Preface

With the recent advances in machine learning theory and algorithms, the design of high-capacity and scalable models such as neural networks, abundant datasets, and sufficient computing resources, machine learning (ML), or more broadly, artificial intelligence (AI), has been transforming our industry and society at an unprecedented speed.

While we are anticipating positive impacts enabled by machine learning technology, we may often overlook potential negative effects, which may bring considerable ethical concerns and even setbacks due to law regulations and catastrophic failures, especially for mission-critical and high-stakes decision making tasks. Therefore, beyond accuracy, trustworthy machine learning is the last milestone for ML-based technology to achieve and thrive. Trustworthy machine learning encompasses a broad set of essential topics such as adversarial robustness, fairness, explainability, accountability, and ethics.

This book focuses on fulfilling the endeavor of evaluating, improving, and leveraging adversarial robustness of machine learning algorithms, models, and systems toward better and more trustworthy versions. Exploiting untrusted machine learning as vulnerabilities create unattended gateways for intended parties to manipulate machine predictions while evading human's attention to gain their own benefits. No matter what one's role is in ML, as a model developer, a stakeholder, or a user, we believe it is essential for everyone to understand adversarial robustness for machine learning, just like knowing the capabilities and limitations of your own vehicle before driving. For model developers, we advocate proactive in-house robustness testing of your own models and systems for error inspection and risk mitigation. For stakeholders, we advocate acknowledgment of possible weaknesses in products and services, as well as honest and thorough risk and threat assessment in a forward-thinking manner to prevent revenue/reputation loss and catastrophic damage to the society and environment. For users using machine learning byproducts, we advocate active understanding of their limitations for safe use and gaining awareness about possible misuses. These aspects related to adversarial robustness, along with the available techniques and tools, are elucidated in this book.

Generally speaking, adversarial robustness centers on the study of the *worst-case* performance in machine learning, in contrast to the standard machine learning practice, which focuses on the *average* performance, e.g.,

prediction accuracy on a test dataset. The notion of worst-case analysis is motivated by the necessity of ensuring robust and accurate predictions for machine learning against changes in the training environments and deployed scenarios. Specifically, such changes can be caused by natural occurrences (e.g., data drifts due to varying lighting conditions) or by malicious attempts (e.g., hackers aiming to compromise and gain control over the system/service based on machine learning). Consequently, instead of asking "How well can machine learning perform on this given dataset/task?", in adversarial robustness, we ask "How robust and accurate can machine learning be if the dataset or the model can undergo different quantifiable levels of changes?" This interventional process often involves introducing a virtual adversary in machine learning for robustness assessment and improvement, which is a key ingredient in adversarial machine learning.

This book aims to offer a holistic overview of adversarial robustness spanning the lifecycle of machine learning, ranging from data collection, model development, to system integration and deployment. The contents provide a comprehensive set of research techniques and practical tools for studying adversarial robustness for machine learning. This book covers the following four research thrusts in adversarial robustness: (i) *Attack* – Finding failure modes for machine learning; (ii) *Defense* – Strengthening and safeguarding machine learning; (iii) *Certification* – Developing provable robustness performance guarantees; and (iv) *Applications* – Inventing novel use cases based on the study of adversarial robustness.

We summarize the contents of each part in this book as follows. In Part 1, we introduce preliminaries for this book, connect adversarial robustness to adversarial machine learning, and provide intriguing findings to motivate adversarial robustness. In Part 2, we introduce different types of adversarial attacks with varying assumptions in attackers' capabilities in the lifecycle of machine learning, knowledge of the target machine learning system, realizations in digital and physical spaces, and data modalities. In Part 3, we introduce certification techniques for quantifying the level of provable robustness for neural networks. In Part 4, we introduce defenses for improving the robustness of machine learning against adversarial attacks. Finally, in Part 5, we present several novel applications inspired from the study of adversarial robustness for machine learning.

Pin-Yu Chen and Cho-Jui Hsieh