# Poisoning Attack Defense and Gradient-Shaping Implementation in Python

Selma Emekci* (selmahaticeemekci@gmail.com), Eilan Tang* (eilantang@gmail.com), Shreya Kochar (shreya.kochar@columbia.edu), Aneesha Sreerama (sreerama.a@northeastern.edu), Andrew Wang (andrew.wang.2@stonybrook.edu), Anirudh Sreerama (anirudh.s@berkeley.edu)

**Abstract**

Machine learning models are susceptible to adversarial attacks that manipulate data to undermine model performance. While there is existing research on adversarial attacks against machine learning models, the vulnerability of linear regression models in particular and the development of effective defense strategies remain understudied. In this paper, we propose a defense approach that uses the FRIENDS mechanism. This defense methodology combines accuracy-friendly perturbation and random varying noise, utilizing a gradient-based approach to minimize prediction errors and simulating poisoning attacks by appending malicious data and fine-tuning the model. Our experimental results show how effective this defense method is, which was achieved by adjusting noise and clip threshold values. Our program offers an open-source and accessible solution for defending linear regression models against poisoning attacks, contributing to the security of AI systems.

*Keywords:* Adversarial Machine Learning, Python, Poisoning Attacks, Machine Learning

## 1. Introduction

Machine learning models, including linear regression, are vulnerable to various adversarial attacks. Linear regression is a widely used machine learning technique, however, it is susceptible to adversarial attacks. Specifically, poisoning attacks involve manipulating the data to add malicious data that can mislead the model during the learning process. These attacks aim to undermine the model's performance and make it generalize poorly on unseen data. Furthermore, gradient masking is a technique used by attackers to make their attacks harder to detect by altering the gradients of the poisoned data points. Existing research has focused on adversarial attacks primarily in classification tasks, such as deep neural networks. However, understanding the impact of poisoning attacks on linear regression models and finding effective defense mechanisms needs to be studied further. We seek to explore and develop potential defense strategies to protect linear regression models from poisoning attacks, where attackers deliberately inject malicious data into the training set to manipulate a model's behavior. Adversarial attacks can poison models, leading to harmful outcomes and these attacks pose a significant threat to the integrity of models. Finding effective defense strategies is crucial to building trustworthy ML systems. First, we plan to demonstrate the vulnerability of linear regression models to poisoning attacks and the potential consequences of these attacks. Second, we propose a comprehensive defense approach that effectively mitigates the impact of poisoning attacks while also maintaining model performance. Prior work has proposed a defense mechanism called FRIENDS to combat poisoning attacks. It utilizes two components to break poisons: accuracy-friendly perturbation and random varying noise. This combination has made this defense mechanism highly effective, leading us to implement it into our algorithm.

## 2. Literature Review

In a paper that proposed a defense strategy named "gradient shaping" which aims to constrain gradient magnitudes, Hong et al. (2020). employed differentially private stochastic gradient descent (DP-SGD) as a representative tool. The research presented in this paper focuses on two specific scenarios, highlighting the impact of poisoning attacks on model training and re-training. These scenarios represent common instances of indiscriminate and targeted poisoning attacks, respectively, each showing distinct aspects of the poisoning phenomenon. Although the paper's introduction of ”gradient shaping" as a defense approach demonstrates innovation in countering poisoning attacks, we realize that it does not fully capture the complexity of real-world poisoning attacks, potentially limiting the generalizability of its findings. Further research could explore the integration of real-world factors, such as data quality and model complexity, to further evaluate defense mechanisms. We believe future research could also extend gradient shaping strategies to different machine learning models and tasks, enhancing our understanding of its applicability and limitations. A similar paper introduced a defense mechanism named FRIENDS that defends against various types of poisoning attacks on deep learning models. Liu et al. (2022) demonstrated that this defense mechanism has high effectiveness in breaking various invisible poisoning attacks with minimal impact on model performance. It utilizes two components to break poisons: accuracy-friendly perturbation and random varying noise. This combination has made this defense mechanism highly effective, leading us to implement it into our algorithm. We aim to demonstrate the effectiveness of FRIENDS as part of a