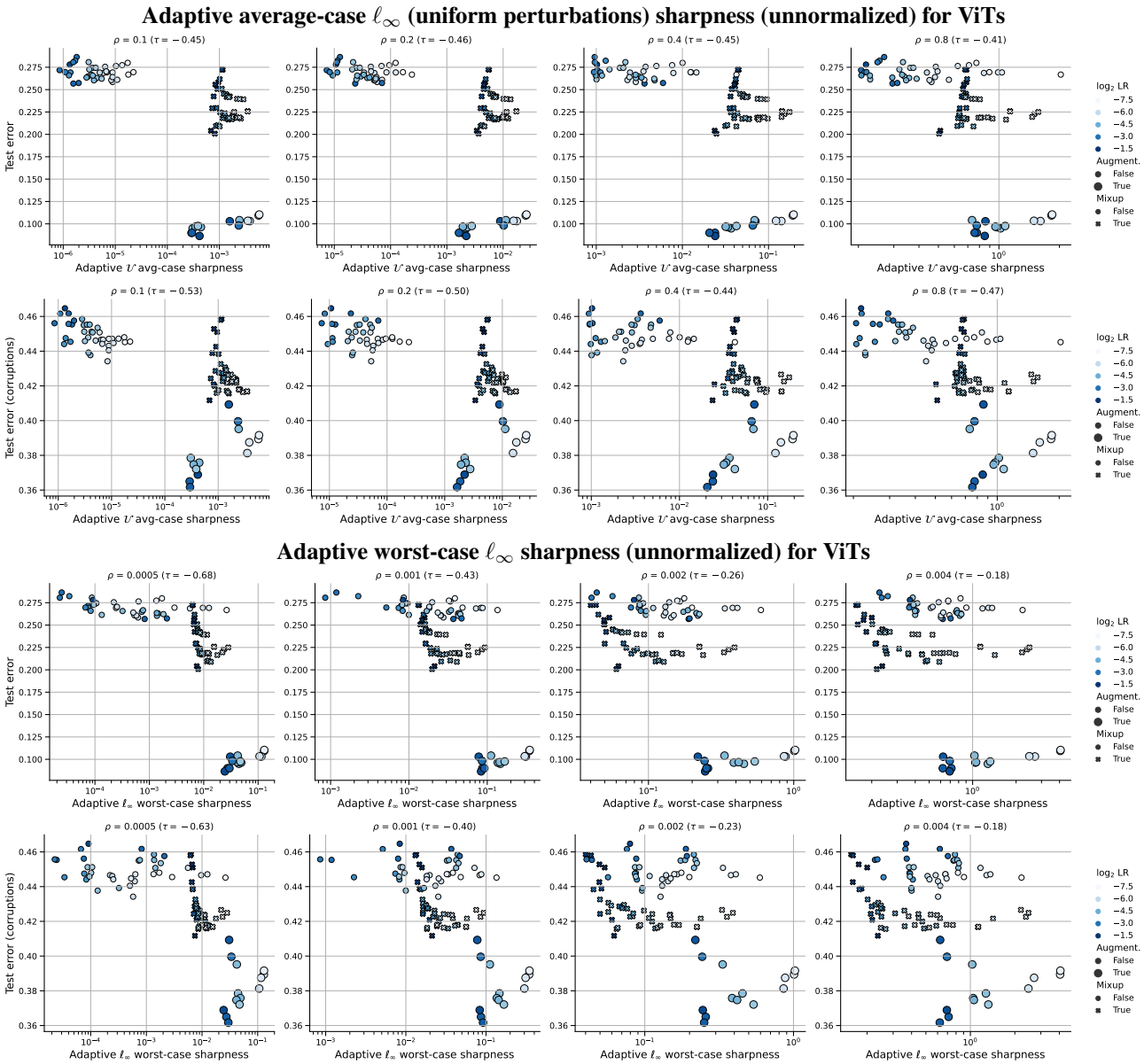


Figure 43: Average and worst-case ℓ_∞ adaptive sharpness definitions (unnormalized) and OOD test error (common corruptions) on CIFAR-10 for ViTs for different radii ρ.

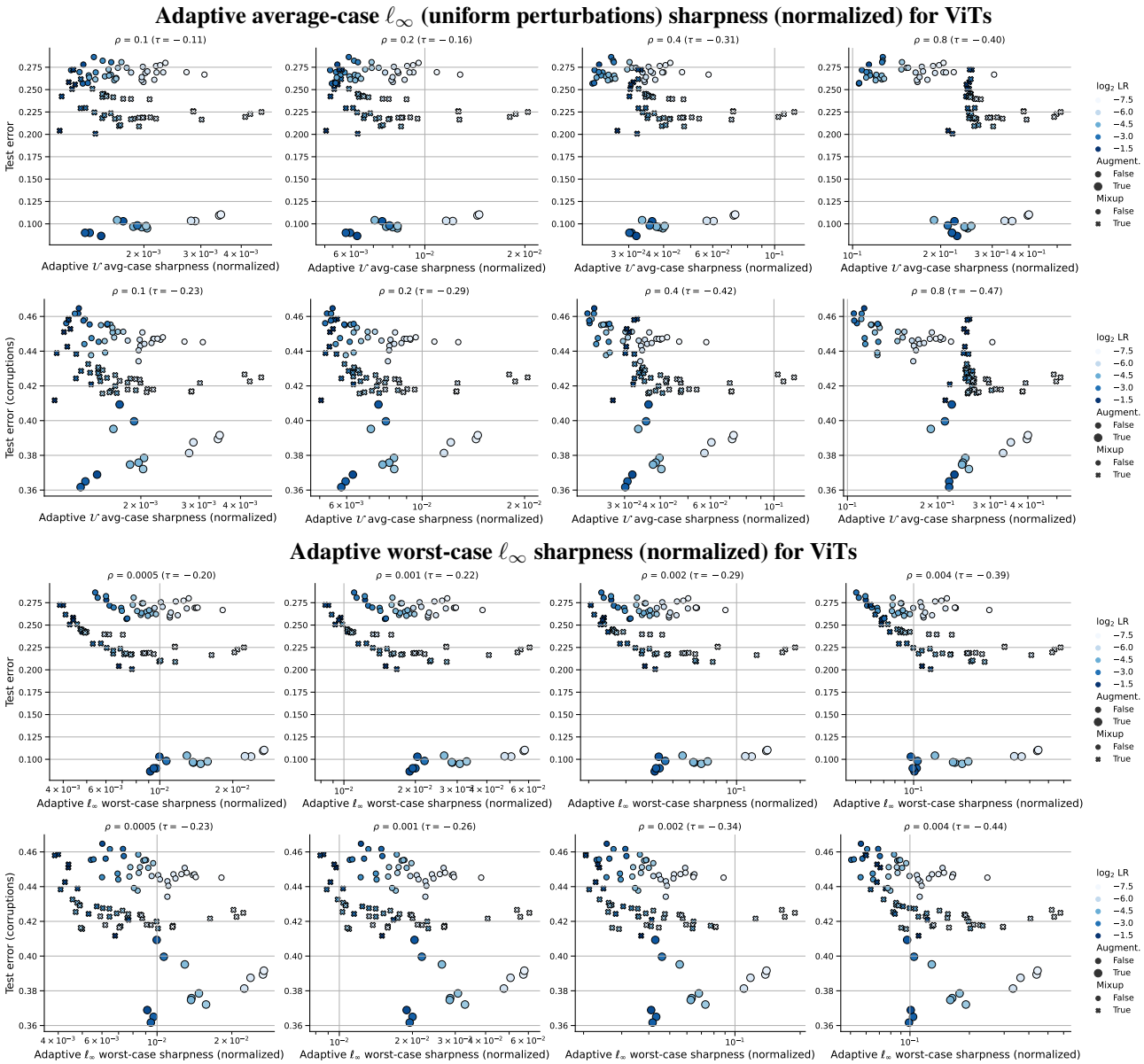


Figure 44: Average and worst-case ℓ_∞ adaptive sharpness definitions (normalized) vs. test error and OOD test error (common corruptions) on CIFAR-10 for ViTs for different radii ρ.