The Myth of Robust Classes: How Shielding Skews Perceived Stability

Ananyapam De Benjamin Säfken

Institut für Mathematik, TU Clausthal

TU Clausthal
Clausthal University of Technology

github.com/Ananya-pam7

Problem of Perceived Robustness

Conventional robustness metrics rely on Euclidean perturbations that ignore the manifold structure of a classifier's output probabilities. We revisit robustness from an information-geometric perspective and show that high "robustness" can emerge because surrounding decision regions shield a class, not because its boundary is genuinely stable. Our contributions:

- Develop a Fisher–Rao (FR) margin framework that measures how far an input must travel on the predictive manifold before the predicted class changes.
- Expose *class shielding*: intermediate classes deflect natural-gradient trajectories, inflating perceived robustness for the shielded class.
- Provide empirical evidence on CIFAR-10 that the classes considered "robust" are also those that most frequently shield neighbouring categories.

Fisher–Rao Margins

Let $p(y \mid \mathbf{x}; \theta)$ be the classifier and $J(y, \mathbf{x}; \theta) = -\log p(y \mid \mathbf{x}; \theta)$. The data-space Fisher Information Matrix (FIM) at $\mathbf{x}$ is

$$\mathbf{G}_{\mathbf{x}} = \sum_c p_c(\mathbf{x}; \theta)\, [\nabla_{\mathbf{x}} J(c, \mathbf{x}; \theta)][\nabla_{\mathbf{x}} J(c, \mathbf{x}; \theta)]^\top,$$

which equips the predictive manifold with a Riemannian metric capturing how output probabilities react to perturbations. Distances measured with $\mathbf{G}_{\mathbf{x}}$ respect the model's intrinsic geometry, unlike conventional ℓ_p norms.

Second-order (local) approximation. For small perturbations δ, the FR length satisfies

$$D_{\mathrm{FR}}(\mathbf{x}, \mathbf{x} + \delta) \approx \sqrt{\delta^\top \mathbf{G}_{\mathbf{x}} \delta} + \mathcal{O}(\|\delta\|^2).$$

For a path $\gamma(t)$ in input space, the FR length is

$$D_{\mathsf{FR}}(\mathbf{x}_0, \mathbf{x}_1) = \int_0^1 \sqrt{\dot\gamma(t)^\top \mathbf{G}_{\gamma(t)}\, \dot\gamma(t)}\, dt.$$

The FR margin of input $(\mathbf{x}, y)$ is the minimum FR distance required to change the predicted class:

$$\mathrm{margin}(\mathbf{x}, y) = \min_{\mathbf{x}'} \left\{ D_{\mathsf{FR}}(\mathbf{x}, \mathbf{x}') : \arg\max_c p_c(\mathbf{x}'; \theta) \neq y \right\}.$$

This metric reveals whether robustness stems from meaningful logit stability or merely from the geometry of surrounding decision regions.

Natural-Gradient Margin Tracing

We approximate FR margins by following unit-length natural-gradient steps along the logit gap $L_k(\mathbf{x}) = \log p_y(\mathbf{x}; \theta) - \log p_k(\mathbf{x}; \theta)$.

Require: Input $\mathbf{x}_0$, true class y, target class k, damping λ
Ensure: Estimated margin D_{FR} and intervening class

1: $\mathbf{x} \leftarrow \mathbf{x}_0$, $D_{\mathsf{FR}} \leftarrow 0$
2: **while** $\arg\max_c p_c(\mathbf{x}; \theta) = y$ **do**
3: Compute $g = \nabla_{\mathbf{x}} L_k(\mathbf{x})$ where $L_k(\mathbf{x}) = \log p_y(\mathbf{x}; \theta) - \log p_k(\mathbf{x}; \theta)$
4: Solve $(\mathbf{G}_{\mathbf{x}} + \lambda I)v = g$ using conjugate gradients
5: Compute step $\delta = v / \sqrt{g^\top v}$
6: Update $\mathbf{x} \leftarrow \mathbf{x} + \delta$
7: Accumulate $D_{\mathsf{FR}} \leftarrow D_{\mathsf{FR}} + 1$
8: **end while**
9: **return** D_{FR}, $\arg\max_c p_c(\mathbf{x}; \theta)$

Key theoretical statements

Shielding Necessitates Intermediate Mass. Consider a smooth classifier on a compact domain. If a class A shields a class B along natural-gradient geodesics from C to B, then there exists an open set where the predictive mass of A is bounded below by a positive constant. In particular, shielding is a geometric–probabilistic statement about volumetric occupancy of the predictive manifold.

FR vs Euclidean vulnerability ranking. Let $\mathcal{D}$ be a dataset and let $\mathrm{rank}_{\mathrm{FR}}$ and rank_{ℓ_2} be vulnerability ranks computed with FR margins and Euclidean norms respectively. Then, under the empirical Fisher approximation and mild regularity,

$$\mathrm{rank}_{\mathrm{FR}} = \mathrm{rank}_{\ell_2} + o(1)$$

as the sample size grows, but constant offsets can appear due to anisotropic score-variances.

Experiments

Setup. We analyze 3,659 successful class-transition paths generated from 800 CIFAR-10 test images under a pretrained ResNet classifier. Each source image spawns natural-gradient trajectories towards all nine alternative targets. We record the cumulative FR distance, the first class hit, and whether shielding occurs.

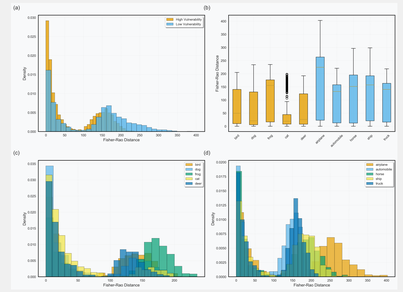


The path matrix shows that animal classes (bird, cat, dog, frog) rarely retain their predictions, whereas vehicles and ship remain dominant self-labels.

Class	Shielding %	Mean FR margin	Mean $\|\delta\|_2$	Vuln rank
airplane	58.2	3.21	0.98	1
automobile	40.5	2.76	0.84	3
bird	12.1	1.05	0.32	9
cat	9.8	0.98	0.28	10
deer	18.4	1.48	0.52	7
dog	14.2	1.12	0.35	8
frog	21.0	1.78	0.60	6
horse	27.5	2.05	0.70	5
ship	62.7	3.45	1.12	2
truck	51.7	2.98	0.95	4

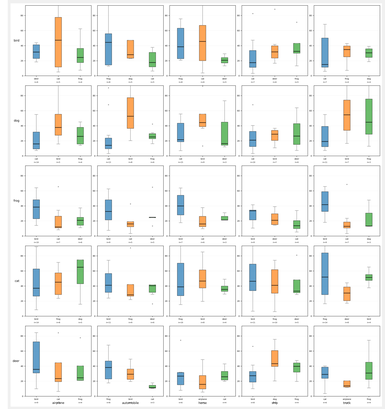
Columns: Shielding % = fraction of successful paths for which the class was the first intervening label; Mean FR margin = average FR distance required to flip from source to any other target; Mean $\|\delta\|_2$ = average Euclidean norm at flip; Vuln rank = vulnerability (1 = most robust).

Attack Effort in FR Space

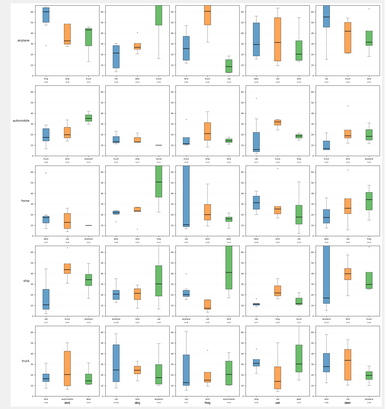


Vulnerable classes require small FR distances to flip, aligning with classical adversarial findings. Airplane and ship demand the largest FR shifts, suggesting that their probability mass covers extensive regions of the predictive manifold.

Shielding Frequency Landscape



High to Low Shielding



Low to High Shielding

Vehicle-style classes (truck, ship, airplane) dominate as shields when moving from robust to fragile classes (left). The reciprocal analysis (right) shows that fragile classes almost never shield their robust counterparts, underscoring the directional nature of shielding.

References

- Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251-276.
- Goodfellow, I. J., et al. (2015). Explaining and harnessing adversarial examples. *ICLR*.
- Moosavi-Dezfooli, S. M., et al. (2016). DeepFool: a simple and accurate method to fool deep neural networks. *CVPR*.
- Madry, A., et al. (2019). Towards deep learning models resistant to adversarial attacks. *ICLR*.
- Zhao, C., et al. (2019). The adversarial attack and detection under the Fisher information metric. *AAAI*.
- Kunstner, F., et al. (2020). Limitations of the empirical Fisher approximation for natural gradient descent. *ICML*.