# DATA PROFILING FOR ADVERSARIAL TRAINING

**Chengyu Dong & Jingbo Shang**
University of California, San Diego
{cdong,jshang}@eng.ucsd.edu

**Liyuan Liu**
University of Illinois at Urbana-Champaign
llychinalz@gmail.com

## ABSTRACT

Multiple intriguing problems hover in adversarial training, including robustness-accuracy trade-off, robust overfitting, and gradient masking, posing great challenges to both reliable evaluation and practical deployment. Here, we show that these problems share one common cause—low quality samples in the dataset. We first identify an intrinsic property of the data called *problematic score* and then design controlled experiments to investigate its connections with these problems. Specifically, we find that when problematic data is removed, robust overfitting and gradient masking can be largely alleviated; and robustness-accuracy trade-off is more prominent for a dataset containing highly problematic data. These observations not only verify our intuition about data quality but also open new opportunities to advance adversarial training. Remarkably, simply removing problematic data from adversarial training, while making the training set smaller, yields better robustness consistently with different adversary settings, training methods, and neural architectures.

## 1 INTRODUCTION

Adversarial training (Goodfellow et al., 2015; Huang et al., 2015; Kurakin et al., 2017; Madry et al., 2018) is arguably the most effective way (Athalye et al., 2018; Uesato et al., 2018) to establish the robustness of deep neural networks against adversarial perturbations (Szegedy et al., 2014; Goodfellow et al., 2015). Nevertheless, intriguing problems and properties hover in adversarial training, including but never limited to (1) Robustness-accuracy Trade-off (Papernot et al., 2016; Su et al., 2018; Tsipras et al., 2019; Zhang et al., 2019); (2) Robust Overfitting (Rice et al., 2020); and (3) Gradient Masking (Papernot et al., 2017; Athalye et al., 2018; Uesato et al., 2018; Engstrom et al., 2018; Mosbach et al., 2018).

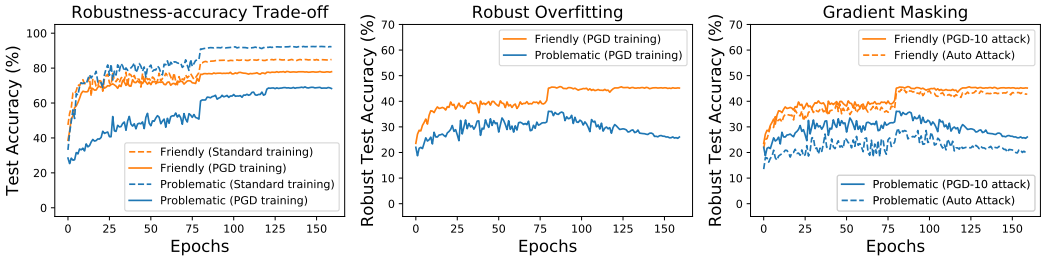



Figure 1: Despite being equal-size and class-balanced, the friendly partition and problematic partition of the CIFAR-10 training set have drastically different behaviors in adversarial training. On friendly partition, there is only minor robustness-accuracy trade-off, almost no robust overfitting, and minimal gradient masking, which is sharp contrast to problematic partition. Here the model is fixed as pre-activation ResNet-18 (He et al., 2016).

In this work, we show that these problems are all interconnected with the characteristics of data itself. We profile the training data according to their distinct behaviors in adversarial training and identify an intrinsic property of the data which we refer as *problematic score*. We show that data with high problematic scores can be one common cause to the three aforementioned problems in adversarial training. To the best of our knowledge, this is the first time these problems are investigated

in a holistic and systematic manner. As a demonstration, we partition the training set of CIFAR-10 (Krizhevsky, 2009) into two exclusive subsets with equal size and balanced classes based on the problematic score, then employ both standard training and adversarial training with Projected Gradient Descent (PGD) (Madry et al., 2018). As shown in Figure 1, adversarial training on the friendly partition yields high robustness, minor robustness-accuracy trade-off, almost no robust overfitting, and minimal gradient masking, which is in sharp contrast to that on the problematic partition. Yet, standard training can perform better on the problematic partition in terms of the standard accuracy.

To further explore the interconnection between problematic score and existing problems in adversarial training, we conduct extensive controlled experiments with different sample sizes, training methods and neural architectures (see Appendix E, F and G).

Our major findings are as follows.

- The level of robustness-accuracy trade-off is positively correlated with the problematic score. Adversarial training can achieve comparable standard accuracy as standard training when conducted on friendly data only.
- Robust overfitting originates from the adversarial training on the problematic data.
- Problematic data can cause gradient masking, due to a conflict of optimization targets in the inner maximization.

Due to the troublesome role of problematic data in adversarial training, we propose to simply remove these data from training, which can yield consistently better adversarial robustness for different training methods, model capacities, and neural architectures[1].

## 2 Related work

Here we briefly review the existing work investigating robust learning from a data perspective. We will discuss more works specifically related to robustness-accuracy trade-off, robust overfitting, and gradient masking in Appendix E, F, and G, respectively. More broadly related works are discussed in Appendix I.

In order to understand the fundamental difficulty in robust learning, increasing attentions have been paid to the data itself. Compared to standard learning, robust learning is highly sensitive to the property of the data. Schmidt et al. (2018) showed that robust learning requires much higher sample complexity than standard learning. They also showed that such sample complexity requirement is highly sensitive to the data distribution. Ding et al. (2019) further observed that adversarial robustness achieved by adversarial training is sensitive to semantically-lossless distribution shift. Shafahi et al. (2019) shows that the adversarial robustness achieved by any classifier is fundamentally limited by the properties of the data distribution.

Training data may have diverse behaviors in adversarial training. Wang et al. (2020) showed that misclassified examples, where adversarial examples are not well defined, have greater impact on adversarial robustness. Zhang et al. (2020b) observed that some examples in the training data are more robust to adversarial attack.

More data will not always yield better adversarial robustness. Ding et al. (2019) observed that the robust test accuracy plateaus after training size is sufficiently large on MNIST (LeCun et al., 1998). Introducing additional data unrelated to the original dataset is shown to be detrimental to the robustness (Uesato et al., 2019; Gowal et al., 2020). Yang et al. (2020a) also showed that dataset pruning can effectively enhance the robustness of non-parametric classifiers.

## 3 Problematic Data in Adversarial Training

### 3.1 Estimation of the Problematic score

We conjecture the existence of an intrinsic property of the data referred as *problematic score*, which plays a critical role in adversarial training. Since it is not possible to visually check every example and quantify such a property, we motivate from learning stability to estimate the problematic score.

[1]Code available at `https://github.com/shwinshaker/RobustDataProfiling`

Specifically, we note that during training, an example correctly classified at some epoch will not necessarily be correctly classified ever since (Toneva et al., 2019). The fraction of epochs that an example is correctly classified throughout the training can be used to describe its stability of being learned. We conjecture such learning stability is monotonically correlated with its problematic score, namely a more problematic example will always be learned more unstably. The rank of the examples based on the problematic score can thus be estimated, which we refer as the problematic rank. In the following analyses, we will use the problematic rank as a surrogate to investigate the effect of problematic score in adversarial training.

Towards a more accurate estimation, we calculate the problematic rank based on the ensemble of multiple experiments. Figure 3 and Figure 4 show samples of the most problematic examples and the most friendly examples in the CIFAR-10 training set, respectively (see Appendix A.2). Note that other motivations to estimate problematic rank such as prediction probability, minimum perturbation, and learning order are feasible and will lead to similar results (see Appendix B).

### 3.2 Learning on friendly data only

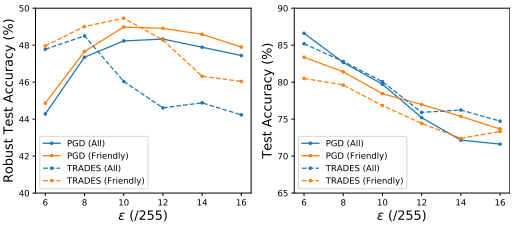
We propose to remove data with the highest problematic rank from adversarial training. We find that with a proper amount of problematic data being removed, state-of-the-art adversarial training methods such as PGD and TRADES (Zhang et al., 2019) combined with early stopping (Rice et al., 2020) can achieve better robustness when other settings are held constant, as shown in Table 1. We evaluate the robustness against AutoAttack (AA) (Croce & Hein, 2020b), one of the strongest adversarial attacks. We note that the optimal fraction of data being removed is higher than the misclassification rate of the best model in terms of robustness, which implies there is indeed a portion of the misclassified examples contributing to the robustness. Besides, less amount of data would save more computational cost during the model training.

Table 1: Performance of adversarially training a pre-activation ResNet-18 on CIFAR-10 using PGD and TRADES. Robust accuracy is evaluated by the standard version of AutoAttack. We search the perturbation radius yielding the best robust accuracy on the entire dataset for each method ($12/255$ for PGD and $8/255$ for TRADES) on a basis of $2/255$. "All" denotes the adversarial training on the entire training set of CIFAR-10, while "Friendly" denotes the adversarial training on a friendly training subset consisting of $k$ examples with the smallest problematic ranks. We search $k$ yielding the best robust accuracy for each method ($k = 40 \times 10^3$ for PGD and $k = 45 \times 10^3$ for TRADES) on a basis of $5 \times 10^3$. Other training hyperparameters, evaluation metrics and hardware settings are fixed for both cases (see Appendix C). We repeat every experiment $5$ times and report the average and standard deviation.

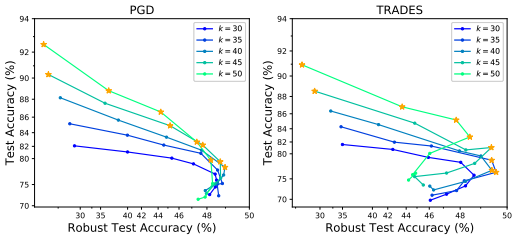
| Method | Dataset | Standard Acc (%) | Robust Acc (%) | Wall time (mins) |
|---|---|---|---|---|
| PGD | All | $75.58 \pm 0.53$ | $48.37 \pm 0.22$ | $888.20 \pm 6.76$ |
| PGD | Friendly | $76.69 \pm 0.40$ | $48.98 \pm 0.26$ | $768.60 \pm 3.88$ |
| TRADES | All | $82.54 \pm 0.22$ | $48.33 \pm 0.24$ | $1519.00 \pm 11.80$ |
| TRADES | Friendly | $80.61 \pm 0.26$ | $49.21 \pm 0.12$ | $1408.00 \pm 7.40$ |

Furthermore, in Figure 2a, we show that removing problematic data can yield better robustness consistently across different perturbation radii $\varepsilon$. Here we report the robustness against a custom version of AutoAttack (see Appendix C) due to computational constraints. Note that PGD training can obtain better robustness with a slightly larger perturbation radius, and close the gap to TRADES, whereas TRADES cannot benefit from a larger perturbation radius likewise. Similar observations are also reported by Gowal et al. (2020). Nevertheless, we show that with problematic data being removed, TRADES can also enjoy the benefit from a larger perturbation radius.

One may note that when training on the friendly subset only, the standard accuracy degrades compared to the training on the entire training set. One can acquire more accuracy at the cost of robustness by incorporating more problematic data into the training. This in fact introduces a new trade-off hyper-parameter $k$, namely the number of friendly examples in the training set. Compared to other trade-off hyper-parameters such as the perturbation radius $\varepsilon$, $k$ extends the current performance of adversarial training methods to the far robustness side, as shown in Figure 2b. For future work, this new trade-off curve can serve as a performance baseline.




(a) As the perturbation radius $\varepsilon$ varies from 6/255 to 16/255, and for both PGD and TRADES, the robustness obtained on the friendly subset consistently outperforms that on the entire training set. Here the friendly subset consists of $40 \times 10^3$ training examples with the smallest problematic ranks. The model is fixed as pre-activation ResNet-18. Other details of the experimental settings are mentioned in Appendix C.




(b) The trade-off curve (robust test accuracy v.s. standard test accuracy) by varying the perturbation radius $\varepsilon$ on a basis of $2/255$, for different size of the friendly subset $k$ used in training. $k = 50$ indicates the training conducted on the entire training set. Other experimental settings are mentioned in Appendix C. Here $k$ amounts to a new hyperparameter that extends the trade-off to the robustness side. Stars denote the Pareto frontier. The axes are in log scale to show the highest accuracy and robustness clearly.

In Appendix D, we show that training on friendly data only can consistently improve the robustness for models with different capacities and neural architectures including Wide ResNet (Zagoruyko & Komodakis, 2016).

## 4 Discussions

Towards understanding the difficulty of robust learning, our data profiling perspective complements the existing analyses on the impact of data properties such as distribution and sample size. Robust learning through adverarial training can exhibit diverse levels of difficulty on individual examples in the same benchmark dataset. The ambiguity of the data reflected by our profiling further implies that adversarial training might be susceptible to complex and indistinct image conditions, thus challenging the feasibility of adversarial training in a realistic scenario where ambiguous data may prevail. One may note that achieving robustness through adversarial training on ImageNet (Russakovsky et al., 2015) is already shown to be much harder (Rozsa et al., 2016; Kurakin et al., 2017).

In practice, our profiling of the data provides a new methodology to analyze the effectiveness of adversarial training methods. The learning behaviour of a method on problematic data reveals the potential reason why it may work or not work. For example, TRADES employs an alternative adversarial loss that penalizes the difference between the prediction probability from an adversarial example and its clean counterpart, instead of the true label, which will impose weaker supervision on problematic examples since problematic examples typically have low prediction probabilities (see Appendix B.1). As a result, TRADES can underfit problematic data thus gaining robustness compared to PGD training under the same perturbation setting, which resembles the effect of directly removing problematic data.

Furthermore, our analyses encourage a new perspective to utilize the benchmark dataset for adversarial training. First, the fact that simply removing problematic data improves robustness implies that adversarial training cannot achieve the optimal robustness on the entire dataset, especially when the model capacity is relatively low. Careful pruning of the training data might thus be necessary for future adversarial training practices to excel in the robustness. In addition, a more promising direction to boost the robustness is to design methods able to learn problematic data properly such that it can also contribute to the robustness. Nevertheless, We empirically find that no existing adversarial training method, without resort to larger model capacity, can provably achieve better robustness through proper learning of problematic data. Whether or not such a method exists remains an interesting problem.

Lastly, as the introduction of additional labeled or unlabeled data becomes increasingly popular (Hendrycks et al., 2019a; Uesato et al., 2019; Carmon et al., 2019; Najafi et al., 2019), extra effort needs to be paid to picking high-quality data, which echos the practical suggestions made in previous works (Uesato et al., 2019; Gowal et al., 2020).

## REFERENCES


Maksym Andriushchenko, F. Croce, Nicolas Flammarion, and M. Hein. Square attack: a query-efficient black-box adversarial attack via random search. *ArXiv*, abs/1912.00049, 2020.

D. Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and S. Lacoste-Julien. A closer look at memorization in deep networks. *ArXiv*, abs/1706.05394, 2017.

Anish Athalye, Nicholas Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *ArXiv*, abs/1802.00420, 2018.

D. Ayala. The theory and practice of item response theory. 2008.

Y. Balaji, T. Goldstein, and Judy Hoffman. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. *ArXiv*, abs/1910.08051, 2019.

M. Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116:15849 – 15854, 2019.

Yoshua Bengio, J. Louradour, Ronan Collobert, and J. Weston. Curriculum learning. In *ICML '09*, 2009.

N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017a.

Nicholas Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017b.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, W. Brendel, Jonas Rauber, D. Tsipras, Ian J. Goodfellow, A. Madry, and A. Kurakin. On evaluating adversarial robustness. *ArXiv*, abs/1902.06705, 2019.

Y. Carmon, Aditi Raghunathan, L. Schmidt, Percy Liang, and John C. Duchi. Unlabeled data improves adversarial robustness. *ArXiv*, abs/1905.13736, 2019.

Haw-Shiuan Chang, E. Learned-Miller, and A. McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *NIPS*, 2017.

L. Chen, Y. Min, Mingrui Zhang, and Amin Karbasi. More data can expand the generalization gap between adversarially robust and standard models. *ArXiv*, abs/2002.04725, 2020.

P. Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and C. Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. *ArXiv*, abs/1709.04114, 2018.

Minhao Cheng, Qi Lei, Pin-Yu Chen, I. Dhillon, and C. Hsieh. Cat: Customized adversarial training for improved robustness. *ArXiv*, abs/2002.06789, 2020.

F. Croce and M. Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *ICML*, 2020a.

F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020b.

Jiequan Cui, Shu Liu, L. Wang, and J. Jia. Learnable boundary guided adversarial training. *ArXiv*, abs/2011.11164, 2020.

G. W. Ding, Kry Yik Chau Lui, Xiaomeng Jin, Luyu Wang, and Ruitong Huang. On the sensitivity of adversarial robustness to input data distributions. *ArXiv*, abs/1902.08336, 2019.

G. W. Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Max-margin adversarial (mma) training: Direct input space margin maximization through adversarial training. *ArXiv*, abs/1812.02637, 2020.

Elvis Dohmatob. Limitations of adversarial robustness: strong no free lunch theorem. *ArXiv*, abs/1810.04065, 2018.

S. Embretson and S. Reise. Item response theory for psychologists. 2000.

L. Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *ArXiv*, abs/1807.10272, 2018.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT*, 1995.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy A. Mann, and P. Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *ArXiv*, abs/2010.03593, 2020.

Guy Hacohen, Leshem Choshen, and D. Weinshall. Let's agree to agree: Neural networks share classification order on real datasets. *arXiv: Learning*, 2019.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *ArXiv*, abs/1603.05027, 2016.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *ICML*, 2019a.

Dan Hendrycks, K. Zhao, Steven Basart, J. Steinhardt, and D. Song. Natural adversarial examples. *ArXiv*, abs/1907.07174, 2019b.

Lang Huang, C. Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. *ArXiv*, abs/2002.10319, 2020.

Ruitong Huang, B. Xu, Dale Schuurmans, and Csaba Szepesvari. Learning with a strong adversary. *ArXiv*, abs/1511.03034, 2015.

Andrew Ilyas, Shibani Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019.

A. Javanmard, M. Soltanolkotabi, and H. Hassani. Precise tradeoffs in adversarial training for linear regression. In *COLT*, 2020.

Harini Kannan, A. Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *ArXiv*, abs/1803.06373, 2018.

G. Katz, C. Barrett, D. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. *ArXiv*, abs/1702.01135, 2017.

A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

M. Kumar, Ben Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010.

A. Kurakin, Ian J. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *ArXiv*, abs/1611.01236, 2017.

Y. LeCun, L. Bottou, Yoshua Bengio, and P. Haffner. Gradient-based learning applied to document recognition. 1998.

Mingchen Li, M. Soltanolkotabi, and S. Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *ArXiv*, abs/1903.11680, 2020.

A. Madry, Aleksandar Makelov, L. Schmidt, D. Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018.

Takeru Miyato, S. Maeda, Masanori Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1979–1993, 2019.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2582, 2016.

Marius Mosbach, Maksym Andriushchenko, T. A. Trost, M. Hein, and D. Klakow. Logit pairing methods can fool gradient-based attacks. *ArXiv*, abs/1810.12042, 2018.

A. Najafi, S. Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. *ArXiv*, abs/1905.13021, 2019.

Preetum Nakkiran. Adversarial robustness may be at odds with simplicity. *ArXiv*, abs/1901.00532, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, B. Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *ArXiv*, abs/1912.02292, 2020.

Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018. URL `https://arxiv.org/pdf/1807.01069`.

Nicolas Papernot, P. McDaniel, Arunesh Sinha, and Michael P. Wellman. Towards the science of security and privacy in machine learning. *ArXiv*, abs/1611.03814, 2016.

Nicolas Papernot, P. McDaniel, Ian J. Goodfellow, S. Jha, Z. Y. Celik, and A. Swami. Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2017.

Camilo Pestana, N. Akhtar, W. Liu, David Glance, and A. Mian. Adversarial perturbations prevail in the y-channel of the ycbcr color space. *ArXiv*, abs/2003.00883, 2020a.

Camilo Pestana, Wei Liu, David Glance, and A. Mian. Defense-friendly images in adversarial attacks: Dataset and metrics for perturbation difficulty. *ArXiv*, abs/2011.02675, 2020b.

R. Prudêncio, J. Hernández-Orallo, and A. Martinez-Usó. Analysis of instance hardness in machine learning using item response theory. 2015.

Chongli Qin, J. Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and P. Kohli. Adversarial robustness through local linearization. In *NeurIPS*, 2019.

Aditi Raghunathan, Sang Michael Xie, F. Yang, John C. Duchi, and P. Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *ICML*, 2020.

Leslie Rice, Eric Wong, and J. Z. Kolter. Overfitting in adversarially robust deep learning. *ArXiv*, abs/2002.11569, 2020.

Kevin Roth, Yannic Kilcher, and T. Hofmann. Adversarial training is a form of data-dependent operator norm regularization. *arXiv: Learning*, 2020.

Andras Rozsa, M. Günther, and T. Boult. Are accuracy and robustness correlated. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 227–232, 2016.

Olga Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, Michael S. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

L. Schmidt, Shibani Santurkar, D. Tsipras, Kunal Talwar, and A. Madry. Adversarially robust generalization requires more data. In *NeurIPS*, 2018.

A. Shafahi, W. Huang, Christoph Studer, S. Feizi, and T. Goldstein. Are adversarial examples inevitable? *ArXiv*, abs/1809.02104, 2019.

K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

M. Smith and T. Martinez. Improving classification accuracy by identifying and removing instances that should be misclassified. *The 2011 International Joint Conference on Neural Networks*, pp. 2690–2697, 2011.

M. Smith, T. Martinez, and C. Giraud-Carrier. An instance level analysis of data complexity. *Machine Learning*, 95:225–256, 2013.

Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R. Venkatesh Babu. Guided adversarial attack for evaluating and enhancing adversarial defenses. *ArXiv*, abs/2011.14969, 2020.

David Stutz, M. Hein, and B. Schiele. Disentangling adversarial robustness and generalization. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6969–6980, 2019.

D. Su, Huan Zhang, H. Chen, Jinfeng Yi, P. Chen, and Yupeng Gao. Is robustness the cost of accuracy? - a comprehensive study on the robustness of 18 deep image classification models. In *ECCV*, 2018.

Christian Szegedy, W. Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2014.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and G. Gordon. An empirical study of example forgetting during deep neural network learning. *ArXiv*, abs/1812.05159, 2019.

Florian Tramèr, A. Kurakin, Nicolas Papernot, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *ArXiv*, abs/1705.07204, 2018.

Florian Tramèr, N. Carlini, W. Brendel, and A. Madry. On adaptive attacks to adversarial example defenses. *ArXiv*, abs/2002.08347, 2020.

D. Tsipras, Shibani Santurkar, L. Engstrom, A. Turner, and A. Madry. There is no free lunch in adversarial robustness (but there are unexpected benefits). *ArXiv*, abs/1805.12152, 2018.

D. Tsipras, Shibani Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv: Machine Learning*, 2019.

Jonathan Uesato, Brendan O'Donoghue, A. Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *ArXiv*, abs/1802.05666, 2018.

Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Robert Stanforth, Alhussein Fawzi, and P. Kohli. Are labels required for improving adversarial robustness? *ArXiv*, abs/1905.13725, 2019.

Yisen Wang, Difan Zou, Jinfeng Yi, J. Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020.

Yuxin Wen, Shuai Li, and Kui Jia. Towards understanding the regularization of adversarial robustness on neural networks. *ArXiv*, abs/2011.07478, 2020.

Tsui-Wei Weng, Huan Zhang, P. Chen, Jinfeng Yi, D. Su, Yupeng Gao, C. Hsieh, and L. Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *ArXiv*, abs/1801.10578, 2018.

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, A. Yuille, and Quoc V. Le. Adversarial examples improve image recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 816–825, 2020.

Y. Yang, Cyrus Rashtchian, Yizhen Wang, and K. Chaudhuri. Robustness for non-parametric classification: A generic attack and defense. In *AISTATS*, 2020a.

Y. Yang, Cyrus Rashtchian, Hongyang Zhang, R. Salakhutdinov, and K. Chaudhuri. A closer look at accuracy vs. robustness. *arXiv: Learning*, 2020b.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *ArXiv*, abs/1605.07146, 2016.

Hongyang Zhang, Yaodong Yu, J. Jiao, E. Xing, L. Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. *ArXiv*, abs/1901.08573, 2019.

Jingfeng Zhang, Xilie Xu, B. Han, Gang Niu, Li zhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. *ArXiv*, abs/2002.11242, 2020a.

Jingfeng Zhang, Jianing Zhu, Gang Niu, B. Han, M. Sugiyama, and M. Kankanhalli. Geometry-aware instance-reweighted adversarial training. *ArXiv*, abs/2010.01736, 2020b.

Tianyi Zhou, S. Wang, and J. Bilmes. Curriculum learning by dynamic instance hardness. In *NeurIPS*, 2020.

# A MORE ABOUT PROBLEMATIC SCORE

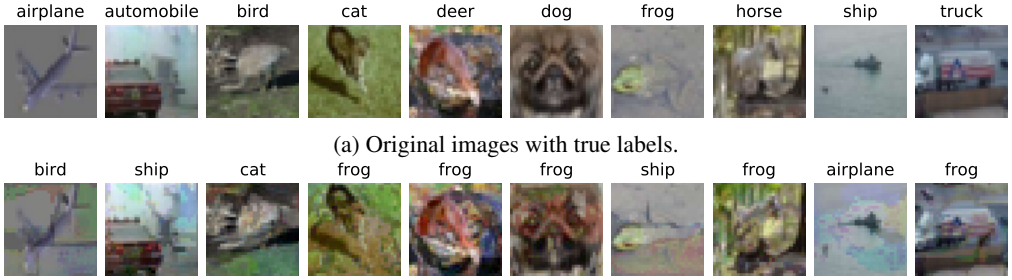
In this section, we introduce more details about the estimation and calculation of problematic score.

## A.1 ESTIMATION OF THE PROBLEMATIC SCORE

In the contexts of adversarial training, how stable a training example will be correctly predicted by the classifier under adversarial attacks is likely to reflect its problematic score. Considering that an example correctly classified at some epoch not necessarily means it will be correctly classified ever since (Toneva et al., 2019), we use the fraction of epochs that an example is correctly classified throughout the training to describe its stability of being learned. Specifically, given an example $x$, the model $f$ and the optimizer $\omega$, we define its learning instability $s(x; f, \omega)$ as

$$s(x; f, \omega) = 1 - \frac{|\{t | f(t, x + \delta) = y(x),\ t \in \{1, \cdots, T\}\}|}{T}, \tag{1}$$

where $\delta$ is a $\ell_\infty$ norm-bounded adversarial perturbation. $f(t, \cdot)$ denotes the classifier at epoch $t$, $y(x)$ denotes the true label of example $x$, and $T$ is the total number of epochs the model will be trained.




(a) Original images with true labels.




(b) Corresponding adversarially perturbated images with predicted labels.

Figure 3: Samples of the most *problematic* examples identified by problematic score from the training set of CIFAR-10.


(a) Original images with true labels.


(b) Corresponding adversarially perturbated images with predicted labels.

Figure 4: Samples of the most *friendly* examples identified by problematic score from the training set of CIFAR-10.

We conjecture that the learning instability of an example is monotonically correlated with its problematic score, namely a more problematic example will always be learned more unstably. Specifically,

$$s(x; f, \omega) = \mathcal{G}_{f,\omega}(\rho(x)), \tag{2}$$

where $\mathcal{G}_{f,\omega}$ is a monotonically-increasing response function associated with model $f$ and optimizer $\omega$. Since we cannot probe into the analytical form of $\mathcal{G}_{f,\omega}$, it is not feasible to directly solve $\rho(x)$ from the observation $s(x; f, \omega)$. Nevertheless, due to the monotonicity, the order of a dataset $X = \{x^{(0)}, x^{(1)}, \cdots, x^{(N)}\}$ should be preserved by $\mathcal{G}_{f,\omega}$, namely

$$\text{rank}(x; \rho(x)) = \text{rank}(x; s(x; f, \omega)), \tag{3}$$

where the ranking function based on a valuation function $s(x)$ is defined as

$$\text{rank}(x; s(x)) = \left|\{x' | s(x') < s(x), x' \in X\}\right|. \tag{4}$$

We refer $\text{rank}(x; \rho(x))$ as *problematic rank*. A large problematic rank indicates an example is highly problematic compared to other examples.

### A.2 Calculation of the problematic rank

To output a relatively accurate estimation of the problematic rank, we synthesize the results of multiple experiments. Specifically, we adversarially train a pre-activation ResNet-18 using PGD-10 on the entire CIFAR-10 training set for $160$ epochs with learning rate initialized as $0.1$ and decayed at epoch $80$ and $120$ with a factor of $10$. We repeat the training $10$ times and average the problematic ranks calculated on each example, which yields a relatively stable estimation[2].

Figure 3 and Figure 4 show samples of the most problematic examples and the most friendly examples in the CIFAR-10 training set, respectively. One can find that compared to friendly examples, problematic examples are intrinsically ambiguous, which cannot easily be identified even by humans. We will later show that such intrinsic ambiguity causes many problems in adversarial training (see Appendix E, F, and G).

### A.3 Problematic score is a property only associated with data

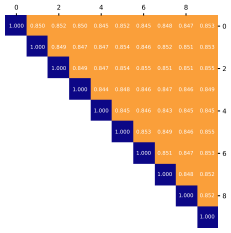
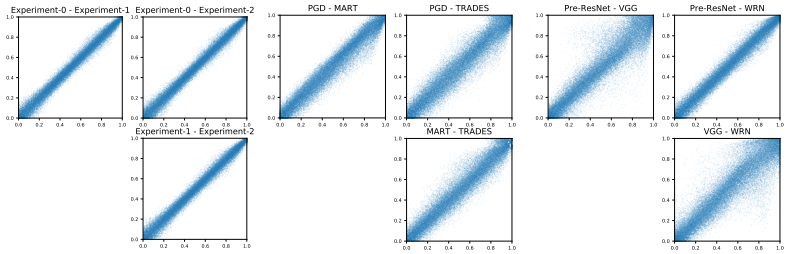


Figure 5: Pairwise IoUs of top problematic examples identified from 10 random initializations with the same other settings.




(a) Random initializations (b) Different methods (c) Different models

Figure 6: Scatter plots of the normalized problematic ranks (i.e. problematic ranks scaled by the number of total examples) estimated by different training settings. The problematic rank is consistent across random initializations, different methods and models.

We empirically show that problematic rank is shared across different random initializations[3], models and training methods. The problematic score is thus likely to be an property intrinsic to the data.

In Figure 5, we use a pre-activation ResNet-18 adversarially trained by PGD-10 to estimate the problematic rank. One can find that the Intersection-over-Unions (IoUs) are consistently high across the subsets consisting of top $10^4$ problematic examples [4] identified by $10$ randomly initialized experiments, which means the problematic examples identified by one experiment is also mostly identified by another experiment. Figure 6a shows the scatter plots of problematic ranks estimated by 3 randomly initialized experiments, which are highly consistent with each other. The problematic rank is also consistent across different adversarial training methods including PGD, TRADES[5],

[2]Due to computational constraints, this estimation process only provides about $1600$ unique values for problematic rank, which are not enough to differentiate all the $5 \times 10^4$ training examples in CIFAR-10. However, in above analyses, we at most partition the training set into $10$ subsets with different problematic levels, which only results in about $0.6\%$ indistinguishable examples between any two adjacent subsets. One can seek more accurate estimation of problematic rank by incorporating the results from more experiments.

[3]Random initialization of model parameters and random shuffling of the data set

[4]$20\%$ of the entire training set

[5]TRADES employs a different method to generate adversarial perturbation compared to PGD. To ensure a consistent definition of learning stability, we implement an additional procedure to generate adversarial perturbation with PGD-10 for TRADES.

MART (Wang et al., 2020) and different models including VGG (Simonyan & Zisserman, 2015), pre-activation ResNet and Wide ResNet (WRN), as shown in Figure 6b and Figure 6c.

# B BROAD MOTIVATIONS OF THE PROBLEMATIC SCORE

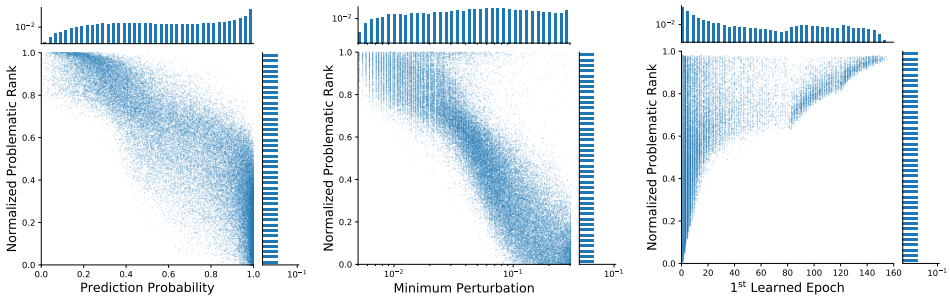
In this section, we show that it is possible to estimate the problematic rank motivated from multiple measurements including prediction probability, minimum perturbation and learning order. Each measurement is itself consistent across different training settings, and is also correlated with the problematic rank estimated in Appendix A.1. This further demonstrates the existence of the problematic score that is intrinsic to the data and profoundly influences the adversarial training. We also briefly review existing adversarial training methods leveraging these measurements in their designs either concerning the inner maximization or outer minimization, which implies that their improved performance is potentially due to the different treatment of examples with different problematic levels.

## B.1 PREDICTION PROBABILITY

Based on the standard adversarial training (Madry et al., 2018), multiple variants pivot on the utilization of soft output in the adversarial loss function. Here we specifically refer the soft output as the either the output before the softmax function, namely logit, or that after the softmax function, namely prediction probability. Adversarial logit pairing (ALP) (Kannan et al., 2018) is a method explicitly penalizing the difference between the logits from an clean example and its adversarial counterpart on top of the standard cross-entropy loss for adversarial training. BGAT (Cui et al., 2020) instead collects the logits of clean examples from an auxiliary clean model. GAT (Sriramanan et al., 2020) adopts a similar loss function penalizing probability difference instead of logit difference. VAT (Miyato et al., 2019) and TRADES (Zhang et al., 2019) propose a loss function matching the probability from an adversarial example with its clean counterpart, instead of the true label. Self-adaptive training (Huang et al., 2020) and MART (Wang et al., 2020) use the probabilities from clean examples to weight the examples in the loss function, although the former focuses more on examples with high probabilities while the latter focuses more on examples with low probabilities.

Prediction probability is also related to whether an example is correctly classified or not. Low probability of the true label indicates an example is likely to be misclassified, which might be troublesome in adversarial training because adversarial examples of misclassfied examples are "undefined" (Wang et al., 2020). Several methods are thus motivated to treat correctly classified and misclassified examples differently. MMA (Ding et al., 2020) employs adverarial training only on correctly classified examples, leaving misclassified examples to standard training. MART (Wang et al., 2020) introduces a term associated with the probability of the true label in the loss function to encourage the learning on misclassified examples. A summary of the loss functions employed in these methods can be found in Table 1 by Wang et al. (2020).

Recently, adversarial training with additional data becomes increasingly popular. Prediction probability is used to identify high-quality data that is more relevant to the original distribution (Gowal et al., 2020).




(a) Prediction probability (b) Minimum perturbation (c) Learning order

Figure 7: Correlation between normalized problematic rank and other measurements

Here we show that the prediction probability of an example is correlated with its problematic rank. These sophisticated methods are thus likely to achieve robustness gain by treating examples with different problematic ranks differently. We specifically refer the *prediction probability* as the probability corresponding to the true label from a clean input, and use the best model in terms of robustness throughout training to estimate it. For every example, we average the probabilities obtained by the same $10$ experiments introduced in Appendix A.2. As shown in Figure 7a, prediction probability is inversely correlated with problematic rank. A friendly example is inclined to be correctly classified by the model with high probability.

### B.2 MINIMUM PERTURBATION

Standard adversarial training often sets a perturbation radius universal to all training examples. However, it has been widely noticed that individual examples may have different levels of robustness against adversarial attacks. It might be helpful to customize the perturbation for each example during adversarial training. MMA (Ding et al., 2020) proposes a method to estimate proper individual perturbation for each example based on its distance to the decision boundary. The perturbation is determined by a line search along the perturbation direction initialized by a norm-constrained PGD attack. In a similar vein, IAAT (Balaji et al., 2019) performs a dynamic update of individual perturbation throughout the training. CAT (Cheng et al., 2020) further incorporates individual label smoothing based on the estimated perturbation level. Instead of customizing individual perturbation radius for each example, FAT (Zhang et al., 2020a) customizes the number of attack iterations for each example, such that the perturbation is just enough to fool the model. GAIRAT (Zhang et al., 2020b) further adopts a weighted loss function based on such individual attack iterations to focus more on those examples far from the decision boundary. Nevertheless, none of these works detailedly analyzes the impact of such individual perturbation on the adversarial training.

Here, we show that profiling the data based on individual perturbation radius leads to similar result as that motivating from learning stability introduced in Appendix A.1. We denote the *minimum perturbation* of an example as the smallest perturbation radius required to change a model's prediction on it. Ideally, the minimum perturbation of an example amounts to its minimum distance to the decision boundary. We use an untargeted attack based on I-FGSM (Iterative Fast Gradient Sign Method) (Goodfellow et al., 2015; Kurakin et al., 2017) with step size $1/255$ based on the implementation in Adversarial Robustness Toolbox (ART) (Nicolae et al., 2018). We empirically found the step size $1/255$ is often small enough to ensure a converged minimum perturbation. One can also employ other attacks including but not limited to DeepFool (Moosavi-Dezfooli et al., 2016), CW (Carlini & Wagner, 2017b), EAD (Chen et al., 2018) and FAB (Croce & Hein, 2020a) in estimation. But note that all these methods cannot obtain the true minimum perturbation, but only the upper bound of it (Weng et al., 2018). Estimation of the minimum perturbation through optimization is known as a NP-hard problem (Katz et al., 2017).

In Figure 7b, we use the best model in terms of robustness obtained in training to estimate the minimum perturbation, and average the results obtained by the same $10$ experiments mentioned in Appendix A.2. One can find that the minimum perturbation is inversely correlated with the problematic rank, which means that a problematic example is more likely to reside near the decision boundary. This suggests the sophisticated methods mentioned above are likely to treat examples with different problematic ranks differently. It also suggests that examples with different amounts of minimum perturbation will influence the adversarial training differently in terms of the contributions to robustness and aforementioned problems.

### B.3 LEARNING ORDER

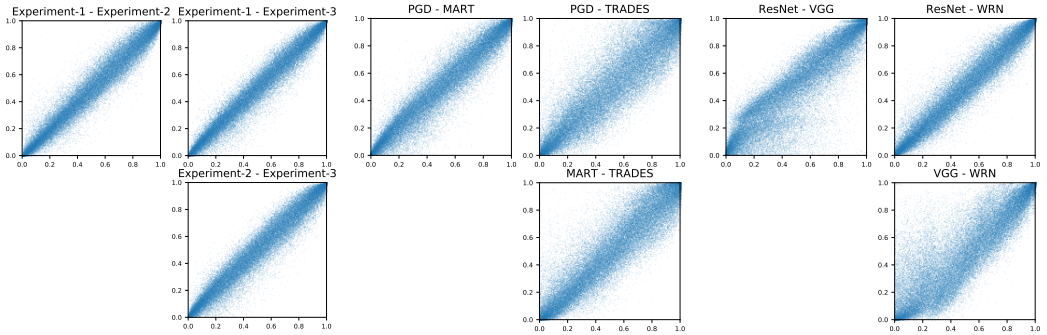
Learning order refers to the phenomenon that Deep Neural Networks (DNNs) learn the examples in a similar order, which is shared by different random initializations and neural architectures. Such phenomenon is observed widely in standard training and training with noisy inputs and labels (Arpit et al., 2017; Li et al., 2020). It is demonstrated that the learning order originates from the coupling between DNNs and benchmark datasets (Hacohen et al., 2019), since DNNs learn synthetic datasets without a specific order and classifiers other than DNNs such as AdaBoost (Freund & Schapire, 1995) learn benchmark datasets without a specific order.

We show that learning order still exists in adversarial training. We denote the $1^{st}$ *learned epoch* as the first epoch when a training example is classified correctly under adversarial attack. We show there is a correlation between the $1^{st}$ learned epochs of an example across different training settings (see Appendix B.4). Furthermore, we show that learning order is correlated with problematic rank. In Figure 7c, we average the $1^{st}$ learned epochs of an example obtained by the same experiments as mentioned in Appendix A.2. One can find the $1^{st}$ learned epoch is positively correlated with its problematic rank, which means problematic examples are likely to be learned late during training.

Note that if we pick the best epoch[6] as a boundary and partition the training examples based on their $1^{st}$ learned epochs, the resulting two subsets correspond to exactly the correctly classified and misclassified examples. This implies that those adversarial training methods treating examples differently based on whether they are correctly classified or not, as mentioned in Appendix B.1, are likely to be a special case of treating examples differently based on their problematic ranks, from yet another perspective.

### B.4 Consistency of the motivation

We show that each motivation mentioned above is itself consistent across different training settings. Therefore it is possible to estimate the problematic rank similarly from each motivation. In Figure 8, 9, 10, we use the same experiments mentioned in Appendix A.2 to show that the prediction probability, minimum perturbation and learning order are all consistent across random initializations, different training methods, and neural architectures. Nevertheless, one may find that these estimations are relatively less consistent compared to the problematic rank estimated from learning stability, which is the major reason that we estimate the problematic rank from learning stability in the main text.




(a) Random initializations (b) Different methods (c) Different models

Figure 8: Scatter plots of the normalized ranks of training examples based on prediction probabilities obtained by different training settings. The prediction probability of an example is consistent across random intializations, different methods and models.

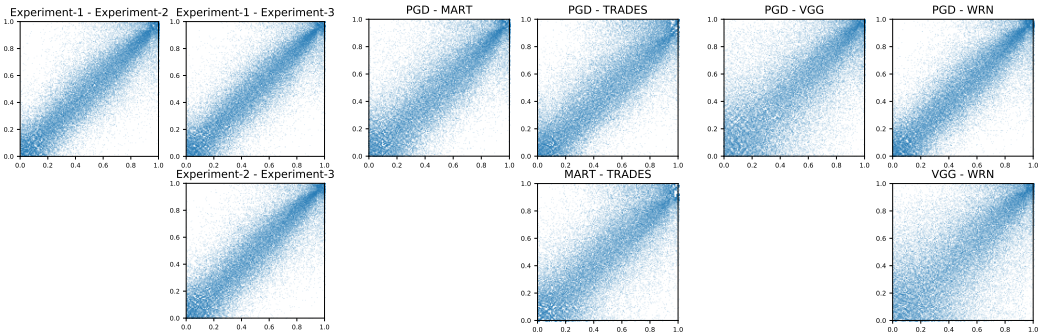
## C Experiment details

We adopt the following setting for all experiments unless otherwise noted.

**Robustness evaluation.** We consider the robustness against $\ell_\infty$ norm-bounded adversarial attack with perturbation radius $8/255$. We evaluate the robustness using *AutoAttack* (Croce & Hein, 2020b), which is the strongest adversarial attack to the best of our knowledge. Throughout the paper, two versions of AutoAttack are used.

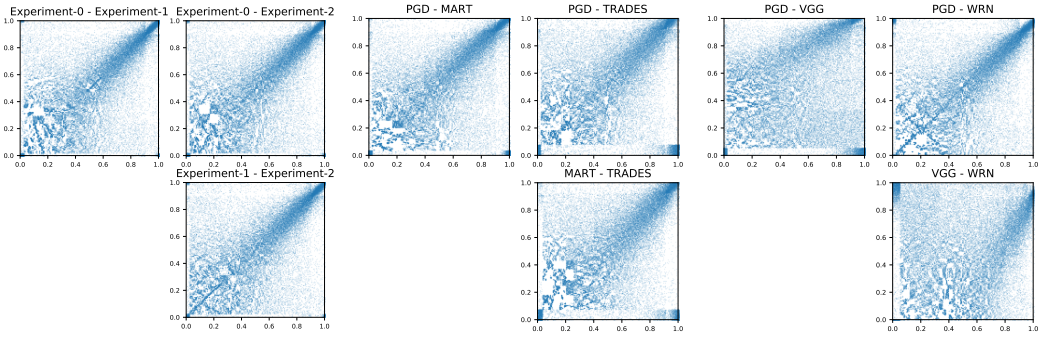
- Standard version, which consists of $\text{APGD}_{\text{CE}}$ (Croce & Hein, 2020b), $\text{APGD}_{\text{DLR}}$ (Croce & Hein, 2020b), FAB (Croce & Hein, 2020a) and Square Attack (Andriushchenko et al., 2020). This is the default version introduced in Croce & Hein (2020b). We use this version to report the numeric results in tables.

[6]The epoch when the best model in terms of robustness is obtained




(a) Random initializations (b) Different methods (c) Different models

Figure 9: Scatter plots of the normalized ranks of training examples based on minimum perturbations obtained by different training settings. The minimum perturbation of an example is consistent across random intializations, different methods and models.




(a) Random initializations (b) Different methods (c) Different models

Figure 10: Scatter plots of the normalized ranks of training examples based on $1^{\text{st}}$ learned epochs obtained by different training settings. The $1^{\text{st}}$ learned epoch refers to the first epoch when an example is classified correctly under adversarial attack, which is consistent across random intializations, different methods and models.

- Custom version, which consists of $\text{APGD}_{\text{CE}}$ and $\text{APGD}_{\text{DLR}}$. This version is faster and consumes less computation resources, and is often sufficient since the FAB-T and Square Attack in the standard version rarely attack any new examples successfully based on our experiments[7]. We use this version to report the results in figures, which typically cover a number of experiments.

**Adversary setting.** We conduct adversarial training with $\ell_\infty$ norm-bounded perturbations. We employ standard PGD training and TRADES as base methods, which can yield state-of-the-art robustness (Gowal et al., 2020). For both methods, the number of attack iterations is fixed as $10$, and the perturbation step size is fixed as $2/255$.

**Training setting.** We employ SGD as the optimizer. The momentum and weight decay are set as $0.9$ and $0.0005$ respectively, which is aligned with the common practice. For pre-activation ResNet, we conduct the training for $160$ epochs, with the learning rate starting at $0.1$ and reduced by a factor of $10$ at epoch $80$ and $120$. For Wide ResNet, we conduct the training for $120$ epochs, with the learning rate starting at $0.1$ and reduced by a factor $10$ at epoch $100$ and $110$. We adopt early stopping (Rice et al., 2020) as a default strategy and report the best robustness obtained throughout the training.

**Dataset.** We conduct experiments on the CIFAR-10 dataset, without additional data.

[7]We occasionally find 1 out of $10^4$ examples in the test set of CIFAR-10 will be attacked successfully by FAB-T or Square Attack while not being attacked successfully by $\text{APGD}_{\text{CE}}$ and $\text{APGD}_{\text{DLR}}$.

**Neural architecture.** We conduct experiments with pre-activation ResNet-18 in the main text, and Wide ResNet in the supplemental results.

**Hardware.** We conduct experiments on a single NVIDIA GeForce GTX 1080 Ti.

## D MORE EXPERIMENT RESULTS

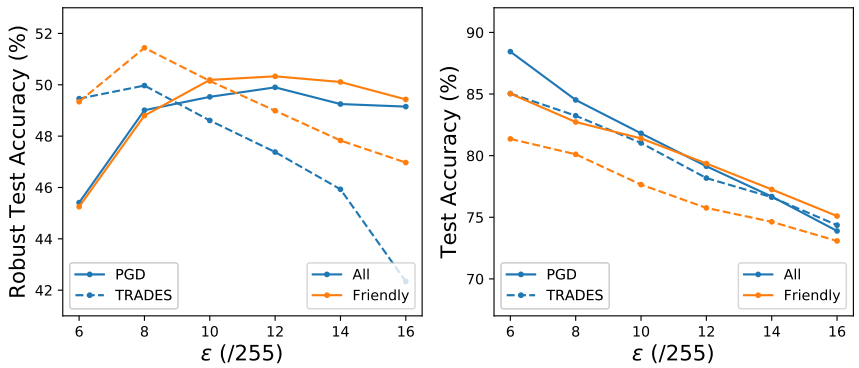



Figure 11: As $\varepsilon$ varies from 6/255 to 16/255, and for both PGD and TRADES, the robustness obtained on the friendly subset mostly outperforms that on the entire training set. Here the size of the friendly subset is fixed as $40 \times 10^3$. The model is WRN-16-10.

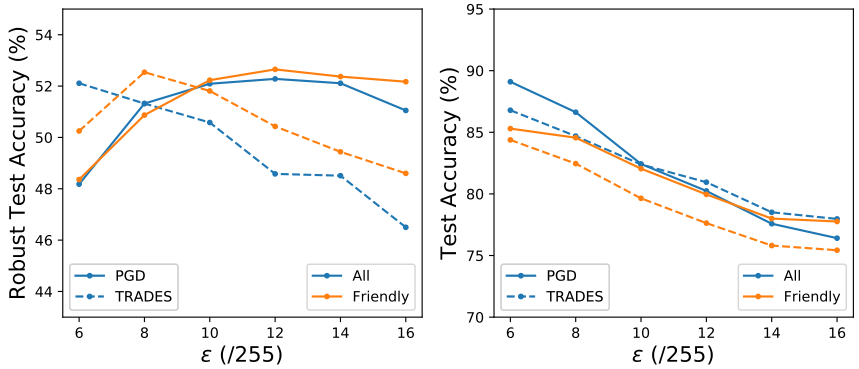



Figure 12: Same as Figure 11, but with a larger model WRN-28-10.

Here we conduct adversarial training on Wide ResNet (WRN) to validate our method. We use the custom version of AutoAttack mentioned in Section C to evaluate the robustness. We choose the size of the friendly subset $k = 40 \times 10^3$ for all experiments. Due to computational constraints, $k$ is not fine-tuned for Wide ResNet, which leaves space for further improvements. Note that when sampling the friendly subset, we still use the problematic ranks estimated based on pre-activation ResNet as mentioned in Section A.2. The effectiveness of our method thus further proves the problematic score is a property only associated with data.

Figure 11 shows the robustness obtained with WRN-16-10, namely the Wide ResNet with depth 16 and widen factor 10. Training on friendly data only can improve the robustness consistently for different methods and perturbation radii. Figure 12 shows the robustness obtained with WRN-28-10. Training on friendly data only can still improve the robustness for such a large model, but the improvement shrinks. The reason may lie in the fact that large network can learn problematic data relatively better. This shows the model capacity is indeed helpful for robust learning, whereas a deep understanding is still worth exploring.

## E ROBUSTNESS-ACCURACY TRADE-OFF

Starting from this section, we investigate the interconnection between problems in adversarial training and the problematic scores of training examples. We use the normalized problematic rank (i.e. problematic rank scaled by the total number of examples in the dataset) as a surrogate for problematic score and conduct extensive analyses based on controlled experiments. Before diving into the first problem, i.e., robustness-accuracy trade-off, we introduce the experimental setup. The same setting also applies to other problems unless otherwise specified.

**Experimental setup.** We conduct experiments on the CIFAR-10 dataset with pre-activation ResNet-18. We employ PGD and TRADES as the adversarial training methods with 10 attack iterations, perturbation radius $8/255$, and perturbation step size $2/255$. To reliably evaluate the robustness, we use a custom version of AutoAttack (see Appendix C), unless otherwise noted. The same experiment settings are applied to Appendix F and G as well.

**The problem and related work.** It is widely observed that adversarially training the model comes at the cost of standard accuracy, which is typically referred as the *robustness-accuracy trade-off* (Papernot et al., 2016; Su et al., 2018; Tsipras et al., 2019; Zhang et al., 2019). For certain learning problems, this trade-off might be inevitable either because no optimal classifier exists (Tsipras et al., 2019; Zhang et al., 2019) or the hypothesis class is not expressive enough (Nakkiran, 2019). The trade-off is shown to be sensitive to the data. Specifically, Tsipras et al. (2019) observed that the adversarial training is actually beneficial to standard accuracy when the sample size is small and insufficient to train the model. Ding et al. (2019) showed that the level of trade-off depends on the data distribution.

**Robustness-accuracy trade-off is correlated with the problematic rank.** We show that robustness-accuracy trade-off is more nuanced—it is sensitive to the individual data examples themselves. On friendly data, this trade-off is not significant, while on problematic data, it is prominent.

Following previous work (Chen et al., 2020), we define the *cross-generalization gap* as the difference between the standard test accuracy obtained by standard training and that obtained by adversarial training. In Figure 1, one may already note that the cross-generalization gap on the friendly partition is significantly smaller than that on the problematic partition, and is only noticeable in the late stage of training after the learning rate is reduced.

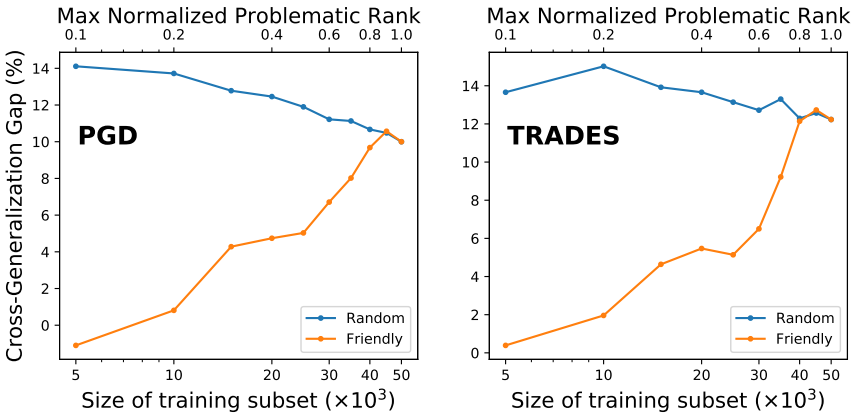



Figure 13: Cross-generalization gap obtained by adversarial training on various sizes of subsets sampled from the training set of CIFAR-10. Compared to a randomly-sampled equal-size subset, a friendly subset consisting of examples with problematic ranks below a threshold produces a consistently smaller cross-generalization gap.

We further verify the correlation between problematic rank and trade-off by experimenting on subsets of training examples with problematic ranks below different thresholds. Figure 13 shows that, compared to a randomly-sampled, equal-size subset, the cross-generalization gap on the friendly training subset is consistently smaller for both PGD and TRADES. Moreover, as the maximum problematic rank of the friendly training subset increases, the cross-generalization gap gradually expands. In contrast, for randomly-sampled subsets, the cross-generalization gap generally decreases.

One may also note that, when the maximum normalized problematic rank is $0.1$ (i.e., the most friendly subset in Figure 13), the cross-generalization gap induced by PGD training is negative, which means that adversarial training is actually beneficial to the standard accuracy. Similar phenomena were only observed on MNIST or extremely small training sets (Tsipras et al., 2019).

**Implication.** Our identified correlation between robustness-accuracy trade-off and problematic rank complements the existing observations of the impact of data distribution on robustness. Besides the training set size (Tsipras et al., 2019) and distribution shift (Ding et al., 2019), the individual characteristics of the training data can also influence this trade-off. Adversarial training can be beneficial to the standard accuracy on a relatively large subset[8] as long as the examples are friendly enough, which implies that the trade-off is not necessarily inevitable at least for a subset of the learning problems. Meanwhile, problematic examples dominate the trade-off potentially because they are intrinsically ambiguous as shown in Figure 3, thus preventing the learning of robust features on them. This is aligned with the understanding of trade-off from the feature learning perspective (Tsipras et al., 2019; Ilyas et al., 2019). More detailed explanation motivated from this perspective will be discussed in Appendix H.1.

In addition, it is the common understanding that robust learning requires a much larger sample complexity compared to standard learning (Schmidt et al., 2018; Raghunathan et al., 2020). We do observe that more data will reduce the trade-off gap when the data is randomly sampled. However, if the extra data is not properly selected, which is likely to happen in practice, the cross-generalization gap will actually expand as shown in Figure 13. Similar findings are made on toy datasets theoretically (Chen et al., 2020).

**Mitigate robustness-accuracy trade-off.** In light of the differentiation of the data in adversarial training, researchers start to customize the attack (Balaji et al., 2019; Ding et al., 2020; Cheng et al., 2020; Zhang et al., 2020a) or loss (Huang et al., 2020; Wang et al., 2020; Zhang et al., 2020b) for each individual example (see more in Appendix B). Although these methods claim the mitigation of trade-off by increasing robustness or accuracy while maintaining the other, the effectiveness is unclear under state-of-the-art robust certification methods (Croce & Hein, 2020b).

## F Robust Overfitting

**The problem and related work.** In standard training, increasing model complexity induced by either larger over-parameterized models or longer training will not hurt the generalizability, which is typically referred as the "double descent" phenomenon (Belkin et al., 2019; Nakkiran et al., 2020). However, it is observed that overfitting does occur in adversarial training (Rice et al., 2020). Specifically, the robust test accuracy will constantly decrease after a certain point in adversarial training, resulting in inferior final performance. Such *robust overfitting* occurs consistently across different datasets, training settings, adversary settings, and neural architectures, and cannot be completely eliminated other than using early stopping (Rice et al., 2020).

**Problematic data causes robust overfitting.** We show that the robust overfitting results from the problematic examples, and the timing of its occurrence during training hinges on the learning order.

Figure 1 already shows that adversarial training on the friendly partition induces almost no robust overfitting. This suggests that robust overfitting is against our conventional understanding of overfitting in the sense that it can be mitigated by a smaller training sample size.

We further extend the experiment by sampling subsets with different maximum problematic rank values. Figure 14 shows that when the maximum normalized problematic rank of the training set is smaller than a particular threshold[9], there is no robust overfitting. As the maximum value surpasses this threshold and increases further, the robust overfitting arises and even exacerbates. In contrast, for randomly sampled subsets, the robust overfitting is prominent consistently across different training sizes. This suggests that problematic examples cause the robust overfitting, and more problematic examples will cause more severe robust overfitting.

[8] Therefore it is parallel to the similar phenomenon observed when the training size is extremely small (Tsipras et al., 2019)

[9] about $0.7$ for PGD and $0.5$ for TRADES

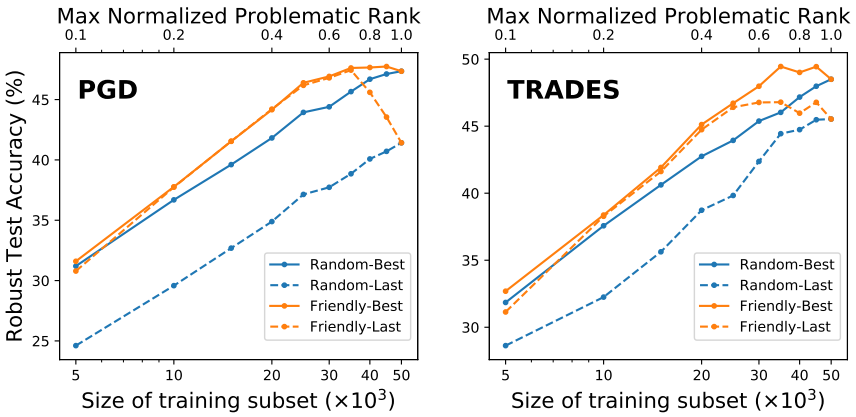
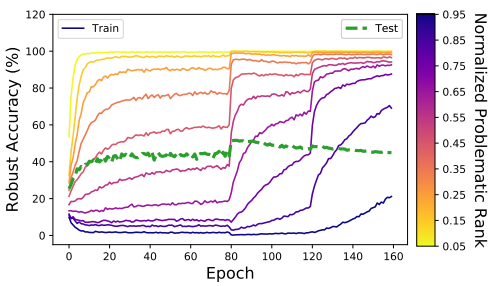


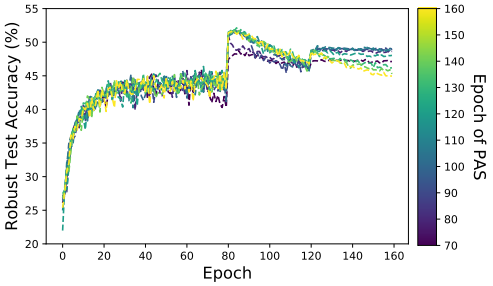

Figure 14: Best and last robust test accuracy obtained when training on various sizes of subsets either sampled randomly ("Random") or selected based on problematic rank ("Friendly"). The robust overfitting emerges only when the maximum problematic rank of the training set is larger than a particular threshold, implying that the problematic examples causes the robust overfitting.

Motivated from the feature learning perspective, we conjecture that the problematic examples cause robust overfitting because robust features might be lost when the problematic examples are adversarially learned. More detailed explanation can be found in Appendix H.2

**Learning order of problematic data determines the timing of robust overfitting.** We demonstrate that robust overfitting emerges in the late stage of training due to the coupling between adversarial training on problematic examples and the learning order (Hacohen et al., 2019). Learning order refers to the phenomenon that the training examples are learned in a similar order throughout the training across different experiments. Learning order also holds in adversarial training, and the problematic examples are always learned late (see Appendix B.3). Therefore, in the late stage of training, as the model is already robust, and starts to adversarially learn the problematic examples, the robust features are ruined, giving rise to the consistent decrease of the robust test accuracy. We visualize this effect by evaluating the robust training accuracy on subsets with different problematic levels. In Figure 15a, we evenly partition the training set into 10 subsets based on problematic rank, and record the robust training accuracy of every subset along with the robust test accuracy. One can find that before the robust overfitting, the robust training accuracies of problematic subsets remain almost constant, while these of friendly subsets increase steadily. However, when the robust overfitting emerges, the robust training accuracies of problematic subsets start to increase drastically; at the same time, those of friendly subsets largely remain the same.




(a) Robust training accuracies of subsets with different problematic ranks. The robust test accuracy (PGD-10) (green, dashed) is also plotted to indicate the robust overfitting. The aggressive learning on highly problematic training examples happens at the same time when the robust overfitting emerges.




(b) Robust test accuracies (PGD-10) when partially early stopping (PAS) the adversarial training at different epochs. The degree of robust overfitting is proportional to the timing of partial early stopping.

Figure 15: Coupling between problematic rank and learning order determines the timing of robust overfitting throughout adversarial training

We further use partial early stopping to reveal the impact of learning order on robust overfitting. Partial early stopping refers to our strategy that the training is early stopped at some epoch, but only for those examples that are not adversarially learned at that specific epoch. The training is conducted on the rest of the examples till the end. In Figure 15b, one can find that the partial early stopping at epoch $80$, which is right before the learning rate decay, induces almost no robust overfitting in the later training. The later the partial early stopping is performed, the more severe the overfitting will be. This phenomenon is due to the fact that problematic examples are gradually learned because of the learning order, and every portion of the problematic examples contributes to part of the robust overfitting.

**Mitigate robust overfitting.** Based on the above understanding, to mitigate robust overfitting, one may need to intentionally prevent the model learning problematic examples. Removing problematic examples from the training set and the early stopping (Rice et al., 2020) are the most straightforward methods to prevent such learning. Beyond that, from the perspective of learnability, possible solutions include reducing the model complexity, adopting heavier regularization, and avoiding learning rate decay, which are partially mentioned in previous work (Rice et al., 2020).

## G GRADIENT MASKING

**The problem and related work.** *Gradient masking* refers to the problem that the gradients of the model are removed, suppressed, or being noisy, which hinders the gradient-based adversary from successfully attacking the model (Papernot et al., 2017; Tramèr et al., 2018; Athalye et al., 2018; Uesato et al., 2018; Engstrom et al., 2018; Mosbach et al., 2018). Broadly speaking, gradient masking can be referred to the general problem of spuriously high robustness against weak adversaries, which frequently appears in existing adversarial defense methods as discussed in a series of reviews (Carlini & Wagner, 2017a; Carlini et al., 2019; Tramèr et al., 2020).

For adversarial training specifically, gradient masking can originate from the improper design of the method. By measuring the local linearity of the loss landscape, Qin et al. (2019) observed that insufficient optimization iteration in the inner maximization will cause gradient masking. Gowal et al. (2020) mentioned that using a margin-based loss (Uesato et al., 2018) will exacerbate gradient masking.

**Problematic data causes gradient masking.** We show that different parts of the data can contribute to the gradient masking differently even for the same method. We use AutoAttack to provably evaluate the robustness. Although AutoAttack not necessarily reflects the true robustness (Tramèr et al., 2020), it serves as a strong baseline compared to PGD attack and is sufficient to reveal the significance of gradient masking relatively.

Figure 1 already shows that adversarial training on the friendly partition yields minimal gradient masking since the robust test accuracy evaluated against PGD-10 attack is close to that evaluated against AutoAttack[10] throughout the training, which is in sharp contrast to the training on the problematic partition.

We further extend the experiments by sampling training subsets with different maximum problematic rank values. We report the best robust test accuracy throughout the training to rule out the effect of robust overfitting. We denote the gradient masking gap as the difference between the robust test accuracies evaluated against PGD-10 attack and AutoAttack. Figure 16 shows that the gradient masking gap expands substantially when the problematic rank of the subset approaches the upper limit (in other words, when more problematic data is added to the training). As a result, although adding more problematic data can improve the robustness against PGD-10, it can hurt the robustness against stronger adversaries such as AutoAttack. Note that when the subset is sampled randomly, the gradient masking gap barely varies and is consistently larger than that of subsets without many problematic data. This demonstrates that the problematic data contributes to gradient masking and causes difficulty for reliable robustness evaluation.

Appendix H.3 shows that the problematic data triggers gradient masking due to a mechanism we refer as "competing", a conflict of optimization targets in the inner maximization.

[10]Here (In Figure 1 only) we conduct our custom version of AutoAttack on a subset composed of $10^3$ randomly sampled test examples due to its heavy computational load to evaluate in every epoch.

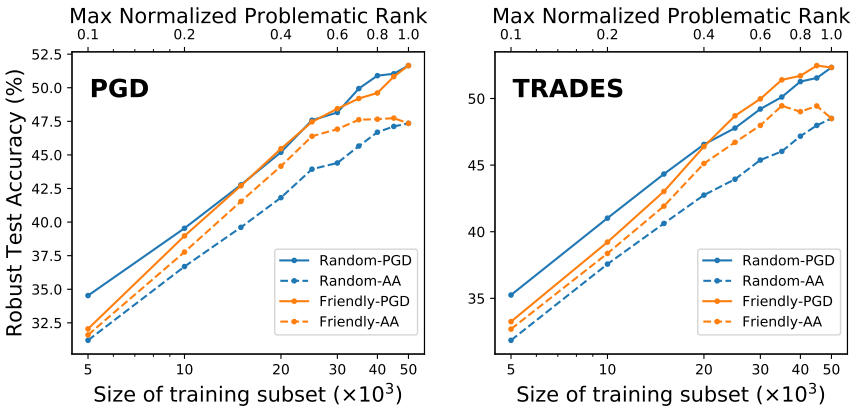



Figure 16: Best robust test accuracy obtained on the subsets sampled either randomly ("Random") or selected based on problematic rank ("Friendly"). We demonstrate the gradient masking by the discrepancy between robustness evaluated against PGD-10 attack and AutoAttack. The gradient masking becomes to exacerbate as highly problematic data is added to the training set.

**Implication.** Since problematic data can cause gradient masking, it can be intentionally exploited to create spuriously high robustness against weak adversaries. Indeed, compared with the standard PGD training, we empirically find that several existing methods will incur significantly higher robust training accuracy on the problematic data, due to their inadequately designed attacks or loss functions that provoke more aggressive learning on the problematic data. As a consequence, notable robustness improvements are often claimed against weak adversaries. However, when it comes to stronger attacks, especially those non-gradient-based ones, there is in fact no significant improvement and even deterioration. Similar gradient masking issues of some methods have been reported in a recent comprehensive study (Gowal et al., 2020).

## H Explanation

In this section, we move one step further to probe into the ruins of problematic data in adversarial training. In previous works, it has been proven that standard classifier and robust classifier learn fundamentally different features (Tsipras et al., 2019). Useful features prevail in the dataset, but are not necessarily robust and comprehensible to human (Ilyas et al., 2019). In light of such analyses of robust/non-robust features, we motivate from a feature learning perspective and try to uncover the potential mechanisms of how problematic data is interconnected with the existing difficulties in adversarial training including robustness-accuracy trade-off, robust overfitting and gradient masking. Under the assumption that similar features will be recognized in a specific class of examples, we selectively conduct the adversarial training on either one or a few classes of examples. In this way, we can isolate the impact of each set of similar features and analyze the effects of problematic data on the learning of features.

### H.1 Robustness-accuracy trade-off

Towards understanding the correlation between robustness-accuracy trade-off and problematic score, we conjecture that problematic data will cause the loss of useful features in adversarial training.

As shown in Figure 3a, problematic examples are intrinsically ambiguous from a human perspective. Therefore, if the perturbation radius is relatively large, the adversarial attack may generate reasonable images of classes other than the true class. Indeed, as shown in 3b, the adversarial counterparts of problematic examples catch salient characteristics of the classes that the model predicts. In contrast, the adversarial counterparts of friendly examples are still explicit images of the true classes. Here, we adversarially attack an example using PGD-20 with perturbation radius $\varepsilon = 16/255$ based on the best model obtained through a PGD training. The perturbation radius $16/255$ is exactly twice the perturbation radius commonly used in adversarial training, which implies that adversarial coun-

terparts generated on the problematic examples during the training might be marginal cases from a human perspective.

Based on the above observation, we suspect that the adversarial training might not be suitable for problematic examples. It is explained in Ilyas et al. (2019) that adversarial training works because the adversary can exploit non-robust features of classes other than the true class, thus forcing the model to rely only on the robust features of the true class. However, due to the ambiguity, such "distracting" features generated on a problematic example might be too prominent such that it overwhelms the regular features of other classes when it is forced to be classified into the true class of this example, thus significantly damage the recognizability of other classes.

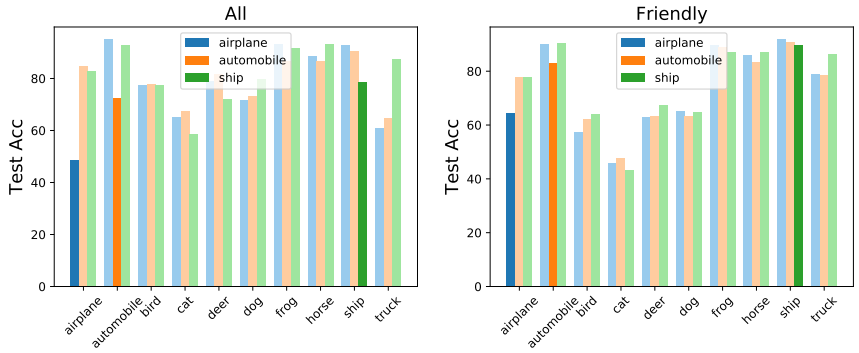



Figure 17: Fine tune the model obtained by standard training with adversarial examples generated by targeted attack on either all the training examples or the friendly subset among them. When the attack in adversarial training is primarily targeted to "Airplane", the test accuracy of "Airplane" after fine-tuning is significantly lower, while the test accuracies of other classes are comparable. Instead, when the attack is targeted to "Automobile", the test accuracy of "Automobile" after fine-tuning is significantly lower. In contrast, when fine tuning the model only on the friendly subset, such difference is not significant.

We manifest such loss of recognizability by robustly fine tuning the model obtained by standard training. Instead of untargeted attack, we use targeted attack in adversarial training[11] to isolate the effect of recognizability loss. Specifically, for the PGD attack employed in adversarial training, we replace the inner maximization by a minimization towards a target class $c$, except for those examples have class $c$ as their true labels, where the target is directed to another class $c''$. We refer this special case as $c - c''$ adversarial training. Figure 17 shows the average standard test accuracy produced by fine-tuning for 30 epochs using Airplane-Truck, Automobile-Truck and Ship-Truck adversarial training. These classes are selected because the model produces highest standard test accuracy on them. One can find that, when the attack in adversarial training is primarily targeted to "Airplane", the test accuracy of "Airplane" after fine-tuning is significantly lower, while the test accuracies of other classes are comparable. Instead, when the attack is targeted to "Automobile", the test accuracy of "Automobile" after fine-tuning is significantly lower. In contrast, when fine tuning the model only on the friendly examples, such difference is not significant. This reflects the detail of how problematic examples in adversarial training hurts the learning of useful features.

### H.2 Robust overfitting

In previous section, we mentioned the problematic examples may damage the recognizability in adversarial training. Here, we further show that the problematic examples will also damage the robust recognizability. In adversarial training, the gradient-based adversary may generate robust features of another class on problematic examples due to their intrinsic ambiguity, especially when the model is already relatively robust. Consequently, the model may lose the robust recognizability of other classes because it is forced to classify such robust feature to the original class, which leads to robust overfitting.

[11] It is not common to use targeted attack in adversarial training, only for demonstration here.

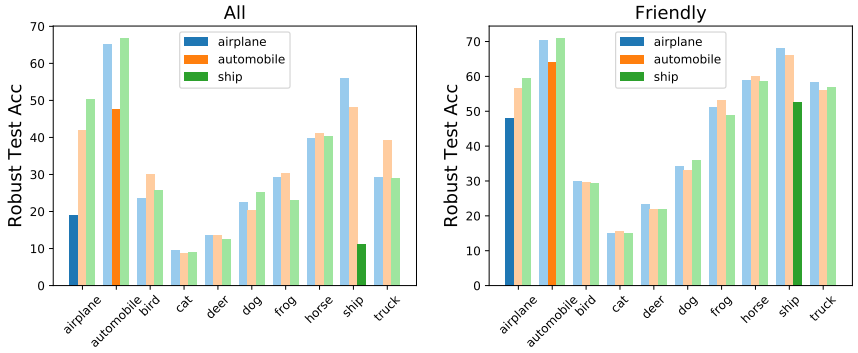



Figure 18: Fine tune the best model obtained in adversarial training with targeted attack on either all the training examples or the friendly subset among them. When the attack in adversarial training is primarily targeted to "Airplane", the robust test accuracy of "Airplane" after fine-tuning is significantly lower, while the test accuracies of other classes are comparable. Instead, when the attack is targeted to "Automobile", the robust test accuracy of "Automobile" after fine-tuning is significantly lower. In contrast, when fine tuning the model only on the friendly subset, such difference is not significant.

We manifest the loss of recognizability by fine-tuning the best model obtained by a regular adversarial training. Similarly, we use targeted attack to isolate the effect. Figure 18 shows the average robust test accuracy produced by fine-tuning the best model for 30 epochs using Airplane-Truck, Automobile-Truck and Ship-Truck adversarial training. These classes are selected because the model produces highest robust test accuracy on them. One can find similar results that, when the attack in adversarial training is primarily targeted to "Airplane", the robust test accuracy of "Airplane" after fine-tuning is significantly lower. Instead, when the attack is targeted to "Automobile", the robust test accuracy of "Automobile" after fine-tuning is significantly lower. In contrast, when fine tuning the model only on the friendly data, such difference is not prominent.

### H.3 Gradient masking

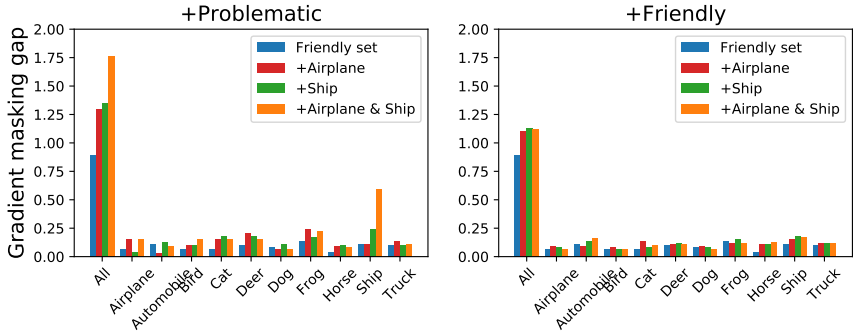



Figure 19: The gradient masking gap, namely the difference between PGD-10 and Auto Attack evaluation, generated by adding examples of two competing classes into a friendly subset.

We show that the problematic data causes gradient masking through a mechanism which we refer as "competing". In Figure 19, we sample a subset by adding 500 additional problematic examples[12] to a class-balanced friendly subset of size $10^4$, adversarially train the model on it, and show the gradient masking gap. One can find that when we only add examples of one class either "Airplane" or "Ship", the gradient masking gap increases, but not significantly compared to the gradient masking gap of the original friendly subset. However, if we add the examples of two "competing" classes "Airplane" and "Ship" at the same time, the gradient masking gap increases substantially, and mostly attributes to the class "Ship". Here, "competing" classes means these two classes contain images that are likely

[12] 10% of all the training examples in one class

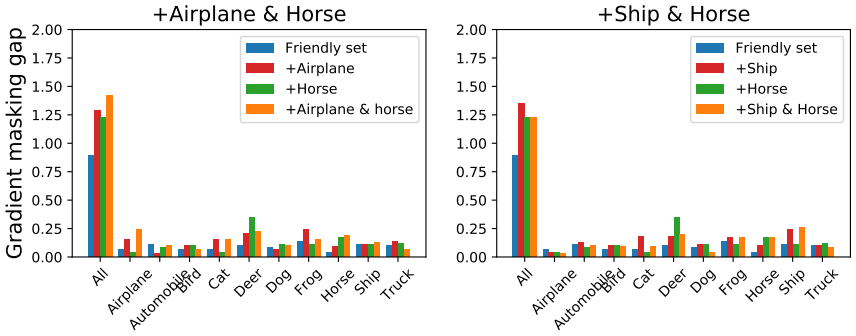



Figure 20: The gradient masking gap generated by adding examples of two non-competing classes into a friendly subset.

to have similar features[13]. The gradient masking gap is not significant if the additional examples are not from two "competing" classes, as shown in Figure 20 where we add additional examples from either "Airplane" and "Horse", or "Ship" and "Horse". The gradient masking gap is also not significant if the additional examples are not problematic, as shown in Figure 19 where we add additional friendly examples, even if they are from competing classes.

Towards understanding this mechanism, we focus on the inner maximization since gradient masking mainly indicates weakened generation of adversarial perturbation, while previously we assume this process is ideal. Recall that adversarial attack works because in the inner maximization, the adversary optimizes towards classes other than the original class and thus can exploit distracting features to fool the model (Ilyas et al., 2019). However, since the ambiguous problematic examples from different classes tend to contain similar features, the adversary may optimize towards different classes starting from similar features, subsequently damage its capability to exploit distracting features of these classes.

Considering from a geometrical perspective, the mechanism is more aligned with the common understanding of gradient masking. As problematic examples locate close to the decision boundary (see Appendix B), their $\ell_p$ perturbation balls may overlap with each other, thus severely complicate the decision boundary and obfuscate the gradients. We will discuss more from the geometrical perspective in the future work.

# I More related work

## I.1 Robustness-accuracy trade-off

We note that there exists an abundant body of works that focuses on the robustness-accuracy trade-off in robust learning. In additional to the related works discussed in the main paper, here we review some works that attack this problem from other perspectives. This is by no means an exhaustive review.

It has been argued that the robustness-accuracy trade-off is inherent to the data distribution and is thus inevitable for any classifier. Tsipras et al. (2018; 2019) and Zhang et al. (2019) theoretically show that no optimal classifier can achieve both robustness and accuracy on toy problems. Dohmatob (2018) formalizes this into a "No Free Lunch" problem, and further proves the inevitability of trade-off under mild assumptions of the data distribution. Nakkiran (2019) shows that the trade-off is inevitable because the hypothesis class is not expressive enough. Javanmard et al. (2020) shows that the adversarial training may improve generalization in an over-parameterized regime, but hurt it in under-parameterized regime.

On the contrary, some works argue that the robustness-accuracy trade-off is not necessarily inevitable in a realistic setting. Raghunathan et al. (2020) shows that the trade-off stems from the over-parameterization of the hypothesis class. Robust self-training (Carmon et al., 2019; Najafi

[13] One can refer to the sample images we showed in Figure 3

et al., 2019; Uesato et al., 2019), overcoming the sample complexity leveraging additional unlabeled data, thus can effectively mitigate the trade-off. Yang et al. (2020b) shows that the trade-off in practice is a result of either the model failing to impose local Lipschiztness, or not generalizing sufficiently. Wen et al. (2020) and Roth et al. (2020) show that adversarial training is essentially a form of operator norm regularization, thus hurting the generalizability if not properly configured.

There are more works implying that the robustness and accuracy may not be contradict by showing that adversarial examples can benefit generalization either through different perturbation generation strategies (Stutz et al., 2019) or different adversarial training strategies (Xie et al., 2020).

### I.2 Data profiling in standard learning

In classical machine learning, data profiling is important because algorithms may be sensitive to noise and outliers. By measuring the degree of class overlapping and skewness in a dataset, Smith et al. (2013) proposes a generic definition of instance hardness representing how likely an example will be misclassified. Prudêncio et al. (2015) motivates from item response theory (IRT) (Embretson & Reise, 2000; Ayala, 2008) to characterize instance hardness. Smith & Martinez (2011) shows that properly removing hard examples in the dataset improves performance for a variety of learning algorithms including but not limited to Decision Tree and Support Vector Machine.

In deep learning regime, models with large capacity are typically more robust to outliers. Nevertheless, data examples can still exhibit diverse levels of difficulties. Arpit et al. (2017) finds that data examples are not learned equally when injecting noisy data into training. Toneva et al. (2019) shows that certain examples are forgotten frequently during training, which means that they can be first classified correctly then incorrectly. Model performance can be largely maintained when removing those forgettable examples from training. Zhou et al. (2020) proposes to dynamically estimate instance hardness during training and encourage the model to focus on those hard examples from a curriculum learning (Bengio et al., 2009) perspective, which can improve both the performance and efficiency for a wide range of datasets and neural architectures. More generally, under a self-paced learning (Kumar et al., 2010) framework, diverse methods have been proposed to mine hard examples on the fly (Chang et al., 2017).

### I.3 Data profiling in robust learning

In robust learning regime, the model is required to learn features that are robust to perturbations. Such task is generally more difficult, where the data examples are thus more likely to differentiate in terms of their behaviours in learning or contribution to the model performance. Plenty of works have analyzed the diverse behaviours of data during adversarial training, and proposed a variety of methods that treating the data examples differently. A detailed review has been made in Section 2, and Section E, F, G that focus on existing problems in adversarial training specifically.

Here we mainly review works that investigate adversarial examples from a data perspective. Hendrycks et al. (2019b) collects a set of unperturbed images, known as the "natural adversarial examples", that significantly degrades the performance of state-of-the-art image classifiers. Pestana et al. (2020a) observes that the adversarial perturbations concentrate in the Y-channel of the YCbCr space, which is believed to contain more shape and texture related information. Pestana et al. (2020b) reports the existence of a set of images that are particularly robust to adversarial perturbations. Including such data in validation significantly limits the reliability of the robustness evaluation.