Robustness of Classifiers to Uniform ℓ_p and Gaussian Noise

Jean-Yves Franceschi[1], Alhussein Fawzi[2], Omar Fawzi[1]

[1]Ecole Normale Supérieure de Lyon – LIP, UMR 5668
[2]UCLA Vision Lab (now at DeepMind)

jean-yves.franceschi@ens-lyon.fr, hussein.fawzi@gmail.com, omar.fawzi@ens-lyon.fr

1. Adversarial and Random Perturbations

- Goal: **Relate robustnesses to adversarial and random noise.**
- Classifier $f : \mathbb{R}^d \to \mathbb{R}^L$, $g = \operatorname{argmax} f$.
- Adversarial perturbation w.r.t. norm ℓ_p:

$$\boldsymbol{r}_p^*(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{r}} \left\{ \|\boldsymbol{r}\|_p \text{ s.t. } g(\boldsymbol{x}+\boldsymbol{r}) \neq g(\boldsymbol{x}) \right\}.$$

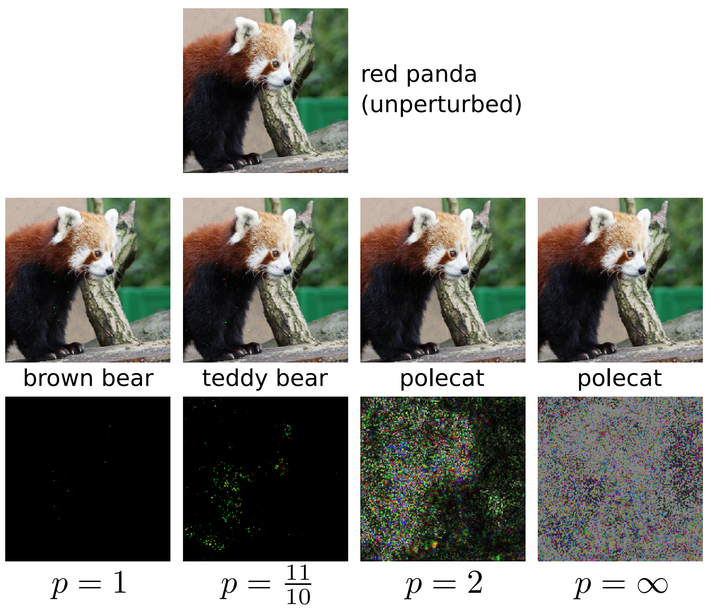

Figure 1: *Adversarial perturbations of an image for VGG-19 with different p-norms.*

- Robustness to random perturbation, for $\boldsymbol{v} \sim \nu$:

$$r_{\nu,\varepsilon}(\boldsymbol{x}) = \min_{\alpha} \left\{ |\alpha| \text{ s.t. } \mathbb{P}_{\boldsymbol{v}} \left\{ g(\boldsymbol{x}+\alpha\boldsymbol{v}) \neq g(\boldsymbol{x}) \right\} \geq \varepsilon \right\}.$$

- Goal: **Derive lower and upper bounds as well as an estimate on** $\frac{r_{\nu,\epsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p}$.

2. Bounds on Linear Classifiers

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + \mathrm{b}$$

2.1 Uniformly Distributed Noise in the ℓ_p Ball

Theorem 1. *Let $p, p' \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{p'} = 1$. Then, for ε small enough:*

$$\zeta_1(\varepsilon) d^{1/p} \frac{\|\boldsymbol{w}\|_{p'}}{\|\boldsymbol{w}\|_2} \leq \frac{r_{p,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_p^*(\boldsymbol{x})\|_p} \leq \zeta_2(\varepsilon) d^{1/p} \frac{\|\boldsymbol{w}\|_{p'}}{\|\boldsymbol{w}\|_2}.$$

- Depends on the choice of orthonormal basis if $p \neq 2$.
- For a typical $\boldsymbol{w}$, each bound if of the form:

$$C(\varepsilon, p)\sqrt{d}.$$

Proof sketch of simple special case. *Lower bound for $p = \infty$:*

- $\boldsymbol{r}_p^*(\boldsymbol{x}) = \frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|_1}$ *and* $\mathbb{P}_{\boldsymbol{v}} \left\{ g(\boldsymbol{x}+\alpha\boldsymbol{v}) \neq g(\boldsymbol{x}) \right\} = \mathbb{P}_{\boldsymbol{v}} \left\{ |\alpha| \sum_{i=1}^d w_i v_i \geq \|\boldsymbol{w}\|_1 |f(\boldsymbol{x})| \right\}$.
- $(v_i)_i$ *are i.i.d. and uniform over $[-1, 1]$: can apply Hoeffding and derive a lower bound.*

Ideas for the general case.

- *Lower bound: use Markov's inequality and estimates of $\mathbb{E}_{\boldsymbol{v}} \left[\left(\boldsymbol{w}^T\boldsymbol{v} \right)^k \right]$.*
- *Upper bound: use Paley-Zygmund's inequality and the previous estimates.*

2.2 Gaussian Noise

Theorem 2. *Let Σ be a $d \times d$ positive semidefinite matrix. Then, for ε small enough:*

$$\zeta_1'(\varepsilon) \frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2} \leq \frac{r_{\Sigma,\varepsilon}(\boldsymbol{x})}{\|\boldsymbol{r}_2^*(\boldsymbol{x})\|_2} \leq \zeta_2'(\varepsilon) \frac{\|\boldsymbol{w}\|_2}{\|\sqrt{\Sigma}\boldsymbol{w}\|_2}.$$

- For a typical $\boldsymbol{w}$, each bound if of the form:

$$C'(\varepsilon, p)\sqrt{d}.$$

- Σ may depend on $\boldsymbol{x}$.

3. Extension to Locally Approximately Flat (LAF) Classifiers

- A classifier is LAF at some point $\boldsymbol{x}$ if the decision boundary can be approached by a plane in a given ball centered at $\boldsymbol{x}$.
- In this setting, the results on linear classifiers are still valid, up to some constants in the bounds.
- The normal vector $\boldsymbol{w}$ may be naturally replaced by the gradient of f at at the closest point $\boldsymbol{x}^*$ on the decision boundary.
- Experiments indicate that the LAF assumption is reasonable for state-of-the-art deep neural networks.

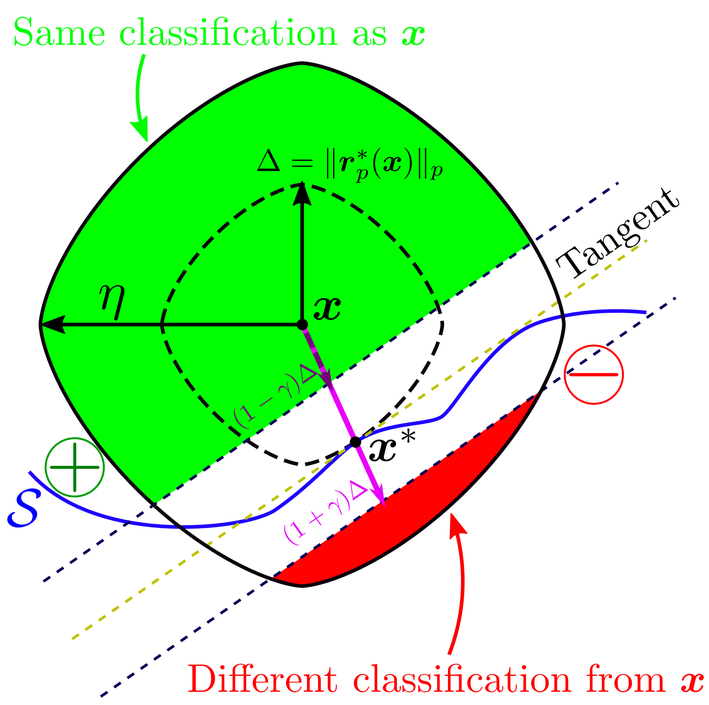

Figure 2: *Illustration of the LAF model.*

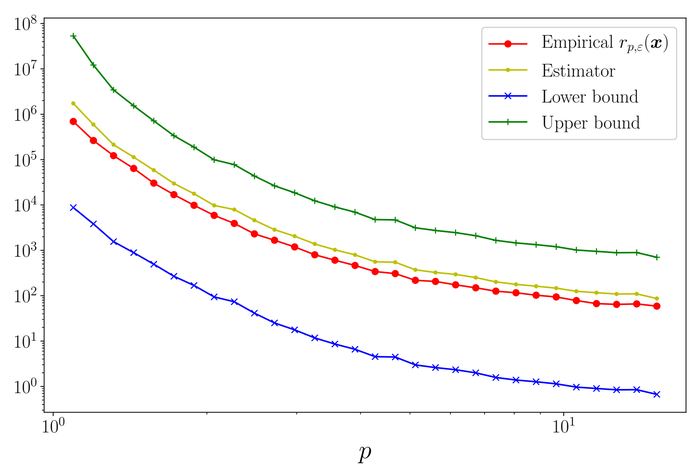

Figure 3: *Experiments on a deep neural network (VGG-19, ImageNet dataset).*

4. Applications: Robustness to Quantization

- Image quantization (discretization of the range of pixel values): $\boldsymbol{x} \mapsto Q(\boldsymbol{x})$.
- Can assume $Q(\boldsymbol{x}) \sim \mathcal{U}\left(\mathcal{B}_\infty\left(\boldsymbol{x}, \frac{\Delta}{2}\right)\right)$.
- Δ: quantization step size.
- $L_q = \frac{255}{\Delta}$: number of quantization levels corresponding to Δ.
- Our results allow us to estimate the level of quantization needed so that a quantized image is still classified correctly.

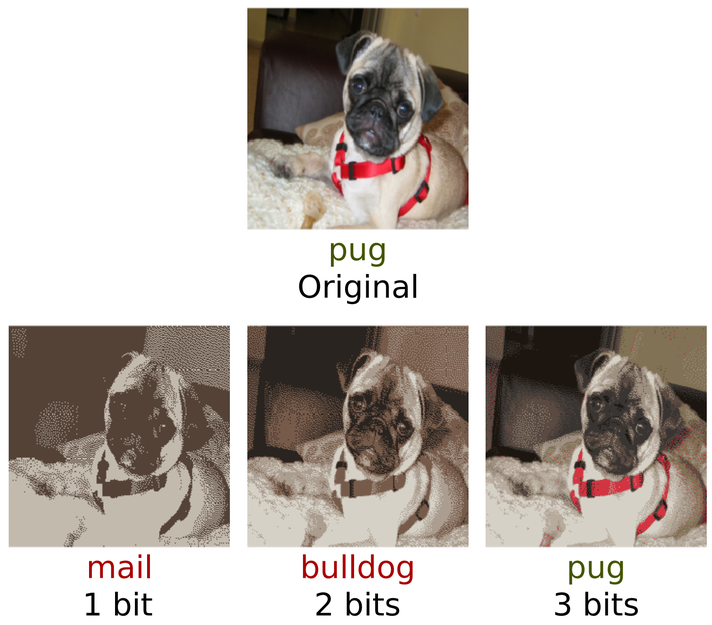

Figure 4: *Illustration of the effects of quantization.*

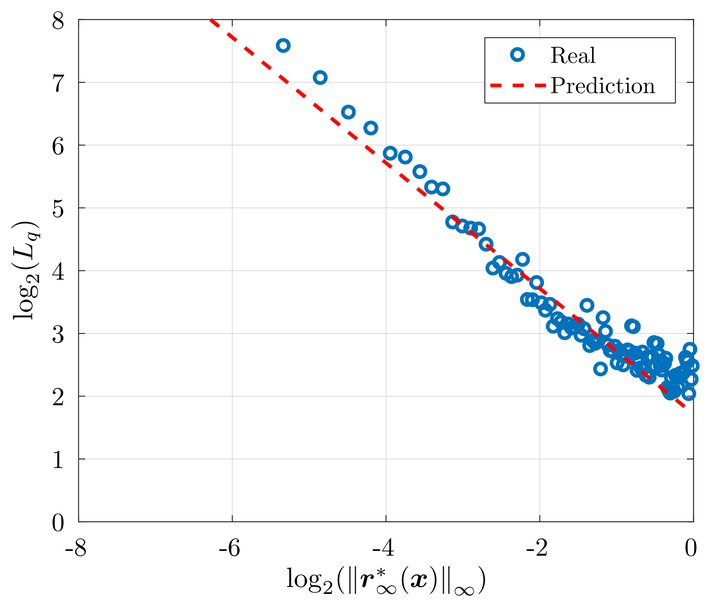

Figure 5: *Experiments on quantization (VGG-19, ImageNet dataset).*

References

Barthe, F., Guédon, O., Mendelson, S., and Naor, A. (2005). A probabilistic approach to the geometry of the ℓ_p^n-ball. *The Annals of Probability*, 33(2).

Fawzi, A., Moosavi-Dezfooli, S., and Frossard, P. (2016). Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems (NIPS) 29*.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

21st International Conference on Artificial Intelligence and Statistics – AISTATS 2018, Playa Blanca, Spain