IMBALANCED GRADIENTS: A NEW CAUSE OF OVERESTIMATED ADVERSARIAL ROBUSTNESS

Linxi Jiang[1]*, **Xingjun Ma**[2]*†, **Zejia Weng**[1], **James Bailey**[3], **Yu-Gang Jiang**[1]
[1]Fudan University, Shanghai, China
[2]Deakin University, Geelong, Australia
[3]The University of Melbourne, Victoria, Australia

ABSTRACT

Evaluating the robustness of a defense model is a challenging task in adversarial robustness research. In this paper, we identify a more subtle situation called *Imbalanced Gradients* that can also cause overestimated adversarial robustness. The phenomenon of imbalanced gradients occurs when the gradient of one term of the margin loss dominates and pushes the attack towards a suboptimal direction. To exploit imbalanced gradients, we formulate a *Margin Decomposition (MD)* attack that decomposes a margin loss into individual terms and then explores the attackability of these terms separately via a two-stage process. We examine 12 state-of-the-art defense models, and find that models exploiting label smoothing easily cause imbalanced gradients, and on which our MD attacks can decrease their PGD robustness (evaluated by PGD attack) by over 23%. For 6 out of the 12 defenses, our attack can reduce their PGD robustness by at least 9%. The results suggest that imbalanced gradients need to be carefully addressed for more reliable adversarial robustness. Our code is available at `https://github.com/Jack-lx-jiang/MD_attacks`.

1 INTRODUCTION

Deep neural networks (DNNs) are vulnerable to adversarial examples, which are input instances crafted by adding small adversarial perturbations to natural examples. (Szegedy et al., 2014; Goodfellow et al., 2015). A number of defenses have been proposed to overcome this vulnerability. However, a concerning fact is that many defenses have been quickly shown to have undergone incorrect or incomplete evaluation (Carlini and Wagner, 2017; Athalye et al., 2018; Engstrom et al., 2018; Uesato et al., 2018; Mosbach et al., 2018; He et al., 2018). In this work, we identify a new situation called *Imbalanced Gradients* that exists in several state-of-the-art defense models and can cause highly overestimated robustness.

Imbalanced gradients is a new type of gradient masking effect where the gradient of one loss term dominates that of other terms. This causes the attack to move toward a suboptimal direction. Different from obfuscated gradients, imbalanced gradients are more subtle and are not detectable by the detection methods used for obfuscated gradients. To exploit imbalanced gradients, we propose a novel attack named *Margin Decomposition (MD)* attack that decomposes the margin loss into two separate terms, and then exploits the attackability of these terms via a two-stage attacking process. We derive MD variants of traditional attacks like PGD and MultiTargeted (MT) (Gowal et al., 2019), and deploy these MD attacks to re-examine the robustness of 12 adversarial training-based defense models. We find that 6 of them are susceptible to imbalanced gradients, and their robustness originally evaluated by the PGD attack drops significantly against our MD attacks.

*Equal contribution.
†Corresponding authors.

2 IMBALANCED GRADIENTS AND MARGIN DECOMPOSITION ATTACK

We denote a clean sample by $\mathbf{x}$, its class by $y \in \{1, \cdots, C\}$ with C the number of classes, and a DNN classifier by f. The probability of $\mathbf{x}$ being in the i-th class is computed as $\mathbf{p}_i(\mathbf{x}) = e^{\mathbf{z}_i} / \sum_{j=1}^{C} e^{\mathbf{z}_j}$, where $\mathbf{z}_i$ is the logits for the i-th class. The goal of adversarial attack is to find an adversarial example $\mathbf{x}_{adv}$ that can fool the model into making a false prediction (*e.g.* $f(\mathbf{x}_{adv}) \neq y$), and is typically restricted to be within a small ϵ-ball around the original example $\mathbf{x}$ (*e.g.* $\|\mathbf{x}_{adv} - \mathbf{x}\|_\infty \leq \epsilon$).

Imbalanced Gradients. The gradient of the margin loss (*e.g.* $\ell_{margin}(\mathbf{x}, y) = \mathbf{z}_{max} - \mathbf{z}_y$) is the combination of the gradients of its two individual terms (*e.g.* $\nabla_{\mathbf{x}}(\mathbf{z}_{max} - \mathbf{z}_y) = \nabla_{\mathbf{x}}\mathbf{z}_{max} + \nabla_{\mathbf{x}}(-\mathbf{z}_y)$). *Imbalanced Gradients* is the situation where the gradient of one loss term dominates that of other term(s), pushing the attack towards a suboptimal direction.

Toy Example. Consider a one-dimensional classification task and a binary classifier with two outputs $\mathbf{z}_1$ and $\mathbf{z}_2$ (like logits of a DNN), Figure 1 illustrates the distributions of $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_2 - \mathbf{z}_1$ around $x = 0$. The classifier predicts class 1 when $\mathbf{z}_1 \geq \mathbf{z}_2$, otherwise class 2. We consider an input at $x = 0$ with correct prediction $y = 1$, and a maximum perturbation constraint $\epsilon = 2$ (*e.g.* perturbation $\delta \in [-2, +2]$). The attack is successful if and only if $\mathbf{z}_2 > \mathbf{z}_1$. In this example, imbalanced gradients occurs at $x = 0$, where the gradients of the two terms $\nabla_x \mathbf{z}_2$ and $\nabla_x(-\mathbf{z}_1)$ have opposite directions, and the attack is dominated by the $\mathbf{z}_1$ term as $\nabla_x(-\mathbf{z}_1)$ is significantly larger than $\nabla_x \mathbf{z}_2$. Thus, attacking x with the margin loss will converge to +2, where the sample is still correctly classified. However, for a successful attack, x should be perturbed towards -2. In this particular scenario, the gradient $\nabla_x \mathbf{z}_2 < 0$ alone can provide the most effective attack direction. Note that this toy example was motivated by the loss landscape of DNNs when imbalanced gradient occurs.

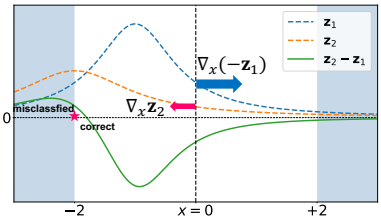


Figure 1: A toy illustration of *imbalanced gradients* at $x = 0$: the gradient of margin loss ($\mathbf{z}_2 - \mathbf{z}_1$) is dominated by its $-\mathbf{z}_1$ term, pointing to a suboptimal attack direction towards +2, where x is still correctly classified.

2.1 MARGIN DECOMPOSITION ATTACK

The above observations motivate us to exploit the individual terms in the margin loss so that the imbalanced gradients situation can be circumvented. Specifically, we propose Margin Decomposition (MD) attack that decomposes the attacking process with a margin loss into two stages: 1) alternately attacking the two individual terms (*e.g.* $\mathbf{z}_{max}$ or $-\mathbf{z}_y$) at different restarts; then 2) attacking the full margin loss. Formally, our MD attack and its loss functions in each stage is defined as follows:

$$\mathbf{x}_{k+1} = \Pi_\epsilon(\mathbf{x}_k + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \ell_k^r(\mathbf{x}_k, y))), \tag{1}$$

$$\ell_k^r(\mathbf{x}_k, y) = \begin{cases} \mathbf{z}_{max} & \text{if } k < \frac{K}{2} \text{ and } r \bmod 2 = 0 \\ -\mathbf{z}_y & \text{if } k < \frac{K}{2} \text{ and } r \bmod 2 = 1 \\ \mathbf{z}_{max} - \mathbf{z}_y & \text{if } k \geq \frac{K}{2}, \end{cases}$$

where, Π is the projection operation that projects the perturbed sample back within ϵ-ball, $k \in \{1, \ldots, K\}$ is the perturbation step, $r \in \{1, \ldots, n\}$ is the r-th restart, mod is the modulo operation for alternating optimization, and ℓ_k^r defines the loss function used at the k-th step and r-th restart. The loss function switches from the individual terms back to the full margin loss at step $\frac{K}{2}$. The first stage exploits individual loss terms to rebalance the imbalanced gradients, while the second stage ensures that the final objective (*e.g.* maximizing the classification error) is achieved. Note that, not all defense models have the imbalanced gradients problem. A model is susceptible to imbalanced gradients if there is a substantial difference between robustness evaluated by PGD attack and that by our MD attack. In addition, to increase attack's diversity, we initialize the perturbation in the first stage by perturbing one step with size $2 \cdot \epsilon$ along the opposite direction of the other loss terms that are left unexplored. We also propose a Margin Decomposition Multi-Targeted (MDMT) attack, a multi-targeted version of our MD attack. Like the MT attack (Gowal et al., 2019), MDMT will attack each possible target class one at a time, then select the strongest adversarial example at the end. The complete algorithms of MD and MDMT can be found in Appendix A, and an ablation study can be found in Appendix C.

3 EXPERIMENTS

We apply our MD attacks to evaluate the robustness of 12 state-of-the-art defense models. We focus on adversarial training models, which are arguably the strongest defense approaches to date (Athalye et al., 2018; Croce and Hein, 2020). All the models are WideResNet variants (Zagoruyko and Komodakis, 2016) and are trained against perturbation $\epsilon = 8/255$ on CIFAR-10. For each defense model, we either download their shared models or retrain the models using the official implementations, unless explicitly stated. Further details about the models can be found in Appendix B. We apply current state-of-the-art attacks and our MD attacks to evaluate the robustness of these models in a white-box setting.

Baseline Attacks and Settings. Following the current literature, we consider 6 existing attacks: 1) FGSM, 2) PGD, 3) L_∞ version of CW attack (Madry et al., 2018; Wang et al., 2019), 4) MultiTargeted (MT) attack and two concurrently proposed attackss 5) AutoAttack (AA), and 6) Output Diversified Initialization (ODI). The evaluation is done under the same maximum perturbation $\epsilon = 8/255$ for training. As AA is an ensemble of four different attacks, we compare our MD attacks with these four attacks independently. We show the best result of individual attacks from AA in Table 1 and the full result of AA are reported in Appendix E. For attacks of AA and ODI, we use the official implementation and parameter setting. For regular iterative attacks, we set the step size to $\alpha = \epsilon/4$ and the total perturbation steps to $K = 40$. For our MD and MDMT, we use a large step size $\alpha = 2 \cdot \epsilon$ in the first stage for a better exploration and $\alpha = \epsilon/4$ in the second stage to ensure a stable optimization for the final objective. For regular attacks PGD, CW and our MD, we use 2 random restarts, while for more powerful attacks ODI, MT and MDMT, we use 20 restarts (MT attacks require more restarts to explore multiple target classes). A parameter analysis of our MD attack can be found in Appendix D. Adversarial robustness is measured by the model accuracy on adversarial examples crafted by these attacks on CIFAR-10 test images.

3.1 EVALUATION RESULTS

Table 1 reports the full evaluation result, where RST, UAT and TRADES are the top 3 best defenses. The SAT defense demonstrates $\sim 45\%$ robustness consistently against either PGD or stronger attacks such as MT, attacks of AA, ODI and our MD attacks. This indicates that SAT does not have imbalanced gradients and indeed brings consistent robustness, which is in line with other studies about SAT (Athalye et al., 2018; Croce and Hein, 2020; Uesato et al., 2018). While the rest 11 defense models are all developed based on SAT, they exhibit quite different robustness. Only 4 defenses including RST, UAT, TARDES and MART are indeed improved over SAT, while the other 7 defense models are actually not as robust as SAT, according to our MD or MDMT attacks. For the 4 improved defenses, their PGD robustness (*e.g.* robustness evaluated by PGD attack) can still be reduced by stronger attacks MT, attacks of AA, ODI or our MD attacks. Considering that their robustness drops against our MD attacks are within 5%, their drops may be caused by sufficient explorations such as more random restarts or better initialization rather than imbalanced gradients. Indeed, MT, attacks of AA, and ODI with more random restarts, multiple target classes, and better initialization can also reduce their robustness to the same level as our MD attacks.

Out of the 7 unimproved defenses, our MDMT attack can reduce the PGD robustness of 6 models (*e.g.* MMA, Bilateral, Adv-Interp, FeaScatter, Sense, and JARN-AT11) by at least 9%. On all 7 unimproved defenses, our MD attacks are always the most effective attacks compared to either classic attacks FGSM, PGD, CW, or more recent attacks MT, attacks of AA and ODI. Note that, for 4 (*e.g.* MMA, Bilateral, Adv-Interp, and Sense) out of the 7 unimproved defenses, even state-of-the-art attacks MT or attacks of AA evaluate them to be more robust than SAT, which is not necessarily the case according to our MD attacks. Particularly, against the MT attack, the robustness of SAT is 45.34%, while the robustness of Bilateral, Adv-Interp and Sense are 55.07%, 61.22% and 46.22%, respectively. For the MMA defense, best attack from AA evaluates its robustness to be 47.38%, which is slightly higher than SAT's 45.26%. However, under our MD attacks, all 4 models show much lower robustness than SAT (3%-10% lower). Next, we will investigate the imbalanced gradients problem in the unimproved defenses.

Table 1: Robustness (%) of 12 defense models evaluated by different attacks. The attacks are divided into 2 groups: 1) traditional attacks for robustness evaluation and our MD (column 3-6); and 2) more recent attacks and our MDMT (column 7-10). The defenses are also divided into 2 groups: 1) SAT or improved defenses (top rows); and 2) those that are not improved over SAT (bottom rows). Results in $(\cdot)$ in the MDMT column show the robustness decrease compared to the PGD attack.

Defense	Clean	FGSM	PGD	CW	**MD**	MT	Best of AA	ODI	**MDMT**
RST	89.69	69.60	62.09	60.87	**60.17**	**59.80**	60.62	59.93	59.86 (-2.23)
UAT	86.46	68.31	61.08	62.11	**59.36**	56.72	58.20	57.98	**56.65** (-4.43)
TRADES	84.92	60.87	55.00	53.69	**53.10**	**52.67**	53.82	52.68	52.78 (-2.22)
MART	83.09	61.43	56.10	53.02	**51.84**	51.12	51.55	51.15	**51.07** (-5.03)
SAT	86.83	56.88	45.94	45.73	**45.64**	45.34	46.38	45.26	**45.25** (-0.69)
Dynamic	85.35	55.19	46.36	45.53	**43.93**	42.75	43.64	43.03	**42.69** (-3.67)
MMA	84.62	61.85	51.09	52.05	**45.63**	42.62	47.38	43.00	**41.92** (-9.17)
Bilateral	90.73	71.10	60.95	57.82	**39.82**	55.07	41.36	38.65	**37.21** (-23.74)
Adv-Interp	90.25	77.94	72.48	67.92	**45.33**	61.22	40.60	41.43	**37.59** (-34.89)
FeaScatter	89.98	77.40	68.64	57.10	**43.12**	43.10	40.84	39.61	**36.86** (-31.78)
Sense	91.51	72.71	59.86	57.67	**40.64**	46.22	38.88	38.15	**35.25** (-24.61)
JARN-AT1	81.96	61.48	42.50	27.46	**15.03**	16.01	37.25	14.90	**14.60** (-27.90)

Table 2: Robustness (%) of WideResNet-34-10 models trained with/without label smoothing.

Defense	FGSM	PGD	MD
SAT	56.88	46.47	**45.71**
+ Label Smoothing	59.10	51.15	**44.54**
Natural	26.41	**0.00**	**0.00**
+ Label Smoothing	48.09	10.86	**0.00**

3.2 Label Smoothing Causes Imbalanced Gradients.

The PGD robustness of Bilateral, FeaScatter, and Adv-Interp decrease the most (*e.g.* $23\% - 34\%$) against our MDMT attack, which indicates that these defenses may have caused imbalanced gradients. All three defenses use label smoothing as part of their training scheme to improve adversarial training, which we suspect is one common cause of imbalanced gradients. Given a sample $\mathbf{x}$ with label y, label smoothing encourages the model to learn an uniform logits or probability distribution over classes $j \neq y$. This tends to smooth out the input gradients of $\mathbf{x}$ with respect to these classes, resulting in smaller gradients. In order to confirm label smoothing indeed causes imbalanced gradients, we train a WideResNet-34-10 model using natural training ('Natural') and SAT with or without label smoothing (smoothing parameter 0.5). We report their robustness in Table 2. The PGD robustness of the naturally-trained model also "increases" to 10.86%, which is still 0% under our MD attack. Using smoothed labels in SAT defense also "increases" PGD robustness by almost 5%, which in fact, decreases by 1%. These evidences confirm that label smoothing indeed causes imbalanced gradients, leading to overestimated robustness if evaluated by regular attacks like PGD. Interestingly, it appears that adversarial training can inhibit moderately the imbalanced gradients problem of label smoothing. This is because the adversarial examples used for adversarial training are specifically perturbed to the $j \neq y$ classes, thus helping avoid uniform logits over classes $j \neq y$ to some extent.

4 Conclusion

In this paper, we identify *Imbalanced Gradients*, a new situation where traditional attacks such as PGD can fail and produce overestimated adversarial robustness. We also proposed a new attack called Margin Decomposition (MD) attack to leverage imbalanced gradients via a two-stage attacking process. By evaluating 12 state-of-the-art defense models, we find that 6 of them are susceptible to imbalanced gradients and their PGD robustness suffers a significant drop against our MD attacks. We also identified label smoothing as a possible cause of imbalanced gradients. Our results indicate that future defenses should avoid causing imbalanced gradients to obtain more reliable adversarial robustness.

REFERENCES

Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In *NeurIPS*, 2019.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *S&P*, 2017.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C. Duchi, and Percy Liang. Unlabeled data improves adversarial robustness. In *NeurIPS*, 2019.

Alvin Chan, Yi Tay, Yew Soon Ong, and Jie Fu. Jacobian adversarially regularized networks for robustness. In *ICLR*, 2020.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *arXiv:2003.01690*, 2020.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Max-margin adversarial (MMA) training: Direct input space margin maximization through adversarial training. *arXiv preprint arXiv:1812.02637*, 2018.

Logan Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272*, 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Sven Gowal, Jonathan Uesato, Chongli Qin, Po-Sen Huang, Timothy A. Mann, and Pushmeet Kohli. An alternative surrogate loss for pgd-based adversarial testing. *arXiv preprint arXiv:1910.09338*, 2019.

Warren He, Bo Li, and Dawn Song. Decision boundary analysis of adversarial examples. In *ICLR*, 2018.

Jungeum Kim and Xiao Wang. Sensible adversarial learning, 2020. URL `https://openreview.net/forum?id=rJlf_RVKwr`.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Marius Mosbach, Maksym Andriushchenko, Thomas Alexander Trost, Matthias Hein, and Dietrich Klakow. Logit pairing methods can fool gradient-based attacks. *arXiv preprint arXiv:1810.12042*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

Jonathan Uesato, Brendan O'Donoghue, Pushmeet Kohli, and Aäron van den Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *ICML*, 2018.

Jianyu Wang and Haichao Zhang. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *ICCV*, 2019.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *ICML*, 2019.

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

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In *NeurIPS*, 2019.

Haichao Zhang and Wei Xu. Adversarial interpolation training: A simple approach for improving model robustness, 2020. URL https://openreview.net/forum?id=Syejj0NYvr.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.

A ALGORITHMS OF MT AND MDMT ATTACKS

We also propose a Margin Decomposition Multi-Targeted (MDMT) attack, a multi-targeted version of our MD attack. The loss terms used by MDMT at different attacking stages are defined as follows:

$$\ell_k^r(\mathbf{x}_k, y) = \begin{cases} \mathbf{z}_t & \text{if } k < \frac{K}{2} \text{ and } r \bmod 2 = 0 \\ -\mathbf{z}_y & \text{if } k < \frac{K}{2} \text{ and } r \bmod 2 = 1 \\ \mathbf{z}_t - \mathbf{z}_y & \text{if } k \geq \frac{K}{2}, \end{cases} \tag{2}$$

where, $\mathbf{z}_t$ is the logits of the target class $t \neq y$. Like the MT attack, MDMT will attack each possible target class one at a time, then select the strongest adversarial example at the end. That is, the target class $t \neq y$ will be switched to a different target class at each restart.

Algorithm 1 and Algorithm 2 below describe the complete attacking procedure of our Margin Decomposition (MD) attack and its Multi-Targeted (MDMT) version.

Algorithm 1 Margin Decomposition Attack

Input: clean sample $\mathbf{x}$, label y, model f.
Output: adversarial example $\mathbf{x}_{adv}$
Parameters: Perturbation bound ϵ, step size α, number of restarts n, number of steps K.
$\mathbf{x}_{adv} \leftarrow \mathbf{x}$
for $r \in \{1, ..., n\}$ **do**
  Initialize $\mathbf{x}_0$ by one step of perturbation along the opposite direction of gradients.
  for $k \in \{1, ..., K\}$ **do**
    Update $\mathbf{x}_k$ by Eq. (1)
    if $\ell(\mathbf{x}_{adv}) < \ell(\mathbf{x}_k)$ **then**
      $\mathbf{x}_{adv} \leftarrow \mathbf{x}_k$
    end if
  end for
end for
return $\mathbf{x}_{adv}$

Algorithm 2 Margin Decomposition MultiTargeted attack

Input: clean sample $\mathbf{x}$, class label y, class set $\mathcal{T}$, model f.
Output: adversarial example $\mathbf{x}_{adv}$
Parameters: Perturbation bound ϵ, PGD step size α, number of restarts n, number of steps K.
$n_{\mathrm{r}} \leftarrow \lfloor n/|\mathcal{T}| \rfloor$, $\mathbf{x}_{adv} \leftarrow \mathbf{x}$
for $r \in \{1, ..., n_{\mathrm{r}}\}$ **do**
  for $t \in \mathcal{T}$ **do**
    Initialize $\mathbf{x}_0$ by one step of perturbation along the opposite direction of gradients.
    for $k \in \{1, ..., K\}$ **do**
      Update $\mathbf{x}_k$ by Eq. (2)
      if $\ell(\mathbf{x}_{adv}) < \ell(\mathbf{x}_k)$ **then**
        $\mathbf{x}_{adv} \leftarrow \mathbf{x}_k$
      end if
    end for
  end for
end for
return $\mathbf{x}_{adv}$

B 12 EXAMINED DEFENSE MODELS

We focus on adversarial training models, which are arguably the most effective defense models to date. The 12 selected defense models are as follows. The standard adversarial training (SAT) (Madry et al., 2018) trains models on adversarial examples generated by PGD attack. Dynamic adversarial training (Dynamic) (Wang et al., 2019) trains on adversarial examples with gradually increased convergence quality. Max-Margin Adversarial training (MMA) (Ding et al., 2018) trains on adversarial examples

with gradually increased margin (*e.g.* the perturbation bound ϵ). For MMA, we evaluate the released "MMA-32" model. Jacobian Adversarially Regularized Networks (JARN) adversarially regularize the Jacobian matrices, and can be combined with 1-step adversarial training (JARN-AT1) to gain additional robustness (Chan et al., 2020). For JARN, we only evaluate the JARN-AT1 as JARN has already been completely broken in (Croce and Hein, 2020). We implement JARN-AT1 on the basis of their released implementation of JARN. Sensible adversarial training (Sense) (Kim and Wang, 2020) trains on loss-sensible adversarial examples (perturbation stops when loss exceeds certain threshold). Bilateral Adversarial Training (Bilateral) (Wang and Zhang, 2019) trains on PGD adversarial examples with adversarially perturbed labels. For Bilateral, we mainly evaluate its released strongest model "R-MOSA-LA-8". Adversarial Interpolation (Adv-Interp) training (Zhang and Xu, 2020) trains on adversarial examples generated under an adversarial interpolation scheme with adversarial labels. Feature Scattering-based (FeaScatter) adversarial training (Zhang and Wang, 2019) crafts adversarial examples using latent space feature scattering, then trains on these examples with label smoothing. TRADES (Zhang et al., 2019) replaces the CE loss of SAT by the KL divergence for a better trade-off between robustness and natural accuracy. Based on TRADES, RTS (Carmon et al., 2019) and UAT (Alayrac et al., 2019) improve robustness by training with $10\times$ more unlabeled data. Misclassification Aware adveRsarial Training (MART) (Wang et al., 2020) further improves the above three methods with a misclassification aware loss function.

C Ablation of the Proposed MD Attacks

In this section, we investigate the influence of three factors to our MD attack: 1) initialization method, 2) the second attacking stage, and 3) the stage ordering. We use AdvInterp as our target model, and conduct the following attack experiments on CIFAR-10 test data.

Initialization Method. We compare the success rates of our MD attacks using random initialization versus the opposite direction initialization (see Algorithm 1 and Algorithm 2). The results are reported in Table 3. As can be observed, the opposite direction initialization demonstrates a clear advantage over random initialization. Particularly, for MD attack, using opposite direction initialization can improve the attack success rate by 8%, while for MDMT attack, the success rate can also be improved.

The Second Attacking Stage. We further investigate the importance of the second stage of attacking with the full margin loss in our MD attacks. Here, we fix the initialization method to the opposite direction initialization. The attack success rates with or without the second stage are also reported in Table 3. We highlight that attacking the full margin loss via the second attacking stage can consistently increase the success rate. Especially for MD attack, a 4.99% improvement can be achieved by the second attacking stage.

The Ordering of the Stages. To verify that the ordering of the two stages is suitable for MD attacks, we evaluate a new version of our MD attacks with the two stages are switched: the first stage optimizes the full margin loss and the second stage explores the individual loss terms. The results are reported in Table 3 (the last two columns). As can be observed, MD attacks become much less effective when the two stages are switched. This is because

Table 3: Attack success rates (%) of our MD and MDMT attacks with 1) different initialization methods, 2) with/without the second attacking stage, and 3) with/without stages being switched. Experiments are conducted on defense model AdvInterp and dataset CIFAR-10.

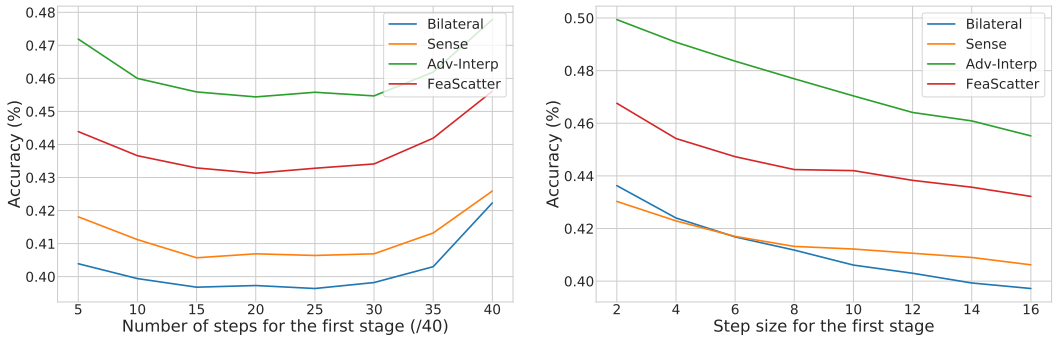
Attacks	Initialization		Second Attacking Stage		Switching Stage	
	Random	Opposite	without	with	Yes	No
MD	46.32	**54.67**	49.68	**54.67**	48.41	**54.67**
MDMT	61.07	**62.41**	61.82	**62.41**	60.62	**62.41**

D Parameter Analysis of the Proposed MD Attack

We further investigate the sensitivity of our MD attack to two parameters: 1) the number of perturbation steps, and 2) the step size. Here, we focus on the first attacking stage as the second stage is a typical PGD attack, which has been thoroughly investigated in (Wang et al., 2019).

Number of Steps for the First Stage. The total number of perturbation steps is set to $K = 40$. When we vary the perturbation steps of the first stage, the remaining steps will be given to the second stage. MD attack will reduce to the regular PGD attack if the perturbation steps of the first stage is set to 0. Here, we vary the steps from 5 to 40 in a granularity of 5. The step size is set to 8/255 and 2/255 for the first and second attacking stage, respectively. The robustness of 4 defense models including Bilateral, Adv-Interp, FeaScatter and Sense are illustrated in Figure 2a. As can be observed, the performance of our MD attack tends to drop at both ends, and the best performance is achieved at $[20, 30]$. Therefore, we suggest to simply use half of the perturbation steps for the first stage (*e.g.* switching to the second stage at the $\frac{K}{2}$-th step).

Step Size for the First Stage. We vary the step size used for the first stage from 2/255 to 16/255 in a granularity of 2/255. Following the above experiments, here we fix the number of steps in each stage to 20. The evaluated robustness (or model accuracy on the generated attacks) of defense models Bilateral, Adv-Interp and FeaScatter are illustrated in Figure 2b. A clear improvement of using large step size in the first stage can be observed. Therefore, we suggest to use a large step size for the first stage of exploration.



(a) (b)

Figure 2: Parameter analysis of MD attack: (a) the accuracies of 5 defense models under MD attacks with different number of perturbation steps in the first stage; (b) the accuracies of 5 defense models under MD attacks with different step sizes in the first stage.

E Comparison to the Four Individual Attacks in AutoAttack

In this section, we compare the model robustness evaluated by the individual attacks in the AA ensemble with our MD attacks. These experiments follow the same setting as in Section 3. The results are shown in Table 4. As can be observed, our MDMT attack demonstrates a superior performance across all the defense models. Moreover, our MD attack which is as efficient as PGD attack can even achieve better performance than all individual attacks on 6 out of 12 models.

Table 4: Attack success rates (%) of the 4 individual attacks (column 2-6) in AA attack and our MD attacks (column 6-7). The best results are highlighted in **bold**. The second best results are highlighted in underline

Defense	APGD$_{CE}$	APGD$_{DLR}$	FAB	Square	MD	MDMT
RST	61.47	60.64	60.62	66.63	60.17	**59.86**
UAT	59.86	62.03	58.20	66.37	59.36	**56.65**
TRADES	55.08	54.04	53.82	59.48	53.10	**52.78**
MART	55.52	52.51	51.55	57.45	51.84	**51.07**
SAT	46.40	46.56	46.38	53.13	45.64	**45.25**
Dynamic	45.81	45.86	43.64	53.49	43.93	**42.69**
MMA	49.40	50.18	47.38	55.48	45.63	**41.92**
Bilateral	58.26	43.11	41.55	59.33	39.82	**37.21**
Adv-Interp	69.36	49.43	40.60	66.87	45.33	**37.59**
FeaScatter	62.03	48.96	40.84	59.12	43.12	**36.86**
Sense	54.80	48.41	38.88	61.31	40.64	**35.25**
JARN-AT1	37.25	67.55	67.48	75.32	15.03	**14.60**