Learner Knowledge Levels in Adversarial Machine Learning

Sihui Dai [1] **Prateek Mittal** [1]

Abstract

For adversarial robustness in a practical setting, it is important to consider realistic levels of knowledge that the learner has about the adversary's choice in perturbations. We present two levels of learner knowledge, (1) full knowledge which contains the majority of current research in adversarial ML and (2) partial knowledge which captures a more realistic setting where the learner does not know how to mathematically model the true perturbation function used by the adversary. We discuss current literature within each category and propose potential research directions within the setting of partial knowledge.

1. Motivation

Currently, the majority of adversarial ML research addresses the problem of defending against an adversary with a well-defined threat model such as ℓ_p perturbations of bounded radius. In practice, however, the threat models that we would like to defend against may not be so well-defined. For example, in tasks such as image classification, we may be interested in defending against the set of imperceptible perturbations, but we currently do not have a good model of imperceptibility. In fact, defining perceptual distance metrics itself is a research area (Zhang et al., 2018; Wang et al., 2004).

Thus, to improve the practicality of the current defenses, it is crucial for us to consider more realistic levels of knowledge that the learner may have about the adversary's threat model. We divide learner knowledge into 2 categories: (1) full knowledge, the setting in which the learner has unlimited access to the true perturbation function used by the adversary, and (2) partial knowledge, the setting in which the learner has access to a set of approximations of the adversary's threat model. We outline current research in these categories and propose additional research directions within the category of partial knowledge.

[1]Electrical and Computer Engineering, Princeton University, Princeton, USA. Correspondence to: Sihui Dai <sihuid@princeton.edu>.

Adversarial Machine Learning Frontiers Workshop at ICML 2022

2. Background and Notations

Let $\mathcal{D} = X \times Y$ denote the data distribution and $\mathcal{H}$ denote the learner's hypothesis class. The problem of adversarial ML can be modeled as a game between the learner and attacker (Huang et al., 2011). The defender first chooses a learning algorithm $\mathcal{A}$ for obtaining a robust model in $\mathcal{H}$ based on training data D_{train} and a set K of knowledge that the learner may have about the attacker's threat model. The attacker then chooses an attack procedure $P : X \times Y \times \mathcal{H} \to X$ to apply during test-time (potentially with knowledge of $\mathcal{A}$). One example of an attack procedure is an ℓ_2 attack with bound 0.5: $P(x, y, h) = \arg\max_{x', ||x'-x||_2<0.5} \ell(h(x'), y)$.

During training, the defender obtains a model by applying their learning algorithm on the training set and knowledge set: $h = \mathcal{A}(D_{\text{train}}, K)$. During testing, for every test data $(x, y) \sim \mathcal{D}$, the defender evaluates the performance of their h on the perturbed input $P(x, y, h)$. This performance is assessed using a loss function $\ell : X \times Y \to \mathbb{R}$.

Formally, the learner's objective is to define a learning algorithm $\mathcal{A}$ such that with high probability over the training samples, $h = A(D_{\text{train}}, K)$ achieves $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(P(x, y, h)), y)] \leq \epsilon$ for small $\epsilon \in \mathbb{R}$.

3. Learner Knowledge Levels

We now present two levels of learner knowledge, full knowledge and partial knowledge, and discuss prior works within each category. We also discuss potential research directions within the category of partial knowledge.

3.1. Full Knowledge

Most existing defenses against adversarial examples fall under the category of a learner with full knowledge. With full knowledge, the learner knows both the exact constraints of the adversary and has a model for generating these adversarial examples. In this setting, the learner's knowledge set K contains P, the true perturbations used by the adversary. The learner can use this knowledge in 2 ways: (1) the learner uses information about the mathematical model for P or (2) the learner can make queries to P on any $x \in X$ and $h \in \mathcal{H}$. Defenses that fall under the first form include certified robustness techniques such as randomized smoothing (Cohen

et al., 2019) and interval bound propagation (Yang et al., 2020); these techniques certify robustness for a specific perturbation type (ie. ℓ_2, ℓ_∞). Meanwhile adversarial training (Madry et al., 2018; Zhang et al., 2019; Tramèr & Boneh, 2019; Maini et al., 2020) falls into the second setting. In adversarial training framework, the learner does not need to know exactly what P is, they only need to query P to obtain adversarial examples during training.

3.2. Partial Knowledge

In practice, the learner does not have full knowledge of the adversary. A better model of learner knowledge is partial knowledge, where the learner has an idea of the space of perturbations that can be performed by the adversary, but this space of perturbations is difficult to model. For example, the learner may know that the adversary is restricted to imperceptible perturbations but does not know exactly how to model imperceptibility.

Under the setting of learning with partial knowledge, we assume that the knowledge set K of the learner contains a set of approximations of the true adversarial perturbation P. For instance, we can consider K to be the set of ℓ_p bounded perturbations with radius δ: $K = \{P' \mid P'(x, y, h) = \arg\max_{x', ||x'-x||_p<\delta} \ell(h(x'), y), p \in \mathbb{R}^+\}$. We can also consider settings where P' are noisy versions of P, for example, K may contain $P' = P + \Delta$ where $\Delta \sim \mathcal{N}(0, I)$.

This learner's learning algorithm can then use knowledge of the exact mathematical form of or query access to each $P' \in K$. The goal of the learner is to achieve $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(P(x, h)), y)] \leq \epsilon$. Current approaches to achieving robustness in this setting generally fall under 2 categories: (1) perturbation modeling and (2) enforcing smoothness.

Perturbation modeling Approaches that fall under this category generally consider using a learning algorithm from the full knowledge setting (ie. adversarial training) with the available approximations $P' \in K$. One example of a defense that falls under this category is perceptual adversarial training (PAT) (Laidlaw et al., 2021), which uses perturbations with bounded LPIPS distance (Zhang et al., 2018) from the input during training. Here, the LPIPS-bounded attack is an approximation of the true imperceptible attack used by the adversary and can be considered an approximation $P' \in K$. (Laidlaw et al., 2021) demonstrate that using this approximation with adversarial training improves robustness against multiple adversaries including ℓ_p, spatially transformed (Xiao et al., 2018), and recolor (Laidlaw & Feizi, 2019) adversaries. On the theoretical front, Montasser et al. (2021) explore the setting of learning with an approximation P' of P and show that it is possible to generalize to attacks generated with P'. However, it is unclear whether it is possible to generalize to P.

The direction of perturbation modeling opens several areas of research in both algorithms and theory. Algorithmically, how can we improve upon approximations for imperceptibility? One potential direction is to utilize generative models to learn to generate realistic perturbations. This idea has been explored in Wong & Kolter (2020) and Madaan et al. (2020) for robustness with full knowledge, but not for the setting of partial knowledge. In the direction of theory, we can ask the question: for what types noise present in the approximations $P' \in K$ is learning to be robust to P feasible?

Enforcing smoothness Another line of works addressing the partial knowledge by enforcing smoothness. Specifically, these works look at methods (mainly regularization) to bias the learning algorithm to select models for which the loss increases more gradually when tested on P that lies outside of K using only $P' \in K$. Dai et al. (2022) provide a theoretical framework for reasoning about what types of learning algorithms produce models which have better generalization to unforeseen perturbations and introduce a regularization term called variation regularization to enforce this. Similarly, Jin & Rinard (2020) propose regularizing Hamming distance between activation patterns and ℓ_2 Lipschitzness of the prediction to bias towards smooth models.

The direction of enforcing smoothness also opens several directions for research. First, what is the best definition for smoothness? In the full knowledge setting with ℓ_p perturbations we could use local-Lipschitzness, it is unclear what specific property we would like when the true perturbation function is unknown. Secondly, Dai et al. (2022) and Jin & Rinard (2020) find that applying regularization trades off significant clean accuracy. How can we reduce this trade-off? Another direction is evaluation: how should we evaluate these approaches and how can we compare with approaches using perturbation modeling?

Incorporating additional query knowledge We can also bridge the gap between the partial and full knowledge settings by further allowing the partial knowledge learner to make a limited number of queries to P.

One problem under this category of partial knowledge with limited query access is the problem of adapting robust models to be robust against new perturbation types. For example, consider the case where we might have a robust model and at some point in time discover a new perturbation type that our model is not robust to. In this case, we may be interested in quickly adapting our model (with few examples of the attack) to be robust against the union of the new attack and all previously known attacks instead of retraining a model from scratch. To the best of our knowledge, there are no works investigating at this setting within adversarial ML, and we encourage additional research addressing this learner knowledge regime.

Acknowledgements

This work was supported in part by the National Science Foundation under grants CNS-1553437 and CNS-1704105, the ARL's Army Artificial Intelligence Innovation Institute (A2I2), the Office of Naval Research Young Investigator Award, the Army Research Office Young Investigator Prize, Schmidt DataX award, and Princeton E-ffiliates Award. This material is also based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2039656. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 2019. URL http://proceedings.mlr.press/v97/cohen19c.html.

Dai, S., Mahloujifar, S., and Mittal, P. Formulating robustness against unforeseen attacks. *arXiv preprint arXiv: Arxiv-2204.13779*, 2022.

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., and Tygar, J. D. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, AISec '11, pp. 43–58, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450310031. doi: 10.1145/2046684.2046692. URL https://doi.org/10.1145/2046684.2046692.

Jin, C. and Rinard, M. Manifold regularization for locally stable deep neural networks. *arXiv preprint arXiv:2003.04286*, 2020.

Laidlaw, C. and Feizi, S. Functional adversarial attacks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 10408–10418, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6e923226e43cd6fac7cfe1e13ad000ac-Abstract.html.

Laidlaw, C., Singla, S., and Feizi, S. Perceptual adversarial robustness: Defense against unseen threat models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=dFwBosAcJkN.

Madaan, D., Shin, J., and Hwang, S. J. Learning to generate noise for robustness against multiple perturbations. *CoRR*, abs/2006.12135, 2020. URL https://arxiv.org/abs/2006.12135.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Maini, P., Wong, E., and Kolter, J. Z. Adversarial robustness against the union of multiple perturbation models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6640–6650. PMLR, 2020. URL http://proceedings.mlr.press/v119/maini20a.html.

Montasser, O., Hanneke, S., and Srebro, N. Adversarially robust learning with unknown perturbation sets. In Belkin, M. and Kpotufe, S. (eds.), *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3452–3482. PMLR, 2021. URL http://proceedings.mlr.press/v134/montasser21a.html.

Tramèr, F. and Boneh, D. Adversarial training and robustness for multiple perturbations. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1904.13000.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13 (4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Wong, E. and Kolter, J. Z. Learning perturbation sets for robust machine learning. In *International Conference on Learning Representations*, 2020.

Xiao, C., Zhu, J., Li, B., He, W., Liu, M., and Song, D. Spatially transformed adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net,

2018. URL https://openreview.net/forum?id=HyydRMZC-.

Yang, G., Duan, T., Hu, J. E., Salman, H., Razenshteyn, I., and Li, J. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pp. 10693–10705. PMLR, 2020.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482. PMLR, 2019. URL http://proceedings.mlr.press/v97/zhang19p.html.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.