Relating Adversarially Robust Generalization to Flat Minima

David Stutz\(^1\)    Matthias Hein\(^2\)    Bernt Schiele\(^1\)
\(^1\)Max Planck Institute for Informatics, Saarbrücken, \{dstutz,schiele\}@mpi-inf.mpg.de
\(^2\)University of Tübingen, Tübingen, matthias.hein@uni-tuebingen.de

Abstract

Adversarial training (AT) has become the de-facto standard to obtain models robust against adversarial examples. However, AT exhibits severe robust overfitting: cross-entropy loss on adversarial examples (robust loss) decreases continuously on training examples, while eventually increasing on test examples. This leads to poor robust generalization, i.e., low adversarial robustness on new examples. We study the relationship between robust generalization and flatness of the robust loss landscape in weight space, i.e., whether robust loss changes significantly when perturbing weights. To this end, we propose a metric to measure “robust flatness” and find a strong correlation between good robust generalization and flatness. Throughout training, flatness reduces during overfitting, i.e., early stopping effectively finds flatter minima. Similarly, AT variants such as AT-AWP or TRADES and simple regularization techniques such as AutoAugment or label noise that improve robustness also correspond to flatter minima.

1. Introduction

In order to obtain robustness against adversarial examples [36], adversarial training (AT) [26] augments training with adversarial examples generated on-the-fly. AT is known to require more training data [21, 31], generally leading to generalization problems [11]. Robust overfitting [30] has been identified as the main obstacle: adversarial robustness on test examples eventually starts to decrease, while robustness on training examples continues to increase (cf. Fig. 2). This is typically observed as increasing robust loss (RLoss) or robust test error (RErr), i.e., (cross-entropy) loss and test error on adversarial examples. As a result, the robust generalization gap, i.e., the difference between test and training robustness, tends to be large. [30], uses early stopping as a simple strategy to avoid robust overfitting. Nevertheless, despite recent work [32, 39, 17], it remains an open and poorly understood problem.

In “clean” generalization (i.e., on natural examples), overfitting is well-studied and commonly tied to flatness of the loss landscape in weight space, both visually [24] and empirically [28, 20, 19]. In general, the optimal weights on test examples do not coincide with the minimum found on training examples. Flatness ensures that the loss does not increase significantly in a neighborhood around the found minimum. Therefore, flatness leads to good generalization because the loss on test examples does not increase significantly (i.e., small generalization gap, cf. Fig. 3, right). [24] showed that visually flatter minima correspond to better generalization. [28, 20] formalize this idea by measuring the change in loss within a local neighborhood. Furthermore, explicitly encouraging flatness during training has been shown to be successful in practice [42, 4, 25, 3, 18].

Recently, [39] applied the idea of flat minima to AT: through adversarial weight perturbations, AT is regularized to find flatter minima of the robust loss landscape. This reduces the impact of robust overfitting and improves robust
generalization, but does not avoid robust overfitting. As result, early stopping is still necessary. Unfortunately, flatness is only assessed visually. Similarly, [12] shows that weight averaging [18] improves robust generalization, indicating that flatness might be beneficial in general. This raises the question whether other “tricks” [29, 12], e.g., different activation functions [32], label smoothing [35], or approaches such as AT with self-supervision [15]/unlabeled examples [2] are successful because of finding flatter minima.

Contributions: We study whether flatness of the robust loss (RLoss) in weight space improves robust generalization. To this end, we propose a scale-invariant [8] flatness measures for the robust case and show that robust generalization generally improves alongside flatness and vice-versa: Fig. 1 plots RLoss (lower is more robust, y-axis) against flatness in RLoss (lower is flatter, x-axis), showing a clear relationship. This trend covers a wide range of AT variants on CIFAR10 [39, 40, 37, 15, 2, 1] and various regularization schemes, including AutoAugment [7], label smoothing/noise [35] or weight clipping [33]. Furthermore, we consider hyper-parameters such as learning rate schedule, weight decay or activation functions [9, 27, 14], and methods explicitly improving flatness [3, 18].

This paper is a short version of [34]. It is intended to be self-contained, but we refer to [34] for further discussion.

2. Robust Generalization and Flat Minima

We consider robust generalization and overfitting in the context of flatness of the robust loss landscape in weight space, i.e., w.r.t. changes in the weights. While flat minima have consistently been linked to standard generalization [16, 24, 28, 20], this relationship remains unclear for adversarial robustness. We briefly provide some background and discuss robust overfitting before introducing our flatness measure based on the change in robust loss along ran-

Figure 2: Robust Overfitting: Robust loss (RLoss, left) and robust error (RErr, right) over normalized epochs on CIFAR10. Left: Training RLoss (light blue) reduces continuously throughout training, while test RLoss (dark blue) eventually increases again. Robust overfitting is not limited to incorrectly classified examples (green), but also affects correctly classified ones (rose). Right: Similar behavior, but less pronounced, can be observed considering RErr. We also show RErr obtained through early stopping (red).

Figure 3: Measuring Flatness. Left: Measuring flatness in a random direction (blue) by computing the difference between RLoss ˜L after perturbing weights (i.e., w + ν) and the “reference” RLoss ˜L given a local neighborhood Bξ(w) around the found weights w, see Sec. 2.1. In practice, we average across several random directions. Right: Large changes in RLoss around the “sharp” minimum causes poor generalization from training (black) to test examples (red).
Flatness Throughout Training. Test RLoss (y-axis) plotted against flatness in RLoss (x-axis) during training, showing a clear correlation. AT with self-supervision reduces the impact of robust overfitting (RLoss increases less) and simultaneously favors flatter minima. This behavior is pronounced for AT-AWP, explicitly optimizing flatness, and AT with additional unlabeled examples.

landscape on test examples changes, loss remains small, ensuring good generalization. The contrary case is illustrated in Fig. 3 (right). The easiest way to “judge” flatness is visual inspection, e.g., following [24], where the loss landscape is visualized along random directions after normalizing the weights per-filter. The normalization is important to handle difference scales (cf. Fig. 4), i.e., weight distributions, and allows comparison across models. However, as shown in Fig. 4, judging flatness visually is difficult: Considering random weight directions, AT with Adam [22] or MiSH [27] improves adversarial robustness (lower RErr vs. AutoAttack [6]) but do not result in (visually) flatter minima. In contrast, AT-AWP [39] or Entropy-SGD [3] improve robustness and flatness.

Average-Case Flatness: Thus, to objectively measure and compare flatness, we draw inspiration from [28] and propose an “average-case” flatness measures adapted to

Flatness Across Hyper-Parameters: RLoss (y-axis) vs. flatness (x-axis) for selected methods and hyper-parameters (cf. supplementary material). For example, we consider different strengths of weight decay (rose) or sizes $\xi$ of adversarial weight perturbations for AT-AWP (orange). For clarity, we plot (dotted) lines representing the trend per method. Clearly, improved adversarial robustness, i.e., low RLoss, is related to improved flatness.

the robust loss. Considering random weight perturbations $\nu \in B_\xi(w)$ within the $\xi$-neighborhood of $w$, flatness is computed as

\[
\mathbb{E}_\nu \left[ \max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}(f(x+\delta; w+\nu), y) \right] - \max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}(f(x+\delta; w), y)
\]  

averaged over test examples $x$, $y$, as illustrated in Fig. 3. We define $B_\xi(w)$ using relative $L_2$-balls per layer as in [39]:

\[
B_\xi(w) = \{ w + \nu : \|\nu\|_2 \leq \xi \|w(t)\|_2 \text{ layers } l \}.
\]

Note that the second term in Eq. (1), i.e., the “reference” robust loss, is important to make the measure independent of the absolute loss (i.e., corresponding to the vertical shift in Fig. 3, left). In practice, $\xi$ can be as large as 0.5. We refer to Eq. (1) as flatness in RLoss. By construction, Eq. (2) is scale-invariant as the weight neighborhood is defined relative to the $L_2$ norm of the weights.

3. Experiments

We conduct experiments on CIFAR10 [23], where our AT baseline uses ResNet-18 [13] and is trained using SGD and a multi-step learning rate schedule. For PGD, we use 7 iterations and $\epsilon = 8/255$ for $L_\infty$ adversarial examples. PGD-7 is also used for early stopping on the last 500 test examples. We do not use early stopping by default. For evaluation on the first 1000 (balanced) test examples, we run PGD with 20 iterations, 10 random restarts to estimate RLoss and AutoAttack [6] to estimate RErr. In Eq. (1), we use 10 random weight perturbations with $\xi = 0.5$. We consider various AT variants, hyper-parameters and optimization strategies as summarized in Tab. 1. We also use models from RobustBench [5], obtained using early stopping.
3.1. Robust Generalization and Flatness in RLoss

Recent work [39, 12], and Tab. 1 (fourth column), suggest that robust overfitting can be mitigated using regularization. We hypothesize that this is because strong regularization helps to find flatter minima in the RLoss landscape.

Flatness in RLoss “Explains” Overfitting: Considering Fig. 5, we find that flatness reduces significantly during robust overfitting. Namely, flatness “explains” the increased RLoss caused by overfitting very well. We explicitly plot RLoss (y-axis) against flatness in RLoss (x-axis) across epochs (dark blue to dark red): RLoss and flatness clearly worsen “alongside” each other during overfitting. Methods such as AT with self-supervision, AT-AWP or AT with unlabeled examples avoid both robust overfitting and sharp minima (right). This relationship generalizes to different hyper-parameter choices of these methods: Fig. 6 plots RLoss (y-axis) vs. flatness (x-axis) across different hyper-parameters. Again, e.g., for TRADES or AT-AWP, hyper-parameters with lower RLoss also correspond to flatter minima. In fact, Fig. 6 indicates that the connection between robustness and flatness also generalizes across different methods (and individual models).

Improved Robustness Through Flatness: Indeed, across all trained models, we found a strong correlation between robust generalization and flatness. Here, we mainly consider RLoss to assess robust generalization as improvements in RLoss above ~2.3 have, on average, only small impact on RErr (for 10 classes). Pushing RLoss below 2.3, in contrast, directly translates to better RErr. This is illustrated in Fig. 7 which plots RErr vs. RLoss for all evaluated models. To avoid this “kink” in the dotted red lines around RLoss∼2.3, Fig. 1 plots RLoss (y-axis) against average-case flatness in RLoss (x-axis), highlighting selected models. This reveals a clear correlation between robustness and flatness: More robust methods, e.g., AT with unlabeled examples or AT-AWP, correspond to flatter minima. Similarly, methods improving flatness, e.g., Entropy-SGD, weight decay or weight clipping, improve adversarial robustness. Note that Fig. 1 highlights selected models from literature (colored), e.g., from [5] obtained with early stopping, while the described relationship is mostly observed across models without early stopping and with varying hyper-parameters, cf. Fig. 4. We found that this also translates to RErr, subject to the described bend at RLoss≈2.3. These results are summarized in tabular form in Tab. 1: Grouping methods by good, average, or poor robustness, we find that methods need some degree of flatness to be successful. Overall, flatness in RLoss has clear advantages in terms of robust generalization, i.e., low RLoss on test examples.

4. Conclusion

We studied the relationship between adversarial robustness, also considering robust overfitting [30], and flatness of the robust loss (RLoss) landscape w.r.t. random perturbations in the weight space. We introduced a scale-invariant measure of robust flatness and considered popular adversarial training (AT) variants, e.g., TRADES [40], MART [37], AT-AWP [39] AT with self-supervision [15] or additional unlabeled examples [2]. Our experiments reveal a clear relationship between adversarial robustness and flatness in RLoss: more robust methods predominantly find flatter minima and, vice versa, approaches known to improve flatness help AT improve robustness.
References


[18] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In UAI, 2018. 1, 2, 4


[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. ICLR, 2018. 1, 2


[34] David Stutz, Matthias Hein, and Bernt Schiele. Relating adversarially robust generalization to flat minima. arXiv.org, abs/2104.04448, 2021. 2


