A law of adversarial risk, interpolation, and label noise*

Amartya Sanyal and Daniel Paleka

In supervised learning, it has been shown that label noise in the data can be interpolated without penalties on test accuracy [1, 2]. However, this raises the question whether such interpolation adversely affects the robustness of the trained model.

Our main contribution is the establishment of the first formal relationship between label noise and adversarial risk, applicable universally across data distributions. Consider μ as a distribution on $\mathbb{R}^d$, f^* as a measurable ground truth classifier mapping $\mathcal{C} \to \{0, 1\}$, and $\mathcal{R}_{\mathrm{Adv},\rho}(f, \mu)$ representing the adversarial risk [5] for classifier f.

Theorem 1. *Given a set $\mathcal{C} \subset \mathbb{R}^d$ where $\mu(\mathcal{C}) > 0$, and its covering number N, for a dataset size m sampled independently from μ with each example mislabelled with probability $\eta > 0$, when the sample count meets the condition $m \geq \frac{8N}{\mu(\mathcal{C})\eta} \log \frac{2N}{\delta}$, with likelihood $1 - \delta$, the adversarial risk is given by:*

$$\mathcal{R}_{\mathrm{Adv},\rho}(f, \mu) \geq \frac{1}{4}\mu(\mathcal{C})$$

for any classifier f interpolating the dataset of size m.

It is interesting to note that the above result holds for all dataset sizes m (for a corresponding $\mathcal{C}$) thus providing an adaptive lower bound on the adversarial risk. Our further investigations reveal that this relationship is nearly tight, especially without additional biases on the learning algorithm.

Additionally, our exploration dives deep into the nuances of label noise types, data distribution characteristics, and the learning algorithm's inductive biases. In particular, we discuss non-uniform label noise distributions; and prove a new theorem showing uniform label noise induces nearly as large an adversarial risk as the worst poisoning with the same noise rate. This result is surprising as poisoning adversaries are often shown to have significantly more adverse impact than oblivious adversaries. On the other hand, we provide theoretical and empirical evidence that uniform label noise is more harmful than typical real-world label noise. This is because real world label noise accumulates on the tail of the data which provably hurts adversarial error less than uniform label noise.

Finally, we show how inductive biases amplify the effect of label noise. Through synthetic experiments, we postulate that neural network architectures have inherent biases that drastically minimize the required dataset size for phenomena like Theorem 1 to hold. This underscores the need for future work in this direction.

*This abstract is primarily based on the work Paleka and Sanyal [3] but also includes elements of Sanyal et al. [4].

References

[1] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

[2] Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*, 2021.

[3] Daniel Paleka and Amartya Sanyal. A law of adversarial risk, interpolation, and label noise. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=0_TxFpAsEI`.

[4] Amartya Sanyal, Puneet K. Dokania, Varun Kanade, and Philip Torr. How benign is benign overfitting? In *International Conference on Learning Representations (ICLR)*, 2021.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.