# Adversarial Attack Generation Empowered by Min-Max Optimization

**Jingkang Wang**[1,2*] **Tianyun Zhang**[3*] **Sijia Liu**[4,5]
**Pin-Yu Chen**[5] **Jiacen Xu**[6] **Makan Fardad**[7] **Bo Li**[8]
University of Toronto[1], Vector Institute[2], Cleveland State University[3]
Michigan State University[4], MIT-IBM Watson AI Lab, IBM Research[5]
University of California, Irvine[6], Syracuse University[7]
University of Illinois at Urbana-Champaign[8]

## Abstract

The worst-case training principle that minimizes the maximal adversarial loss, also known as adversarial training (AT), has shown to be a state-of-the-art approach for enhancing adversarial robustness. Nevertheless, min-max optimization beyond the purpose of AT has not been rigorously explored in the adversarial context. In this paper, we show how a general framework of min-max optimization over multiple domains can be leveraged to advance the design of different types of adversarial attacks. In particular, given a set of risk sources, minimizing the worst-case attack loss can be reformulated as a min-max problem by introducing domain weights that are maximized over the probability simplex of the domain set. We showcase this unified framework in three attack generation problems – attacking model ensembles, devising universal perturbation under multiple inputs, and crafting attacks resilient to data transformations. Extensive experiments demonstrate that our approach leads to substantial attack improvement over the existing heuristic strategies as well as robustness improvement over state-of-the-art defense methods trained to be robust against multiple perturbation types. Furthermore, we find that the self-adjusted domain weights learned from our min-max framework can provide a holistic tool to explain the difficulty level of attack across domains. Code is available at https://github.com/wangjksjtu/minmax-adv.

## 1 Introduction

Training a machine learning model that is capable of assuring its worst-case performance against possible adversaries given a specified threat model is a fundamental and challenging problem, especially for deep neural networks (DNNs) [64, 22, 13, 69, 70]. A common practice to train an adversarially robust model is based on a specific form of min-max training, known as *adversarial training* (AT) [22, 40], where the minimization step learns model weights under the adversarial loss constructed at the maximization step in an alternative training fashion. In practice, AT has achieved the state-of-the-art defense performance against $\ell_p$-norm-ball input perturbations [3].

Although the min-max principle is widely used in AT and its variants [40, 59, 76, 65], few work has studied its power in attack generation. Thus, we ask: *Beyond AT, can other types of min-max formulation and optimization techniques advance the research in adversarial attack generation?* In this paper, we give an affirmative answer corroborated by the substantial performance gain and the ability of self-learned risk interpretation using our proposed min-max framework on several tasks for adversarial attack.

[*] Equal contributions.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.