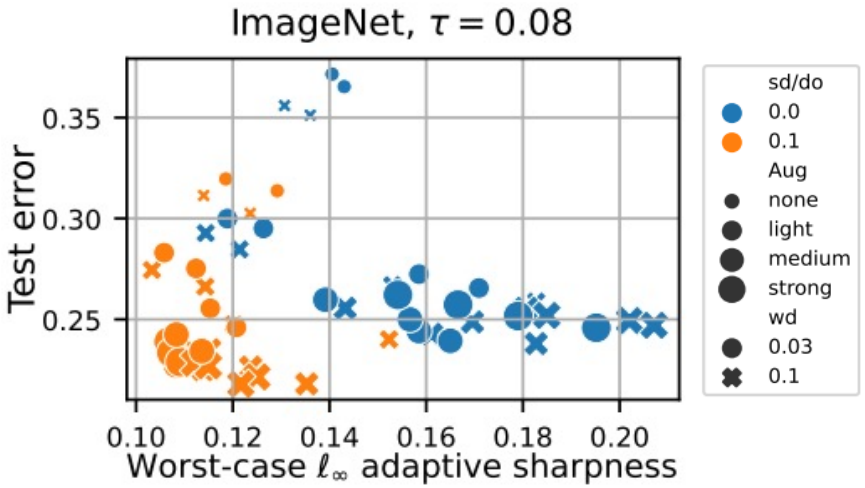
A modern look at the relationship between sharpness and generalization

Maksym Andriushchenko (EPFL), Francesco Croce (U of Tübingen), Maximilian Müller (U of Tübingen), Matthias Hein (U of Tübingen), Nicolas Flammarion (EPFL)

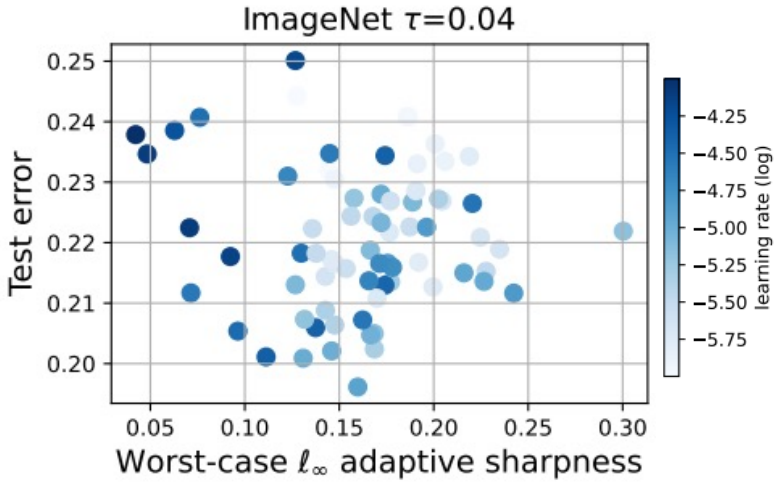
Question: *Can reparametrization-invariant sharpness capture generalization in modern practical settings?*

Training ViT from scratch



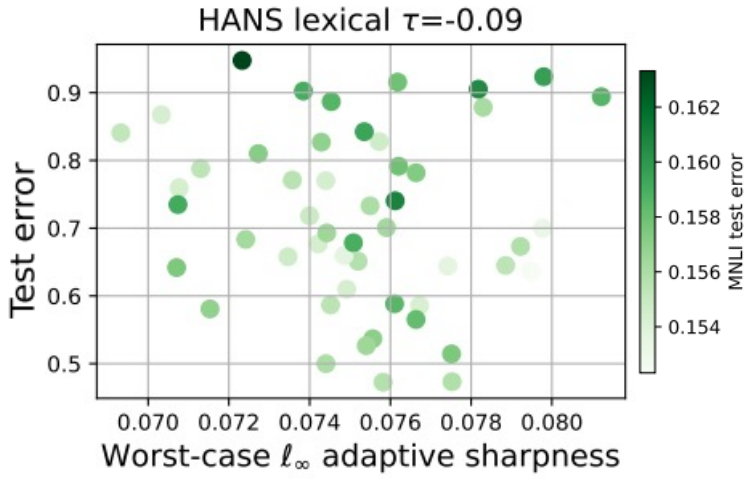
⚠ Correlation $\tau \approx 0$

Fine-tuning CLIP ViT



⚠ Correlation $\tau \approx 0$

Fine-tuning BERT models



⚠ Correlation $\tau \approx 0$

13 March 2023, OOD robustness + generalization reading group (CMU)

Big picture: understanding the generalization puzzle in overparametrized deep learning

- Different global minima can generalize very differently
  e.g., see "*Bad Global Minima Exist and SGD Can Reach Them*" (Liu et al. NeurIPS'19)

- What **measure** *computed on the training set* can distinguish the minima which generalize well?

- Can we figure this measure and optimize it for training? (+ use it as a tool to understand the generalization puzzle)

Prior work: finding such measures is actually not easy!

- **Main ref**: "*Fantastic Generalization Measures and Where to Find Them*" (Jiang et al., ICLR'20) which highlights **sharpness** as a promising measure

- What can we expect from such measure:
  1. **Causal relation**: smaller measure ⇒ better generalization (universally)
  2. **Correlation**: smaller measure ⇒ better generalization (but there may exist counterexamples)
  3. **Sufficient but not necessary**: small measure ⇒ good generalization; large measure ⇒ can't say anything

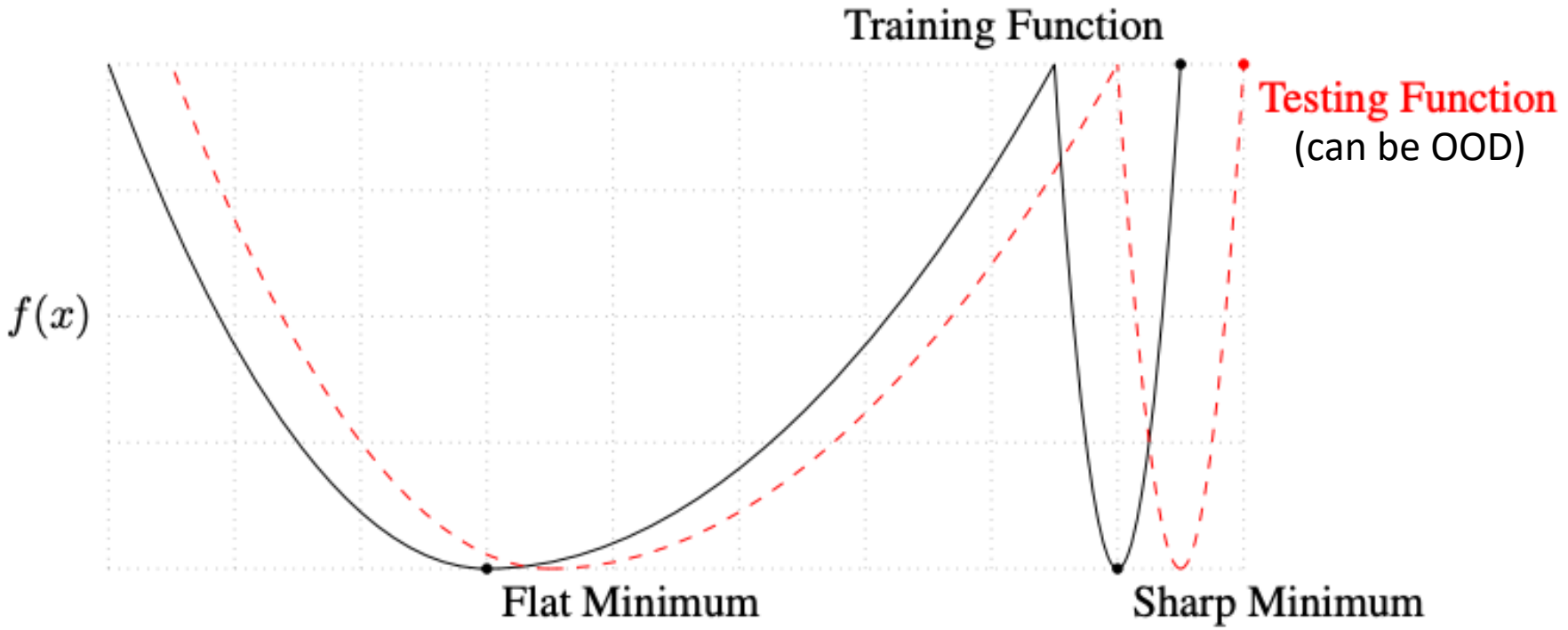
	ref	batchsize	dropout	learning rate	depth	optimizer	weight decay	width	overall τ	Ψ
vc dim	19	0.000	0.000	0.000	−0.909	0.000	0.000	−0.171	−0.251	−0.154
# params	20	0.000	0.000	0.000	−0.909	0.000	0.000	−0.171	−0.175	−0.154
sharpness	51	0.537	−0.523	0.449	0.826	0.221	0.233	−0.004	0.282	0.248
pacbayes	48	0.372	−0.457	0.042	0.644	0.179	−0.179	−0.142	0.064	0.066
sharpness-orig	52	0.542	−0.359	0.716	0.816	0.297	0.591	0.185	0.400	0.398
pacbayes-orig	49	0.526	−0.076	0.705	0.546	0.341	0.564	−0.086	0.293	0.360
frob-distance	40	−0.317	−0.833	−0.718	0.526	−0.214	−0.669	−0.166	−0.263	−0.341
spectral-init	25	−0.330	−0.845	−0.721	−0.908	−0.208	−0.313	−0.231	−0.576	−0.508
spectral-orig	26	−0.262	−0.762	−0.665	−0.908	−0.131	−0.073	−0.240	−0.537	−0.434
spectral-orig-main	28	−0.262	−0.762	−0.665	−0.908	−0.131	−0.073	−0.240	−0.537	−0.434
fro/spec	33	0.563	0.351	0.744	−0.898	0.326	0.665	−0.053	−0.008	0.243
prod-of-spec	32	−0.464	−0.724	−0.722	−0.909	−0.197	−0.142	−0.218	−0.559	−0.482
prod-of-spec/margin	31	−0.308	−0.782	−0.702	−0.907	−0.166	−0.148	−0.179	−0.570	−0.456
sum-of-spec	35	−0.464	−0.724	−0.722	0.909	−0.197	−0.142	−0.218	0.102	−0.223
sum-of-spec/margin	34	−0.308	−0.782	−0.702	0.909	−0.166	−0.148	−0.179	0.064	−0.197
spec-dist	41	−0.458	−0.838	−0.568	0.738	−0.319	−0.182	−0.171	−0.110	−0.257
prod-of-fro	37	0.440	−0.199	0.538	−0.909	0.321	0.731	−0.101	−0.297	0.117
prod-of-fro/margin	36	0.513	−0.291	0.579	−0.907	0.364	0.739	−0.088	−0.295	0.130
sum-of-fro	39	0.440	−0.199	0.538	0.913	0.321	0.731	−0.101	0.418	0.378
sum-of-fro/margin	38	0.520	−0.369	0.598	0.882	0.380	0.738	−0.080	0.391	0.381
1/margin	22	−0.312	0.593	−0.234	−0.758	−0.223	0.211	−0.125	−0.124	−0.121
neg-entropy	23	0.346	−0.529	0.251	0.632	0.220	−0.157	0.104	0.148	0.124
path-norm	44	0.363	−0.190	0.216	0.925	0.272	0.195	0.178	0.370	0.280
path-norm/margin	43	0.363	0.017	0.148	0.922	0.230	0.280	0.173	0.374	0.305
param-norm	42	0.236	−0.516	0.174	0.330	0.187	0.124	−0.170	0.073	0.052
fisher-rao	45	0.396	0.147	0.240	−0.516	0.120	0.551	0.177	0.090	0.160
cross-entropy	21	0.440	−0.402	0.140	0.390	0.149	0.232	0.080	0.149	0.147
$1/\sigma$ pacbayes	53	0.501	−0.033	0.744	0.200	0.346	0.609	0.056	0.303	0.346
$1/\sigma$ sharpness	54	0.532	−0.326	0.711	0.776	0.296	0.592	0.263	0.399	0.406
num-step-0.1-to-0.01-loss	64	−0.151	−0.069	−0.014	0.114	0.072	−0.046	−0.021	−0.088	−0.016
num-step-to-0.1-loss	63	−0.664	−0.861	−0.255	0.440	−0.030	−0.628	0.043	−0.264	−0.279
$1/\alpha'$ sharpness mag	62	0.570	0.148	0.762	0.824	0.297	0.741	0.269	0.484	0.516
$1/\sigma'$ pacbayes mag	61	0.490	−0.215	0.505	0.896	0.186	0.147	0.195	0.365	0.315
pac-sharpness-mag-init	59	−0.293	−0.841	−0.698	−0.909	−0.240	−0.631	−0.171	−0.225	−0.541
pac-sharpness-mag-orig	60	0.401	−0.514	0.321	−0.909	0.181	0.281	−0.171	−0.158	−0.059
pacbayes-mag-init	56	0.425	−0.658	−0.035	0.874	0.099	−0.407	0.069	0.175	0.052
pacbayes-mag-orig	57	0.532	−0.480	0.508	0.902	0.188	0.155	0.186	0.410	0.284
grad-noise-final	66	0.452	0.119	0.427	0.141	0.245	0.432	0.230	0.311	0.292
grad-noise-epoch-1	65	0.071	0.378	0.376	−0.517	0.121	0.221	0.037	0.070	0.098
oracle 0.01		0.579	0.885	0.736	0.920	0.529	0.622	0.502	0.851	0.682
oracle 0.02		0.414	0.673	0.548	0.742	0.346	0.447	0.316	0.726	0.498
oracle 0.05		0.123	0.350	0.305	0.401	0.132	0.201	0.142	0.456	0.236
oracle 0.1		0.069	0.227	0.132	0.223	0.086	0.121	0.093	0.241	0.136
canonical ordering		−0.652	0.969	0.733	0.909	−0.055	0.735	0.171	0.005	0.402
canonical ordering depth		−0.032	0.001	0.033	−0.909	−0.061	−0.020	0.024	−0.363	−0.138

Table 5: Complexity measures (rows), hyperparameters (columns) and the **rank-correlation coefficients** with models trained on **CIFAR-10**.

- We will focus on **2.** and **3.** showing they don't hold for **sharpness** (this will also rule out **1.**)

Flat vs. sharp minima: intuition

- **Popular intuition**: the test loss will be close to the training loss for a flat minimum



Source: "*On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima*" (Keskar et al., ICLR'17)

- Keskar et al., ICLR'17: small-batch SGD converges to flat minima unlike large-batch SGD
- Sharpness also received renewed interest with the **Edge of Stability** phenomenon and the empirical success of **Sharpness-Aware Minimization**

Flat vs. sharp minima: theory

- There are generalization bounds based on **sharpness**

$$\mathbb{E}_{\mathbf{u}\sim\mathcal{N}(u,\sigma^2 I)}\left[L(f_{\mathbf{w}+\mathbf{u}})\right] \leq \mathbb{E}_{\mathbf{u}\sim\mathcal{N}(u,\sigma^2 I)}\left[\hat{L}(f_{\mathbf{w}+\mathbf{u}})\right] + \sqrt{\frac{\frac{\|\mathbf{w}-\mathbf{w}^0\|_2^2}{4\sigma^2} + \log(\frac{m}{\sigma}) + 10}{m-1}}$$

perturbed population loss — perturbed training loss — term that depends on the scale of the predictor

- But they can be often of limited use as illustrated well by Jiang et al., ICLR'20

	overall τ
vc dim	−0.251
# params	−0.175
sharpness	0.282
pacbayes	0.064

τ = rank correlation coefficient:

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

- While there exist networks for which these bounds can be quite tight (Lotfi et al., NeurIPS'22), this doesn't apply to all possible networks ⇒ **these quantities are not necessarily meaningful to solve the generalization puzzle**

Problems with the standard sharpness definitions

$$\max_{|\boldsymbol{\nu}_i| \leq \alpha(|\mathbf{w}_i|+\mathbf{1})} \widehat{L}(f_{\mathbf{w}+\boldsymbol{\nu}}) - \widehat{L}(f_{\mathbf{w}})$$

Keskar et al., ICLR'17
+ many other papers

$$Tr(\nabla_w^2 L(f_w))$$

Keskar et al., ICLR'17,
Damian et al., NeurIPS'21
+ many other papers

$$\lambda_{max}(\nabla_w^2 L(f_w))$$

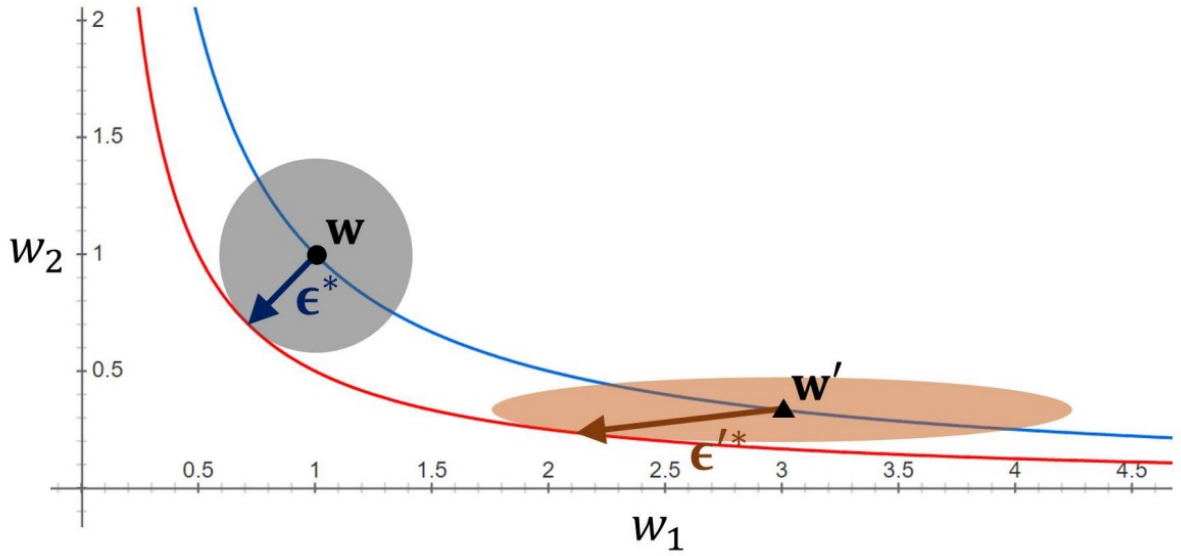
Keskar et al., ICLR'17,
+ many other papers

- Main problem (Dinh et al., ICML'17): lack of invariance to layerwise rescaling: $\frac{1}{\beta}\boldsymbol{V} \cdot \sigma(\beta \boldsymbol{W}\boldsymbol{x}) = \mathbf{V} \cdot \sigma(\boldsymbol{W}\boldsymbol{x})$ (for a homogeneous σ) $\Rightarrow$ same network but with different sharpness!

- However, this is pretty easy to fix: **adaptive sharpness** is invariant to such rescaling and is reported to correlate better with generalization
  "*ASAM: Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks*" (Kwon et al., ICML'21)

Adaptive sharpness: definition

- Average-case sharpness: $$S^{\rho}_{avg}(\boldsymbol{w}, \boldsymbol{c}) \triangleq \mathbb{E}_{\substack{(\boldsymbol{x},\boldsymbol{y})\sim P_1 \\ \boldsymbol{\delta}\sim\mathcal{N}(0,\rho^2 diag(\boldsymbol{c}^2))}} \ell_{\boldsymbol{xy}}(\boldsymbol{w}+\boldsymbol{\delta}) - \ell_{\boldsymbol{xy}}(\boldsymbol{w})$$
- Worst-case sharpness: $$S^{\rho}_{max}(\boldsymbol{w}, \boldsymbol{c}) \triangleq \mathbb{E}_{\mathcal{S}\sim P_m} \max_{\|\boldsymbol{\delta}\odot\boldsymbol{c}^{-1}\|_p \leq \rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w})$$
- Choosing $\boldsymbol{c} \coloneqq |\boldsymbol{w}|$ leads to adaptive sharpness $S(\boldsymbol{w}, |\boldsymbol{w}|)$ which ensures that for any $\boldsymbol{\gamma} \in \mathbb{R}^p$ such that $f(\boldsymbol{w} \odot \boldsymbol{\gamma}) = f(\boldsymbol{w})$:
  $$S(\boldsymbol{w} \odot \boldsymbol{\gamma}, |\boldsymbol{w} \odot \boldsymbol{\gamma}|) = S(\boldsymbol{w}, |\boldsymbol{w}|)$$
- This also covers normalization layers (BatchNorm, LayerNorm) and makes sharpness reparametrization-invariant for the whole modern networks (ResNets / ViTs)

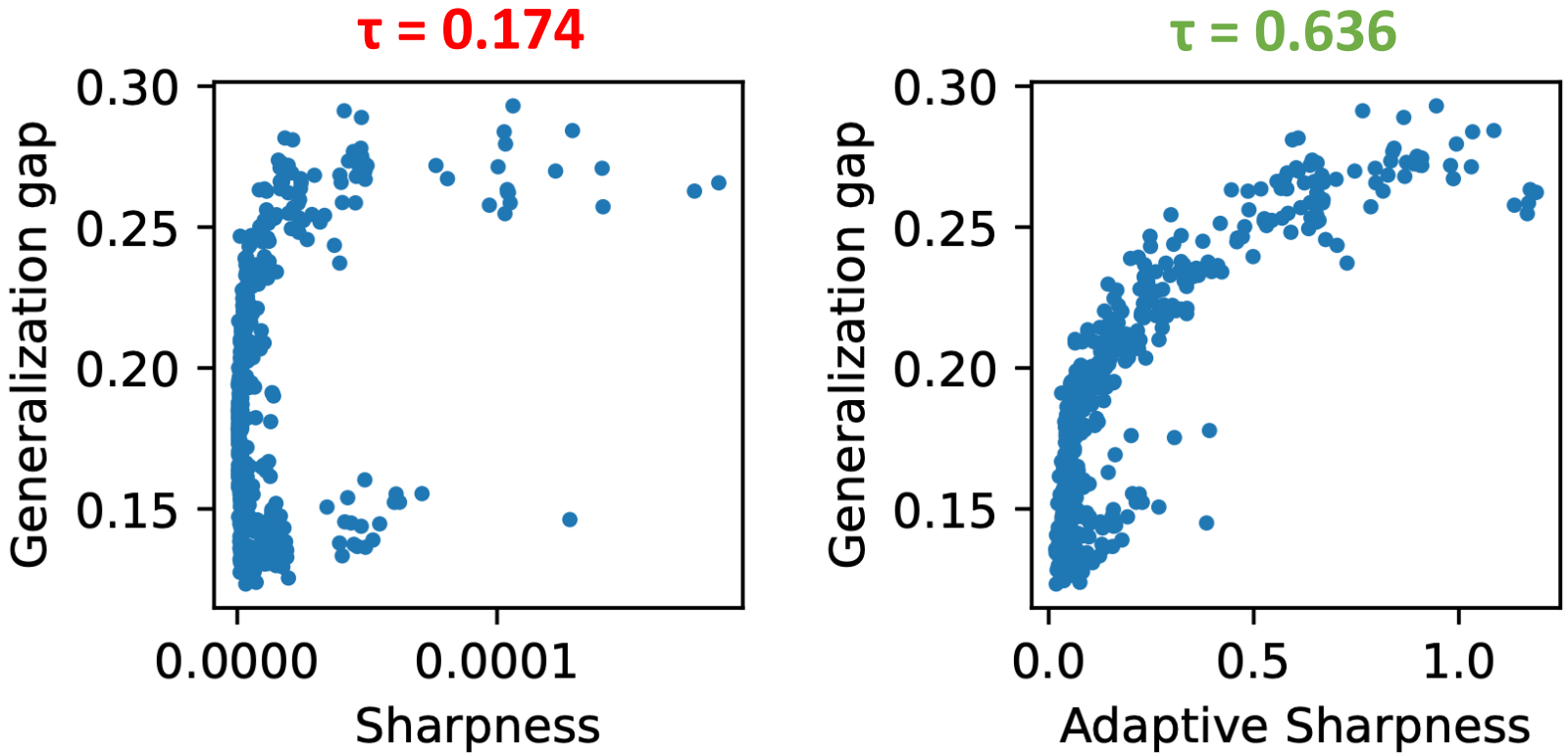


Source: Kwon et al., ICML'21

for a proof: see our paper or Kwon et al.

Adaptive sharpness: better correlation with generalization

- **Adaptive sharpness** is reported to correlate better with generalization (setting: WideResNet-16-8 on CIFAR-10)



"ASAM: Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks" (Kwon et al., ICML'21)

Very nice, no? Is adaptive sharpness the answer to the generalization puzzle?

Motivation of our work

Main question we want to answer: *Can adaptive sharpness capture generalization in modern practical settings?*

What we mean by **modern practical settings**:

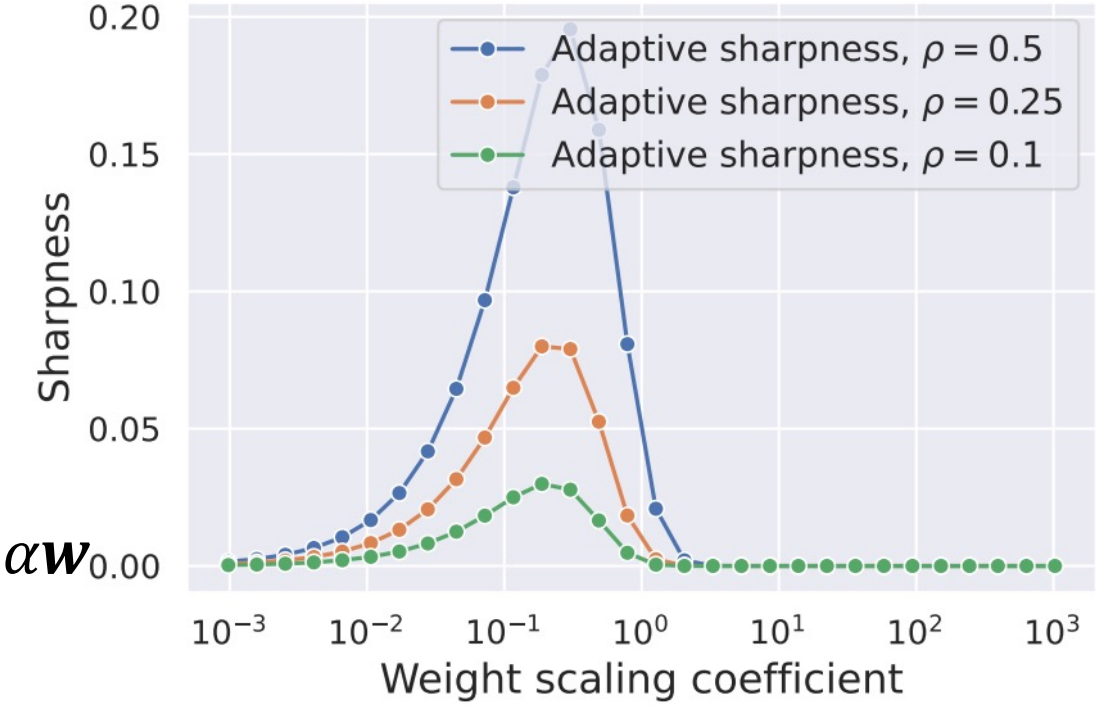
- datasets beyond toyish CIFAR-10 / SVHN,
- vision transformers,
- fine-tuning (totally underexplored),
- out-of-distribution generalization.

We want to have a definite answer about whether sharpness is the right quantity!

Another concern: sensitivity to the scale of the classifier

- Sharpness is strange for classification: scaling the logits by $\alpha \geq 0$ will preserve the classifier but **can arbitrarily change sharpness**
- Adaptive sharpness is no exception: you can keep optimizing the cross-entropy loss and this will drive adaptive sharpness to 0
- This is well illustrated on linear models: $\mathbf{w}' \leftarrow \alpha \boldsymbol{w}$
- Possible solution: **logit normalization**

Linear model that achieves 100% train accuracy



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

We will benchmark all sharpness definitions with and without logit normalization

Setting #1: ViTs trained from scratch on ImageNet

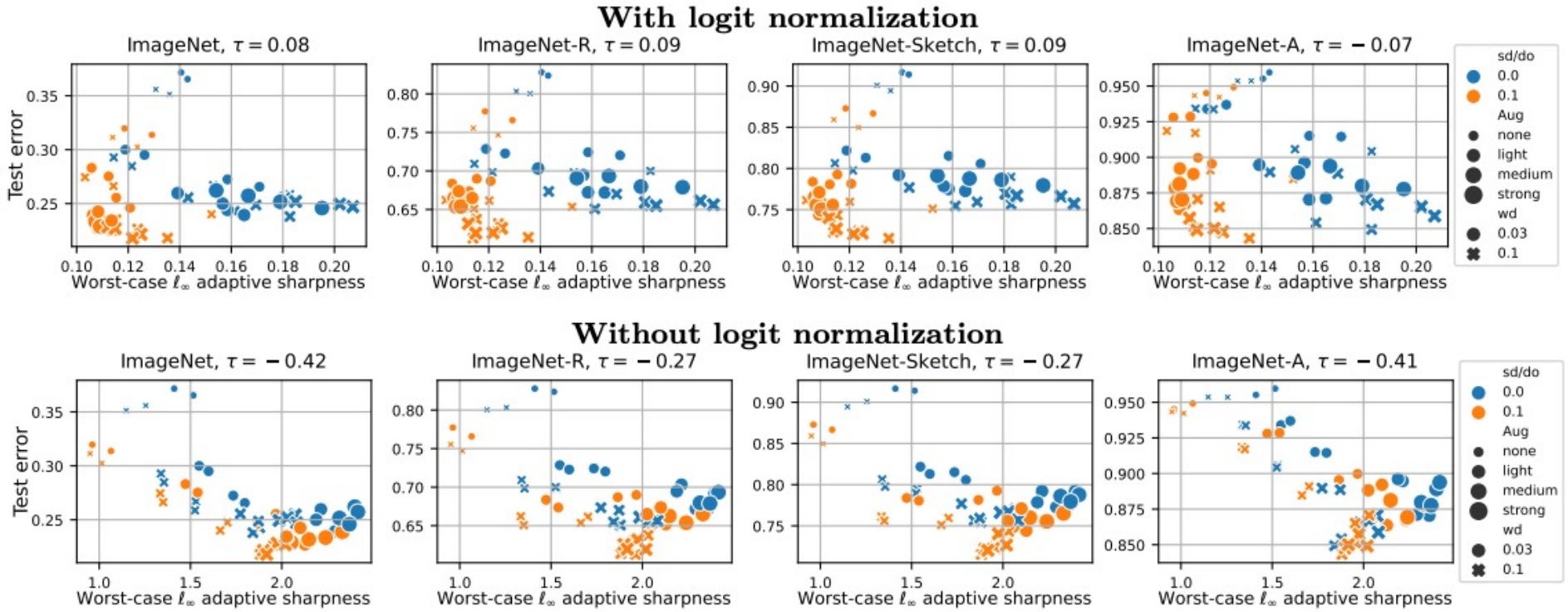


Figure 2: ViT-B/16 trained from scratch on ImageNet-1k. We show for 56 models from Steiner et al. (2021) the test error on ImageNet or its variants (distribution shifts) vs worst-case ℓ_∞ sharpness with (top) or without (bottom) normalization at $\rho = 0.002$. The color indicates whether the networks were trained with stochastic depth/dropout.

The correlation (τ) is either close to 0 or even slightly negative (-0.42 for ImageNet-A)!

Setting #2: ViTs fine-tuned from CLIP on ImageNet

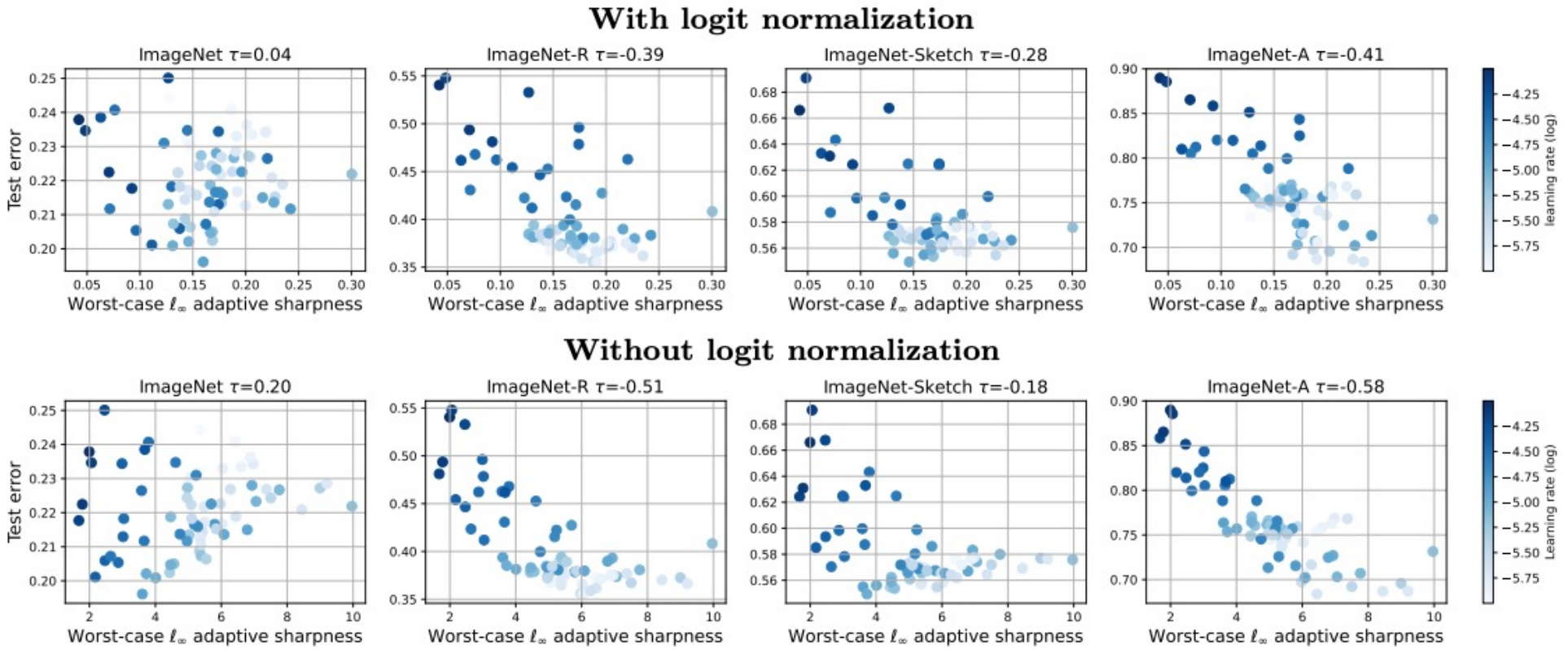


Figure 3: Fine-tuning CLIP ViT-B/32 on ImageNet-1k. We show for 72 models from Wortsman et al. (2022a) the test error on ImageNet or its variants (distribution shifts) vs worst-case ℓ_∞ sharpness with (top) or without (bottom) normalization at $\rho = 0.002$. Darker color indicates larger learning rate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

The correlation is again either **close to 0 or negative**, especially on distribution shifts like ImageNet-R and ImageNet-A (as low as -0.51 and -0.58!)

Setting #3: BERT models fine-tuned on MNLI

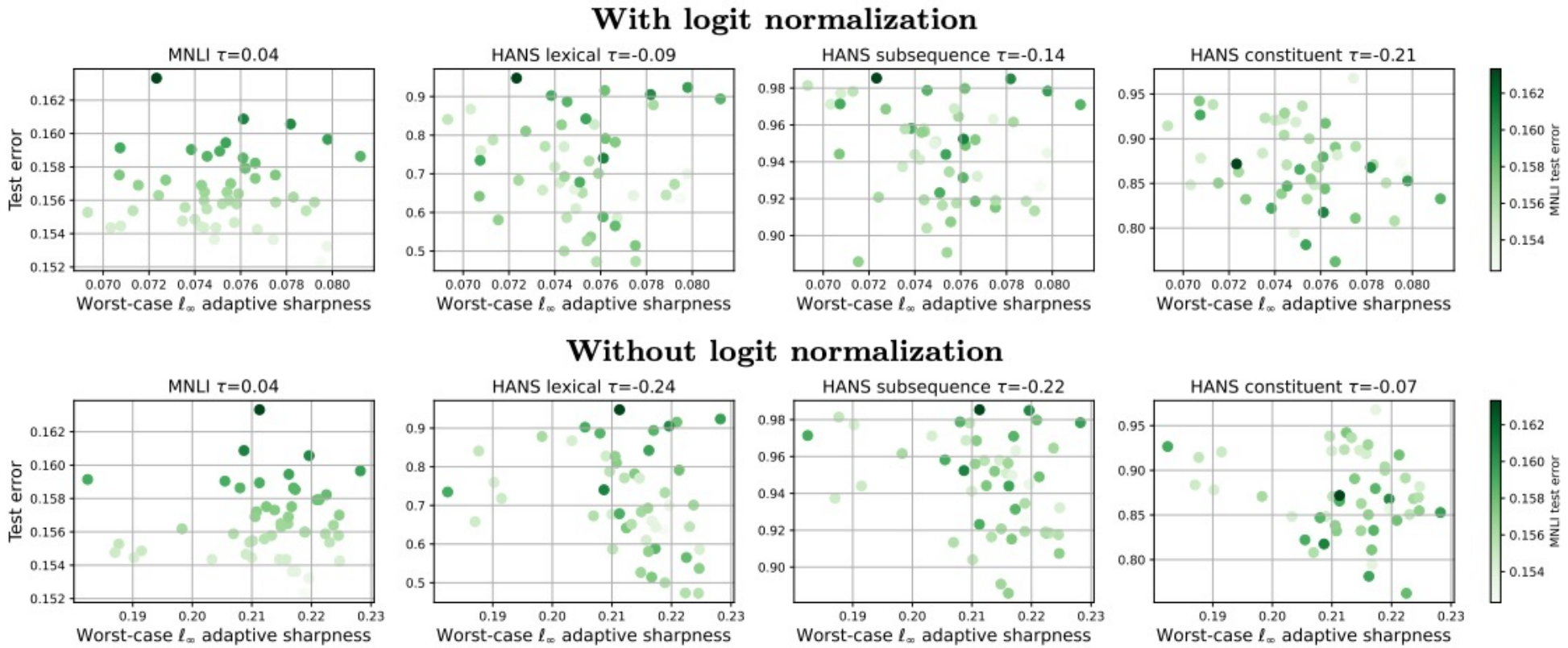
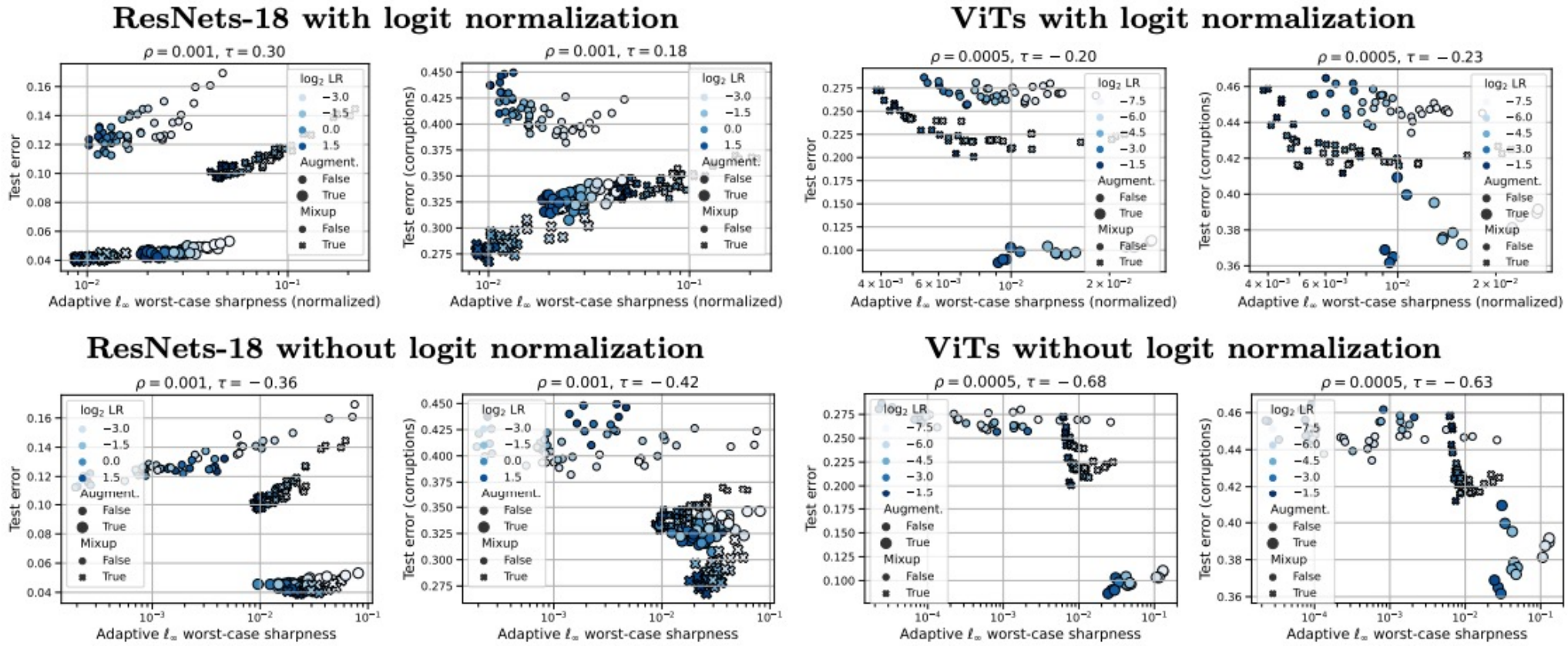


Figure 4: Fine-tuning BERT on MNLI. We show for 50 models the error on MNLI or out-of-distribution domains (HANS subsets) vs worst-case ℓ_∞ sharpness with (top) or without (bottom) normalization at $\rho = 0.0005$. Darker color indicates higher test error on MNLI.

- This case is famous since OOD generalization (see *HANS lexical*) **can be very different**
- However, sharpness **is not helpful** to distinguish which solutions will generalize better for OOD

Setting #4: ResNets and ViTs trained from scratch on CIFAR-10

- Maybe sharpness has to be measured close to a min? select only models ≤1% train error

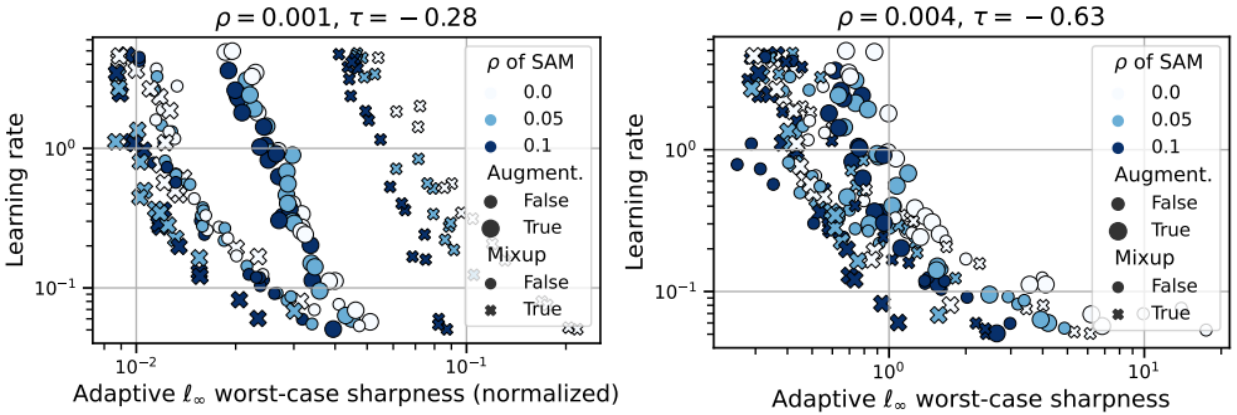


- Positive correlation is present but only within subgroups of models trained with the same augmentations
- Globally, however, correlation is **either close to 0 or negative** (as much as -0.68!)

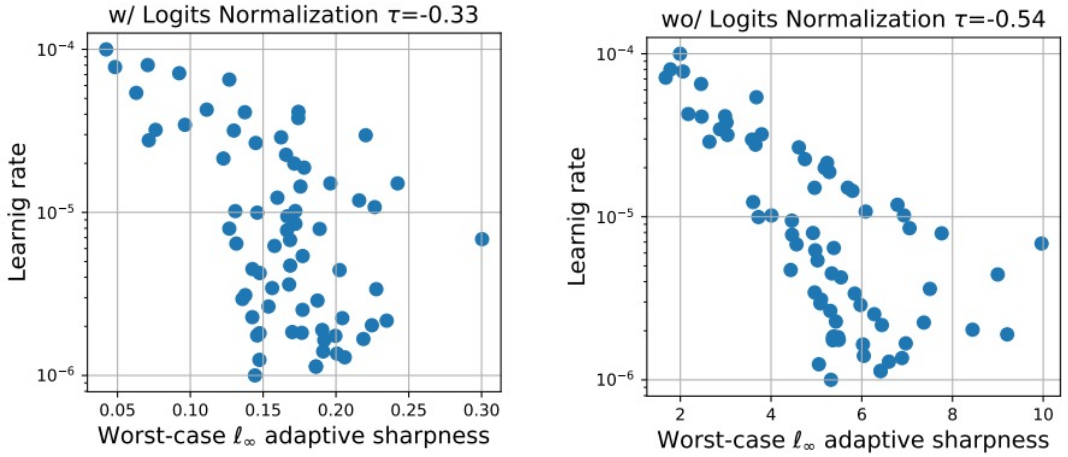
So what does sharpness really capture?

- Overall, we observe that sharpness doesn't correlate well with generalization but rather with some training parameters like the **learning rate**

Training ResNet from scratch



Finetuning ViT on ImageNet



- However, the learning rate can positively or negatively correlate with generalization **depending on the setup**

- Roughly speaking: large LRs are good for pretraining (at least for CNNs), small LRs are good for fine-tuning

Is sharpness the right quantity in the first place? Theoretical insights

- **Simple model**: sparse regression with a diagonal linear network $\boldsymbol{\beta} \coloneqq \boldsymbol{u} \odot \boldsymbol{v}$

  $L(\boldsymbol{w}) := \|\boldsymbol{X}(\boldsymbol{u} \odot \boldsymbol{v}) - \boldsymbol{y}\|_2^2$ for $L(\boldsymbol{w}) = 0$ and $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}$: $\nabla^2 L(\boldsymbol{w}) = \begin{bmatrix} \mathrm{diag}(\boldsymbol{v} \odot \boldsymbol{v}) & \mathrm{diag}(\boldsymbol{u} \odot \boldsymbol{v}) \\ \mathrm{diag}(\boldsymbol{u} \odot \boldsymbol{v}) & \mathrm{diag}(\boldsymbol{u} \odot \boldsymbol{u}) \end{bmatrix}$

- For appropriate adaptive sharpness with

  $c_i = \sqrt{|v_i|/|u_i|}$ for $1 \leq i \leq d$ and $c_i = \sqrt{|u_i|/|v_i|}$ for $d+1 \leq i \leq 2d$

  we get for $\rho \to 0$ that different sharpness definitions capture **totally different quantities**:

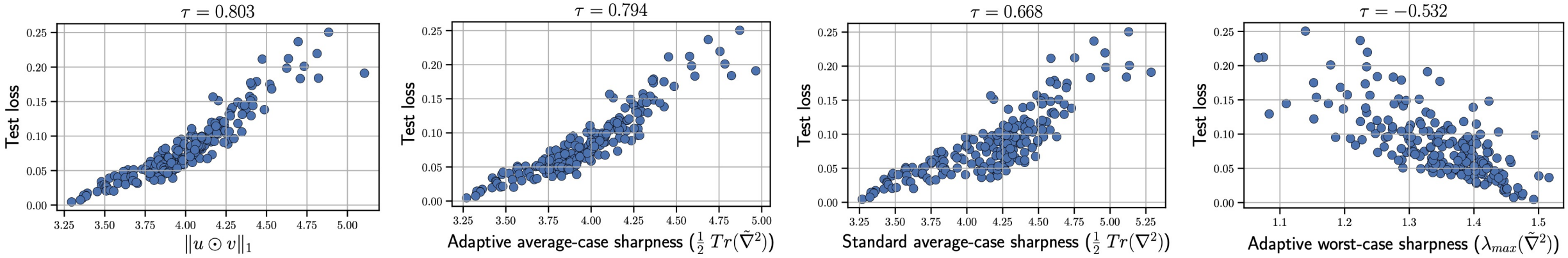
  $$S^\rho_{avg}(\boldsymbol{w}, \boldsymbol{c}) = \frac{1}{2}\sum_{i=1}^{d} u_i^2 |v_i|/|u_i| + \frac{1}{2}\sum_{i=1}^{d} v_i^2 |u_i|/|v_i| = \|\boldsymbol{\beta}\|_1, \qquad S^\rho_{max}(\boldsymbol{w}, \boldsymbol{c}) = \max_{1 \leq i \leq d} |u_i||v_i| = \|\boldsymbol{\beta}\|_\infty$$

- However, we know apriori that for sparse regression only $||\boldsymbol{\beta}||_1$ is the right quantity

- Thus, only a very specific sharpness definition *for this given problem* can explain generalization

What can go wrong with the sharpness definition?

Empirical validation: a bunch of diagonal linear nets trained with different LR and init



- Our analysis suggests that sharpness **can be** the right quantity
- However, choosing the right definition of sharpness requires a **precise understanding of the data and how it interacts with the architecture**
- **This is obviously challenging beyond toy models!**

Note: lots of experiments in the appendix

- We tried many-many sharpness definitions (ℓ_2 vs. ℓ_∞ norms, avg- vs. worst-case, with/without normalization, adaptive vs. non-adaptive sharpness)
- **50+ pages of appendix!**
- We hope we answered the question comprehensively

Appendix

The appendix is organized as follows:

- Sec. A: additional related work.
- Sec. B: omitted derivations for sharpness when $\rho \to 0$, first for the general case and then specifically for diagonal linear networks.
- 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 APGD, 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.

Outlook

- Is it even possible to have a **single measure** that would be causally related to generalization?
- I think it's highly unlikely and too good to be true (as the DLN example illustrates: this depends a lot on the data distribution)
- But: there are some creative proposals like computing disagreement on unlabeled data which correlates pretty well with generalization
- However, for this, we need at least a small amount of *unseen unlabeled data*… then why not assuming that we have unseen *labeled* data?
- Regarding the success of sharpness-aware minimization: it must be a combination of sharpness with some implicit bias of (S)GD that we don't quite understand

Thanks for your attention!

Happy to discuss more :)