IMBALANCED GRADIENTS: A NEW CAUSE OF OVERESTIMATED ADVERSARIAL ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating the robustness of a defense model is a challenging task in adversarial robustness research. Obfuscated gradients, a type of gradient masking, have previously been found to exist in many defense methods and cause a false signal of robustness. 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 to 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.

1 INTRODUCTION

Deep neural networks (DNNs) are vulnerable to adversarial examples, which are input instances crafted by adding small adversarial perturbations to natural examples. Adversarial examples can fool DNNs into making false predictions with high confidence, and transfer across different models (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). One common pitfall in adversarial robustness evaluation is the phenomenon of gradient masking (Papernot et al., 2017; Tramèr et al., 2018) or obfuscated gradients (Athalye et al., 2018), leading to weak or unsuccessful attacks and false signals of robustness. To demonstrate "real" robustness, newly proposed defenses claim robustness based on results of white-box attacks such as PGD (Madry et al., 2018), and at the same time, demonstrate that they are not a result of obfuscated gradients. In this work, we show that the robustness may still be overestimated even when there are no obfuscated gradients. Specifically, 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. Our key contributions are:

- We identify a new type of subtle effect called *imbalanced gradients*, which can cause highly overestimated adversarial robustness and cannot be detected by detection methods for obfuscated gradients. Especially, We highlight that label smoothing is one of the major causes of imbalanced gradients.
- We propose *Margin Decomposition (MD)* attacks to exploit imbalanced gradients. MD leverages the attackability of the individual terms in the margin loss in a two-stage attacking process. We also introduce two variants of MD for existing attacks PGD and MT.
- We conduct extensive evaluations on 12 state-of-the-art defense models and find that 6 of them suffer from imbalanced gradients and their PGD robustness drops by more than 9% against our MD attacks. Our MD attacks exceed state-of-the-art attacks when imbalanced gradients occur.

2 BACKGROUND

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.* $\|\mathbf{x}_{adv} - \mathbf{x}\|_\infty \leq \epsilon$).

Adversarial Attack. Adversarial examples can be crafted by maximizing a classification loss ℓ by one or multiple steps of adversarial perturbations. For example the one-step Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) and the iterative FGSM (I-FGSM) attack (Kurakin et al., 2017). Projected Gradient Descent (PGD) (Madry et al., 2018) attack is another iterative method that projects the perturbation back onto the ϵ-ball centered at $\mathbf{x}$ when it goes beyond. Carlini and Wagner (CW) (Carlini and Wagner, 2017) attack generates adversarial examples via an optimization framework. Whilst there exist other attacks such as Frank-Wolfe attack (Chen et al., 2018a), distributionally adversarial attack (Zheng et al., 2019) and elastic-net attacks (Chen et al., 2018b), the most commonly used attacks for robustness evaluations are FGSM, PGD, and CW.

Several recent attacks have been proposed to produce more accurate robustness evaluations than PGD. This includes Fast Adaptive Boundary Attack (FAB) (Croce and Hein, 2019), MultiTargeted (MT) attack (Gowal et al., 2019), Output Diversified Initialization (ODI) attack (Tashiro et al., 2020), and AutoAttack (AA) (Croce and Hein, 2020). FAB finds the minimal perturbation necessary to change the class of a given input. MT (Gowal et al., 2019) is a PGD-based attack with multiple restarts and picks a new target class at each restart. ODI provides a more effective initialization strategy with diversified logits. AA attack is a parameter-free ensemble of four attacks: FAB, two Auto-PGD attacks, and the black-box Square Attack (Andriushchenko et al., 2019). AA has demonstrated to be one of the state-of-the-art attacks to date (Croce and Hein, 2020).

Adversarial Loss. Many attacks use Cross Entropy (CE) as the adversarial loss: $\ell_{ce}(\mathbf{x}, y) = -\log \mathbf{p}_y$. The other commonly used adversarial loss is the margin loss (Carlini and Wagner, 2017): $\ell_{margin}(\mathbf{x}, y) = \mathbf{z}_{max} - \mathbf{z}_y$, with $\mathbf{z}_{max} = \max_{i \neq y} \mathbf{z}_i$. Shown in (Gowal et al., 2019), CE can be written in a margin form (*e.g.* $\ell_{ce}(\mathbf{x}, y) = \log(\sum_{i=1}^{C} e^{\mathbf{z}_i}) - \mathbf{z}_y$), and in most cases, they are both effective. While FGSM and PGD attacks use the CE loss, CW and several recent attacks such as MT and ODI adopt the margin loss. AA has one PGD variant using the CE loss and the other PGD variant using the Difference of Logits Ratio (DLR) loss. DLR can be regarded as a "relative margin" loss. In this paper, we identify a new effect that causes overestimated adversarial robustness from the margin loss perspective and propose new attacks by decomposing the margin loss.

Adversarial Defense. In response to the threat of adversarial attacks, many defenses have been proposed such as defensive distillation (Papernot et al., 2016), feature/subspace analysis (Xu et al., 2017; Ma et al., 2018), denoising techniques (Guo et al., 2018; Liao et al., 2018; Samangouei et al., 2018), robust regularization (Gu and Rigazio, 2014; Tramèr et al., 2018; Ross and Doshi-Velez, 2018), model compression (Liu et al., 2018; Das et al., 2018; Rakin et al., 2018) and adversarial training (Goodfellow et al., 2015; Madry et al., 2018). Among them, adversarial training via robust min-max optimization has been found to be the most effective approach (Athalye et al., 2018). A number of new techniques have been proposed to further enhance the adversarial training (Wang et al., 2019; Zhang et al., 2019; Carmon et al., 2019; Alayrac et al., 2019; Wang and Zhang, 2019;

Zhang and Wang, 2019; Zhang and Xu, 2020; Wang et al., 2020; Kim and Wang, 2020; Ding et al., 2018; Chan et al., 2020). We will discuss and evaluate these adversarial training-based defenses with our proposed attacks in Section 4.

3 Imbalanced Gradients and Margin Decomposition Attack

We first give a toy example of imbalanced gradients and show how regular attacks can fail in such a situation. We then empirically verify their existence in deep neural networks, particularly for some adversarially-trained models. Finally, we propose the Margin Decomposition attack to exploit the imbalanced gradients. Since CE and margin loss are the two commonly used loss functions for adversarial attack and CE can be written in a margin form (Gowal et al., 2019), here we focus on the margin loss to present the phenomenon of imbalanced gradients.

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.

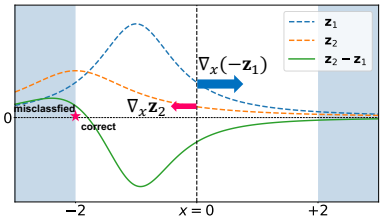


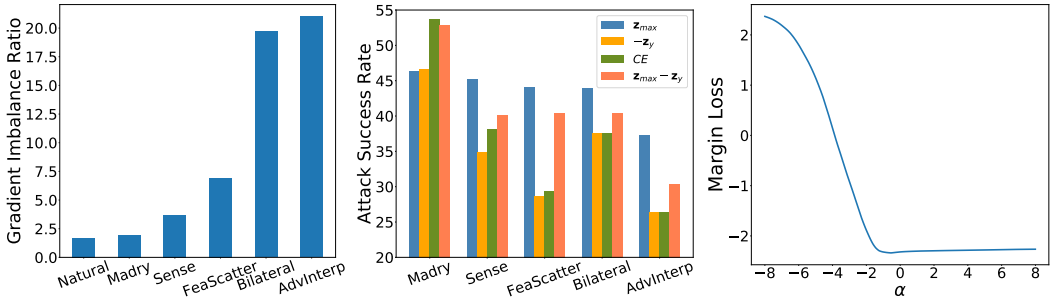
Figure 1: A toy illustration of *imbalanced gradients* at $x = 0$: the gradient of margin loss ($\mathbf{z}_2 - \mathbf{z}_1$) is dominated by its $-\mathbf{z}_1$ term, pointing to a suboptimal attack direction towards +2, where x is still correctly classified.

3.1 Imbalanced Gradients in DNNs

The situation can be extremely complex for DNNs with high-dimensional inputs, as imbalanced gradients can occur at each input dimension. It thus requires a metric to quantitatively measure the degree of gradient imbalance. Here, we propose such a metric named *Gradient Imbalance Ratio* (GIR) to measure the imbalance ratio for a single input $\mathbf{x}$, which can then be averaged over multiple inputs to produce the imbalance ratio for the entire model.

Definition of GIR. To measure the imbalance ratio, we focus on the input dimensions that are dominated by one loss term. An input dimension x_i is dominated by a loss term (*e.g.* $\mathbf{z}_{max}$) means that 1) the gradients of loss terms at x_i have different directions ($\nabla_{x_i}\mathbf{z}_{max} \cdot \nabla_{x_i}(-\mathbf{z}_y) < 0$), and 2) the gradient of the dominant term is larger (*e.g.* $|\nabla_{x_i}\mathbf{z}_{max}| > |\nabla_{x_i}(-\mathbf{z}_y)|$). According to the dominant term, we can split these dimensions into two subsets $\mathbf{x}_{s_1}$ and $\mathbf{x}_{s_2}$ where $\mathbf{x}_{s_1}$ are dominated by the $\mathbf{z}_{max}$ term, while $\mathbf{x}_{s_2}$ are dominated by the $-\mathbf{z}_y$ term. The overall dominance effect of each loss term can be formulated as $r_1 = \left\|\nabla_{\mathbf{x}_{s_1}}(\mathbf{z}_{max} - \mathbf{z}_y)\right\|_1$ and $r_2 = \left\|\nabla_{\mathbf{x}_{s_2}}(\mathbf{z}_{max} - \mathbf{z}_y)\right\|_1$. Here, we use the L_1-norms instead of L_0-norms (*i.e.* the number of dominated dimensions) to also take into consideration the gradient magnitude. To keep the ratio larger than 1, GIR is computed as: $GIR = \max\{\frac{r_1}{r_2}, \frac{r_2}{r_1}\}$.

Note that the GIR metric is not a general measure of imbalance. Rather, it is designed only for assessing gradient imbalance for *adversarial robustness evaluation*. GIR focuses specifically on the imbalanced input dimensions, and uses the L_1 norm to also take into account the influence of these



(a) (b) (c)

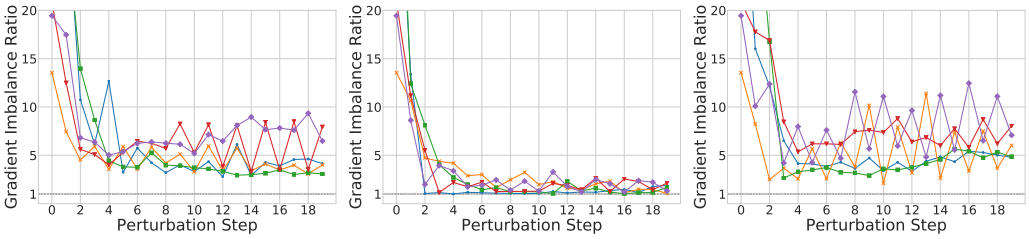
Figure 2: (a): Gradient imbalance ratio of 5 models. (b): Attack success rate of PGD-20 with different losses. (c): The margin loss of the AdvInterp defense model on points $\mathbf{x}^* = \mathbf{x} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}}(-\mathbf{z}_y))$, where $\mathbf{x}$ is a natural sample and $\text{sign}(\nabla_{\mathbf{x}}(-\mathbf{z}_y))$ is the signed gradient of loss term $-\mathbf{z}_y$. All these experiments are conducted on test images of CIFAR-10.

dimensions to the final output. The ratio reflects how far away the imbalance towards one direction than the other.

GIR of both Naturally- and Adversarial-trained DNNs. With the GIR metric, we next investigate 6 DNN models including a naturally-trained (Natural) model and 5 adversarially-trained models using standard adversarial training (Madry et al., 2018) (Madry), sensible adversarial training (Kim and Wang, 2020) (Sense), feature scattering-based adversarial training (Zhang and Wang, 2019) (FeaScatter), bilateral adversarial training (Wang and Zhang, 2019) (Bilateral), and adversarial interpolation training (Zhang and Xu, 2020) (AdvInterp). We present these defense models here because they represent different levels of gradient imbalance (a complete analysis of more models can be found in Appendix F). Natural, Madry and Sense are WideResNet-34-10 models, while others are WideResNet-28-10 models. We train Natural and Madry following typical settings in (Madry et al., 2018) while others use their officially released models. We compute the GIR scores of the 6 models based on 1000 randomly selected test samples, and show them in Figure 2a. One major observation is that some defense models can have a much higher imbalance ratio than either naturally-trained or Madry's model. This confirms that gradient imbalance does exist in DNNs, and some defenses tend to train the model to have highly imbalanced gradients. We will show, in Section 4, that this situation of imbalanced gradients can cause highly overestimated robustness when evaluated using a traditional PGD attack.

Imbalanced Gradients Reduce Attack Effectiveness. When there are imbalanced gradients, the attack can be pushed by the dominant term to produce weak attacks, and the non-dominant term alone can lead to more successful attacks. To illustrate this, in Figure 2b, we show the success rates of PGD attack on the above 5 defense models (Natural has zero robustness against PGD) with different losses: CE loss, margin loss, and the two individual margin terms. We consider 20-step PGD (PGD-20) attacks with step size $\epsilon/4$ and $\epsilon = 8/255$ on all CIFAR-10 test images. Intuitively, the two margin terms could lead to less effective attacks, as they only provide partial information about the margin loss. This is indeed the case for the low gradient imbalance model Madry. However, for highly imbalanced models Sense, FeaScatter, Bilateral and AdvInterp, attacking the $\mathbf{z}_{max}$ term produces even more powerful attacks than attacking the margin loss. This indicates that the gradient of the margin loss is shifted by the dominant term (*e.g.* $-\mathbf{z}_y$ in this case) towards a less optimal direction, which inevitably causes less powerful attacks. Compared between attacking CE loss and attacking $-\mathbf{z}_y$, they achieve a very close performance on imbalanced models. This shows a stronger dominant effect of $-\mathbf{z}_y$ in CE loss ($\ell_{ce}(\mathbf{x}, y) = \log(\sum_{i=1}^{C} e^{\mathbf{z}_i}) - \mathbf{z}_y$). It is worth mentioning that, while both GIR and this individual term-based test can be used to check whether there are significantly imbalanced gradients in a defense model, GIR alone cannot fully reflect the attack success rate. Figure 2c shows an example of how the $-\mathbf{z}_y$ term leads the attack to a suboptimal direction: the margin loss is flat at the $\nabla_{\mathbf{x}}(-\mathbf{z}_y)$ direction, yet increases drastically at an opposite direction. In this example, the attack can actually succeed if it increases (rather than decreases) $\mathbf{z}_y$.

Gradients can be Balanced by Attacking Individual Loss Terms. Here, we show that, interestingly, imbalanced gradients can be balanced by attacking the non-dominant term. Consider the AdvInterp model tested above as an example, the dominant term is $-\mathbf{z}_y$. Figure 3 illustrates the GIR values of 5 randomly selected CIFAR-10 test images by attacking them using PGD-20 with



(a) PGD-20 with loss $\mathbf{z}_{max} - \mathbf{z}_y$ (b) PGD-20 with loss $\mathbf{z}_{max}$ (c) PGD-20 with loss $-\mathbf{z}_y$

Figure 3: Changes in gradient imbalance ratio when apply PGD-20 ($\epsilon = \frac{8}{255}$) attack with the margin loss (a), only the $\mathbf{z}_{max}$ term (b), or only the $-\mathbf{z}_y$ term (c), on the AdvInterp model for 5 CIFAR-10 test images. The imbalance ratio is effectively reduced by attacking a single $\mathbf{z}_{max}$ term.

different margin terms or the full margin loss. As can be observed that, for all three losses, the GIRs are effectively reduced after the first few steps. However, only the non-dominant term $\mathbf{z}_{max}$ manages to stably reduce the imbalance ratio to around 2. This indicates optimizing the individual terms separately can help avoid the situation of imbalanced gradients and the attack can indeed benefit from more balanced gradients (see the higher success rate of $\mathbf{z}_{max}$ in Figure 2b).

3.2 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, $k \in \{1, \ldots, K\}$ is the perturbation step, $r \in \{1, \ldots, n\}$ is the r-th restart, $\bmod$ 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 help escape the flat loss landscape observed in Figure 2c, 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. The loss terms used by MDMT at different attacking stages are defined as follows:

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

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. The complete algorithms of MD and MDMT can be found in Appendix A, and an ablation study can be found in Appendix G.

Initialization Perspective Interpretation of MD Attacks. Previous works have shown that random or logits diversified initialization are crucial for generating strong adversarial attacks (Madry et al., 2018; Tashiro et al., 2020). Compared to random or logits diversified initialization, our MD attacks can be interpreted as a type of *adversarial initialization*, *i.e.*, initialize at the adversarial sub-directions defined by the two terms of the margin loss. Moreover, rather than a single step of initialization, our MD attacks iteratively explore the optimal starting point during the entire first attacking stage.

4 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 E. 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. For 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 H. Adversarial robustness is measured by the model accuracy on adversarial examples crafted by these attacks on CIFAR-10 test images.

4.1 EVALUATION RESULTS

Table 1 reports the full evaluation result, where RST, UAT and TRADES are the top 3 best defenses. The Madry defense demonstrates $\sim 45\%$ robustness consistently against either PGD or stronger attacks such as MT, AA, ODI and our MD attacks. This indicates that Madry does not have imbalanced gradients and indeed brings consistent robustness, which is in line with other studies about Madry (Athalye et al., 2018; Croce and Hein, 2020; Uesato et al., 2018). While the rest 11 defense models are all developed based on Madry, they exhibit quite different robustness. Only 4 defenses including RST, UAT, TARDES and MART are indeed improved over Madry, while the other 7 defense models are actually not as robust as Madry, 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, 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, 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, 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 AA evaluate them to be more robust than Madry, which is not necessarily the case according to our MD attacks. Particularly, against the MT attack, the robustness of Madry 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, AA attack evaluates its robustness to be 45.69%, which is slightly higher than Madry's 45.26%. However, under our MD attacks, all 4 models show much lower robustness than Madry (3%-10% lower). Next, we will investigate the imbalanced gradients problem in the unimproved defenses.

4.2 DEFENSE TECHNIQUES THAT MAY CAUSE IMBALANCED GRADIENTS

Here, we focus on 6 unimproved (compared to Madry) defenses: MMA, Bilateral, Adv-Interp, FeaScatter, Sense, and JARN-AT1. Their PGD-evaluated robustness has been reduced for $> 9\%$ by our MDMT attack.

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) Madry or improved defenses (top rows); and 2) those that are not improved over Madry (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	AA	ODI	**MDMT**
RST	89.69	69.60	62.09	60.87	**60.17**	**59.80**	59.66	59.93	59.86 (-2.23)
UAT	86.46	68.31	61.08	62.11	**59.36**	56.72	56.94	57.98	**56.65** (-4.43)
TRADES	84.92	60.87	55.00	53.69	**53.10**	**52.67**	53.18	52.68	52.78 (-2.22)
MART	83.09	61.43	56.10	53.02	**51.84**	51.12	**51.05**	51.15	51.07 (-5.03)
Madry	86.83	56.88	45.94	45.73	**45.64**	45.34	**45.17**	45.26	45.25 (-0.69)
Dynamic	85.35	55.19	46.36	45.53	**43.93**	42.75	42.88	43.03	**42.69** (-3.67)
MMA	84.62	61.85	51.09	52.05	**45.63**	42.62	45.69	43.00	**41.92** (-9.17)
Bilateral	90.73	71.10	60.95	57.82	**39.82**	55.07	37.96	38.65	**37.21** (-23.74)
Adv-Interp	90.25	77.94	72.48	67.92	**45.33**	61.22	38.58	41.43	**37.59** (-34.89)
FeaScatter	89.98	77.40	68.64	57.10	**43.12**	43.10	38.79	39.61	**36.86** (-31.78)
Sense	91.51	72.71	59.86	57.67	**40.64**	46.22	36.10	38.15	**35.25** (-24.61)
JARN-AT1	81.96	61.48	42.50	27.46	**15.03**	16.01	30.11	14.90	**14.60** (-27.90)

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

Defense	FGSM	PGD	MD
Madry	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**

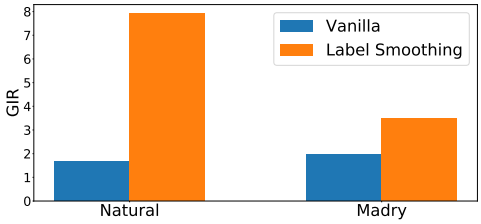


Figure 4: Gradient Imbalance Ratio (GIR) of models trained with/without label smoothing.

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. This indicates that these defenses may have caused imbalanced gradients, as also indicated by their high GIR values in Figure 2a. 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 Madry adversarial training with or without label smoothing (smoothing parameter 0.5). We report their robustness in Table 2, and show their gradient imbalance ratios (GIRs) in Figure 4. According to GIRs, adding label smoothing into the training process immediately increases the imbalance ratio, especially in natural training. 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 Madry 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.

Other Defense Techniques that may Cause Imbalanced Gradients. The other 3 unimproved defenses MMA, Sense and JARN-AT1 adopt different defense techniques to improve robustness. MMA is a margin-based defense that maximizes the shortest successful perturbation for each data point. MMA only perturbs correctly classified clean examples, and the perturbation stops immediately at misclassification (into a $j \neq y$ class). In other words, MMA focuses on examples that are around the decision boundary (*e.g.* $\mathbf{z}_{max} = \mathbf{z}_y$) between class y and all other classes $j \neq y$. During training, the decision boundary margin is maximized by pulling the boundary away from these examples. This

process tries to maximize the distance to the closest decision boundary (*e.g.* towards the weakest class) and finally results in equal distances to all other classes. This tends to generate a uniform prediction over classes $j \neq y$, a similar effect of label smoothing, and causes imbalanced gradients.

Similar to MMA, Sense perturbs training examples until a certain loss threshold is satisfied. While in MMA the threshold is misclassification, in Sense, it is the loss value with respect to probability (*e.g.* $\mathbf{p}_y = 0.7$). This type of training procedures with a specific logits or probability distribution regularization has caused the imbalanced gradients problem for both MMA and Sense. Note that, Sense causes much severe imbalanced gradients than MMA. We conjecture it is because optimizing over a probability threshold is much easier than moving the decision boundary.

JARN-AT1 is also a regularization-based adversarial training method. Different from MMA or Sense, it regularizes the model's Jacobian (*e.g.* input gradients) to resemble natural training images. Such an explicit input gradients regularization tends to reduce the input gradients to a much smaller magnitude and only keep the salient part of input gradients. The input gradients associated with other $j \neq y$ classes will be minimized to cause an imbalance to that associated with class y. This has caused PGD to produce 27.90% more robustness than our MDMT attack. Note that, even the recent AA attack still produces 15.51% overestimated robustness compared to our MDMT.

4.3 An Attack View of Imbalanced Gradients

As shown in Table 1, recent attacks ODI and AA are more effective than traditional attacks PGD and CW against imbalanced gradients. Here, we provide some insights into why these techniques are effective against imbalanced gradients. We consider attacking AdvInterp as an example and show how the gradient imbalance ratio (GIR) changes in different attacking processes.

Logits Diversified Initialization Helps Avoid Imbalanced Gradients. ODI randomly initializes the perturbation by adding random weights to logits at its first 2 steps. The random weights change the gradients' size, thus can also mitigate imbalanced gradients, as shown in Figure 5a in Appendix C. However, initialization only helps the first 2 steps, and the imbalance ratio still jumps in the following steps. Our attack provides a more direct and efficient exploration of imbalance gradients, thus can maintain a low imbalance ratio even after the first few steps (see Figure 5c in Appendix C). As also shown in Table 1, our MDMT attack is consistently more effective than ODI.

Exploration Beyond the ϵ-ball Helps Avoid Imbalanced Gradients. AA is an ensemble of four attacks: two Auto-PGD attacks and two existing attacks FAB and Square. By inspecting the individual attacks, we found that the most effective method is FAB. FAB first finds a successful attack using unbounded perturbation size (*e.g.* $> \epsilon$), then minimizes the perturbation to be within the ϵ-ball. As shown in Figure 5b in Appendix C, the first few steps of exploration outside the ϵ-ball can effectively avoid imbalanced gradients. This is also why our MD attacks use a large step size in the first stage. However, the imbalance ratio tends to increase when FAB attempts to minimize the perturbation (steps 10 - 16). We believe FAB can be further improved following our decomposition strategy.

Imbalanced Gradients are not Easily Detected or Circumvented by Existing Methods. We also show, in Appendix B, that defense models with imbalanced Gradients can still pass the five checking rules of obfuscated gradients, and that many times of restarts with random initialization or momentum method does not help escape imbalanced gradients in Appendix D. This makes imbalanced gradients more subtle and should be carefully checked to avoid overestimated robustness.

5 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 proposed a new metric to investigate the imbalanced gradients problem in current defense models. 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 identified a set of possible causes of imbalanced gradients, and effective countermeasures. 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.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. *arXiv:1912.00049*, 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.

Jinghui Chen, Dongruo Zhou, Jinfeng Yi, and Quanquan Gu. A frank-wolfe framework for efficient and effective adversarial attacks. *arXiv preprint arXiv:1811.10828*, 2018a.

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *AAAI*, 2018b.

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. *arXiv:1907.02044*, 2019.

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

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. Compression to the rescue: Defending from adversarial attacks across modalities. In *KDD*, 2018.

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.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 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.

Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *ICLR*, 2018.

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`.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *ICLR*, 2017.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *CVPR*, 2018.

Qi Liu, Tao Liu, Zihao Liu, Yanzhi Wang, Yier Jin, and Wujie Wen. Security analysis and enhancement of model compressed deep learning systems under adversarial attacks. In *ASPDAC*, 2018.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *ICLR*, 2018.

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.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *S&P*, 2016.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Asia CCS*, 2017.

Adnan Siraj Rakin, Jinfeng Yi, Boqing Gong, and Deliang Fan. Defend deep neural networks against adversarial examples via fixed and dynamic quantized activation functions. *arXiv preprint arXiv:1807.06714*, 2018.

Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *AAAI*, 2018.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *ICLR*, 2018.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013.

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

Yusuke Tashiro, Yang Song, and Stefano Ermon. Output diversified initialization for adversarial attacks. *arXiv:2003.06878*, 2020.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *ICLR*, 2018.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *ICLR*, 2018.

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.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

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.

Tianhang Zheng, Changyou Chen, and Kui Ren. Distributionally adversarial attack. In *AAAI*, 2019.