Incorporating Label Uncertainty in Intrinsic Robustness Measures

UNIVERSITY of VIRGINIA

Xiao Zhang David Evans

Standard Concentration is not Sufficient

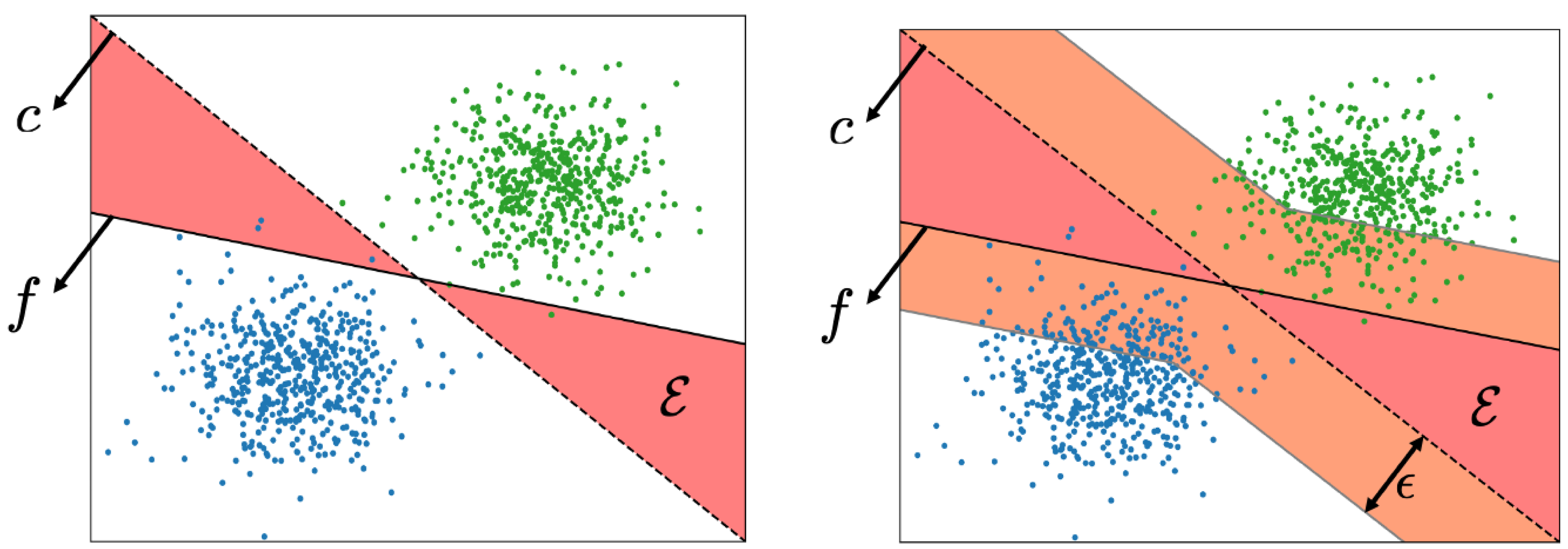
Given metric probability space $(\mathcal{X}, \mu, \Delta)$, concept function $c(\cdot)$, and parameters (α, ϵ):

The problem of **concentration of measure** can be cast as

$$\underset{\mathcal{E}\in\text{pow}(\mathcal{X})}{\text{minimize}}\ \mu(\mathcal{E}_\epsilon) \quad \text{subject to} \quad \mu(\mathcal{E}) \geq \alpha.$$

Mahloujifar et al. (2019) showed it is equivalent to the **intrinsic robustness estimation** problem:

$$\underset{f}{\text{minimize}}\ \text{AdvRisk}_\epsilon(f, c) \quad \text{subject to} \quad \text{Risk}(f, c) \geq \alpha.$$



(a) Risk(f, c) (b) AdvRisk$_\epsilon(f, c)$

In this work, we argue that the standard concentration of measure is *not* sufficient to capture a realistic intrinsic robustness limit for robust classification problem: the labels matter.

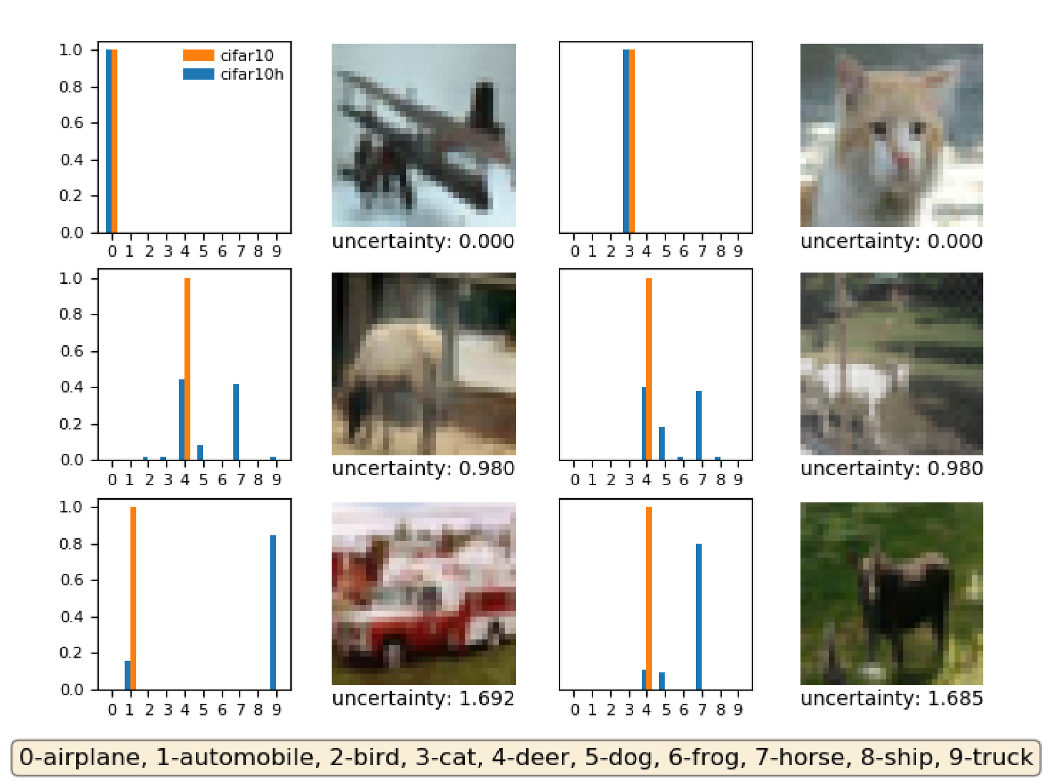
Introducing Label Uncertainty

Define **label uncertainty** with respect to an input region $\mathcal{E}$ as:

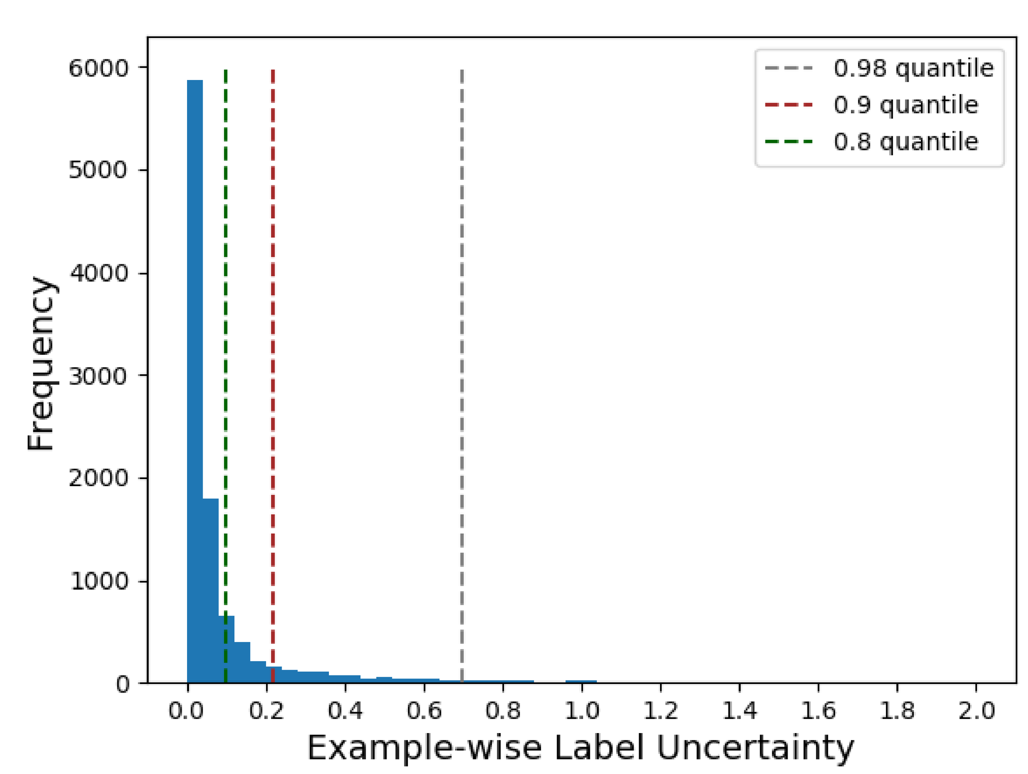
$$\mathbf{LU}(\mathcal{E}; \mu, c, \eta) = \frac{1}{\mu(\mathcal{E})} \int_{\mathcal{E}} \left\{1 - [\eta(\boldsymbol{x})]_{c(\boldsymbol{x})} + \max_{y' \neq c(\boldsymbol{x})} [\eta(\boldsymbol{x})]_{y'}\right\} d\mu,$$

where $\eta(\cdot)$ is the label distribution function, and $[\eta(\boldsymbol{x})]_y$ represents the description degree of y to $\boldsymbol{x}$.

Visualizing CIFAR-10 label uncertainty using the CIFAR-10H dataset (Peterson et al., 2019)



(c) Illustration of Human Label Uncertainty on CIFAR-10



(d) Distribution of CIFAR-10 Label Uncertainty

Concentration with Label Uncertainty Constraint

Standard concentration of measure:

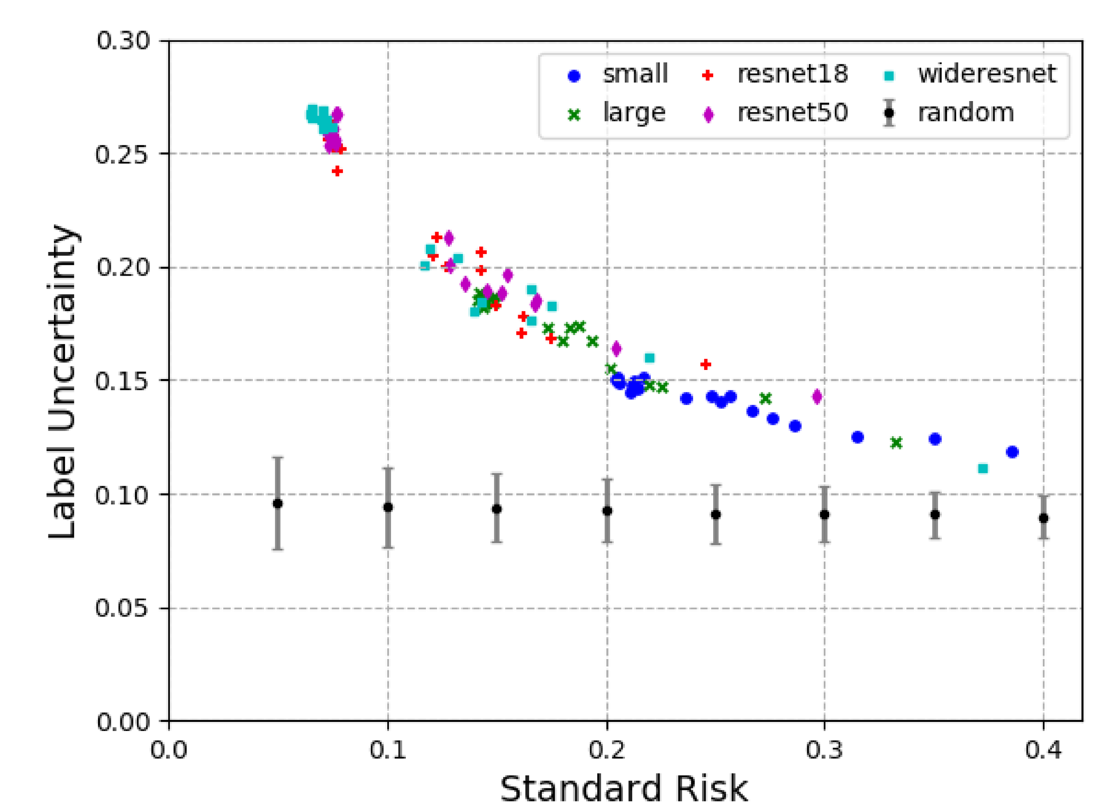
$$\min_{\mathcal{E}\in\text{pow}(\mathcal{X})} \mu(\mathcal{E}_\epsilon) \text{ s.t. } \mu(\mathcal{E}) \geq \alpha$$

Incorporate the label uncertainty information

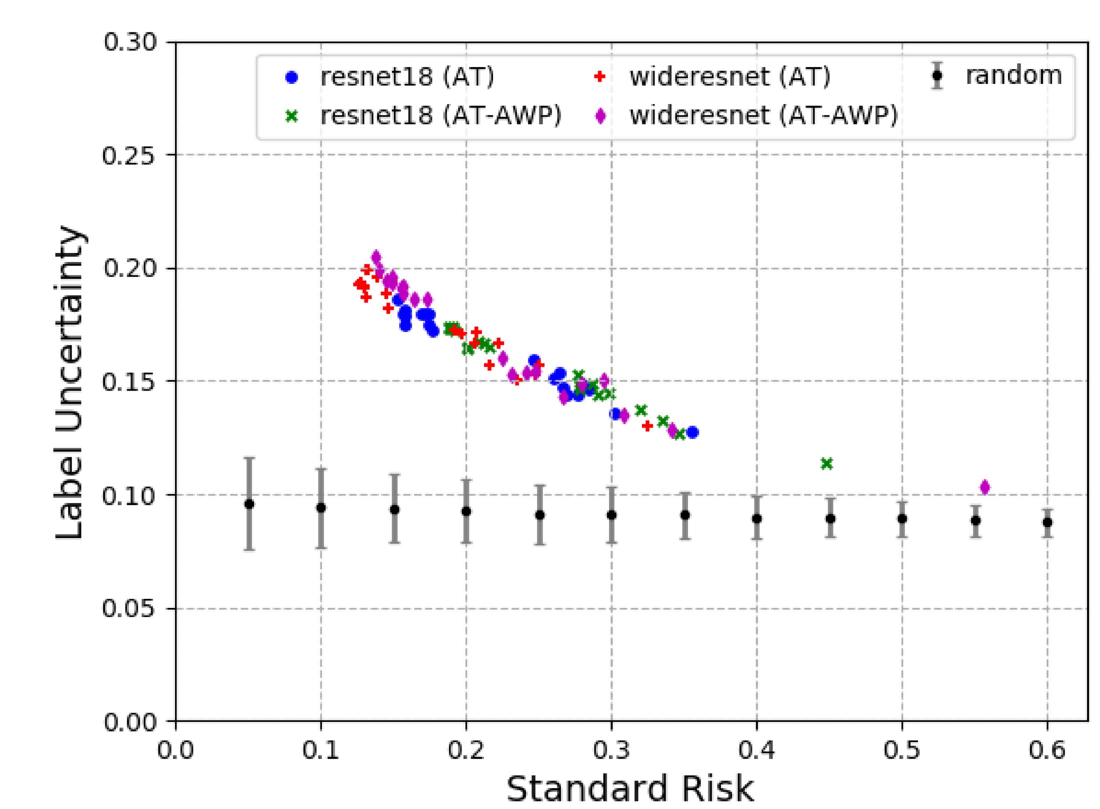
Concentration of measure with label uncertainty constraint:

$$\min_{\mathcal{E}\in\text{pow}(\mathcal{X})} \mu(\mathcal{E}_\epsilon) \text{ s.t. } \mu(\mathcal{E}) \geq \alpha \text{ and } \text{LU}(\mathcal{E}; \mu, c, \eta) \geq \gamma$$

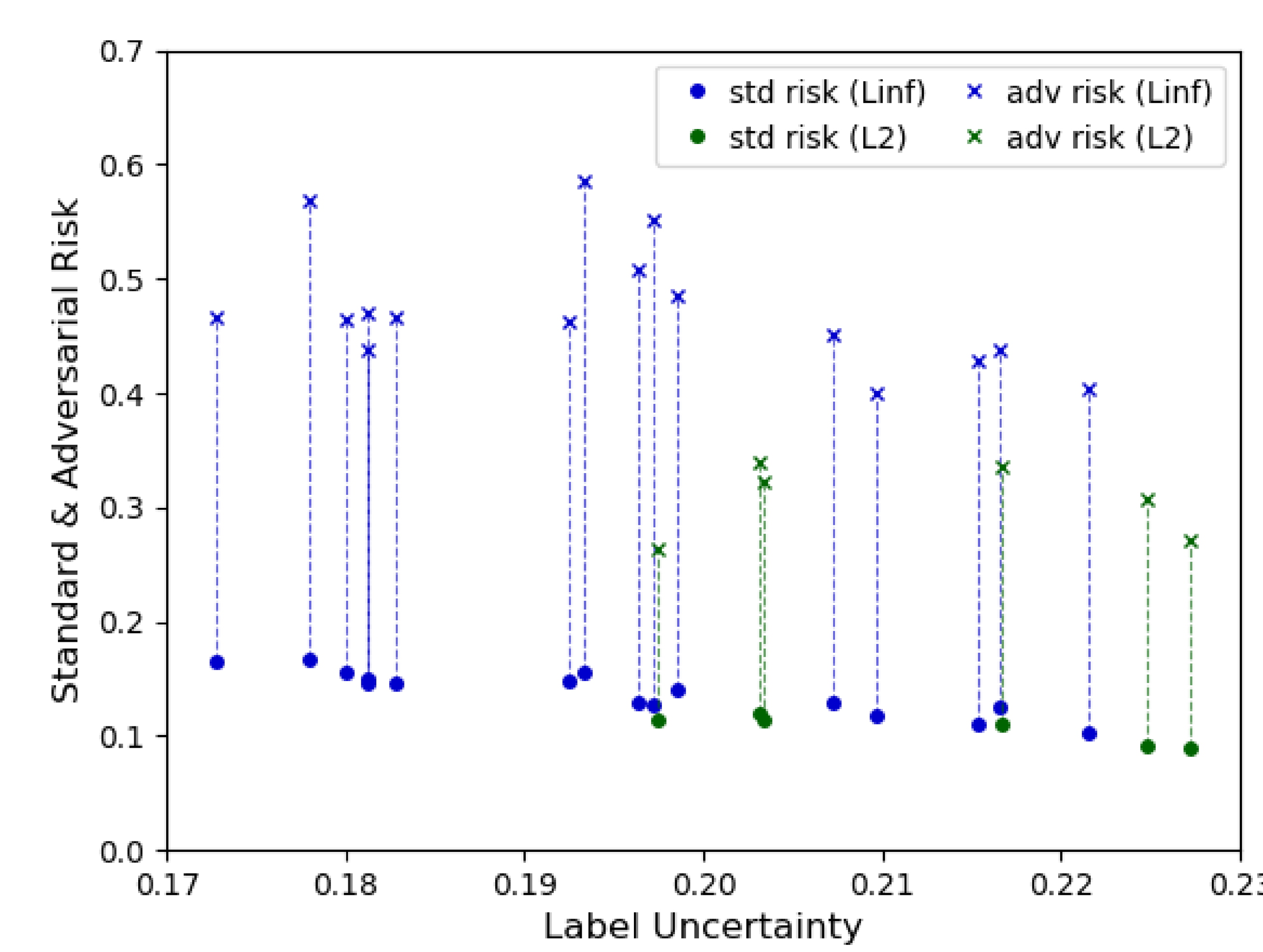
Experiments on CIFAR-10



(e) Standard-trained Models



(f) Adversarially-trained Models



(g) RobustBench Models (Croce et al., 2020)

Regardless of model architecture or training methodology, the error regions of state-of-the-art CIFAR-10 classification models have much higher label uncertainty, compared with randomly selected subsets.