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]

Contributions

Can sharpness predict generalization in modern practical settings?

- Empirical evaluation:
  1. training from scratch on {ImageNet, CIFAR-10} with {transformers, CNNs}
  2. fine-tuning transformers on ImageNet and MNLI

- Observation:
  1. sharpness does not correlate well with generalization
  2. sharpness correlates well with LR

- In some cases, sharper minima can generalize better

- Analysis on toy model:
  1. right sharpness measure for generalization is highly data-dependent

Background

Sharpness definitions

- Adaptive average-case m-sharpness wrt vector c in R^p:

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

- Adaptive worst-case m-sharpness wrt vector c in R^p for radius rho:

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

- Experiments use L_inf worst-case adaptive sharpness with m=256

Is sharpness predictive of generalization?

- Strong hypothesis:
  - low sharpness <=> high generalization (high correlation)
  - causal relation

- Weak hypothesis:
  - low sharpness => high generalization
  - sufficient but not necessary

- Spoiler: neither hypotheses hold empirically

When can we compare sharpness across models?

- Only compare models within **the same loss surface**

- For the same loss surface:
  1. architecture should be the same
  2. set of points to measure sharpness should be the same

Invariances for sharpness

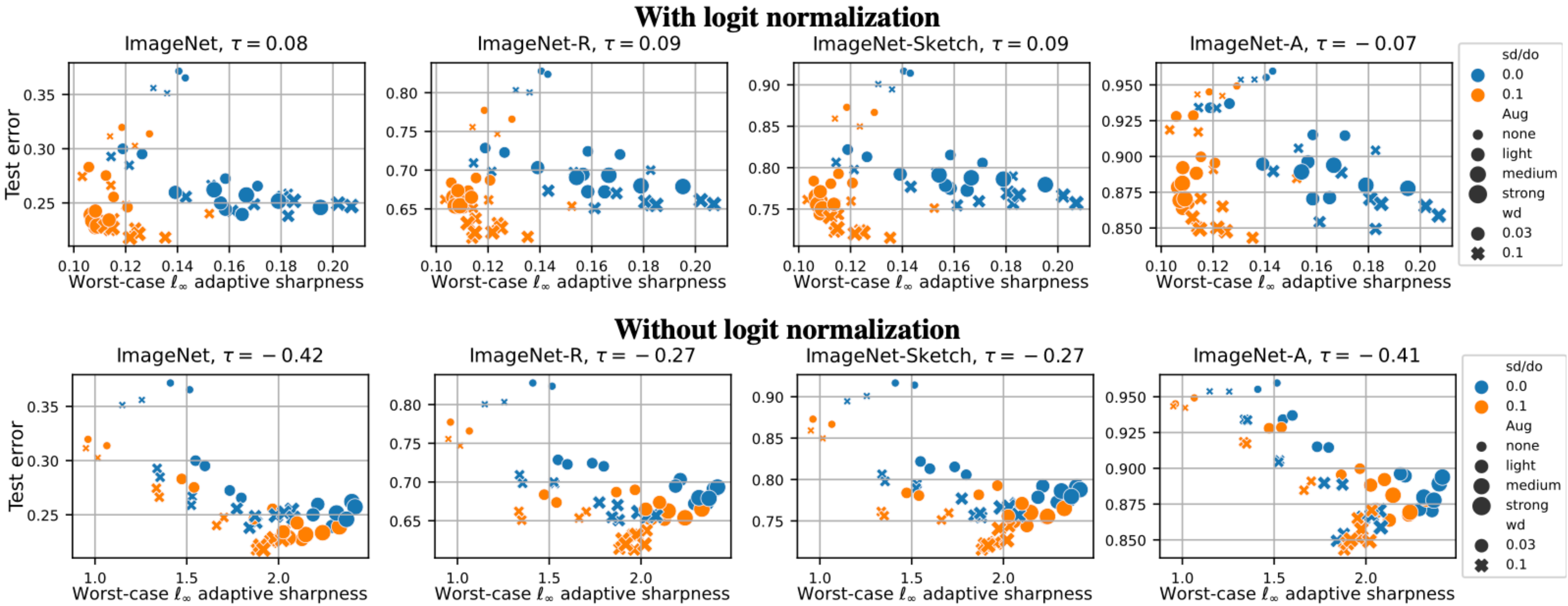
- If T(w) does not change predictions, then it should not change sharpness
- Adaptive sharpness has such invariances
- Need to normalize classification logits to get scale-invariance:

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

Experiments

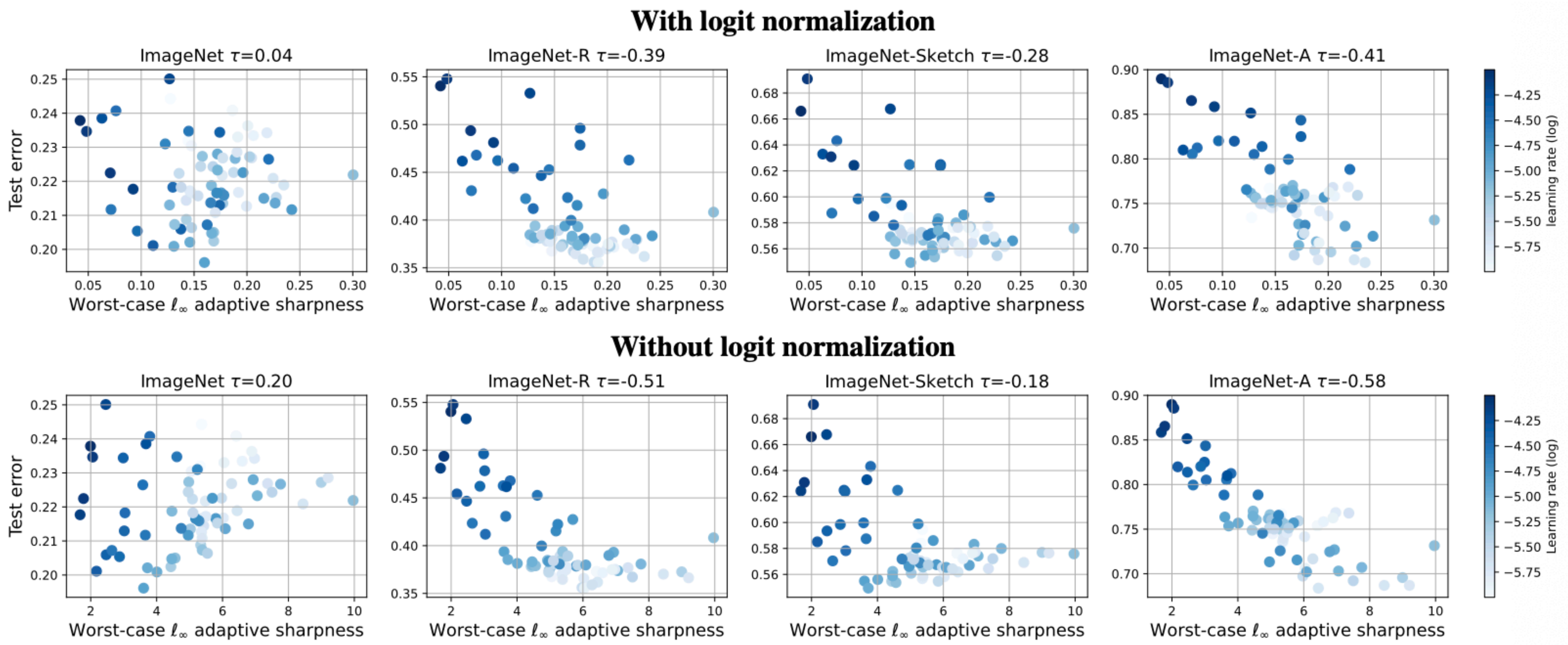
Setting 1: ImageNet training from scratch

- 56 ViT models w different hparams: augmentations, weight decay, dropout, etc
- Test errors from 21.8 % to 37.2 %



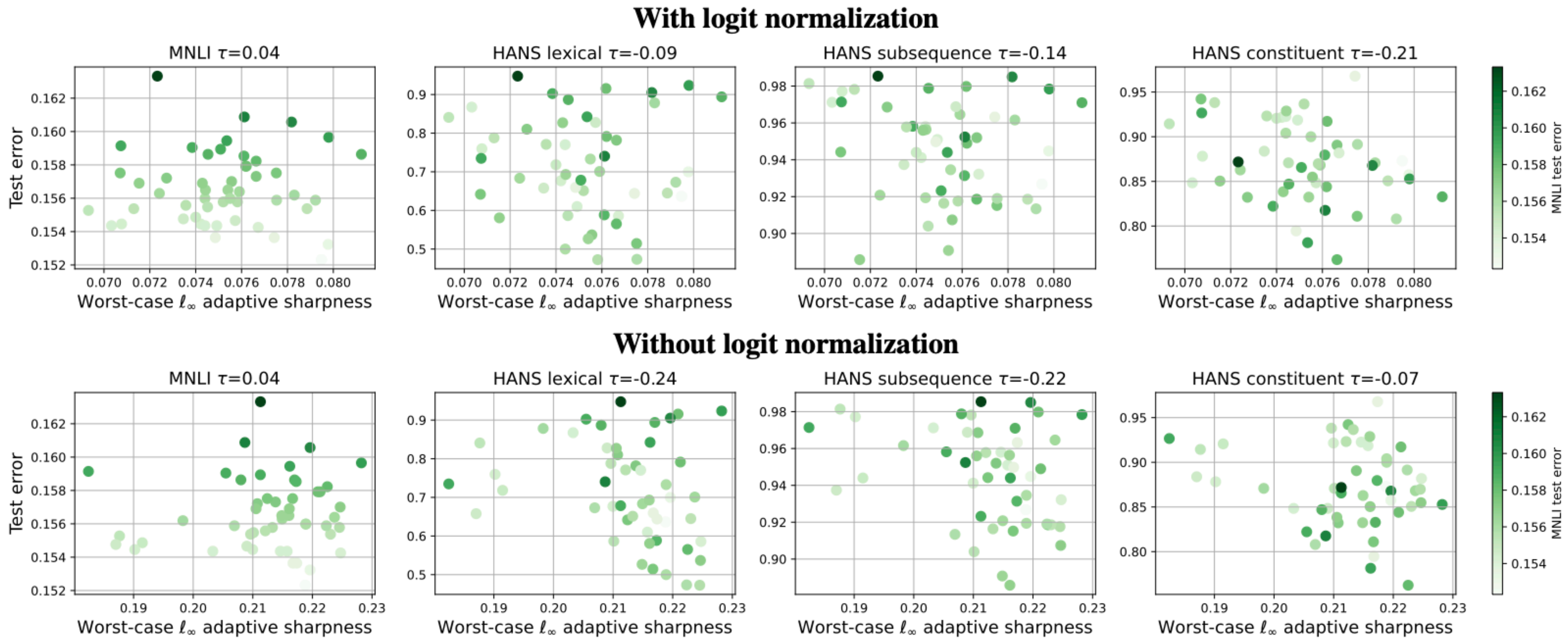
Setting 2: Fine-tuning on ImageNet-1k from CLIP

- 71 fine-tuned CLIP ViT models w hparams: LR, epochs, wt decay, label smoothing, data aug
- Note: higher LR => higher test error. Flatter minima are worse on OOD.



Setting 3: Fine-tuning BERT on MNLI

- 50 fine-tuned BERT models w different seeds: random clf head init, random batching



Analysis

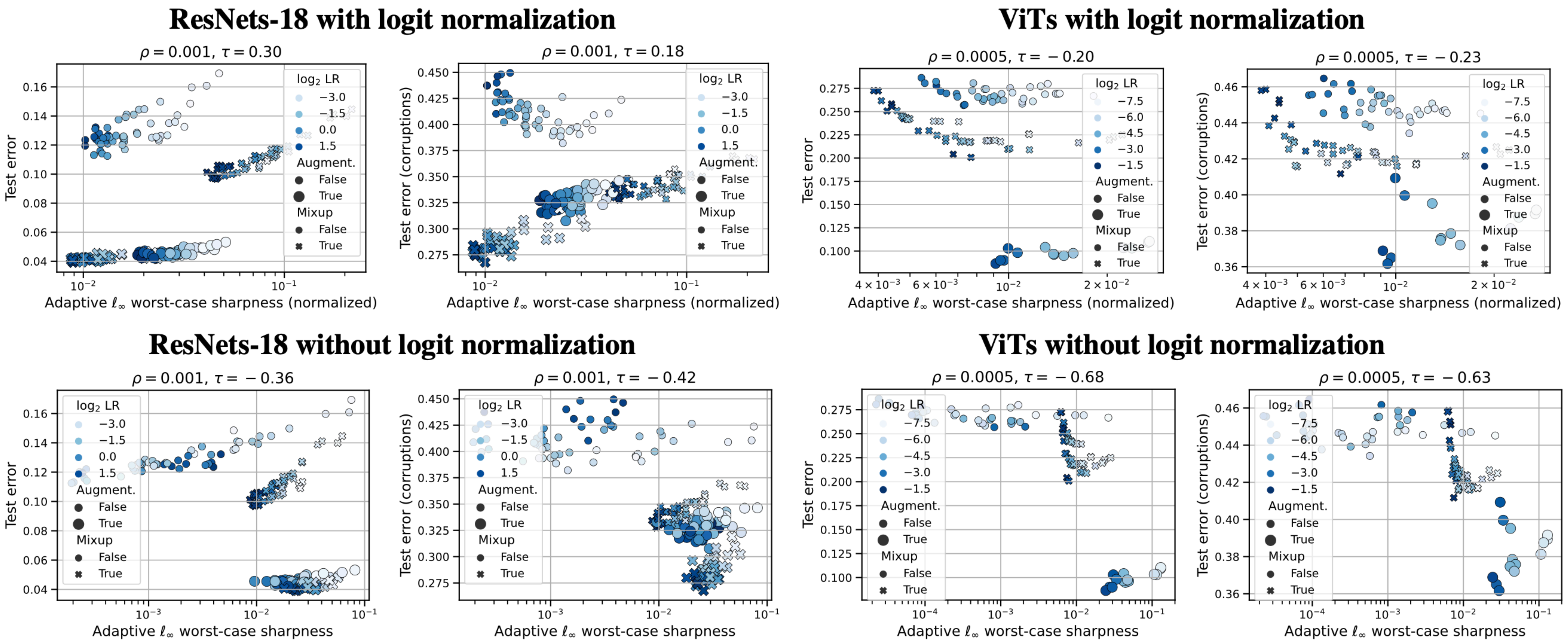
Why are these results counter to prior work?

- Architecture? transformers vs CNNs
- Larger datasets? ImageNet vs CIFAR-10
- Measure sharpness close to a minimum

- New controlled setup: ResNet vs ViTs, on CIFAR-10, trained to ~0% train error.

Observations

- Order of magnitude difference in sharpness, but similar test error
- Still no support for strong or weak hypothesis



Sharpness has strong -ve correlation to LR

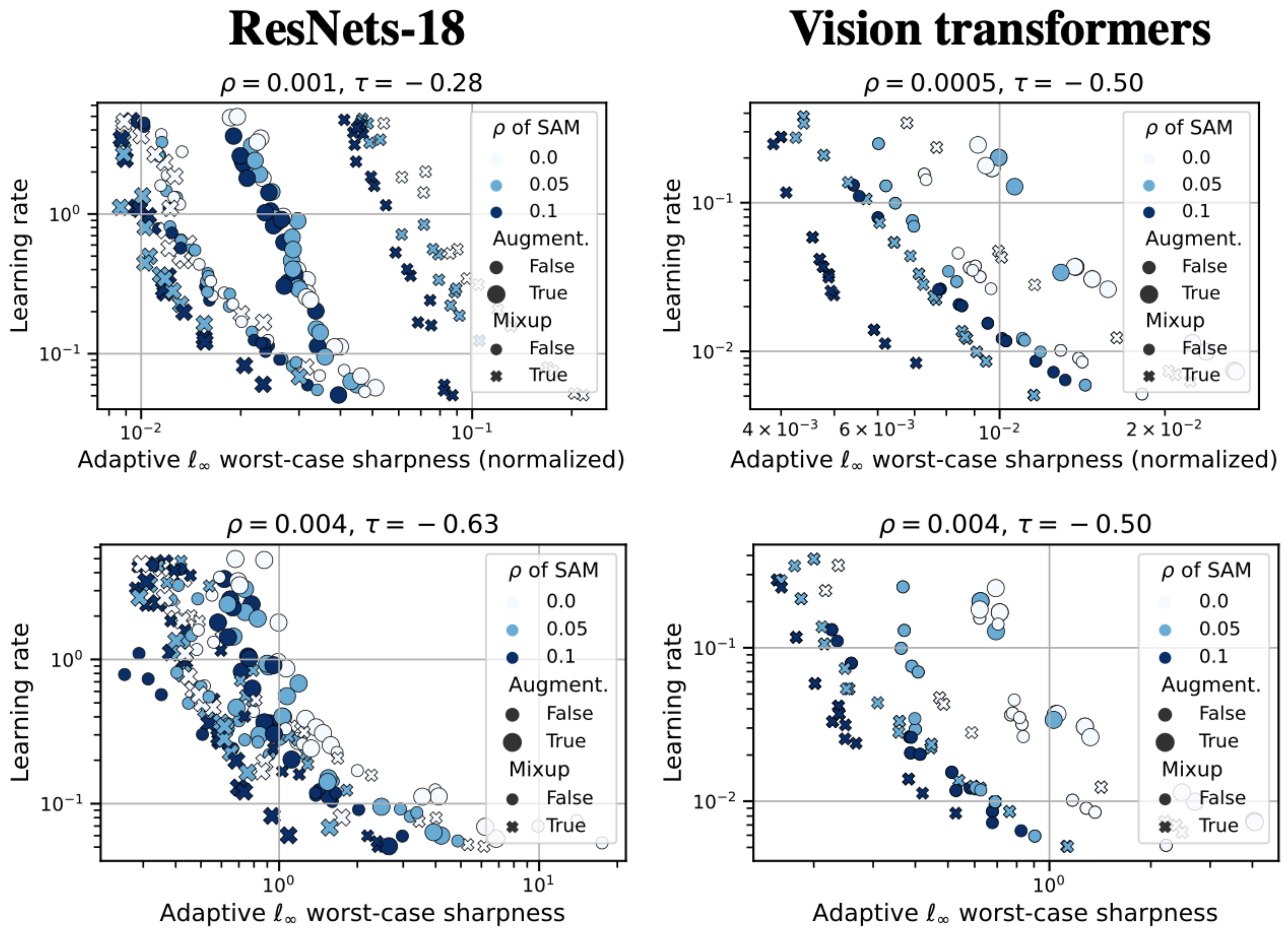


Figure 6: Training from scratch on CIFAR-10. Sharpness negatively correlates with the *learning rate*, especially within each subgroup defined by the same values of `augment` × `mixup`.

Is sharpness even the right measure?

Different sharpness measures have different generalization

- Well understood case of diagonal linear networks

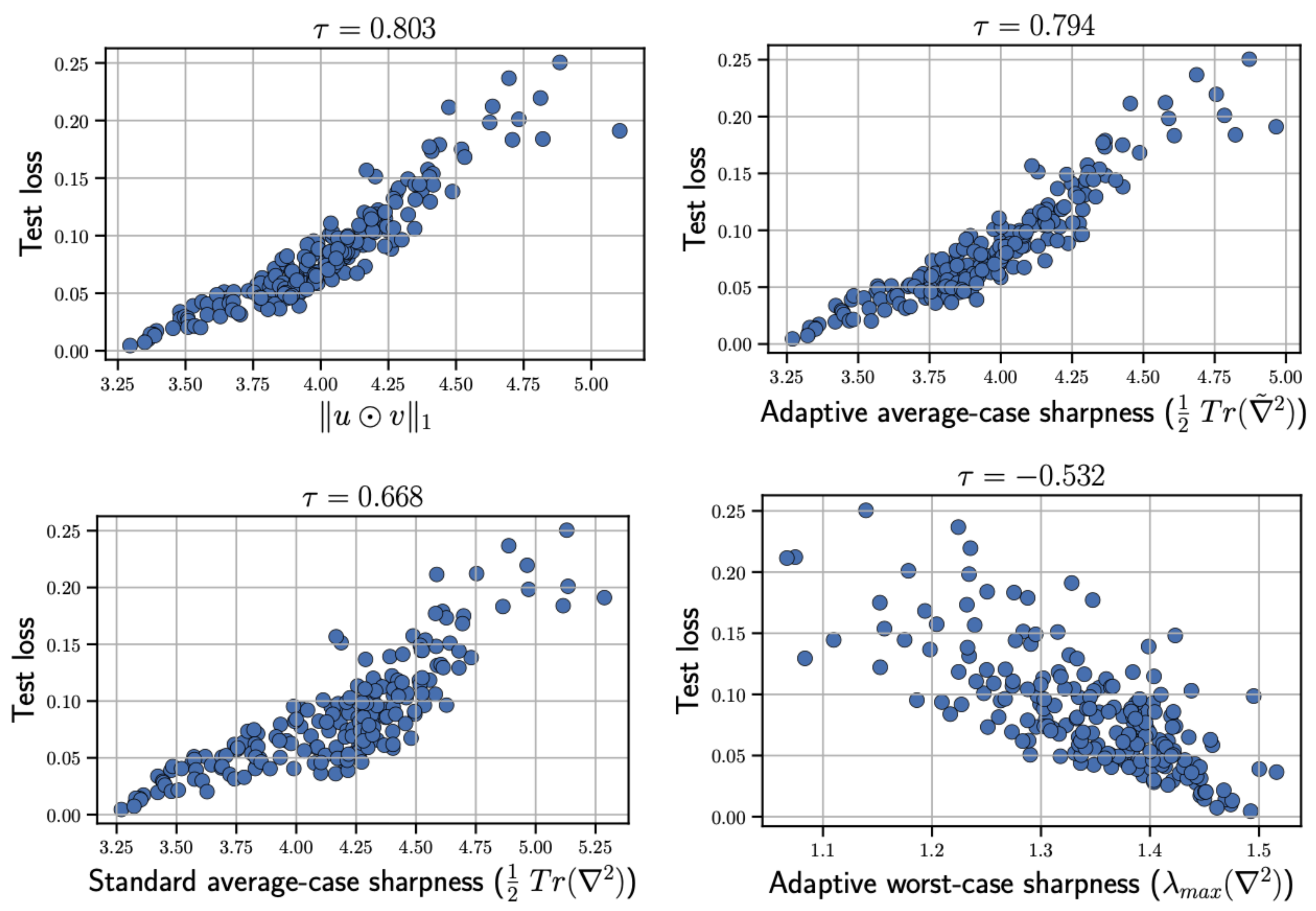


Figure 7: Different generalization measures for diagonal linear networks. $\tilde{\nabla}^2$ denotes the rescaled Hessian corresponding to adaptive sharpness.

Conclusion

- In modern practical settings, sharpness does NOT imply generalization.
- In some setting, sharper minima can generalize better.
- On simple models and data we can understand well,
  there is no universal sharpness definition that predicts generalization.