A Complexity Theoretic Approach to Adversarial Machine Learning

Saeed Mahloujifar

December 2019

Abstract

With the ever increasing applications of machine learning algorithms many new challenges, beyond accuracy, have been raised. Among them, and one of the most important ones, is robustness against adversarial attacks. The persistent impact of these attacks on the security of otherwise successful machine learning algorithms begs a fundamental investigation. My research aims at building a foundation to systematically investigate robustness of machine learning algorithms in the presence of different adversaries.

Two special cases of security threats, which have been the focus of many studies in the recent years, are *evasion attacks* and *poisoning attacks*. Evasion attacks occur during the inference phase and refer to adversaries who perturb the input to a classifier to get their desired output. Poisoning attacks occur in the training phase where an adversary perturbs the training data, with the goal of leading the learning algorithm to choose an *insecure*[1] hypothesis. Following, I will first explain my work on evasion and then poisoning attacks. I will also discuss the implications of my work to randomness extractors and coin tossing protocols. I will conclude by stating my future research plans.

Inference-time Attacks

As mentioned above, evasion attacks are one of the important attacks that happen during inference phase. The usual objective of an evasion attack is to degrade the overall performance of the model by perturbing the test instances. in the literature, there exist various definitions of robustness of classifiers in the presence of evasion attacks. Although all these definitions seem to capture the same phenomenon, in [DMM18], we showed that they sometimes lead to significantly different results. However, Adding an stability assumption over the ground truth, all of these definitions converge to a single definition where the goal of adversary is to push instances to the *error region* of the target classifier. The ground truth is usually stable in the practical applications where evasion attacks are relevant. Thus, we take the error region definition as the default definition of adversarial risk in our studies. Quantifying adversarial risk then leads to identifying the degree of security of a classifier against evasion attacks.

Inherent upper bound on the robustness of classification against evasion attacks Persistence of adversarial examples has raised a serious concern regarding possibility of implementing a robust machine learning algorithm. To investigate this important issue we posed a research question. In particular, we asked, is there an upper bound on the robustness of machine learning algorithms against evasion adversaries? Alternatively, is there a lower bound on the power of evasion adversaries? We attended to this questions in a series of publications [DMM18; MDM19; MZME19]. In [DMM18], we showed an inherent upper bound on achievable robustness in the presence of evasion adversaries, with sublinear perturbation, when the instances are drawn from uniform hypercube $\{0,1\}^n$. In particular, we showed that for any classifier with a constant error rate (e.g. 0.01), there is an adversary who changes only $O(\sqrt{n})$ bit of the inputs increasing the error of the classifier to almost 1. Our bounds, which were based on an isoperimetric inequality for hamming cube, were independent of the structure of the classifier in use.

In a follow up work [MDM19], we generalized this upper bound to many more metric probability spaces. Explicitly, we drew a connection between the robustness of learning algorithms and a well-studied mathematical phenomena known as *concentration of measure*. We showed that if the metric probability space of the underlying input distribution is well concentrated and the trained hypothesis has some non-negligible error, for most of the instances, there exist perturbations with sub-linear magnitude

[1] Insecure could refer to different criteria in different scenarios.

which when applied to that instance, cause the classifier to output a wrong label. The concentration of measure phenomenon, which is tightly related to isoperimetric and optimal transport inequalities, states that for many natural metric probability spaces (e.g. all so called Lévy families) and for any subset with a constant measure, almost every point sampled from the measure has sub-linear distance from that subset. There are many mathematical results on the concentration of natural metric probability spaces, such as product distributions under hamming distance, Gaussian distribution under euclidean distance, product of unit spheres under the euclidean or geodesic distance and more. We showed that any such concentration inequality for a metric probability space will give an upper bound on achievable robustness of any classification problem where instances are coming from that probability space.

Our result in [MDM19] leveraged the fact that the target classifier have some initial error. A natural question that follows is whether we can decrease the error rate of the classifier so that the effect of the adversaries are not as severe. In other words, can we mitigate these attacks by training better classifiers with smaller error rate? To answer this question, in [DMM19], we studied the effect of evasion adversaries on the required sample complexity of learning algorithms. We showed that there exist a learning problem which requires exponential sample complexity to achieve small adversarial risk. However, in the absence of adversaries, the same problem can be learned with polynomial number of samples.

Black-box estimation of concentration: Extending theoretical upper bound to real world distributions. Although our work [MDM19] showed the connection between concentration of measure and adversarial robustness, it was still unclear whether real-world data distributions are concentrated enough to justify the existence of adversarial examples. The next immediate question then was whether we can translate the theoretical upper bounds to the real world applications. In a follow up project [MZME19], we introduced a new method to estimate concentration of measure for an arbitrary metric probability space, using i.i.d. samples. We designed an empirical concentration problem and proved that the solution of this empirical problem converges to the solution of the actual concentration problem asymptotically. We also provided a heuristic algorithm to solve the empirical concentration for estimating concentration of measure on image datasets such as MNIST and CIFAR10. Our results showed that even though there are cases where the robustness of the algorithm in practice is very close to the achievable upper bounds, for some cases, the gap is larger. This finding suggested that concentration of measure alone cannot fully explain the existence of adversarial examples in some of the practical scenarios.

A cryptographic approach: Can computational limitation of adversaries help robustness? An important open question about security of machine learning is whether one can rely on the fact that adversaries are computationally bounded and design secure schemes. This technique is what enables many constructions in cryptography which are provably secure as long as the adversary can only compute bounded number of operations. Inspired by the success of the field of cryptography in exploiting computational limitation of adversaries, I studied the power of computationally bounded evasion and poisoning adversaries in a series of work [MM19; EMM20; GJMM20].

Based on the lower bounds proved in [MDM19], we already knew that information theoretic adversaries are very powerful. However, all those results are *existential* and do not address the *computational* aspects of finding adversarial examples. In [MM19] we showed that if instances are coming from a product distribution, it is computationally feasible to find adversarial examples with $O(\sqrt{n})$ perturbations (under Hamming distance) as long as the adversary has black-box access to the hypothesis. In this paper, we introduced a new notion called *Computational Concentration of Measure* and showed that it is sufficient for getting polynomial time evasion attacks.

This result showed a barrier against leveraging on hardness assumptions to design learning algorithms that are robust against polynomial time adversaries. Yet, there was a gap between the power of algorithmic attacks of [MM19] and the existential attacks of [MDM19]. In a follow up work [EMM20], we designed an algorithm that closed this gap and obtained asymptotically optimal bounds in polynomial time. We also showed how to extend computational concentration of measure to other metric probability spaces by introducing a special type of embedding that preserves computational concentration of measure. In other words, our work shows that the best existing lower bounds on the power of information theoretic and computationally bounded adversaries are equal for certain metric probability spaces.

Although this might sounds disappointing, it does not imply that computational hardness assumptions cannot be helpful. Considering that there is a gap between existing algorithm's robustness and theoretical upper bounds, computational hardness might help closing this gap. In fact, in [GJMM20], we constructed a learning problem which its computational robustness was much higher than its information theoretic robustness.

Training-time Attacks

Another category of attacks are poisoning attacks that occur during training. Poisoning adversaries could have different goals in mind. For instance they could aim for reducing the overall performance of the trained models or they could attempt to change the prediction of the trained model for a specific instance (Aka. indiscriminate attacks and targeted attacks[2]). We modeled the malicious goals of the poisoning adversaries in a unified and abstract way where we assumed an arbitrary *bad* property over the hypothesis space. The objective of the poisoning adversary is to increase the probability of this bad property over the training process. We then studied the power of poisoning adversaries in achieving this objective.

Another aspect of poisoning adversaries that must be defined is their tampering pattern. For instance, one can imagine a poisoning adversary who looks at the training data and changes a fraction of the training examples arbitrarily. Alternatively, a weaker adversary may only add some poisoned data to the original dataset, without knowing the other examples in training set[3]. The adversaries could also be limited in their computational power or the way they label poisoned data. In my research, I studied power of poisoning adversaries based on their tampering pattern and their computational power.

Inherent lower bound on the power of poisoning attacks Throughout my research, I have studied the inherent vulnerability of machine learning against poisoning adversaries with different tampering patterns. Bellow I summarize these results.

- **Adversaries with random tampering locations:** In [MM17], we studied poisoning adversaries who could substitute a random p fraction of a training examples and replace them with other examples[4]. We showed that in this model there are adversaries who increase the probability of an arbitrary bad property by $\Omega(p)$ if the probability of getting the bad property is originally constant. The adversaries in this attack model, called p-tampering model, are very powerful in the sense that they can achieve these bounds with many restrictions such as (only) black-box access to the training algorithm, working online, and using the correct labels for the training examples. In a follow up work [MM19], we further improved our bounds on the power of such adversaries and introduced new attacks. We also proved that our attacks could be implemented in polynomial time, given oracle access to the training algorithm and enough samples from the distribution of instances.

- **Adversaries who choose tampering locations:** One can define a stronger poisoning adversary that has control over the tampering locations as well. In [MDM19], we studied these type of adversaries and proved that there are adversaries who select $\tilde{O}(\sqrt{m})$ (m is the sample complexity of the learning algorithm) number of training examples, replace them with other correctly labeled training examples, and increase the probability of the bad property to almost 1, if the original probability is $1/\text{poly}(m)$. In [MM19], we provided an algorithm for adversary that could achieve almost the same bound in polynomial time, as long as it has oracle access to training algorithm and enough samples from the distribution. However, this polynomial time algorithm could not cover the cases were the probability of the bad property could potentially decrease by sample complexity. Later, using our optimal computational concentration of measure for product spaces, introduced in [EMM20], we improved our poisoning attacks by providing a polynomial time algorithm that could achieve the same bounds of [MDM19] even in the case of vanishing probability of the bad property.

- **Byzantine adversaries in multi-party learning:** Multi-party learning enables distinct parties to combine their data and train a shared model. With the recent advances in collaborative machine learning, it has become very important to study the effect of malicious parties who provide corrupted data. In [MMM19], we introduced a new model of (k, p) poisoning adversaries, in multi-party learning setting, where there are m parties who provide the training data. Among those, k are partially corrupted meaning that for each training example provided (by the partially corrupted parties) there is a probability p that the example is generated by the adversary. For $k = m$, this model becomes the notion of p-tampering poisoning, and for $p = 1$ it coincides with the standard notion of static corruption in multi-party computation. we showed, in this setting, for any m-party learning protocol there exist a computationally bounded (k, p) poisoning adversary that increases the probability of the bad property by $\Omega(p \cdot k/m)$. Our (k, p) poisoning attacks are online and only use correct labels for the corrupted training data. Moreover, we showed that our attack can be implemented in polynomial time as long as it has access to sampling oracle for distributions of all the parties as well as oracle access to the training algorithm.

[2]For a survey on different goals of adversaries see Marco Barreno, Blaine Nelson, Anthony D Joseph, and J Doug Tygar. "The security of machine learning". In: *Machine Learning* 81.2 (2010), pp. 121–148

[3]These two models are known as strong contamination model and Huber's contamination model. For more details see Ilias Diakonikolas and Daniel M Kane. "Recent Advances in Algorithmic High-Dimensional Robust Statistics". In: *arXiv preprint arXiv:1911.05911* (2019).

[4]This adversarial model is tightly related to valiant's malicious noise model. For more details see Michael Kearns and Ming Li. "Learning in the presence of malicious errors". In: *SIAM Journal on Computing* 22.4 (1993), pp. 807–837.

Tampering Attacks in Cryptography

In addition to applications in machine learning, my work on robustness of random process has important implications in Cryptography. Following, I explain the implication of my work in tamperable cryptography and coin tossing.

Blockwise attacks on the randomness of cryptographic primitives. Austrin et. al[5] studied the notion of bitwise p-tampering attacks over randomized algorithms in which an efficient *virus* gets to control each bit of the randomness with independent probability p in an online way. They showed how to break certain *privacy primitives* (e.g., encryption, commitments, etc.) through bitwise p-tampering. Their attacks were heavily relying on the adversary being able to tamper with each bit with independent probability p. However, randomness is usually generated in blocks rather than bits. We generalized the result of Austrin et. al. to the blockwise setting and introduced new p-tampering attacks that could break the semantic security of any encryption scheme [MM17]. Our p-tampering attacks also yield an algorithmic proof on the impossibility of extracting an unbiased random bit form a so called *blockwise* Santha-Vazirani sources of randomness.

Lower bounds for coin tossing protocols. Many of my work in security of machine learning are inspired by, and have direct implication to, coin tossing protocols. Our computational concentration of measure results of [MM19] and [EMM20] had direct implication on the power of adversaries in single-round multi-party coin tossing protocols. We introduced an adversarial model where the adversary observes messages sent by parties one by one and after seeing each message is allowed to perturb with it. We proved that in this model, there is always an adversary who tampers $\sqrt{m}$ number of messages and biases the average of the protocol to almost 1. Surprisingly, we are able to achieve these bounds by a polynomial time attack. Moreover, in [MMM19], we introduced a new model of attack to an m-party coin tossing protocol where an adversary selects k parties prior to the start of the protocol. It then controls each message sent by each of those parties with probability p. This model generalizes both static corruption model (when $p = 1$) and the p-tampering model (when $m = k$). We showed that in the presence of such adversaries, the output of any m-party coin tossing protocol can be biased by $\Omega(k \cdot p/m)$.

Future plans

Provable robustness in machine learning My research has so far developed a theoretical basis for limitation of worst-case robustness for machine learning, based on properties of high dimensional data such as (computational) concentration of measure. I intend to continue my research on understanding limitation of learning algorithms especially those which are due to properties of high dimensional data or computational constraints. Furthermore, I would like to work on designing and improving machine learning algorithms with provable robustness guarantees. I also plan to expand my research to adversarial robustness in different learning settings such as online learning and control.

Other aspects of automated decision making. Today's application of machine learning spans across many different areas from finance to health to infrastructures and even public security. However, this approach raises new concerns such as **fairness** and **privacy** of data for stakeholders. To ensure an unbiased and accountable decision making, machine learning algorithms need to be designed for fairness. Similarly, the privacy of information is another major concern which can limit application of machine learning algorithms in sensitive areas such as health. Currently, these topics are becoming more and more important in the field and I am interested in expanding my research to understanding the limitations of provable fairness and privacy of data as well as the trade-offs between the accuracy, fairness and privacy.

Building more bridges between cryptography and machine learning. I intend to investigate power and limitation of cryptographic primitives, specially those with applications in machine learning. There are cryptographic tools (e.g. Multi party computation, Homomorphic Encryption, Differential Privacy) that can facilitate improvement of "safety" of machine learning. Yet, the applications are limited due to efficiency constraints. Developing new cryptographic tools that are machine learning friendly and machine learning techniques that are cryptography friendly is my other future plan.

[5] Per Austrin, Kai-Min Chung, Mohammad Mahmoody, Rafael Pass, and Karn Seth. "On the impossibility of cryptography with tamperable randomness". In: *Algorithmica* 79.4 (2017), pp. 1052–1101

References

[DMM18] Dimitrios Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. "Adversarial risk and robustness: General definitions and implications for the uniform distribution". In: *Advances in Neural Information Processing Systems*. 2018, pp. 10359–10368.

[DMM19] Dimitrios I Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. "Lower Bounds for Adversarially Robust PAC Learning". In: *arXiv preprint arXiv:1906.05815* (2019).

[EMM20] Omid Etesami, Saeed Mahloujifar, and Mohammad Mahmoody. "Computational Concentration of Measure: Optimal Bounds, Reductions, and More". In: *ACM-SIAM Symposium on Discrete Algorithms* (2020).

[GJMM20] Sanjam Garg, Somesh Jha, Saeed Mahloujifar, and Mohammad Mahmoody. "Adversarially Robust Learning Could Leverage Computational Hardness". In: *Algorithmic Learning Theory* (2020).

[MDM19] Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. "The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 4536–4543.

[MM17] Saeed Mahloujifar and Mohammad Mahmoody. "Blockwise p-tampering attacks on cryptographic primitives, extractors, and learners". In: *Theory of Cryptography Conference (TCC)*. Springer. 2017, pp. 245–279.

[MM19] Saeed Mahloujifar and Mohammad Mahmoody. "Can Adversarially Robust Learning LeverageComputational Hardness?" In: *Algorithmic Learning Theory*. 2019, pp. 581–609.

[MMM19] Saeed Mahloujifar, Mohammad Mahmoody, and Ameer Mohammed. "Universal Multi-Party Poisoning Attacks". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 4274–4283. URL: http://proceedings.mlr.press/v97/mahloujifar19a.html.

[MZME19] Saeed Mahloujifar, Xiao Zhang, Mohammad Mahmoody, and David Evans. "Empirically Measuring Concentration: Fundamental Limits on Intrinsic Robustness". In: *Advances in Neural Information Processing Systems* (2019).