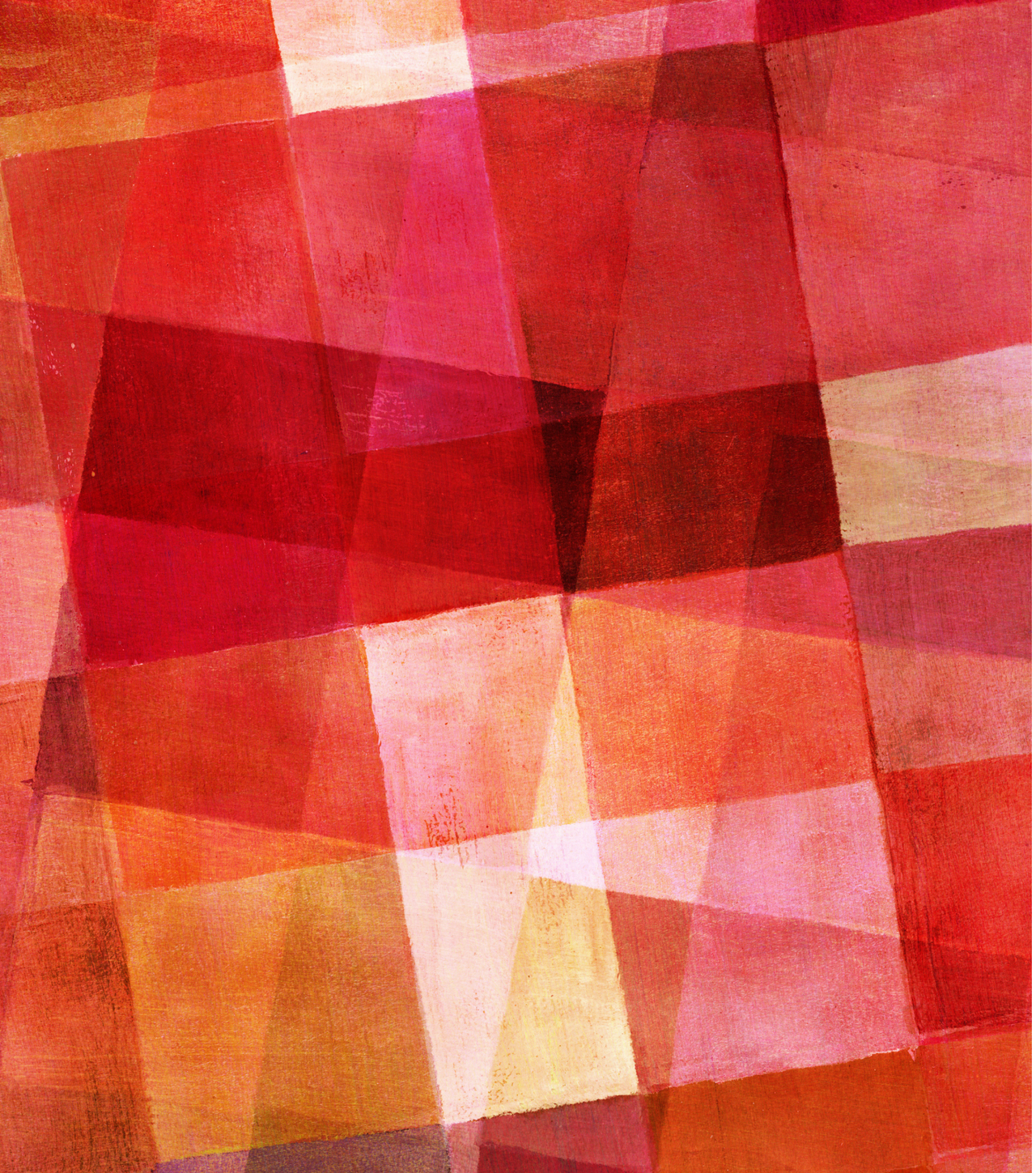
EXPLICIT TRADEOFFS BETWEEN ADVERSARIAL AND NATURAL DISTRIBUTIONAL ROBUSTNESS

Nahal Mirzaie, Amir Ezzati

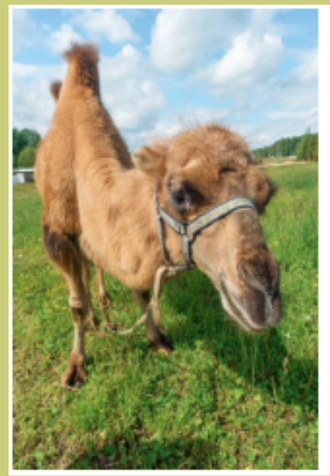


NATURAL DISTRIBUTIONAL ROBUSTNESS

- Ensuring models rely on **core features** (causal) rather than **spurious features** (correlated but non-causal)
- **Spurious features** are features that only hold in specific training distributions but broken under others.
- **Core features** are causal features that remain predictive across all distributions.

Train Distribution

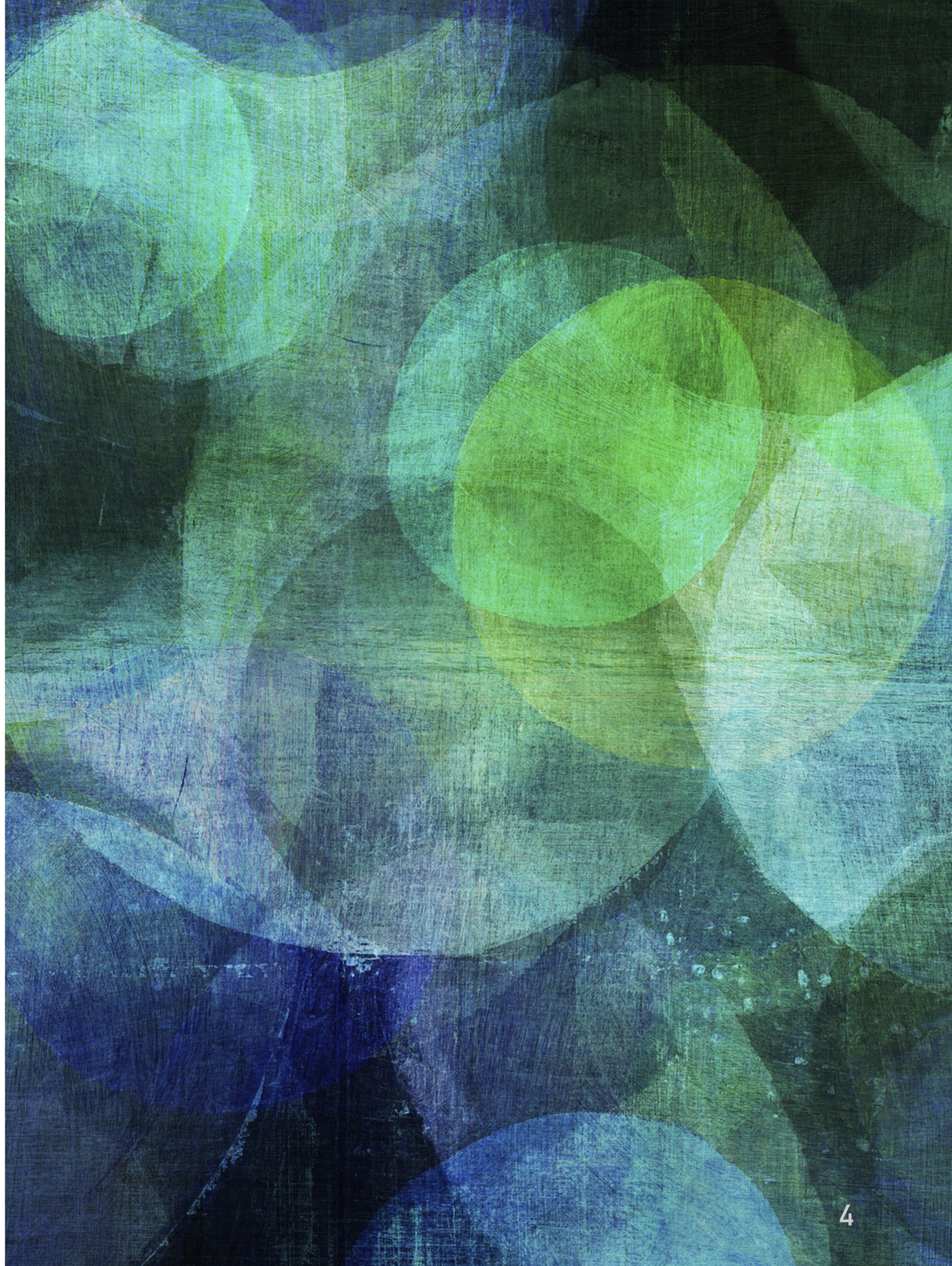
Test Distribution



Background as a shortcut in Camel-Cow classification task

ADVERSARIAL ROBUSTNESS

Robustness to adversarial examples, slight perturbations within the data distribution.



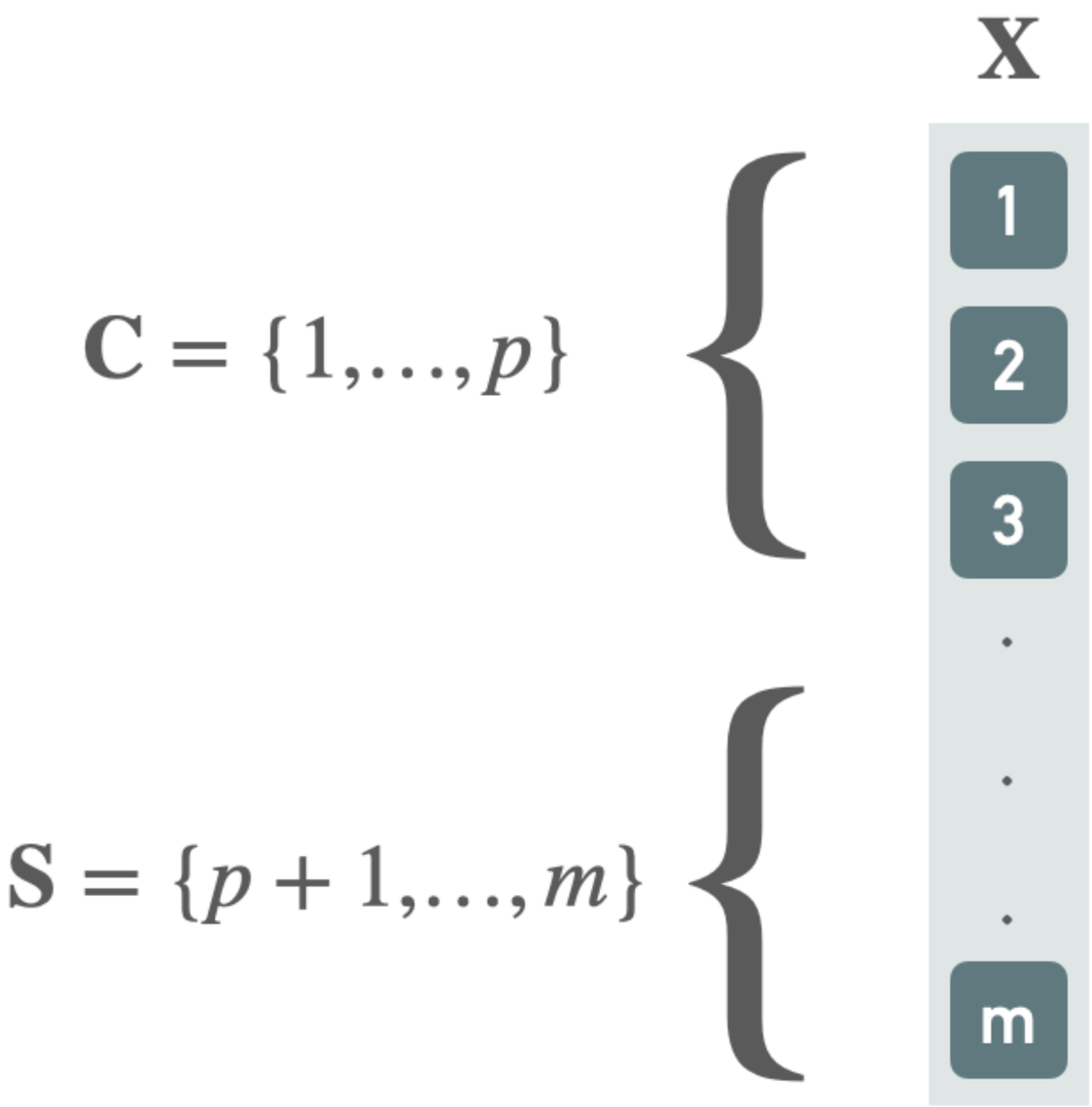
RELIABLE MODELS

- Should have natural distributional robustness
  - Do not rely on spurious features for predictions
- Should be robust against adversarial examples
  - Have the lowest error on adversarial examples

Can a single model achieve both types of robustness simultaneously? Specifically, are the optimal parameters for adversarial robustness the same as those required for natural distributional robustness?

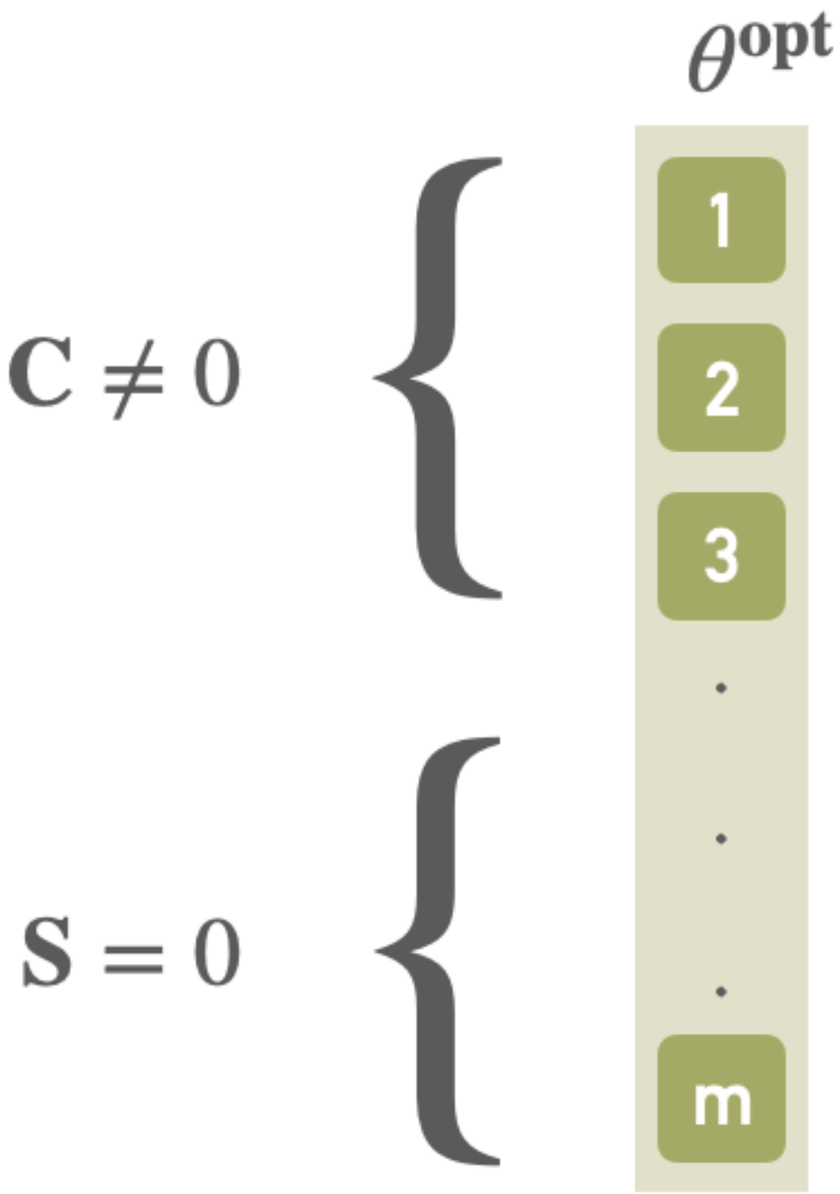
PAPER ORGANIZATION

- **Theoretical Evidence:**
  - In a linear regression model, it has been proven that the parameters ensuring adversarial robustness differ from those ensuring natural distributional robustness, which avoids reliance on spurious features.
- **Empirical Evidence:**
  - In two setups **1) linear models on mock dataset**, and **2) deep models on real datasets** it has been shown:
    - **(Effect)** Adversarial training *reduces* core and *increases* background reliances
    - **(Reverse Effect)** Presence of Spurious Correlations *Improves* Adversarial Robustness



THEORETICAL EVIDENCE

- $\mathbf{X} \in \mathbb{R}^m$
- $\mathbf{X} \sim \mathcal{N}(0,\boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$
- WLOG, $\mathbf{X} = [\mathbf{C}, \mathbf{S}]$
- $\mathbf{Y} = \langle \mathbf{X}, \boldsymbol{\theta} \rangle + \mathbf{W}$
- $\boldsymbol{\theta} \in \mathbb{R}^m$
- $\mathbf{W} \in \mathbb{R}$
- $\mathbf{W} \sim \mathcal{N}(0,\sigma_w^2)$



THEORETICAL EVIDENCE

- Optimal parameters for natural distributional robustness (θ^{opt}) don't rely on spurious features
- This indicates $\mathbf{Y} \perp \mathbf{X}_S \mid \mathbf{X}_C$

THEORETICAL EVIDENCE

- Now consider adversarial loss as follows

$$L_{p,\epsilon}(\boldsymbol{\theta}) = \mathbb{E}\left[\max\left(Y - \langle \mathbf{X} + \boldsymbol{\delta}, \boldsymbol{\theta}\rangle\right)^2\right], \quad \|\boldsymbol{\delta}\|_p \leq \epsilon$$

- We first show that the inner maximization is equal to

$$\max\left(\,|\,Y - \langle \mathbf{X} + \boldsymbol{\delta}, \boldsymbol{\theta}\rangle\,|\,\right)^2 = \max(\,|\,Y - \langle \mathbf{X}, \boldsymbol{\theta}\rangle\,|\, + \epsilon \cdot \|\boldsymbol{\theta}\|_q)^2$$

- And then we show that the adversarial loss can be written as

$$L_{p,\epsilon}(\boldsymbol{\theta}) = c_2 \cdot \sigma_\theta^2 + \left(c_1\sigma_\theta + \epsilon \cdot \|\boldsymbol{\theta}\|_q\right)^2$$

$$\textit{where } \sigma_\theta^2 = (\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{opt}})^\top \boldsymbol{\Sigma} (\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{opt}}) + \sigma_w^2$$

$$\textit{and } c_1 = \sqrt{\frac{2}{\pi}} < 1, \quad c_2 = 1 - c_1^2$$

THEORETICAL EVIDENCE

$$| Y - \langle \mathbf{X} + \boldsymbol{\delta}, \boldsymbol{\theta} \rangle | = | Y - \langle \mathbf{X}, \boldsymbol{\theta} \rangle | + \epsilon \cdot \| \boldsymbol{\theta} \|_q$$

- Proof:

$$\left| Y - \langle \mathbf{X} + \boldsymbol{\delta}, \boldsymbol{\theta} \rangle \right| \leq \left| Y - \langle \mathbf{X}, \boldsymbol{\theta} \rangle \right| + \left| \langle \boldsymbol{\delta}, \boldsymbol{\theta} \rangle \right|$$

$$\left| Y - \langle \mathbf{X} + \boldsymbol{\delta}, \boldsymbol{\theta} \rangle \right| \leq \left| Y - \langle \mathbf{X}, \boldsymbol{\theta} \rangle \right| + \left\| \boldsymbol{\delta}\boldsymbol{\theta} \right\|_1$$

$$\left| Y - \langle \mathbf{X} + \boldsymbol{\delta}, \boldsymbol{\theta} \rangle \right| \leq \left| Y - \langle \mathbf{X}, \boldsymbol{\theta} \rangle \right| + \left\| \boldsymbol{\delta} \right\|_p \left\| \boldsymbol{\theta} \right\|_q$$

$$\left| Y - \langle \mathbf{X} + \boldsymbol{\delta}, \boldsymbol{\theta} \rangle \right| \leq \left| Y - \langle \mathbf{X}, \boldsymbol{\theta} \rangle \right| + \epsilon . \left\| \boldsymbol{\theta} \right\|_q$$

(Hölder's inequality)

$$\|fg\|_1 \leq \|f\|_p \|g\|_q$$

$$\frac{1}{p} + \frac{1}{q} \leq 1$$

THEORETICAL EVIDENCE

- Remember that $L_{p,\epsilon}(\boldsymbol{\theta}) = \mathbb{E}\left[\max\left(Y - \langle \mathbf{X} + \boldsymbol{\delta}, \boldsymbol{\theta}\rangle\right)^2\right], \quad \|\boldsymbol{\delta}\|_p \leq \epsilon$

$$
\begin{aligned}
L_{p,\epsilon} &= \mathbb{E}\left[\left(|Y - \langle X, \theta\rangle| + \epsilon \cdot \|\theta\|_q\right)^2\right] \\
&= \mathbb{E}\left[(Y - \langle X, \theta\rangle)^2\right] + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot \epsilon \cdot \|\theta\|_q \cdot \mathbb{E}\left[|Y - \langle X, \theta\rangle|\right]
\end{aligned}
$$

- We know that $\mathbf{Y} = \langle \mathbf{X}, \boldsymbol{\theta}\rangle + \mathbf{W}$

$$
\overset{(a)}{=} \mathbb{E}\left[\left(\left\langle X, \theta - \theta^{\text{opt}}\right\rangle + W\right)^2\right] + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot \epsilon \cdot \|\theta\|_q \cdot \mathbb{E}\left[\left|\left\langle X, \theta - \theta^{\text{opt}}\right\rangle + W\right|\right]
$$

THEORETICAL EVIDENCE

- Define v_θ as $\langle \mathbf{X}, \boldsymbol{\theta} - \boldsymbol{\theta}_{\text{opt}} \rangle + \mathbf{W}$
- So we can argue that $v_\theta \sim \mathcal{N}(0, \sigma_\theta^2)$

$$L_{p,\epsilon} = \mathbb{E}\left[\left(\langle X, \theta - \theta^{\text{opt}} \rangle + W\right)^2\right] + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot \epsilon \cdot \|\theta\|_q \cdot \mathbb{E}\left[\left|\langle X, \theta - \theta^{\text{opt}} \rangle + W\right|\right]$$

$$= \mathbb{E}\left[v_\theta^2\right] + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot \epsilon \cdot \|\theta\|_q \cdot \mathbb{E}\left[|v_\theta|\right]$$

- We know $\mathbb{E}\left[\,|\,\mathcal{N}(0, \sigma^2)\,|\,\right] = \sqrt{\frac{2}{\pi}} \cdot \sigma$

$$\overset{(a)}{=} \sigma_\theta^2 + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot c_1 \cdot \epsilon \cdot \|\theta\|_q \cdot \sigma_\theta$$

$$= (c_1^2 + c_2) \cdot \sigma_\theta^2 + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot c_1 \cdot \epsilon \cdot \|\theta\|_q \cdot \sigma_\theta \qquad (c_1^2 + c_2 = 1)$$

$$= c_2 \cdot \sigma_\theta^2 + (c_1 \sigma_\theta + \epsilon \cdot \|\theta\|_q)^2$$

THEORETICAL EVIDENCE CONCLUSION

Theorem 1. *Assume that* $Y = \langle X, \theta^{opt} \rangle + W$ *where* $W \sim N(0, \sigma_w^2)$ *is independent of* X *and* $\theta^{opt} \in \mathbb{R}^m$ *is a fixed parameter. Assume further that* X *follows the distribution* $N(0, \Sigma)$ *and define* σ_θ^2 *as* $(\theta - \theta^{opt})^T \Sigma (\theta - \theta^{opt}) + \sigma_w^2$. *The loss function* (1) *is equivalent to*

$$L_{p,\epsilon}(\theta) = c_2 \cdot \sigma_\theta^2 + (c_1 \sigma_\theta + \epsilon \cdot \|\theta\|_q)^2 \tag{2}$$

where $c_1 = \sqrt{\frac{2}{\pi}} < 1$, $c_2 = 1 - c_1^2$ *and* $\|.\|_q$ *is the dual norm of* $\|.\|_p$, *i.e,* $\frac{1}{p} + \frac{1}{q} = 1$. *Furthermore, the above formulation is convex in* θ.

- Theorem 1 shows that the optimal value $\hat{\theta}$ minimizing the adversarial loss $L_{p,\epsilon}(\boldsymbol{\theta})$ *is not* θ^{opt} and in general , may be non-zero on the set of spurious features S.

EMPIRICAL EVIDENCE WITH LINEAR MODELS

- **Setup:**
  - $\mathbf{\Sigma} = \mathbf{Q}\mathbf{Q}^\top$
  - $\mathbf{X} \sim \mathcal{QN}(0, \mathbf{I})$
  - $\sigma_w = 0.1$

$$\widetilde{Q} = \begin{bmatrix} 1 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & 0 \\ \eta & \eta & 1 & 0 & 0 \\ \eta & \eta & 0 & 1 & 0 \\ \eta & \eta & 0 & 0 & 1 \end{bmatrix} \qquad Q_{i,j} = \frac{\widetilde{Q}_{i,j}}{\sqrt{\sum_{i,j'} \widetilde{Q}^2_{i,j'}}}$$

- **Metric**
  - Norm Fraction over Spurious features

$$\mathbf{NFS}(\theta) = \frac{\sum_{i \in S} \theta_i^2}{\sum_j \theta_j^2}$$

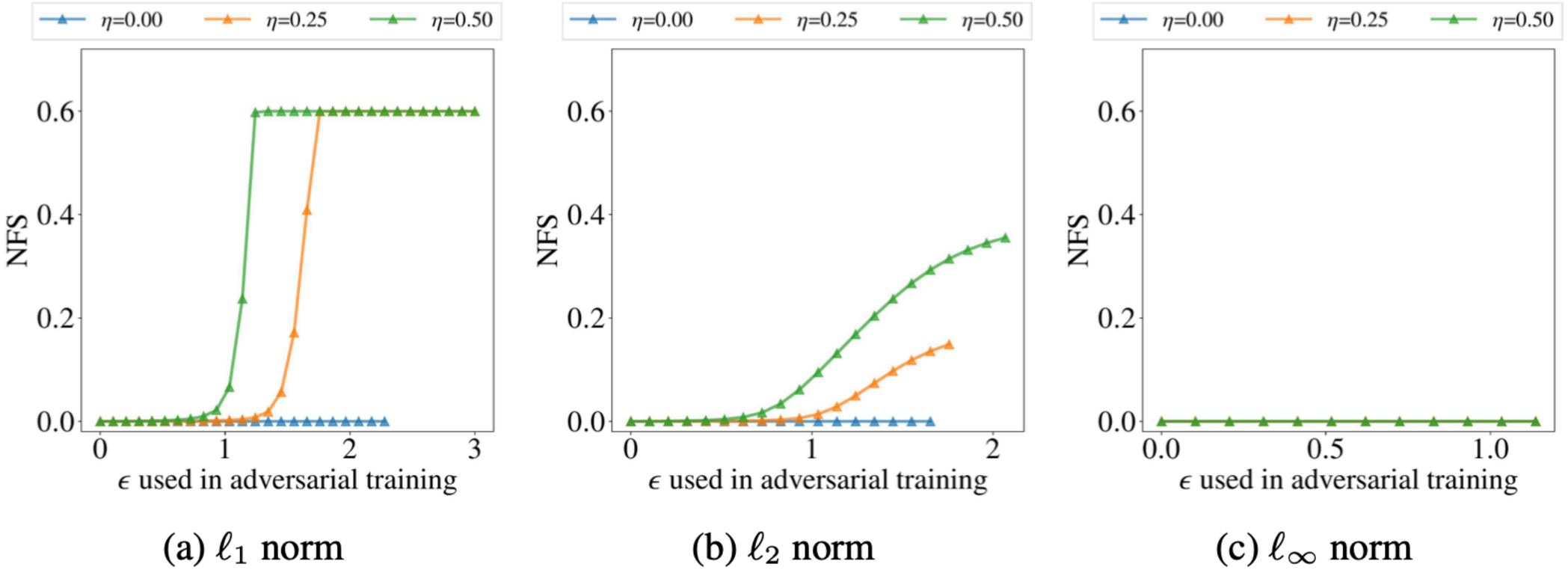


Figure 1) Reliance on Spurious features

Empirical Evidences with linear models

(Reliance on Spurious features)

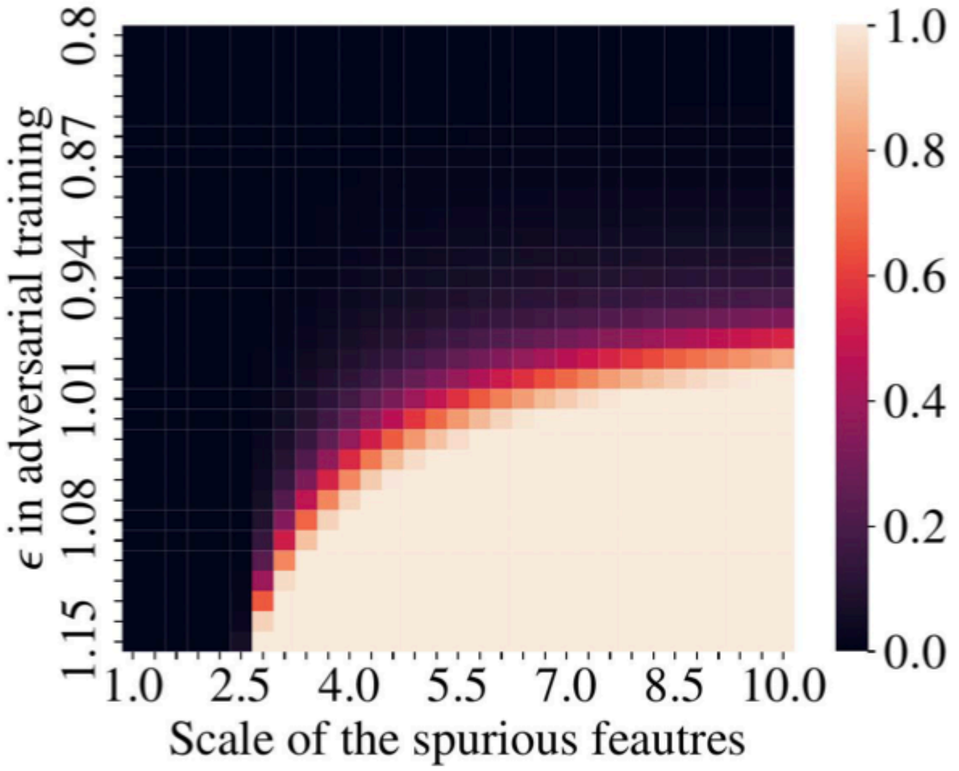


Figure 2) Effect of Scale of spurious features for l_∞

Empirical Evidences with linear models

(Scale of spurious features)

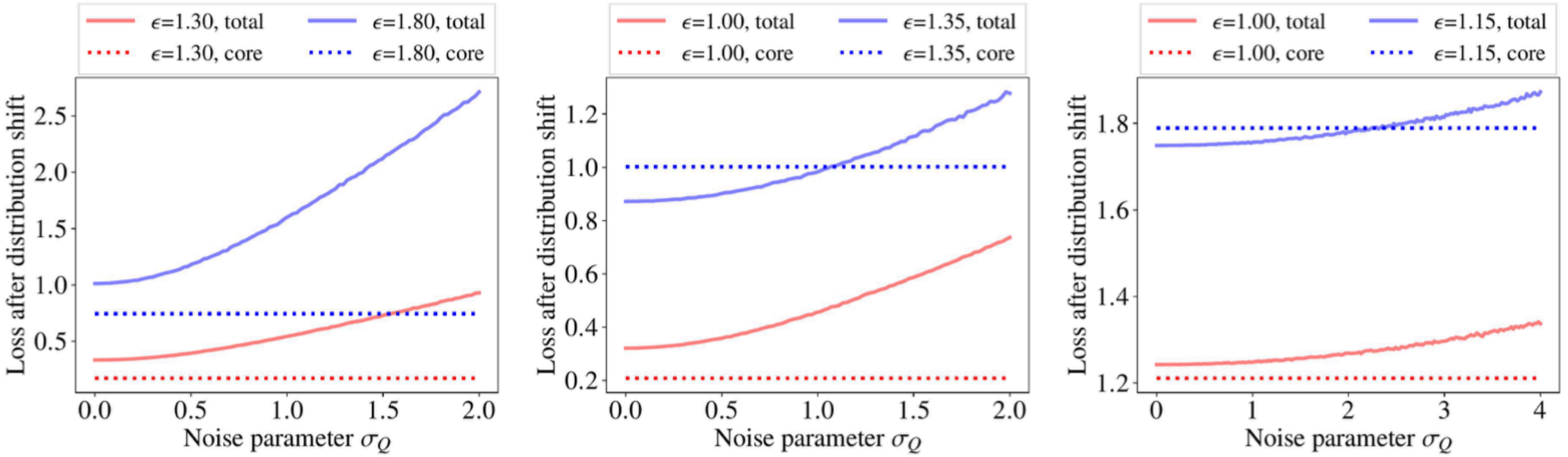


Figure 3) Effect of spurious reliance on distributional robustness

Empirical Evidences with linear models

(Effect of spurious reliance on distributional robustness)

EMPIRICAL EVIDENCE WITH DEEP MODELS

- Models
  - ResNet18, ResNet50
- Norms
  - l_2, l_∞
- Datasets
  - ImageNet-C
  - ObjectNet
  - RIVAL10
  - Salient ImageNet-1M
  - ImageNet-9
  - WaterBird
  - CIFAR10

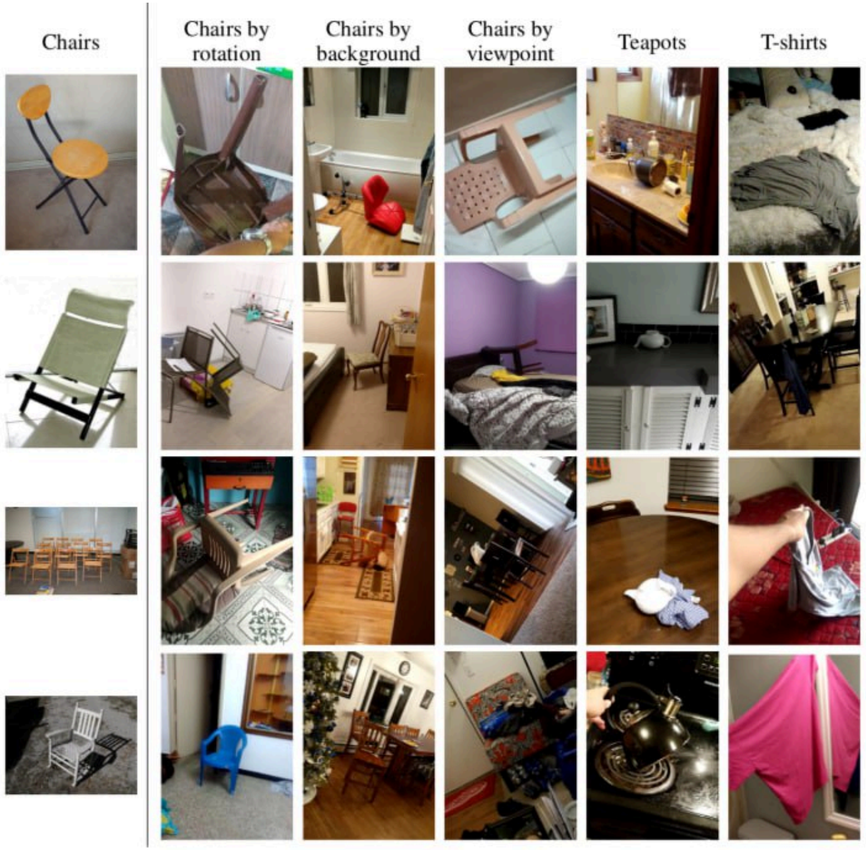


Figure 4) ObjectNet

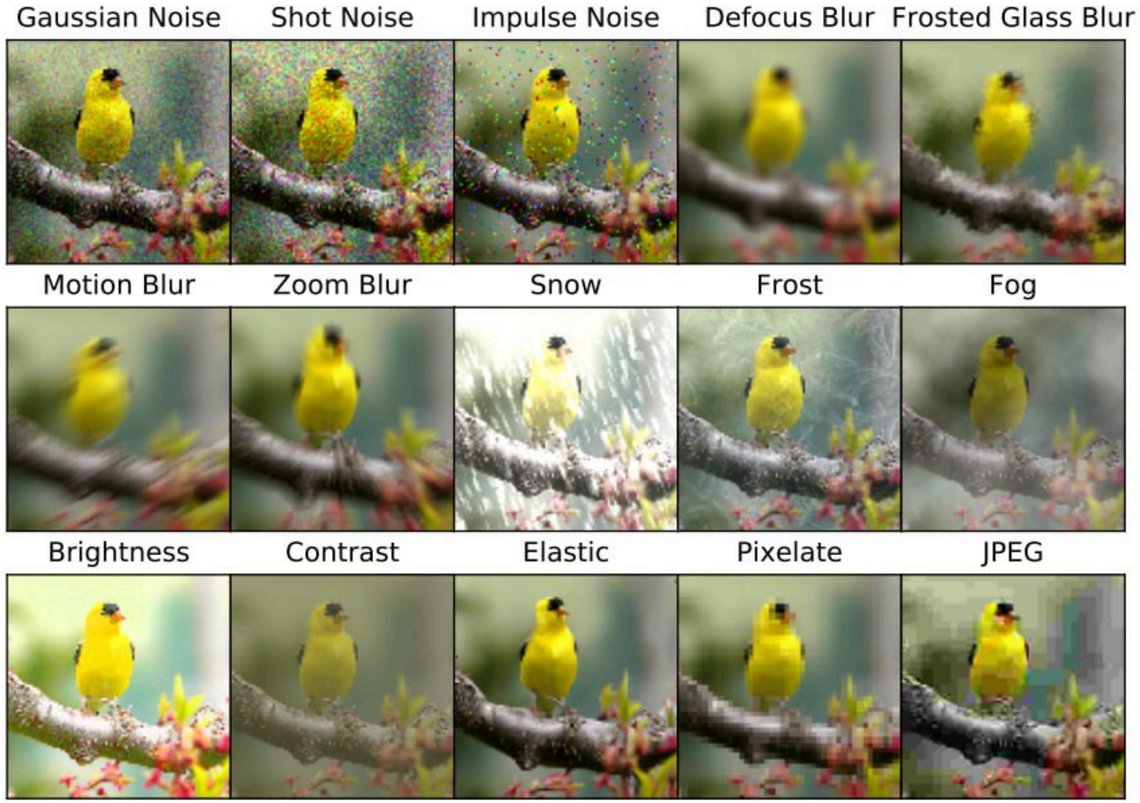


Figure 5) ImageNet-C

ObjectNet vs ImageNet-C datasets

EMPIRICAL EVIDENCE WITH DEEP MODELS

- Adversarial Training hurts Natural Distributional Robustness *only* when Spurious Correlations are broken.

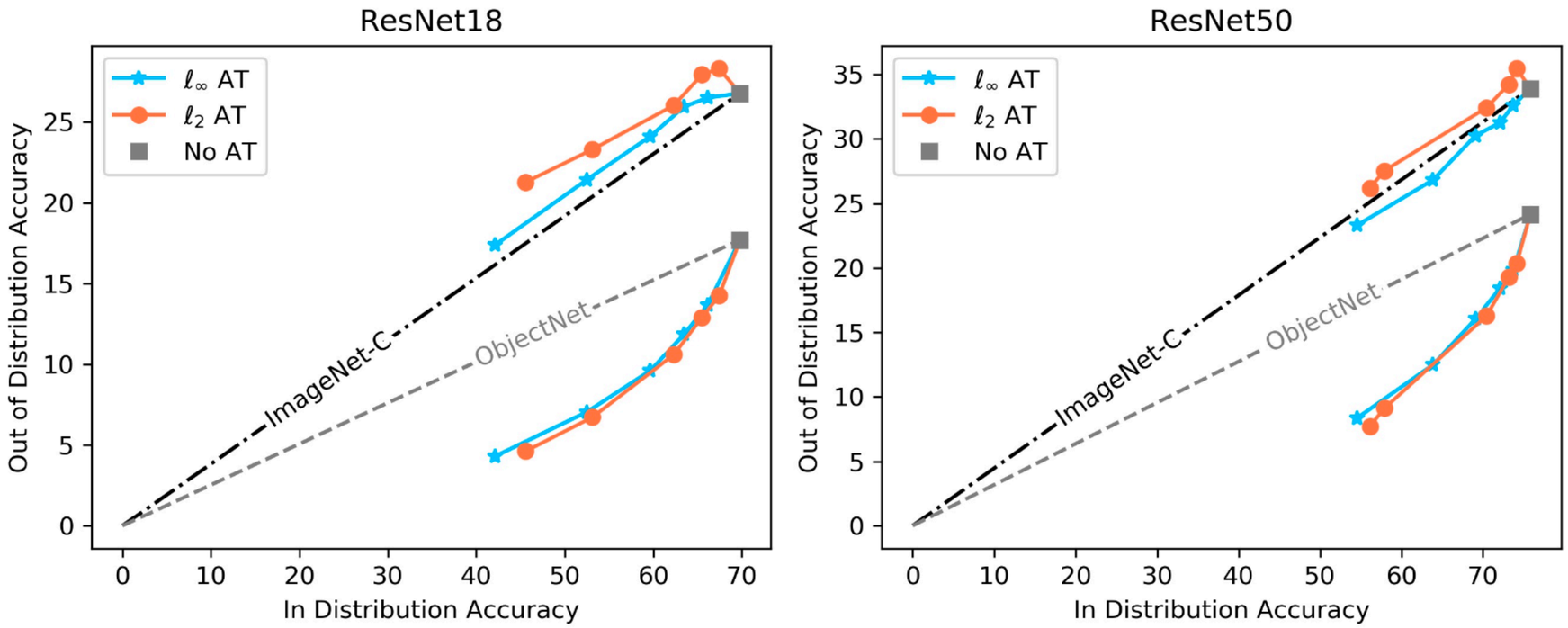


Figure 6) OOD accuracy vs in distribution accuracy

EMPIRICAL EVIDENCE WITH DEEP MODELS

- Reduced Core Sensitivity, and Difference in the Effect of l_2 *and* l_∞Adversarial Training

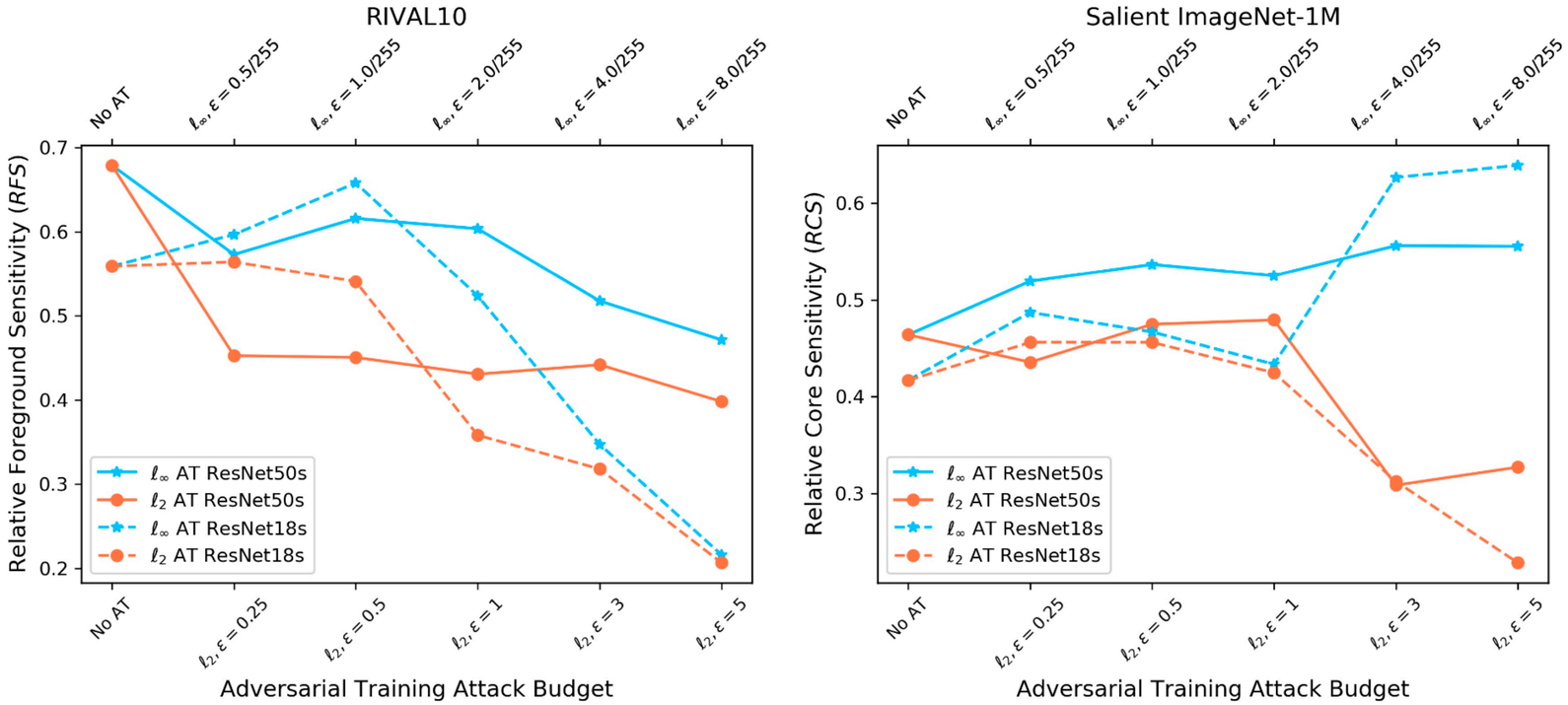


Figure 7) Relative Foreground and Core sensitivity vs ϵ

EMPIRICAL EVIDENCE WITH DEEP MODELS

- Adversarial Training Increases Background Reliance in Synthetic Datasets

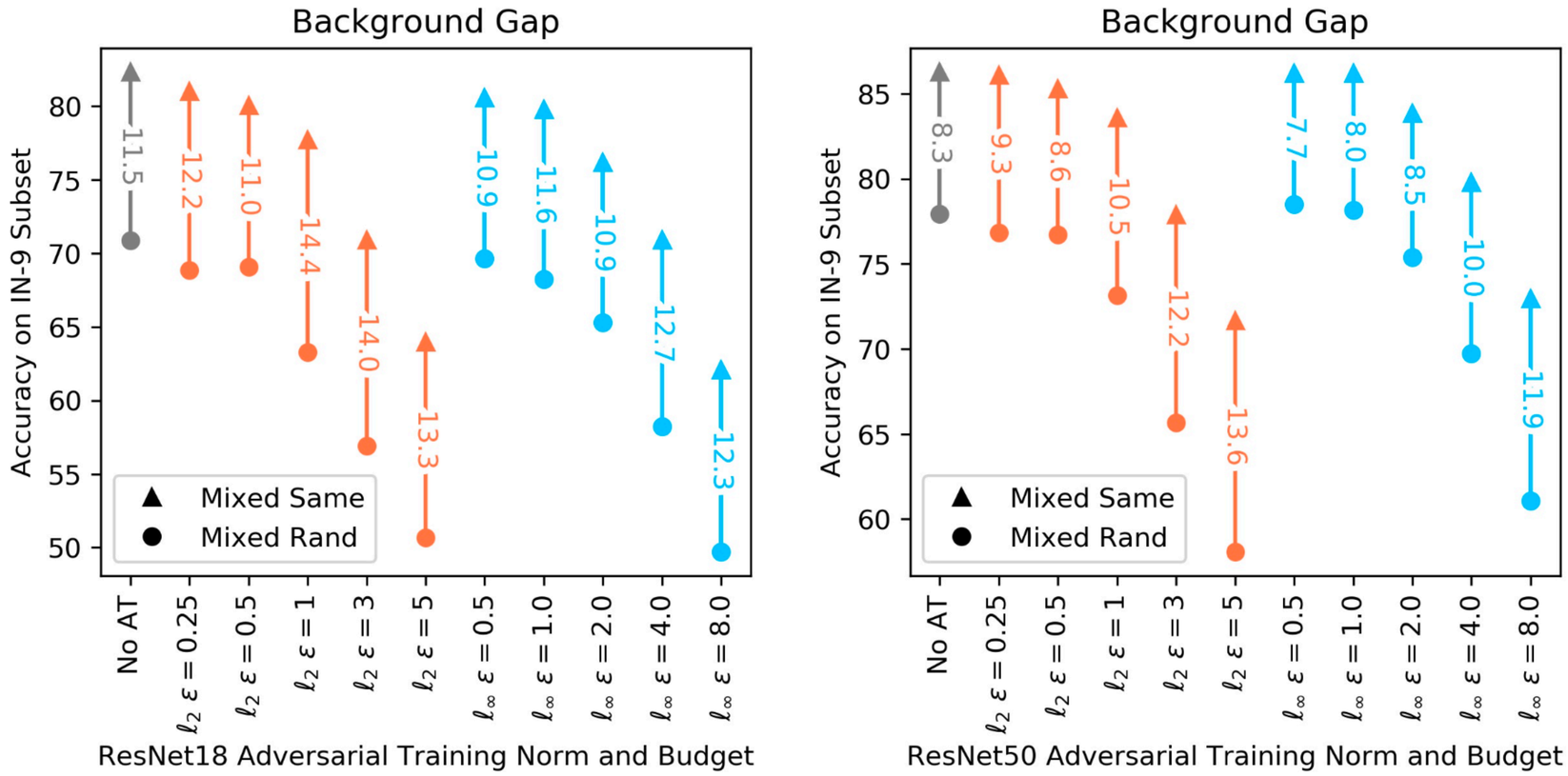


Figure 8) Adversarial Training Increases Background Reliance

EMPIRICAL EVIDENCE WITH DEEP MODELS

- Adversarial Training Increases Background Reliance in Synthetic Datasets

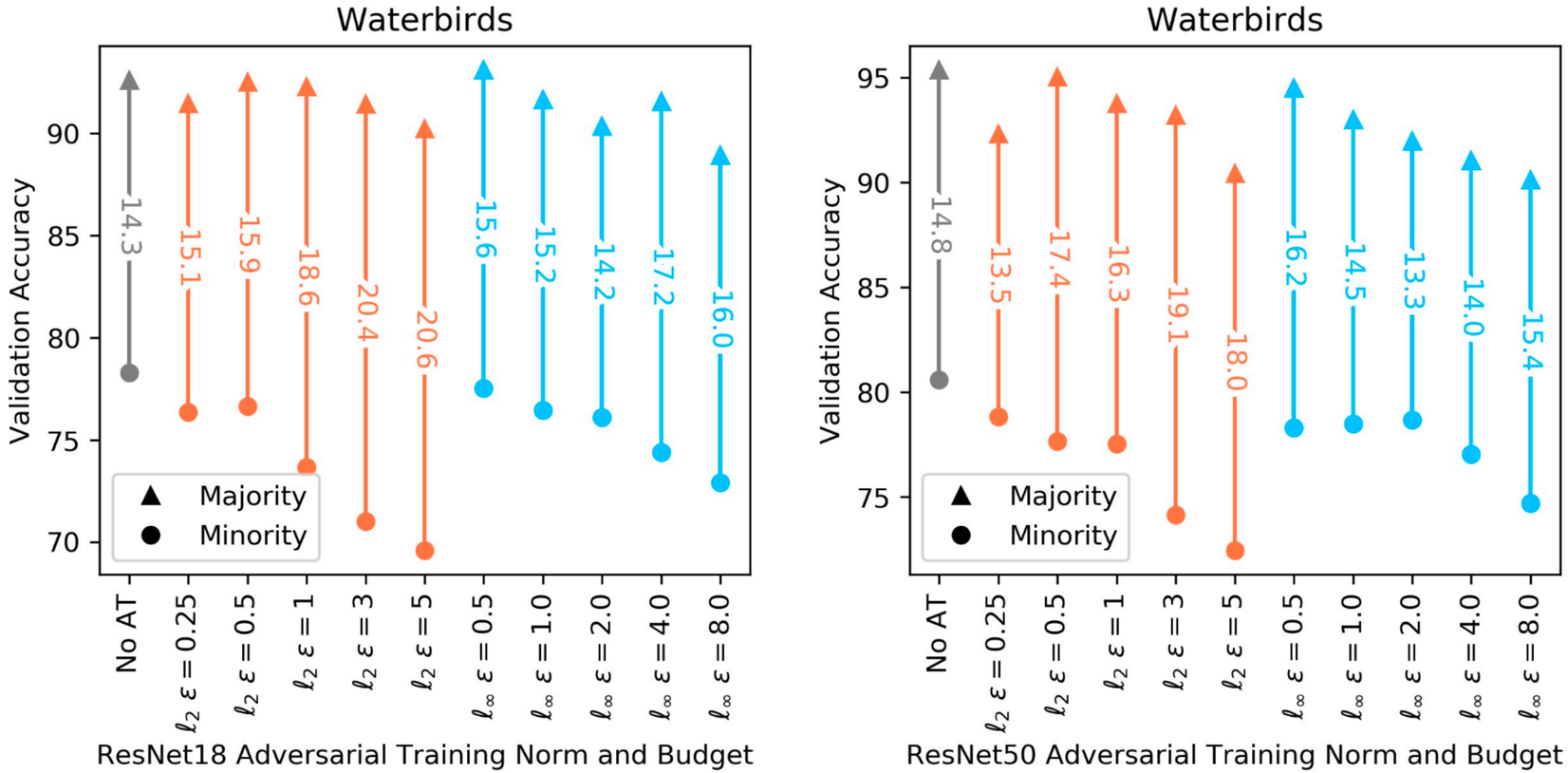


Figure 9) Adversarial Training Increases Background Reliance

EMPIRICAL EVIDENCE WITH DEEP MODELS

- Presence of Spurious Correlations Can *Improve* Adversarial Robustness
  - The majority group consists of red-shifted images from classes $0-4$ and green shifted images from classes $5-9$, while the minority group has reverse color-shifts

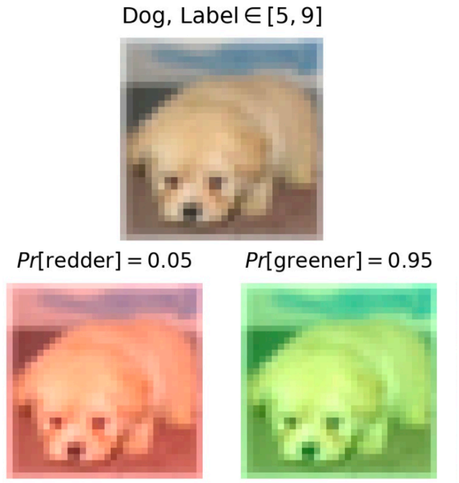


Figure 12: Color Shift, $\rho = 19:1$

Figure 10) CIFAR10 red/ green shifted

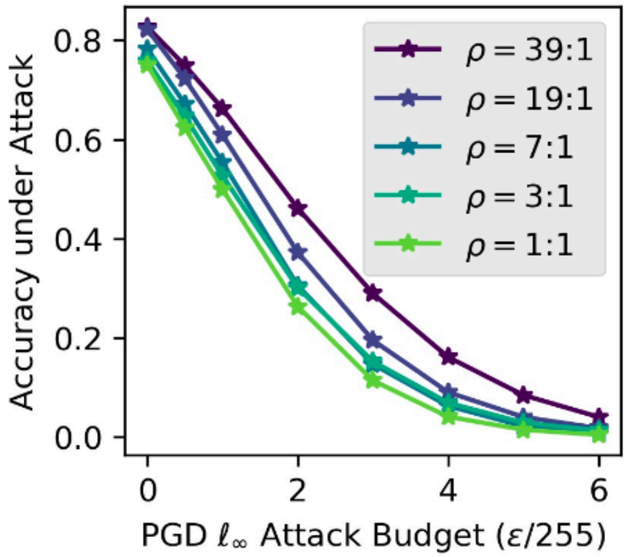


Figure 11) Reverse Effect