Empirically Measuring Concentration: Fundamental Limits on Intrinsic Robustness
Saeed Mahloujifar*  Xiao Zhang*  Mohammad Mahmoody  David Evans

Empirically Measuring Concentration:
Fundamental Limits on Intrinsic Robustness

Main Question
What is the minimum possible Adversarial Risk, given that Risk is at least \( \alpha \)?

Concentration of measure of nice distributions gives lower bound on adversarial risk:

> Spheres under \( \ell_2 \) (Gilmer et al., 2018)
> Gaussian under \( \ell_2 \) (Fawzi et al., 2018)
> Any product distribution under \( \ell_0 \) (Mahloujifar et al., 2018)

Can we estimate concentration of measure for real world distributions (e.g. MNIST)?

Theoretical Results for \( \ell_\infty \)
Let \( (X, \mu) \) be the underlying probability space and \( f' \) be the ground-truth classifier.

- **Risk:**
  \[ \text{Risk}(f, f') = \mathbb{P}_{x \sim \mu} \left[ f(x) \neq f'(x) \right]. \]

- **Adversarial Risk w.r.t. perturbations:**
  \[ \text{AdvRisk}(f, f') = \mathbb{P}_{x \sim \mu} \left[ \exists x' : \|x - x'\|_\infty < \epsilon \implies f(x) \neq f'(x') \right]. \]

**Illustration of the proposed method for finding robust error region under \( \ell_\infty \):**
- Sort all the training data using \( \ell_\infty \) distance to the \( k \)-th nearest neighbors.
- Perform kmeans clustering on the top-\( q \) densest images.
- Obtain \( T \) hyperrectangular image clusters and expand each of them by \( \epsilon \) in \( \ell_\infty \).
- Treat the complement of union of these hyperrectangles as our error region.

**Table:** Summary of the main results using our empirical method under \( \ell_\infty \) perturbations

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( \alpha )</th>
<th>( \epsilon )</th>
<th>( T )</th>
<th>( q )</th>
<th>Risk</th>
<th>Adversarial Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>0.01</td>
<td>0.2</td>
<td>5</td>
<td>0.660</td>
<td>1.23%</td>
<td>5.29%</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.2</td>
<td>10</td>
<td>0.660</td>
<td>1.15%</td>
<td>4.94%</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.2</td>
<td>10</td>
<td>0.660</td>
<td>1.21%</td>
<td>4.56%</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>0.05</td>
<td>4/255</td>
<td>0.665</td>
<td>0.080</td>
<td>5.72%</td>
<td>8.13%</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>255</td>
<td>0.665</td>
<td>0.080</td>
<td>5.94%</td>
<td>8.29%</td>
</tr>
<tr>
<td>SVHN</td>
<td>0.05</td>
<td>0.2</td>
<td>10</td>
<td>0.734</td>
<td>2.78%</td>
<td>12.86%</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.2</td>
<td>10</td>
<td>0.734</td>
<td>2.86%</td>
<td>13.14%</td>
</tr>
</tbody>
</table>

**Implications of our experiments**
- Provide examples of rather robust error regions for real image datasets.
- Suggest the concentration of measure phenomenon is not the sole reason behind vulnerability of the existing classifiers to adversarial examples.
- Suggest the impossibility results, such as Gilmer et al. (2018) and Mahloujifar et al. (2018), should not make the community hopeless in finding more robust image classifiers.

Experimental for \( \ell_\infty \)
- **Tuning for the best parameters**
- **Main experimental results**

**Challenge 1:** We do not know the PDF of the distribution.
**Our solution:** Replace the actual distribution \( \mu \) with empirical distribution \( \hat{\mu} \) based on a set of i.i.d. samples \( S \):

\[ \hat{\mu}(A) = \frac{1}{|S|} \sum_{x \in S} 1_A(x). \]

**Challenge 2:** We cannot search through all possible subsets.
**Our solution:** Limit the search space to a carefully chosen collection of subsets \( \mathcal{G} \).

**Remaining task:** Solve the following optimization problem:

\[ \min_{E \in \mathcal{G}} \hat{\mu}(E) \quad \text{s.t.} \quad \hat{\mu}(E) \geq \alpha. \]

**Fundamental Limits on Intrinsic Robustness**

- **Empirical Framework to Measure Concentration**
  - **Challenge 1:** We do not know the PDF of the distribution.
  - **Our solution:** Replace the actual distribution \( \mu \) with empirical distribution \( \hat{\mu} \) based on a set of i.i.d. samples \( S \).
  - **Challenge 2:** We cannot search through all possible subsets.
  - **Our solution:** Limit the search space to a carefully chosen collection of subsets \( \mathcal{G} \).
  - **Remaining task:** Solve the following optimization problem:

**Theoretical Results for \( \ell_\infty \)**
Let \( \mathcal{G}_T \) be the collection of subsets specified by complement of union of \( T \) hyperrectangles.

\[
\mathcal{E} = X - \left( \bigcup_{i=1}^{T} \mathcal{R}_i \right)
\]

Let \( \hat{\mu} \) be the empirical distribution based on a i.i.d. dataset of size \( T \).

- For any product distribution under \( \mu \), define

\[ c = \min_{E \in \mathcal{G}} \hat{\mu}(E) \quad \text{s.t.} \quad \hat{\mu}(E) \geq \alpha. \]

Also define

\[ c_T = \min_{E \in \mathcal{G}_T} \hat{\mu}(E) \quad \text{s.t.} \quad \hat{\mu}(E) \geq \alpha. \]

**Main Theorem:** With probability 1 over the randomness of training data we have

\[ \lim_{T \to \infty} c_T = c. \]