A ALGORITHMS OF MT AND MDMT ATTACKS

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. (**??**)
      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 IMBALANCED GRADIENTS ARE DIFFERENT FROM OBFUSCATED GRADIENTS

Imbalanced gradients occur when one loss term dominating the attack towards a suboptimal gradient direction, which does not necessarily block gradient descent like obfuscated gradients. Therefore, it does not have the characteristics of obfuscated gradients, and can not be detected by the five checking rules for obfuscated gradients (Athalye et al., 2018). Here, we test all the five rules on the four defense models that exhibited significant imbalanced gradients: Adv-Interp, FeaScatter, Bilateral, and Sense. Note that all these models were trained and tested on CIFAR-10 dataset.

One-step attacks perform better than iterative attacks. When gradients are obfuscated, iterative attacks are more likely to get stuck in a local minima. To test this, we compare the success rate of one-step attack FGSM and iterative attack PGD in Table 3. We see that PGD outperforms FGSM consistently on all the four defense models, i.e., no obvious sign of obfuscated gradients.

Unbounded attacks do not reach 100% success. Increasing distortion bound does not increase success. Larger distortion bound gives the attacker more ability to attack. So, if gradients are not obfuscated, unbounded attack should reach 100% success rate. To test this, we run an "unbounded"

PGD attack with $\epsilon = 1$. As shown in Table 3, all models are completely broken by this unbounded attack. This again indicates that the overestimated robustness is caused by a different effect rather than obfuscated gradients.

Black-box attacks are better than white-box attacks. If a model is obfuscating gradients, it should fail to provide useful gradients in a small neighborhood. Therefore, using a substitute model should be able to evade the defense, as the substitute model was not trained to be robust to small perturbations. To test this, we run black-box transferred PGD attack on naturally trained substitute models. We find that all four defenses are robust to transferred attacks ("Transfer" in Table 3). We also attack the four defense models using gradient-free attack SPSA (Uesato et al., 2018). For SPSA, we use a batch size of 8192 with 100 iterations, and run on 1000 randomly selected CIFAR-10 test images. We confirm that SPSA cannot degrade their performance. None of these results indicate obfuscated gradients.

Random sampling finds adversarial examples. Brute force random search within some ϵ-ball should not find adversarial examples when gradient-based attacks do not. Following (Athalye et al., 2018), we choose 1000 test images on which PGD fails. We then randomly sample 10^5 points for each image from its $\epsilon = 8/255$-ball region, and check if any of them are adversarial. The results (*e.g.* "Random") shown in Table 3 confirms that random sampling cannot find an adversarial example when PGD does not.

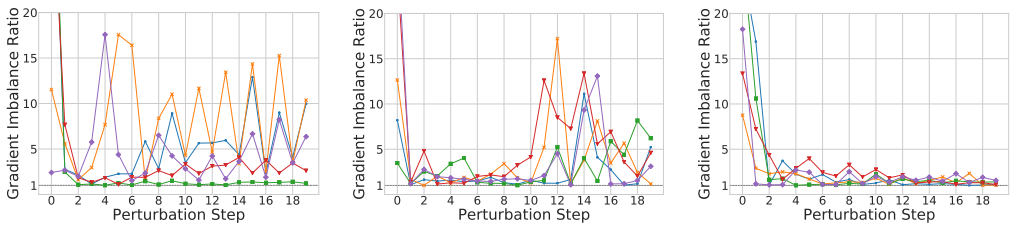
All the above test results lead to one conclusion that the robustness of the four defenses is not a result of obfuscated gradients. This indicates that imbalanced gradients does not share the characteristics of obfuscated gradients, thus cannot be detected following the five test principles for obfuscated gradients. This makes adversarial robustness evaluation more difficult. Therefore, imbalanced gradients should be carefully addressed for more accurate robustness evaluation.

Table 3: Test of obfuscated gradients for four defense models that have significant imbalanced gradients following (Athalye et al., 2018): attack success rate (%) of different attacks. None of the above results indicates a clear sign of obfuscated gradients.

Defense	FGSM	PGD	Unbounded	Transfer	SPSA	Random
Adv-Interp (Zhang and Xu, 2020)	23.06	27.52	100.00	10.89	24.80	0.00
FeaScatter (Zhang and Wang, 2019)	22.60	31.36	100.00	11.11	28.20	0.00
Bilateral (Wang and Zhang, 2019)	28.90	39.05	100.00	9.23	36.00	0.00
Sense (Kim and Wang, 2020)	27.29	40.14	100.00	9.90	37.90	0.00

C Can Logits Diversified Initialization Help Circumvent Imbalanced Gradients?

Figure 5 shows the GIR values of 5 randomly selected CIFAR-10 test images at the first 20 steps of ODI, FAB, or our MDMT attack. The FAB attack is the most effective attack in the AA ensemble.



(a) ODI (b) FAB (c) MDMT (Ours)

Figure 5: Gradient imbalance ratio at the first 20 steps of ODI (a), FAB (b) and our MDMT (c) attacks on the AdvInterp model for 5 randomly selected CIFAR-10 test images.

D CAN RANDOM RESTART OR MOMENTUM HELP CIRCUMVENT IMBALANCED GRADIENTS?

As we discussed in Section 3, many times of random starts can potentially increase the probability of finding an adversarial example. Momentum method is another way to help escape overfitting to local gradients (Sutskever et al., 2013). Here, we test whether random restart or momentum can help avoid imbalanced gradients. For random restart, we run 400-step PGD attack with 100 restarts (PGD$^{100\times400}$). For momentum, we use momentum iterative FGSM (MI-FGSM) (Dong et al., 2018) with 40 steps, 2 restarts and momentum 1.0. For both attacks, we set $\epsilon = 8/255$ and step size $\alpha = 2/255$. We apply the two attacks on 1000 randomly chosen CIFAR-10 test images, and report the robustness in Table 4 for the four defense models checked in Section B. Compared to traditional PGD with 40 steps, the robustness can indeed be decreased by PGD$^{100\times400}$ except Bilateral, an observation consistent with our analysis in Section 3 that more restarts can lower model accuracy. However, the robustness is still highly overestimated compared to that by our MDMT attack. This indicates that imbalanced gradients can exist in wide-spanned input regions, resulting in a low probability for random restart to find successful attacks. To our surprise, MI-FGSM performs even worse than traditional PGD. On three defense models (eg. Adv-Interp, FeaScatter, and Sense), it produces even higher robustness than PGD. This implies that accumulating velocity in the gradient direction can make the overfitting even worse when there are imbalanced gradients. This again confirms that the imbalanced gradients problem should be explicitly addressed to obtain more reliable adversarial robustness.

Table 4: Robustness (%) of four defense models that have significant imbalance gradients against PGD$^{100\times400}$ and MI-FGSM attack.

Defense	PGD	MDMT	PGD$^{100\times400}$	MI-FGSM
Adv-Interp	72.48	**37.59**	70.70	73.25
FeaScatter	68.64	**36.86**	64.10	70.79
Bilateral	60.95	**37.21**	64.08	51.52
Sense	59.86	**35.25**	56.00	62.41

E 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 (Madry) (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 Madry 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.

F GRADIENT IMBALANCED RATIO OF MORE DEFENSE MODELS

In this Section, we provide a complete analysis on the gradient imbalance ratios (GIRs) of all 12 examined defense models and a naturally trained model. The GIR values of these models are shown in Figure 6. One immediate observation is that the GIR value of a defense model is positively correlated with its robustness drop against our MDMT attack in Table 1. Slightly imbalanced defense models Madry, TRADES and RST demonstrate minimum robustness drop, while the PGD-evaluated robustness of highly imbalanced defense models FeaScatter, Bilateral and AdvInterp can drop drastically against our MD attacks. This verifies that higher gradient imbalance can indeed causes more overestimated robustness by regular PGD attack.

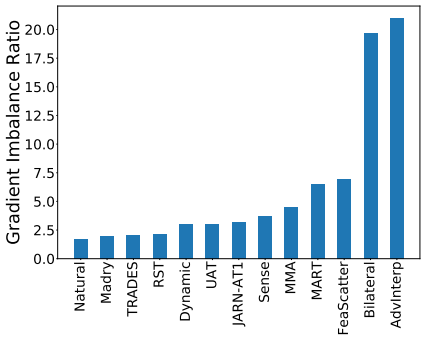


Figure 6: Gradient imbalance ratios (GIRs) of 12 defense models and a naturally trained model ("Natural"). All models are trained on CIFAR-10 dataset.

G ABLATION OF THE PROPOSED MD ATTACKS

In this section, we investigate the influence of two factors to our MD attack: 1) initialization method, and 2) the second attacking stage. 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 5. 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 5. 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.

Table 5: Attack success rates (%) of our MD and MDMT attacks with 1) different initialization methods, and 2) with/without the second attacking stage. Experiments are conducted on defense model AdvInterp and dataset CIFAR-10.

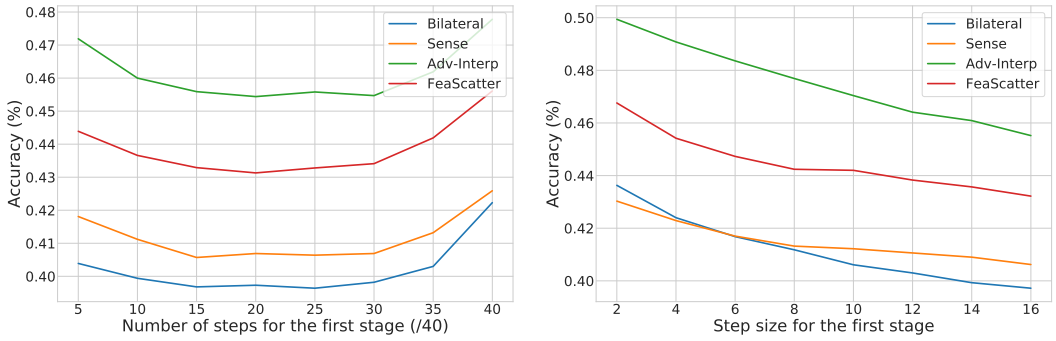
Attacks	Initialization		Second Attacking Stage	
	Random	Opposite	without	with
MD	46.32	**54.67**	49.68	**54.67**
MDMT	61.07	**62.41**	61.82	**62.41**

H 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 7a. 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 7b. 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 7: 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.